Н.В. Лукашевич

Тезаурусы
в задачах информационного поиска

Москва

2010

Лукашевич Наталья Валентинова
Тезаурусы в задачах информационного поиска – М., 2010. – 396 с., ил.

Оглавление

Введение

Область современного информационного поиска чрезвычайно разнообразна. Она включает такие задачи, как собственно поиск информации, фильтрация, рубрикация и кластеризация документов, поиск ответов на вопросы, автоматическое аннотирование документа и группы документов, поиск похожих документов и дубликатов, сегментирование документов и многое другое. Когда подобные операции выполняет человек, ему необходимо выявить основное содержание документа, его основную тему и подтемы, и для этого обычно используется большой объем знаний о языке, мире, организации связного текста.

Абсолютно подавляющее число современных методов обработки неструктурированной информации решают эти задачи на основе минимальных дополнительных предварительных знаний и базируются на моделях текста как набора слов (“bag of words”), предлагая изощренные методы учета частотностей встречаемости слов в предложении, тексте, наборе документов, совместной встречаемости слов и т.п. Пословные модели не учитывают такие языковые явления как синонимия, многозначность, существование лексических отношений между словами.

Недостаток лингвистических и онтологических знаний (знаний о мире), используемых в приложениях информационного поиска и автоматической обработки текстов, приводит к разнообразным проблемам. Нехватка знаний приводит к нерелевантному поиску в тех случаях, если способы формулировки запросов отличаются от способов описания релевантных ситуаций в документах. Эта проблема усугубляется при обработке длинных запросов, при поиске ответов на вопросы в вопросно-ответных системах, а также при поиске информации в специализированных поисковых системах, в которых содержится значительно меньшее число документов, чем в Интернет. Нехватка знаний приводит к снижению качества при автоматической фильтрации и рубрикации документов, к излишним повторам или нарушению связности при автоматическом аннотировании и др.

Еще одним типом обычно не достаточно используемых лингвистических знаний в приложениях информационного поиска является неучет структурных свойств связного текста. Как известно, связный текст имеет сложную иерархическую структуру. Существенным проявлением связности текста является так называемая глобальная связность текста, когда в тексте имеется одна главная тема, а вся остальная информация подчинена изложению этой основной темы. Одним из проявлений глобальной связности текста является его лексическая связность, когда в тексте содержится множество близких по смыслу слов и выражений. Между тем подавляющее большинство подходов рассматривает текст как совокупность независимых друг от друга слов, характеризующихся частотностью встречаемости в документе и коллекции.

В настоящее время знания о языке и мире описываются в таких компьютерных ресурсах как онтологии и тезаурусы. Однако на практике применение тезаурусов и онтологий в промышленных информационных системах, основанных на автоматической обработке текстов, не слишком распространено.

Такая ситуация связана с целым рядом обстоятельств.

Во-первых, если предлагается использовать некоторый лингвистический ресурс, то он должен включать описания десятков тысяч слов и словосочетаний. Процент ошибок ресурса должен быть настолько мал, чтобы не испортить возможные улучшения, получаемые от применения этого ресурса. При этом нужно понимать, что ведение любого лингвистического ресурса всегда будет отставать от развития предметной области, то есть даже наиболее качественный лингвистический ресурс будет всегда неполон.

Во-вторых, применение тезаурусов и онтологий в информационном поиске требует высокого качества разрешения многозначности слов текста. Однако тестирование

качества разрешения лексической многозначности, проводимых на конференциях SemEval и Senseval, показало, что качество разрешения многозначности для всех многозначных слов текста пока не достигает уровня, достаточного для эффективного применения тезаурусов и онтологий в приложениях информационного поиска.

В-третьих, применение отношений тезауруса или онтологии для расширения запросов может столкнуться с проблемой неточно описанных отношений или отношений, которые не соответствуют контексту запроса. Применение таких отношений часто ведет к значительному снижению точности поиска. Так, в последнее время глобальные поисковые системы Яндекс и Google стали активно применять расширение запросов однокоренными словами, что может рассматриваться как минимальный тезаурус, но во многих случаях даже такое минимальное расширение запроса может оказаться нерелевантным.

Наконец, существует мнение, что применяемые статистические методы имплицитно учитывают лингвистическую информацию, что текст – это лишь набор характеристик (features), которые хорошо учитываются статистическими моделями. В качестве примеров моделирования лингвистических подходов статистическими методами Хелен Ворхес (Voorhees, 1999) приводит следующие примеры: морфологический анализ может быть приближен стеммингом, извлечение словосочетаний - выявлением часто встречающихся пар слов, процедуры разрешения многозначности могут быть смоделированы мерами сходства контекстов.

Вместе с тем, как показали эксперименты в рамках конференции по информационному поиску TREC и семинаре «Надежный доступ к информации» (Reliable Information Access), проведенном в 2003 году, существуют типы запросов к поисковым системам, которые являются сложными для современных технологий информационного поиска и, следовательно, качество поиска по этим запросам достаточно низкое. Среди потенциальных методов, которые могли бы улучшить выдачу поисковых систем по таким запросам, указывались методы расширения запросов, в том числе, и с использованием специальных ресурсов – тезаурусов.

При поиске в отличных от Интернета коллекциях документов, таких как профессиональные информационные базы, внутрикорпоративные ресурсы, отличающиеся относительно небольшим (по сравнению с Интернет) размером, возможность несоответствия языка запроса и языка документов считается достаточно серьезной проблемой.

Таким образом, важным является вопрос о том, каково должно быть внутреннее устройство лингвистических ресурсов, содержащих знания о понятиях, терминах, значениях языковых выражений в широких предметных областях, которые не только бы не ухудшали характеристики информационного поиска, а, напротив, сделали его более содержательным. Кроме того, необходимо понять, каким образом описанные в лингвистических ресурсах знания могут быть встроены в современные модели информационного поиска.

Рассмотрим основные направления использования разного рода лингвистических и терминологических ресурсов в информационном поиске.

Как известно, в 1960 – 1980е годы в информационном поиске активно использовались так называемые информационно-поисковые тезаурусы, которые предназначались для описания содержания документов нормализованными ключевыми словами в процессе ручного индексирования людьми-индексаторами.

В то время большинство информационных систем не являлись полнотекстовыми, а хранили достаточно ограниченный набор информации о документе: библиографические данные, реферат. Добавление списка ключевых слов, характеризующих основное содержание документа, существенно расширяло возможности поиска документов. С начала семидесятых годов создаются национальные и международные стандарты разработки информационно-поисковых тезаурусов.

Появление полнотекстовых информационно-поисковых систем, а также возможностей поиска по всем словам текста с помощью методов ранжированного информационного поиска значительно снизило значимость разработки и использования информационно-поисковых тезаурусов, поскольку давало возможность поиска текста неподготовленному пользователю в любых предметных областях без дополнительных посредников в виде специально разработанных тезаурусов и профессиональных индексаторов.

Многочисленные исследования по определению эффективности различных методов представления документов при информационном поиске показали, что эффективность пословного индексирования сравнима с эффективностью поиска, использующего ручное индексирование по качественному информационно-поисковому тезаурусу (Salton, 1986; Sparck Jones, 1981), для создания которого нужно было еще затратить достаточно много средств и усилий,а, кроме того, нужно было еще осуществлять качественное ручное индексирование документов по этому тезаурусу.

Действительно, использование хорошо разработанного тезауруса при ручном индексировании должно снимать проблемы синонимии, близких понятий, многозначности. Однако при этом могут возникнуть существенные различия между понятиями, используемыми в тезаурусе, и информационной потребностью пользователя, когда пользователю трудно сформулировать описание нужных ему текстов посредством понятий тезауруса, или тезаурус действительно не содержит адекватных понятий. В этих случаях пословное индексирование имеет преимущество из-за больших выразительных возможностей в том смысле, что пользователь может сформулировать запрос на естественном языке без всяких дополнительных ограничений.

Кроме того, при ручном индексировании серьезную проблему составляет фактор субъективности, когда приписывание тексту терминов тезауруса зависит от умения и опыта индексаторов, от количества текстов, которые необходимо проиндексировать и т.п.

Тем не менее, и в настоящее время существуют информационные службы, имеющие и разрабатывающие информационно-поисковые тезаурусы, а также имеющие штат профессиональных индексаторов, индексирующих документы на основе тезаурусов. Примерами таких организаций являются Исследовательская служба Конгресса США, индексирующая по тезаурусу Legislative Indexing Vocabulary, Продовольственная и Сельскохозяйственная организация при ООН (ФАО), которая развивает тезаурус AGROVOC, службы Европейского сообщества, использующие для индексирования Европейского законодательства тезаурус EUROVOC. Деятельность таких служб наиболее близка к библиотечной деятельности, в рамках которой книги и документы классифицируются по библиотечным классификаторам типа УДК.

Происходит и процесс обновления стандартов разработки тезаурусов. Так, например, американский национальный стандарт по разработке и ведению контролируемых словарей Z39.19 последний раз обновлялся в 2003 году.

Современные стандарты разработки и использования информационно-поисковых тезаурусов четко ограничивают сферу их применения. Так, например, международный стандарт по разработке одноязычных тезаурусов (ISO 2788) указывает, что стандарт должен применяться в организациях, имеющих людей-индексаторов, которые анализируют содержание документов и описывают основные темы документов с помощью терминов тезауруса. «Применение стандарта не предполагает его применение в тех организациях, которые используют полностью автоматические методы индексирования».

Возникает вопрос, почему существующая парадигма разработки информационно-поисковых тезаурусов не дает возможности использовать созданные ресурсы в автоматических режимах индексирования текста. Как и можно ли создавать тезаурусы для автоматического индексирования? В книге мы рассмотрим, какие особенности

существующей парадигмы разработки информационно-поисковых тезаурусов ограничивают их использование в автоматических режимах.

С 80-х годов 20 века начинает активно обсуждаться парадигма автоматического концептуального индексирования документов, то есть индексирования документов не пословным индексом, а концептуальным, в котором синонимы сведены к одной и той же единице, а многозначные слова и термины разведены к разным концептуальным единицам (Woods, 1997).

Такие системы как SCISSORS (Jacobs, Rau, 1990) и FERRET (Mauldin, 1991) реализуют идею концептуального индексирования для узких предметных областей: используются специальные структуры представления понятий и развиваются специальные алгоритмы для создания концептуального индекса.

С опубликованием в 1995 году ресурса английского языка WordNet, структура которого представляет собой иерархическую сеть лексикализованных понятий английского языка – синсетов, многие исследователи пытались реализовать идею концептуального индексирования на базе этого ресурса.

Однако изначально WordNet не предназначался для приложений автоматической обработки текстов, и исследователи в области компьютерной обработки текстов встретились с многочисленными проблемами, которые затрудняют использование его в таких приложениях. В частности, в большом числе экспериментов по использованию знаний, описанных в WordNet, часто не наблюдалось улучшение характеристик информационного поиска.

Среди наиболее существенных проблем, которыми обычно объясняется такая ситуация, можно отметить следующие: слишком большое количество значений слов, проблемы с автоматическим выбором значения, нехватка отношений между синсетами, другой информации для разрешения многозначности, проблемы собственно описания отношений между синсетами (какие должны быть, по каким правилам устанавливаться и т.п.).

Несмотря на некоторые неудачи использования WordNet в конкретных приложениях, появление этого ресурса вызвало огромный резонанс в мире. На базе WordNet выполнены тысячи экспериментов исследователями из многих стран мира, предложены самые разнообразные алгоритмы. Понимание уровня достигнутых результатов, знакомство с описанными в литературе экспериментами очень важно для исследований в области информационного поиска на базе других тезаурусных и онтологических ресурсов.

Так, мы покажем, что после примерно 10 лет исследования применения WordNet для решения задачи эффективного расширения поискового запроса, в течение которых не удавалось получить устойчивого улучшения качества информационного поиска, в 2004 году в трудах конференции SIGIR было опубликовано исследование, в котором был предложен метод использования информации из WordNet в классическом информационном поиске для расширения запроса, который улучшил показатели поиска по сравнению с достаточно качественной базовой моделью поиска.

Кроме того, формализованное описание лексики английского языка, представленное в WordNet, позволяет в автоматизированном режиме относительно легко строить словари разного назначения, извлекать те или иные классы слов, что в значительной мере облегчает создание различных словарных ресурсов и внутренних словарей информационных систем и систем автоматической обработки текстов.

Исследователи из разных стран начали разработку сходных ресурсов для своих языков. Согласованные усилия для развития wordnet'ов были реализованы в таких европейских проектах как EuroWordNet, BalkaNet, Meaning, в рамках которых были разработаны wordnet'ы для голландского, итальянского, испанского, немецкого и других языков. Было начато и несколько проектов по созданию русского WordNeta.

Разработчики wordnet'ов пытались учесть проблемы так называемого Принстонского WordNet'а, сделать их более приспособленными к компьютерным приложениям, в том числе и в сфере информационного поиска. Многочисленные публикации обсуждают возможности кластеризации различных значений в обобщенные значения, проблемы введения дополнительных отношений в новые ресурсы, появляются дополнения в уже созданные ресурсы. Так, например, исходный Принстонский WordNet обогатился отношениями между разными частями речи, разметкой по тематическим областям, словообразовательными отношениями.

Другие исследователи изучают возможности более смыслового семантического поиска на основе так называемых онтологий – концептуальных описаний знаний о предметных областях и в целом о мире, содержащих совокупности понятий, отношений между ними, правил вывода. Была выдвинута концепция Семантической сети Интернет (Semantic Web), где предполагалось, что качество поиска в Интернет можно значительно улучшить посредством использования таких онтологий.

Существует множество разных определений онтологий. Широкие определения онтологий, позволяющие разные степени формализации описаний, включают в понятие онтологии и упомянутые выше информационно-поисковые тезаурусы, и тезаурусы типа wordnet.

Часть исследователей считает, что онтологии должны описывать знания о мире и быть независимыми от конкретного языка. Однако для того, чтобы применить такого рода независимую от языка онтологию в практических задачах информационных технологий, которые во многом связаны с переработкой неструктурированной информации, текстов, необходимо установить отношения между понятиями языковонезависимой онтологии и значениями лексических единиц конкретного естественного языка. Кроме того, часть исследователей (см. например, (Wilks, 2008)) подвергают сомнению возможность создания большой онтологии совершенно независимо от естественного языка.

Онтологии обычно классифицируются на онтологии верхнего уровня, описывающие наиболее общие знания о мире, и предметные онтологии, описывающие знания о конкретных предметных областях. Так и знания о языке делятся на общеупотребительные («литературный язык») и терминологию конкретной предметной области.

Но какой бы текст, принадлежащий значимой предметной области, мы ни взяли, он всегда включает и общеупотребительные языковые единицы, и термины данной предметной области, а понимание этого текста требует как общих знаний о мире, так и знаний в данной конкретной области. На практике же одни исследователи создают онтологии верхнего уровня, другие создают онтологии предметных областей, общезначимый язык изучается лингвистами, а термины – языковые единицы конкретных предметных областей - исследуются терминологами. Однако лингвистический ресурс, предназначенный для поддержки автоматической обработки текста в рамках современных информационных технологий, должен каким-то образом совмещать эти разные типы знаний.

Более того, для удобства создания того или иного терминологического ресурса, онтологии для некоторой предметной области, исследователи, разработчики считают, что эта область некоторым образом отделима от других предметных областей. Однако современные информационные системы имеют дело со сверхбольшими коллекциями документов, значимая часть которых содержит документы, включающие терминологию разных предметных областей. Так, в экономических документах значимую роль занимает терминология правовой области, а в правовых документах - экономическая терминология, в документах по банковскому делу значимое место занимает терминология налоговой сферы, бухгалтерии, фондового рынка и т.п.

Таким образом, при всем обилии научной литературы по вопросам построения информационно-поисковых тезаурусов, тезаурусов типа WordNet, онтологий открытыми остаются следующие вопросы:

- каким образом в прикладных компьютерных ресурсах оптимально сочетать описание взаимоотношений лексических единиц и описание онтологических знаний о мире,
- какая модель описания неструктурированной широкой предметной области наиболее оптимальна для того, чтобы, с одной стороны, создать ее в разумные сроки и охватить всю важную для специалистов терминологию, с другой стороны, чтобы созданная формализованная модель была полезна в широком круге приложений информационного поиска и автоматической обработки текстов,
- каким образом оптимально сочетать описание общеупотребительной лексики литературного языка и терминологии конкретной предметной области в формализованных моделях, предназначенных для компьютерных приложений.

В данной книге предлагаются подходы к решению вышеперечисленных вопросов. Книга посвящена описанию опыта автора по созданию сверхбольших лингвистических ресурсов для автоматической обработки текстов в рамках современных информационных технологий и сопоставлению созданных ресурсов и технологий с подобными проектами, развиваемыми в мире.

Под руководством и с непосредственным участием автора книги разрабатываются такие онтологические ресурсы как Тезаурус русского языка РуТез, Онтология по естественным наукам и технологиям ОЕНТ, созданы ряд онтологических ресурсов в конкретных областях, таких как компьютерная безопасность, авиация, банковское дело, выборы и др.

Созданные ресурсы применяются в таких технологиях автоматической обработки текстов как автоматическое концептуальное индексирование, расширение поискового запроса, рубрицирование, автоматическое аннотирование отдельных документов и групп тематически близких документов, кластеризация документов.

Исследования, связанные с представлением знаний о языке и предметной области, были поддержаны рядом международных и российских научных грантов: грантами Фонда МакАртуров, Фонда Форда, российских научных фондов РФФИ и РГНФ, стипендиями компании Яндекс.

Созданные ресурсы и технологии использовались в проектах, выполненных для ряда государственных и коммерческих организаций (ФГУП НИИ Восход, Государственная Дума Российской Федерации, Счетная палата Российской Федерации, Банк России, ФСБ, компания Гарант, компания Рамблер Медиа и др.)

Материал, изложенный в книге, частично излагался в спецкурсах, читавшихся в Московском государственном университете на филологическом факультете и факультете ВМиК в 2003-2005 году.

Учебный курс, разработанный на основе предварительных материалов книги, вошел в число победителей:

- открытого конкурса учебных курсов в области разработки программного обеспечения, организованного компанией Microsoft и факультетом вычислительной математики и кибернетики МГУ им. М.В. Ломоносса в 2006 году (http://www.microsoft.com/ Rus/Msdnaa/Curricula/Default.mspx);
- конкурса учебных курсов по информационному поиску «Класс 2006», организиванного компанией Яндекс (http://company.yandex.ru/class/courses/solovyev.xml).

Предварительные материалы данной книги излагались в ряде глав учебного пособия Добров Б.В., Иванов В.В., Лукашевич Н.В., Соловьев В.Д. «Онтологии и тезаурусы: модели, инструменты, приложения». – М., Изд-во Интуит, 2008.

Книга делится на два раздела.

В первом разделе (части 1-3) мы опишем различные подходы к созданию больших лингвистических ресурсов на примере конкретных проектов. Также мы подробно рассмотрим различные алгоритмы и системы, которые используют эти ресурсы для решения различных задач информационного поиска. Описывая алгоритмы, мы будем обращать особое внимание на методы оценки их качества, достигнутые показатели, которые указывают на то, удалось или нет разработчикам ресурсов и алгоритмов достигнуть лучшего качества по сравнению с пословными статистическими методами.

Во второй разделе книги (части 4-6) мы опишем принципы разработки лингвистического ресурса русского языка тезауруса РуТез и наши эксперименты по применению этого тезауруса в различных задачах обработки текстов для приложений информационного поиска. Описывая собственные алгоритмы, мы также уделяем большое внимание экспериментам, которые показывают, насколько качественно удается решать конкретные задачи на базе тезаурусных знаний.

В каждом из двух разделов книги выделяются части, которые подразделяются на главы.

Первая часть первого раздела книги посвящена описанию различных видов тезаурусов, включая тезаурус Роже, информационно-поисковые тезаурусы, тезаурусы типа WordNet.

Во второй части книги мы рассматриваем основные положения современных онтологических исследований, принципы создания онтологических ресурсов.

Особенно подробно рассматриваются принципы установления онтологических отношений, которые нужны для создания ресурсов в различных предметных областях. В это число входят отношения «класс-подкласс», часть-целое, отношения онтологической зависимости.

Следующая, третья часть описывает применение тезаурусов и онтологий в конкретных приложениях информационного поиска. Здесь мы рассматриваем такие системы, как собственно информационный поиск, системы автоматической рубрикации, вопросно-ответные системы, алгоритмы разрешения лексической многозначности, алгоритмы установления лексической связности в тексте, алгоритмы автоматического аннотирования текстов.

Каждая глава этой части строится схожим образом. Сначала описывается общая постановка задачи, некоторые теоретические положения и (или) основные статистические пословные алгоритмы, а также меры измерения качества решения задачи, а далее излагаются методы и результаты применения тезаурусов и онтологий в данной задаче.

Отметим, что среди значимых приложений, относимых к информационному поиску, мы не рассматриваем задачу извлечения информации, в которой могут использоваться онтологические ресурсы. Это связано с тем, что главным предметом нашего интереса являются сверхбольшие плохо структурированные предметные области, и неструктурированные тексты. Задача извлечения информации характеризуется тем, что из текстов извлекается очень небольшое количество типов информации, при этом если используется онтология, то число понятий в ней относительно невелико (Moens, 2006).

С четвертой части начинается второй раздел книги, посвященный рассмотрению наших собственных ресурсов и экспериментов с ними. В этой части будут рассмотрены основные принципы построения Тезауруса русского языка РуТез, методы описания понятий, языковых выражений, тезаурусных отношений, способы отражения разных значений слов, терминов, языковых выражений, описание синонимичности языковых выражений.

В пятой части книги рассматриваются эксперименты и приложения, основанные на знаниях, описанных в Тезаурусе РуТез. В число этих приложений входят: информационный поиск, автоматическая рубрикация текстов, автоматическое аннотирование отдельного текста и совокупности сходных текстов, автоматической разрешение лексической многозначности, построение лексических цепочек и тематического представления связного текста.

В шестой, последней части книги мы рассмотрим основные направления развития тезауруса РуТез, а также технологии разработки других ресурсов, которые были созданы на основе тезауруса РуТез, а именно, принципы устройства и современное состояние Онтологии по естественным наукам и технологиям (ОЕНТ).

Книга предназначена для специалистов, научных работников, аспирантов и студентов, интересующихся вопросами автоматической обработки текстов, применения в информационном поиске лингвистических ресурсов, а также информационным поиском в целом, практическими вопросами применения онтологий.

Для читателей, не знакомых с теориями, применяемыми в компьютерной лингвистике, семантике, с одной стороны, или с теорией и практикой информационного поиска, тестирования информационно-поисковых систем, с другой стороны, мы постарались изложить необходимый для понимания материал, насколько это было возможно в рамках одной книги. Во многих разделах книги имеются специальные подразделы, содержащие такого рода сведения.

Автор благодарит Доброва Б.В. за всемерную поддержку данного исследования; Салий А.Д., Шаталову М.Г., Штернову О.А., Агеева М.С., Сидорова А.В., Штернова С.В. за многолетнее сотрудничество; Юдину Т.Н., Леонтьеву Н.Н., Исакадзе Н. В. за обсуждение результатов работы.

ЧАСТЬ 1. ТЕЗАУРУСЫ

Термин «тезаурус» употребляется по отношению к достаточно различным лингвистическим ресурсам и словарям (Kilgarriff, Yallop, 2000):

1) Во-первых, тезаурусом называется особый вид словарей – идеографический, лексика в которых организуется по тематическому принципу. Первым такого рода словарем явился знаменитый Тезаурус Роже, созданный в 19 веке. Основное назначение таких словарей – помощь в подборе синонимов и близких по смыслу слов при написании текста.

2) Второй тип тезаурусов - информационно-поисковые тезаурусы, описывающие отношения между терминами предметной области – создаются экспертами в некоторой предметной области, и предназначены для помощи при информационном поиске.

3) Тезаурусами также называют относительно недавно появившиеся лингвистические ресурсы типа WordNet и EuroWordNet, описывающие отношения между лексическими значениями естественного языка как иерархическую систему групп синонимов – синсетов.

4) Словосочетание «Ассоциативные тезаурусы» может относиться к двум принципиально разным ресурсам.

С одной стороны, ассоциативным тезаурусом называется словарь описывающий психологические ассоциации между словами, возникающие у людей. Таким словарем, например, является Русский ассоциативный словарь (Караулов, 2002).

Кроме того, термин «ассоциативный тезаурус» употребляется для ссылки на ресурсы, создаваемые автоматически на основе обработке корпусов и показывающие совместную встречаемость пар слов в документах.

Между всеми этими употреблениями термина «тезаурус» есть существенное сходство. В работе (Kilgarriff, Yallop, 2000) дается объединяющее определение тезауруса как ресурса, в котором слова со схожим значением сгруппированы вместе.

Никитина С.Е. (Никитина, 1987, стр. 52) определяет тезаурус как словарь с концептуальным входом и фиксированными семантическими связями между его единицами. Она подчеркивает, что для определения тезауруса существенны оба указанных независимых признака. Например, существуют словари, которые, обеспечивая концептуальный вход, например, по набору синонимов, при этом отношения между словами описывают традиционными толкованиями.

В данной книге рабочим определением тезауруса будет следующее:

> *Тезаурус – это словарь, в котором слова и словосочетания с близкими значениями сгруппированы в единицы, называемые понятиями, концептами или дескрипторами, и в котором явно (в виде отношений, иерархии) указываются семантические отношения между этими понятиями (концептами, дескрипторами).*

Поскольку в данной книге мы рассматриваем, как человеческие знания могут быть описаны в созданных человеком ресурсах и применяться затем в компьютерных приложениях, базирующихся на автоматической обработке текста, то нас прежде всего будут интересовать тезаурусы 2)-3).

Ссылки на использование тезаурусов типа Тезауруса Роже в экспериментах по автоматической обработке текстов можно найти в ряде работ (Kennedy, Szpakowicz, 2008; Jarmasz, Szpakowicz, 2003). Однако такое их использование в компьютерных системах ограничено рядом факторов, которые мы рассмотрим на примере конкретной словарной статьи.

Структура словаря типа Тезауруса Роже (Таб. 1.1.) обычно включает разделение на категории (например, Land – суша, земля) и подкатегории; подразделение подкатегорий обычно производится на основе разделения по частям речи. Слова, следующие за выделенным словом, могут обозначать синонимы, родовые и видовые лексемы по

отношению к предшествующему выделенному слову. Некоторые слова в словарной статье имеют отсылки к другим категориям или подкатегориям тезауруса

Land 342

N. land, earth, ground, dry land, terra firma
continent, mainland, peninsula, chersonese[Fr], delta; tongue of land, neck of land; isthmus, oasis; promontory &c. (projection) 250; highland &c. (height) 206.
coast, shore, scar, strand, beach; playa; bank, lea; seaboard, seaside, seabank, seacoast, seabeach[obs3]; ironbound coast; loom of the land; derelict; innings; alluvium , alluvion[obs3]; ancon.
riverbank, river bank, levee
soil, glebe, clay, loam, marl, cledge, chalk, gravel, mold, subsoil, clod, clot; rock, crag.
V. land, come to land, set foot on the soil, set foot on dry land; come ashore, go ashore, debark
Adj. earthy, continental, midland, coastal, littoral, riparian,; alluvial; terrene &c. (world) 318; landed, predial, territorial; geophilous; ripicolous
Adv. ashore; on shore, on land,

Таблица 1.1 Фрагмент словарной статьи Тезауруса Роже (Roget, 1982)

Обычно отмечаются следующие особенности словарей типа тезаурусов Роже, препятствующие применению таких тезаурусов в автоматической обработке текстов.

Во-первых, в структуре такого тезауруса, в отсылках между категориями заключено большое разнообразие различных типов отношений, которые явным образом не указаны, что затрудняет их использование в приложениях.

Во-вторых, существенным фактором является отсутствие выделенных значений слов. В тех случаях, когда то или иное слово упоминается в разных разделах тезауруса, то это может происходить как из-за того, что в разные разделы попали разные значения слов, так и из-за того, что одно и то же значение слова может быть отнесено в разные категории.

Кроме того, отмечаются проблемы классификации, связанные с жесткой заданностью древесной структуры категорий тезауруса. Возникают вопросы по поводу последовательности решений разделения на категории: какие именно признаки выделять в категории тезауруса, а какие нет (Морковкин, 1970).

В связи с перечисленными проблемами тезаурусов типа тезаурус Роже и в связи с тем, что число публикаций по применению таких тезаурусов для автоматической обработки текстов сравнительно невелико, мы далее не будем подробно рассматривать эксперименты, базирующиеся на использовании такого рода тезаурусов.

Глава 1. Информационно-поисковые тезаурусы

Информационно-поисковые тезаурусы появились в 60-е годы 20 века. В это время большинство информационно-поисковых систем не являлись полнотекстовыми, а хранили достаточно ограниченный набор информации о документе: библиографические данные, реферат. Добавление списка ключевых слов, характеризующих основное содержание документа, существенно расширяли возможности поиска документов. С начала семидесятых годов создаются национальные и международные стандарты разработки информационно-поисковых тезаурусов.

В соответствии с определениями стандартов информационно-поисковый тезаурус – это нормативный словарь, явно указывающий отношения между терминами и предназначенный для описания содержания документов и поисковых запросов.

Основными целями разработки информационно-поисковых тезаурусов являются следующие:

- обеспечение перевода естественного языка документов и пользователей на один и тот же словарь, используемый для индексирования и поиска, таким образом, различия в лексическом составе документа и запроса пользователя сводились к одним и тем же единицам тезауруса,
- обеспечение последовательного использования единиц индексирования,
- обеспечение отношений между терминами - отношения между единицами тезауруса позволяют найти оптимальный термин для описания документа или запроса,
- использование как поискового средства при поиске документов.

Информационно-поисковые тезаурусы создавались как инструмент для ручного описания документов специалистами-индексаторами. Поисковый запрос также предполагалось формулировать на основе единиц тезауруса.

Появление полнотекстовых информационно-поисковых систем, а также возможностей поиска по всем словам текста с помощью методов ранжированного информационного поиска (см. раздел 11.1) значительно снизило значимость разработки и использования информационно-поисковых тезаурусов, поскольку давало возможность поиска текста неподготовленному пользователю в любых предметных областях, без предварительных затрат на разработку тезаурусов.

Многочисленные исследования по определению эффективности различных методов представления документов при информационном поиске показали, что эффективность пословного индексирования сравнима с эффективностью поиска, использующего ручное индексирование по качественному тезаурусу (Salton, 1986; Sparck Jones, 1981), для создания которого нужно было еще затратить достаточно много средств и усилий, кроме того, нужно было еще осуществлять качественное ручное индексирование документов по этому тезаурусу.

Эксперименты по автоматическому индексированию документов и запросов на базе информационно-поисковых тезаурусов не привели к практическому использованию созданных информационно-поисковых тезаурусов в процессе автоматической обработки текстов.

В данной главе мы рассмотрим основные структурные особенности информационно-поисковых тезаурусов, методы их создания и использования, а также обсудим, как эти особенности ограничивают применение информационно-поисковых тезаурусов в процессе автоматической обработки текстов.

1.1. Единицы информационно-поисковых тезаурусов

Основными единицами тезаурусов являются термины предметной области.

Большинство версий стандартов по информационно-поисковым тезаурусам указывают на связь терминов с понятиями предметной области. Американский стандарт указывает, что термин является одним или большим числом слов, обозначающих понятие. Стандарт ISO-2788 подчеркивает, что индексирующий термин - это представление понятия предпочтительно в форме существительного или именной группы.

При этом понятие рассматривается как единица мысли, формируемая мысленно для отражения всех или некоторых свойств конкретного или абстрактного, реально существующего или мысленного объекта. Понятия существуют как абстрактные сущности, независимо от терминов, которые их выражают.

Российский ГОСТ рассматривает понятие как форму мышления, отражающую существенные свойства, связи и отношения предметов и явлений, а термином в определении ГОСТа является слово или словосочетание, являющееся точным обозначением определенного понятия какой-либо области знания.

При этом, определяя единицы тезауруса, ГОСТ 7.74-96 не опирается на определение термина, а определяет единицы тезауруса как лексические единицы информационно-поискового языка – то есть обозначения отдельного понятия, принятые в информационно-поисковом языке и неделимые в этой функции.

Стоит отметить, что не все разработчики тезаурусов четко разделяли понятия и термины. Так, разработчики тезауруса AGROVOC характеризуют его как термино-ориентированный (term-oriented), что находит свое проявление в том, что к термину невозможно добавить синонимы. Эта особенность тезауруса рассматривается авторами как недостаток, который необходимо исправить (Soergel и др., 2004).

Таким образом, разработчики тезаурусов предполагают, что понятие предметной области обычно имеет несколько возможных вариантов лексического представления в тексте, которые рассматриваются как синонимы. Среди таких синонимов выбирается дескриптор – термин, который рассматривается как основной способ ссылки на понятие в рамках тезауруса. Другие термины из синонимического ряда, включенные в тезаурус, называются аскрипторы или недескрипторы. Они используются как вспомогательные элементы, текстовые входы, помогающие найти подходящие дескрипторы.

Поскольку информационно-поисковые тезаурусы обычно создаются для конкретных предметных областей, то их построение существенным образом базируется на таких сущностях как «понятие» и «термин», под которым обычно понимается слово или словосочетание, номинирующее понятие определенной области знания или деятельности (Суперанская и др., 2003, Гринев, 1993; Лейчик, 1994; Володина, 1996).

Именно такое понимание термина является основанием рассматривать информационно-поисковые тезаурусы как вид онтологических ресурсов (см. раздел 4.1).

1.1.1. Дескрипторы информационно-поискового тезауруса

Дескрипторы тезауруса должны соответствовать выбранной предметной области тезауруса. Каждый дескриптор, внесенный в тезаурус, должен представлять отдельное понятие данной области. Дескриптор может быть однословным или многословным. Поскольку часто достаточно трудно понять, представляет ли отдельное понятие многословное словосочетание, многие тезаурусы и руководства уделяют особое внимание основным принципам включения в тезаурус в качестве дескрипторов многословных терминов.

Набор дескрипторов тезауруса должен удовлетворять следующим требованиям:

- посредством выделенных дескрипторов должно быть возможно описать темы абсолютного большинства текстов предметной области;

- для уменьшения субъективности индексирования множество дескрипторов не должно включать совокупности близких дескрипторов, формируются классы условной эквивалентности, когда совокупности близких, но различных понятий сводятся к одному дескриптору (LIV, 1994);
- дескриптор должен быть сформулирован однозначно, его подразумеваемое в рамках тезауруса значение должно быть понятно пользователю. Если однозначный и ясный дескриптор подобрать не удается, термин, взятый в качестве дескриптора, снабжается релятором (краткой пометой) или комментарием.

Стандарт z39.19 рекомендует использовать реляторы для имен дескрипторов даже в тех случаях, когда дескриптор звучит однозначно внутри заданной предметной области, но имеет другие значения в общеупотребительном языке или других значимых областях. «Это облегчает поиск по нескольким базам данных и сопоставление дескрипторов различных предметных областей». Например, предлагается вводить дескриптор *Shells(structures)* для инженерной предметной области, поскольку слово *shell* имеет много значений в английском языке.

Комментарий к дескриптору может серьезно направлять, ограничивать индексатора по использованию того или иного дескриптора для описания текстов.

Так, в тезаурусе LIV (LIV, 1994), который используется для индексирования документов в Исследовательской Службе Конгресса США, имеется дескриптор BUILDING CONSTRUCTION (СТРОИТЕЛЬСТВО), который снабжен следующим комментарием:

> *Используется для публикаций о процессе строительства. Для публикаций по строительному бизнесу, описывающих финансы, планирование, управление, используется дескриптор Construction industries. Публикации о типах производимых работ индексируются дескриптором Construction workers.*

При наличии нескольких кандидатов на роль дескриптора факторами, влияющими на выбор дескриптора, могут быть (ГОСТ 7.25; Герд, с.159-160, 2005):

- соответствие стандартам и рекомендациям по научно-технической терминологии, - в стандарте Z39.19 такое соответствие называется “literary warrant” – литературный мандат,
- краткость и понятность (ГОСТ 7.25),
- соображения частотности (Герд, 2005; ГОСТ 7.25; Z39.19) – частотность в текстах и запросах позволяет приблизить язык тезауруса к языку пользователей и документов (Z39.19)
- выбор наиболее стилистически нейтрального термина. Например, стандарт Z39.19 рекомендует в качестве дескриптора предпочесть термин developing nations (развивающиеся страны), а не underdeveloped countries (недоразвитые страны). Следует избегать в качестве дескрипторов неологизмов, жаргонных и сленговых выражений.
- Герд А.С. (Герд, 2005) указывает, что при выборе дескрипторов важно учитывать лексическую структуру иерархически подчиненных дескрипторов (из двух или большего числа синонимов дескриптором считался тот термин, лексическая структура которого повторяется в подчиненных ему терминах). Например, из двух терминов *возникновение дислокации* и *зарождение дислокации* в качестве дескриптора выбирается второй, так как подчиненные термины *гетерогенное зарождение дислокации*, *гомогенное зарождение дислокации*, дублируют лексему *зарождение*.

Задача выделения дескрипторов из набора близких по значению терминов тесно связана с широкой проблемой стандартизации и унификации терминологии. В идеале тезаурусу должны предшествовать целостное лингвистическое описание языка науки,

данной отрасли знания, работа по стандартизации соответствующей терминологии. Однако связь должна быть и обратной: слова и словосочетания, выделенные в качестве дескрипторов в лучших тезаурусах, следует рекомендовать и в качестве стандартных терминов для тех или иных понятий (Герд, 2005).

1.1.2. Критерии ввода многословных дескрипторов

Поскольку в текстах предметной области может встречаться достаточно много частотных словосочетаний, то обычно стандарты на тезаурусы вводят правила включения терминологических словосочетаний в тезаурусы

Так, ГОСТ 7.25 указывает, что допускается включать словосочетания в словник тезауруса, если в качестве опорного слова они содержат существительное, и если выполнено одно из следующих условий:

- значение словосочетания не выводится из значений его компонентов, например,
- ЧЕРНЫЙ ЯЩИК, АБСОЛЮТНО ЧЕРНОЕ ТЕЛО, ЦАРСКАЯ ВОДКА;
- хотя бы один из компонентов словосочетания не употребляется в составе других сочетаний или употребляется всегда в другом смысле, например, ТОРГОВЛЯ НА ВЫНОС, ЛЕГКАЯ ПРОМЫШЛЕННОСТЬ;
- для данного словосочетания в словнике тезауруса существуют полные синонимы, например, НАТРИЯ ХЛОРИД = ПОВАРЕННАЯ СОЛЬ;
- данное словосочетание является устойчивым словосочетанием с именем собственным:. ТАБЛИЦА МЕНДЕЛЕЕВА, ЗАКОН БОЙЛЯ-МАРИОТТА;
- отдельные слова словосочетания имеют слишком широкое значение, например, слово *машины* в словосочетаниях: СТРОИТЕЛЬНЫЕ МАШИНЫ, ЭЛЕКТРИЧЕСКИЕ МАШИНЫ;
- для данного словосочетания в словнике тезауруса существует общепринятая аббревиатура, например,:
  ПОВЕРХНОСТНО-АКТИВНЫЕ ВЕЩЕСТВА - ПАВ,
  УНИВЕРСАЛЬНАЯ ДЕСЯТИЧНАЯ КЛАССИФИКАЦИЯ - УДК,
  ИНФОРМАЦИОННО-ПОИСКОВЫЙ ТЕЗАУРУС - ИПТ,
  ЭЛЕКТРОННО-ВЫЧИСЛИТЕЛЬНАЯ МАШИНА - ЭВМ;
- разбиение словосочетаний на отдельные компоненты приводит к потере важных для поиска семантических связей. Так, разбиение языкового выражения ЯЗЫК ПРОГРАММИРОВАНИЯ не позволяет установить связи с такими языковыми выражениями как «АЛГОЛ», «КОБОЛ», «ФОРТРАН».

Словосочетания, которые не удовлетворяют перечисленным условиям, разбивают на компоненты.

Американский стандарт Z39.19 помимо вышеперечисленных случаев приводит также критерий общепринятости термина профессиональным сообществом, например, *data processing – обработка данных.*

Кроме того, этот стандарт указывает, что введение многословного дескриптора позволяет избегать ложных корреляций, например, разбиение термина Library science (наука о библиотеках = библиотековедение), может привести к нахождению документов о научных библиотеках (science library).

В работе (Шемакин, 1974) на примере терминов научно-технического тезауруса (Шемакин, 1972) приводятся следующие принципы ввода многословных дескрипторов:

- значение одного из терминов изменилось бы в результате комбинации. Например, термин ПОСАДОЧНЫЕ ПЛОЩАДКИ нельзя представить комбинацией терминов ПОСАДКА и ПЛОЩАДКА, поскольку это привело бы к несвойственному в данном случае использованию термина ПОСАДКА;

- термин-словосочетание обозначает некоторую физическую целостность или специфическое вещество (например, ЦИФРОВЫЕ ВЫЧИСЛИТЕЛЬНЫЕ МАШИНЫ, ПЕРЕКИСЬ ВОДОРОДА);
- термин-словосочетание имеет один или несколько синонимов на уровне словосочетания, или синонимия существует только на уровне словосочетания, а не на уровне отдельных слов, образующих словосочетания (ПОЛУПРОВОДНИКОВЫЕ ТРИОДЫ – ТРАНЗИСТОРЫ);
- термин-словосочетание употребляется только в единственном или множественном числе (например, АВТОМАТИЧЕСКИЙ ПЕРЕВОД, АНГЛИЙСКИЙ ЯЗЫК, СТРОИТЕЛЬНЫЕ МАТЕРИАЛЫ);
- для термина словосочетания существует общепринятая аббревиатура, составленная из первых букв компонентов словосочетания (ЭЛЕКТРОННЫЕ-ЦИФРОВЫЕ ВЫЧИСЛИТЕЛЬНЫЕ МАШИНЫ – ЭЦВМ);
- для некоторых элементов термина-словосочетания мала вероятность использования вне данного словосочетания (например, ОБЗОР ВЕЕРНЫМ ЛУЧОМ, ЭТАЖЕРОЧНЫЕ МИКРОМОДУЛИ);
- один из элементов термина-словосочетания снимает омонимию другого: АВТОМАТЫ: АВТОМАТЫ ДОЗИРОВАНИЯ, АВТОМАТЫ КУРСА;
- словосочетания являются единственным способом уменьшения информационного шума, то есть выдачи документов, не соответствующих запросу. Например, если разложить термины-словосочетания ПРЕОБРАЗОВАТЕЛИ ПОСЛЕДОВАТЕЛЬНОГО КОДА В ПАРАЛЛЕЛЬНЫЙ и ПРЕОБРАЗОВАТЕЛИ ПАРАЛЛЕЛЬНОГО КОДА В ПОСЛЕДОВАТЕЛЬНЫЙ на составляющие их терминологические элементы, то без указателей роли и связи между этими терминоэлементами их невозможно различить.

1.1.3. Аскрипторы

Некоторое понятие может быть выражено с помощью двух разных или большего количества терминов, один из которых выбирается в качестве основного термина – дескриптора. Дескриптор фактически рассматривается как представитель терминов, выражающих такое же или почти такое же понятие, то есть устанавливается отношение эквивалентности между терминами.

Отношение эквивалентности между терминами включает три подтипа:

- собственно синонимы,
- лексические варианты,
- квазисинонимы.

Основными видами синонимов, включаемыми в тезаурусы, являются следующие:

- термины различного происхождения,
- общеупотребительные слова и научные термины,
- общеупотребительные термины и жаргонные или диалектные выражения и др.

Лексические варианты отличаются от синонимов тем, что они представляют собой некоторую модификацию одного и того же выражения, например, различное написание, аббревиатуры, и т.п.

В качестве аскрипторов часто могут использовать квазисинонимы, то есть такие термины, значения которых, вообще говоря, различаются, но которые рассматриваются как эквиваленты для целей тезауруса, например, как квазисинонимы часто рассматриваются антонимы (*ядерная опасность – ядерная безопасность*).

Другим частым видов квазисинонимов является случай, когда в качестве дескриптора рассматривается некий обобщающий тип, а его подвиды описываются как аскрипторы к этому дескриптору.

Аскрипторы, не совпадающие по значению, вводятся по ГОСТу в следующих случаях: относительными синонимами (если случаи несовпадения значений несущественны для задач ИПТ):

СТОЛ = ДИЕТА = ПИТАНИЕ,

БЮРО = КОНТОРА = ФИРМА,

ВИНТ = БОЛТ.

Допускается установление эквивалентности также между единицами, различными по значению, но семантически связанными, в тех случаях, когда отождествление этих понятий полезно для функционирования информационной системы:

УСТОЙЧИВОСТЬ = НЕУСТОЙЧИВОСТЬ,

ТОРГОВЛЯ == ПРОДАЖА,

РЕКА = РУЧЕЙ,

МАСЛО = СМАЗКА.

Например, в тезаурусе LIV Исследовательской службы Конгресса США статья дескриптора Transplantation of organs, tissues ets. (ТРАНСПЛАНТАЦИЯ ОРГАНОВ, ТКАНЕЙ и др.) содержит такие аскрипторы как *medical transplantation, organ transplantation, Skin grafting, Surgical tranplantation, Tissue transplantation*, некоторые из которых соответствуют объемлющему понятию ТРАНПЛАНТАЦИИ, а некоторые представляют видовые понятия (Skin grafting) (LIV, 1994).

В этом же тезаурусе термин *deflation (дефляция*) включено в качестве аскриптора в тезаурусную статью дескриптора *inflation (инфляция*), поскольку разработчики считают, что это разные проявления одного и того же более общего понятия.

Как правило, авторы тезаурусов предпочитают вводить квазисинонимы для понятий, которые рассматриваются как периферийные по отношению к основной области разрабатываемого тезауруса.

1.2. Отношения в информационно-поисковых тезаурусах

ГОСТ 7.25 указывает, что основными типами отношений, обычно отражаемых в информационно-поисковых тезаурусах являются следующие:

- род—вид,
- часть — целое,
- причина — следствие,
- сырье — продукт,
- административная иерархия,
- процесс — объект,
- функциональное сходство,
- процесс — субъект,
- свойство — носитель свойства,
- антонимия.

Такие содержательные типы связей между дескрипторами, чаще всего, не отражаются в подробном перечне отношений тезауруса, а записываются с помощью небольшого набора отношений, которые обычно разделяются на два класса: иерархические и ассоциативные. Иерархические отношения обычно рассматриваются как несимметричные и транзитивные.

1.2.1. Иерархические отношения

Иерархические отношения в тезаурусе могут использоваться в трех логически различных и взаимно исключающих ситуациях, а именно для установления следующих отношений:

- родовидовое отношение
- отношение часть-целое
- отношение пример-класс.

Американский стандарт на разработку тезаурусов (Z39.19) приводит общую рекомендацию для установления иерархических отношений:

> *каждый вышестоящий дескриптор должен относиться к тому же базисному семантическому типу, что и нижестоящий дескриптор, например, оба (нижестоящий и вышестоящий) дескрипторы могут обозначать предмет, действие, свойство и т.п.*

Например, *АНАТОМИЯ (ДИСЦИПЛИНА)* и *ЦЕНТРАЛЬНАЯ НЕРВНАЯ СИСТЕМА* относятся к разным типам понятий, поэтому они не могут быть соединены иерархическими отношениями.

Дескрипторы *ЦЕНТРАЛЬНАЯ НЕРВНАЯ СИСТЕМА* и *МОЗГ* относятся к органам живого организма и поэтому могут быть соединены иерархически**.**

Некоторые авторы обсуждают необходимость ограничения иерархических уровней в тезаурусе. Так, Герд А.С. (Герд, 2005), указывает, что практический опыт показывает, что иерархическая глубина тезауруса не должна превышать некоторого порога, иначе он будет громоздким и неудобным в эксплуатации. Авторы работы (Методика, 1973) также подчеркивают, что не рекомендуется использовать более 9 уровней иерархии.

Ограничение числа уровней иерархии достигается исключением слишком конкретных для данной предметной области дескрипторов. Считается, что для отражения таких конкретных дескрипторов индексатор может выбрать и более общий дескриптор.

Кроме того, учитывается еще и фактор субъективности: чем больше уровней иерархии в тезаурусе, тем больше вероятность, что для отражения одного и того же содержания индексаторы могут выбрать дескрипторы с разных уровней иерархии.

1.2.1.1. Отношение Выше-Ниже

Многие руководства и стандарты (Z39.19; Will, 2004) подчеркивают, что иерархические отношения в информационно-поисковых тезаурусах должны устанавливаться в тех случаях, когда отношения истинны независимо от контекста, - только в таких случаях дескрипторы информационно-поискового тезауруса могут быть организованы в иерархии. Эта рекомендация связана с тем, что обычно в информационном поиске очень трудно четко определить контекст употребления термина и понять, применимо ли в данном контексте то или иное отношение.

Так, в (Will 2004) указывается, что для мышей можно указать, что они грызуны, поскольку это внутренняя характеристика мышей. В то же время неправильно указывать, что мыши – вредители, поскольку имеются лабораторные мыши и домашние мыши, которые не являются вредителями.

Американский стандарт на информационно-поисковые тезаурусы (Z39.19) предлагает при описании родовидовых отношений использовать тест «все-некоторые». Например, *все мыши являются грызунами, но некоторые мыши являются вредителями.*

Шемакин Ю.И. (Шемакин, 1974) также подчеркивает, что одна из наиболее распространенных ошибок при построении классификационных схем заключается в том, что ассоциативная связь между понятиями, основанная на возможном применении или использовании кого-то одного свойства, принимается за родовидовую связь. Так, например, ошибкой является, если в тезаурусе понятие НИТРОГЛИЦЕРИН связано

родовидовой связью с понятием ВЗРЫВЧАТЫЕ ВЕЩЕСТВА, хотя как химическое соединение оно находит применение и в других областях (например, в медицине). Родовая связь между понятиями в аналогичных ситуациях сохраняется лишь тогда, когда данный предмет (процесс) имеет только одно применение (например, ТРОТИЛ применяется только как взрывчатое вещество и поэтому может быть связан родовидовой связью только с БРИЗАНТНЫМИ ВЗРЫВЧАТЫМИ ВЕЩЕСТВАМИ).

1.2.1.2. Отношение Часть-Целое

Отношение Часть-Целое относится к иерархическим отношениям тезауруса. Это отношение используется в информационно-поисковых тезаурусах значительно реже, чем родовидовое отношение ВЫШЕ-НИЖЕ. В конкретных тезаурусах часто принимается решение описывать отношение ЧАСТЬ-ЦЕЛОЕ как обобщенное отношение ВЫШЕ-НИЖЕ (Мдивани, 2004), или как отношение АССОЦИАЦИИ (Методика, 1973) (см. раздел 1.2.2).

Американский стандарт z39.19 подчеркивает, что отношение ЧАСТЬ-ЦЕЛОЕ в тезаурусах должно устанавливаться в тех случаях, когда одно понятие включено в другое понятие независимо от контекста, тогда дескрипторы могут быть организованы в иерархии. Обычно приводится следующий список независимых от контекста отношений ЧАСТЬ-ЦЕЛОЕ, впрочем, список не считается исчерпывающим:

1) системы и органы тела:
нервная система
центральная нервная система
мозг

2) географические объекты:
Россия
Ростовская область
Ростов-на-Дону

3) дисциплины и сферы деятельности:
наука
биология
ботаника
зоология

4) иерархически организованные общественные, политические военные структуры:
батальон
рота
взвод

В тех случаях, когда имеется множественная принадлежность части к целому, то между такими терминами не должно устанавливаться иерархическое отношение. Между такими дескрипторами может быть установлено отношение ассоциации. Например, карбюраторы являются частями не только автомобилей. Поэтому дескрипторы *КАРБЮРАТОР* и *АВТОМОБИЛЬ* не должны быть связаны отношением ЧАСТЬ-ЦЕЛОЕ в тезаурусе.

Но даже так строго ограничиваемое установление отношений ЧАСТЬ-ЦЕЛОЕ может иметь проблемы. Так, в работе (Smith и др., 2004) указывается на проблемы в установлении отношений ЧАСТЬ-ЦЕЛОЕ в онтологии генов GO, которая, по сути, является информационно-поисковым тезаурусом.

Авторы указывают на три разных интерпретации отношения ЧАСТЬ_ЦЕЛОЕ в этом ресурсе:

1) *А является частью В* означает, что А иногда является частью В в том смысле, что каждый пример А в некоторый момент своего существования становится частью В, понимаемой как часть-целое между конкретными сущностями, то есть в некоторые момент своего существования А является частью В, а в другие не является.
2) *А является частью В* означает, что А может быть частью В в смысле независимого от времени отношения между классами: класс А является частью класса В, если существует класс С, являющийся подклассом В, для которого все примеры А являются частями С и все примеры С содержат в качестве частей примеры А.
3) *А является частью В* означает, что словарь *А* включен в словарь *В, например*, онтология клеточных компонентов включается в онтологию генов.

1.2.1.3. Обобщенные отношения ВЫШЕ-НИЖЕ

Несмотря на то, что современный стандарт на разработку тезаурусов Z39.19 рекомендует описывать иерархические отношения так, чтобы семантические классы вышестоящего понятия и нижестоящего понятия совпадали, на практике разработчики тезаурусов часто использовали обобщенное отношение ВЫШЕ-НИЖЕ, нарушающее это требование. Например, в тезаурусе EUROVOC (EUROVOC, 2001) можно найти следующие примеры отношений ВЫШЕ-НИЖЕ, в которых вышестоящее понятие представляет собой сферу деятельности или процесс, а нижестоящее понятие имеет другой семантический тип.

АВИАЛИНИИ
ВЫШЕ *ВОЗДУШНЫЙ ТРАНСПОРТ*

АВТОСТОЯНКИ
ВЫШЕ *КОММУНАЛЬНОЕ ХОЗЯЙСТВО*

ЗЕМЛИ ПОД ПАРОМ
ВЫШЕ *АГРОТЕХНИКА*

АДМИНИСТРАТИВНАЯ ОТВЕТСТВЕННОСТЬ
ВЫШЕ *АДМИНИСТРАТИВНОЕ ПРАВО*

ОБЪЕКТЫ АКВАКУЛЬТУРЫ
ВЫШЕ *АКВАКУЛЬТУРА*

БАЗЫ ДАННЫХ
ВЫШЕ *ОБРАБОТКА ДАННЫХ.*

Также в тезаурусах в качестве обобщенного отношения ВЫШЕ-НИЖЕ может быть представлены отношения ЧАСТЬ-ЦЕЛОЕ, например, в тезаурусе AGROVOC находим следующие пример такого отношения:

МОЛОЧНЫЙ ЖИР
ВЫШЕ *МОЛОКО.*

1.2.2. Отношения ассоциации

Основным назначением установления ассоциативных отношений между дескрипторами информационно-поискового тезауруса является то, что установление такой связи может указать дополнительные дескрипторы, полезные при индексировании или поиске.

Отношение ассоциации является неиерархическим. Ассоциативное отношение наиболее трудно определить. Российский стандарт на создание информационно-

поисковых тезаурусов указывает, что «ассоциативное отношение является объединением отношений, не входящих в иерархические отношения или в отношения синонимии. Допускается включать в ассоциативное отношение все виды отношений, кроме синонимии и отношения род — вид» (ГОСТ 7.25-2001).

Другие источники стараются изложить более подробные принципы установления ассоциативных отношений, поскольку в противном случае отношение будет проставляться непоследовательно (Aitchinson, Gilchrist, 1987).

Американский стандарт описывает наиболее общее правило установления ассоциативного отношения между дескрипторами таким образом, что это отношение стоит устанавливать между двумя дескрипторами, если при употреблении одного термина другой термин как бы подразумевается. Более того, один термин часто есть необходимый элемент определения другого термина, например, термин *клетка* составляет необходимую часть определения термина *цитология*.

Более конкретно типы ситуаций, в которых необходимо установить ассоциативные отношения по версии Американского стандарта, могут быть следующими.

Если дескрипторы принадлежат одной иерархии, то ассоциативные отношения устанавливаются в следующих случаях:

- между видами одного и того же понятия, когда их значения пересекаются, например, английские слова ship и boat, которые не являются эквивалентными, но в то же время во многих контекстах являются взаимозаменимыми.
- между понятиями, одно из которых происходит от другого, например, поскольку известно, что мул – это помесь осла и лошади, то ассоциативное отношение должно быть установлено между дескрипторами мул – осел и мул – лошадь.

Случаи, в которых необходимо установить отношения, между дескрипторами, принадлежащими разным иерархиям, являются достаточно разнообразными:

1) научная дисциплина – объект изучения или специалист в этой дисциплине:
математика – математик
неврология - нервная система

2) операции или процессы и их агент или инструмент:
контроль температуры – термостат
охотник – охота

3) объекты или процессы и их контрагенты:
растения – гербициды

4) действия и их продукты:
ткачество – ткань
слезоотделение – слеза

5) действия и их цели:
переплетное дело – книга

6) объекты и вещества и их свойства (уникальные свойства – unique):
яды – токсичность
жидкость – поверхностное натяжение

7) понятия, связанные причинно-следственной связью:
смерть – оплакивание

8) понятия и единицы их измерения
электрический ток - ампер

Авторы конкретных тезаурусов могут вводить свои правила описания ассоциативных отношений. Так, в тезаурусе EUROVOC ассоциативные отношения

устанавливаются в случаях, когда между дескрипторами существуют следующие отношения (EUROVOC, 2001):

- причина;
- инструмент;
- иерархические отношения, когда полииерархия возможна, но запрещена и поэтому заменяется на ассоциацию;
- отношения сопутствия, следования во времени или пространстве;
- материал;
- свойство, особенность;
- предмет действия, процесса, отрасли знаний;
- локализация
- сходство, подобие
- антонимия.

При такой расплывчатости отношения АССОЦИАЦИИ возникает вопрос, на какие источники можно опереться при описании этих отношений. При создании тезауруса конкретной предметной области может использоваться несколько различных источников ассоциативных отношений.

Во-первых, используются тексты данной предметной области. Анализ таких текстов позволяет вскрыть реальные типы смысловых отношений, характерных для данной предметной области. При таком подходе ассоциативные отношения, выделяемые в информационно-поисковом тезаурусе, будут соответствовать смысловым отношениям, существующим в тексте (Герд, 2005).

По текстам также может изучаться совместная встречаемость различных терминов в конкретных текстах, что не является достаточно надежным критерием установления правильных ассоциативных отношений (Мдивани, 2004).

Второй путь – это установление ассоциативных отношений через обращение к энциклопедиям, терминологическим словарям, справочникам для логического анализа определений терминов. Вместе с тем в случаях широких предметных областей, неустоявшихся терминологий, определения терминов могут значительно отличаться, отражать разные особенности концепций авторов словарей, что может привести к появлению ложных ассоциативных связей в тезаурусе.

Серьезной проблемой является также и то, что словарная статья термина в словаре, энциклопедии может упоминать достаточно много других терминов. Возникает вопрос, какие из них нужно ассоциировать с заглавным термином.

Третий путь – обращение к специалистам, которые могут дать обоснованную оценку отношениям между терминами.

Наконец, на основе всех источников может быть создан четкий перечень наиболее важных смысловых отношений данной предметной области. Как ассоциативные отношения могут рассматриваться лишь те отношения, которые соответствуют составленному списку.

В целом, можно отметить, что установление отношения АССОЦИАЦИИ, несмотря на все попытки ограничить установление этого отношения, являются наиболее субъективными (Мдивани, 2004), часто встречается искусственное и надуманное конструирование ассоциативных отношений (Герд, 2005). Особые проблемы установления ассоциативных отношений возникают при создании тезауруса для больших, гетерогенных областей, не позволяющих четко зафиксировать типы смысловых отношений, соответствующих ассоциативным отношениям тезауруса.

Из приведенных примеров также очевидно, что хотя отношения ассоциации рассматриваются как симметричные, по сути, многие типы упомянутых отношений явно не симметричны, по крайней мере, в тех случаях, когда в качестве определяемого термина служит один из этих терминов.

1.3. Основные принципы разработки тезаурусов

При разработке информационно-поисковых тезаурусов первой задачей является отбор терминов для включения в тезаурус. Существует несколько возможных источников терминов для разработки информационно-поисковых тезаурусов.

Прежде всего, должны быть изучены существующие тезаурусы в близких предметных областях. Они могут содержать значимое количество полезных терминов для нового тезауруса.

Термины - кандидаты для внесения в тезаурус могут быть предложены экспертами предметной области.

Кроме того, термины тезауруса могут быть получены из текстов предметной области применением автоматизированных методов или ручной обработки документов. При ручной обработке документов сначала некоторое время индексаторы индексируют поступающие документы наиболее релевантными ключевыми словами, которые затем сводятся в единый список, которые и может служить основой для тезауруса (Архангельская, Базарнова, 2001; Z39.19).

После того, как список терминов-кандидатов получен, из него исключаются слишком частотные термины, поскольку предполагается, что они являются малоинформативными для различения отдельных документов. Относительно малочастотные термины могут быть удалены из списка или представлены как аскрипторы более общих или более частотных понятий.

Слишком конкретные термины также могут быть исключены из списка терминов-кандидатов, поскольку считается, что если тезаурус содержит слишком много уровней иерархии, то им трудно управлять, возрастает субъективность индексирования, так как индексаторы могут использовать для индексирования документов дескрипторы разного уровня (Z39.19; Герд, 2005).

Если в списке обнаруживается несколько близких по смыслу терминов, то из них выделяется наиболее представительный термин, остальные термины могут быть частично исключены и переведены в аскрипторы (Архангельская, Базарнова, 2001).

Шемакин Ю.И. (Шемакин, 1974, стр. 41) подчеркивает, что из синонимических рядов тезауруса исключаются явные синонимичные термины, например, РАЗВЕДЫВАТЕЛЬНАЯ ИНФОРМАЦИЯ – РАЗВЕДЫВАТЕЛЬНЫЕ СВЕДЕНИЯ, НАВИГАЦИОННЫЕ СИСТЕМЫ – СИСТЕМЫ НАВИГАЦИИ. Такая рекомендация связана с тем, что эта информация очевидна для человека-индексатора, большое количество таких синонимических терминов в тезаурусе может затруднять работу человека-индексатора.

Разработчики тезауруса LIV Исследовательской службы Конгресса США (LIV, 1994) описывают правила включения терминов в тезаурус следующим образом:

- термины тезауруса должны представлять понятия, которые реально упоминаются в литературе, и должны отбираться из соображений эффективности их использования в поиске документов;
- важным фактором включения термина является частотность его упоминания в текстах, которую необходимо периодически проверять;
- включение новых терминов в тезаурус должно происходить с учетом уже включенных тезаурусных терминов. Термины-кандидаты должны проверяться на предмет соответствия их общности/специфичности к другим терминам тезауруса. Также должно проверяться, представляет ли термин-кандидат отдельное понятие, которому нет соответствий среди существующих терминов тезауруса. Необходимо избегать включения терминов, чьи значения пересекаются со значениями уже существующих тезаурусных терминов настолько, что индексаторам и пользователям будет трудно различать между ними и др.

Таким образом, разработка хорошего информационно-поискового тезауруса представляет собой достаточно сложный, многоэтапный процесс, в котором необходимо найти «золотую середину». С одной стороны, набор дескрипторов тезауруса должно быть достаточен для описания произвольного документа предметной области, с другой стороны, дескрипторов не должно быть слишком много, поскольку слишком большая величина тезауруса повышает субъективность индексирования и затрудняет развитие и использование тезауруса.

Не случайно, значительная доля информационно-поисковых тезаурусов в самых широких областях включает не более 10 тысяч терминов и 6-7 тысяч дескрипторов. Широко известным исключением являются Тезаурус по архитектуре и искусству (Тезаурус AAT), содержащий более 30 тысяч дескрипторов, что, видимо, связано со спецификой соответствующей предметной области, когда нужно индексировать не столько документы, сколько конкретные музейные предметы.

Другим известным исключением, сверхбольшим тезаурусом является тезаурус по медицине MeSH, что связано с гетерогенностью области медицины, состоящей из множества подобластей с собственной терминологией.

1.4. Конкретные тезаурусы

Рассмотрим принципы устройства и функционирования некоторых известных информационно-поисковых тезаурусов. Специфика предметной области каждого тезауруса находит отражение в его структуре.

1.4.1 Тезаурус Европейского союза EUROVOC

Многоязычный тезаурус EUROVOC разработан специально для содержательной обработки и поиска документов по всем направлениям деятельности институтов ЕС. Последнее третье издание тезауруса на девяти языках было опубликовано в 1995 году. Тезаурус составлен в соответствии с международными стандартами ИСО 2788-1986 и ИСО 5964-1985 и имеет стандартную структуру информационно-поискового тезауруса, предназначенного для ручного индексирования:

- термины тезауруса разделены на дескрипторы, которые используются для индексирования документов и аскрипторы (условные синонимы), которые входят в классы условной эквивалентности дескрипторов;
- установлены иерархические отношения между дескрипторами (объединяют отношения «род-вид» и «часть целое»);
- установлены ассоциативные отношения между дескрипторами;
- дескрипторы объединены в более широкие тематические классы, называемые микротезаурусами.

В 2001 году Парламентская библиотека Российской Федерации подготовила русскую версию тезауруса EuroVoc, которая содержит переводы всех дескрипторов тезауруса EUROVOC, а также более 5 тысяч понятий, отражающих российскую специфику (EUROVOC, 2001).

1.4.2. Тезаурус исследовательской службы Конгресса США

Тезаурус Legislative Indexing Vocabulary (далее тезаурус LIV) используется для индексирования и поиска законов, законопроектов, политической литературы в исследовательской службе Конгресса США (LIV, 1994).

Разработка тезауруса была начата в 1967 году. Последняя версия тезауруса LIV была подготовлена в 1995 году и включает более 10 тысяч терминов, среди которых около 5 тысяч дескрипторов.

Тезаурус включает термины из широкой области общественной жизни, включая как социальные науки, так и социальные аспекты естественных и прикладных наук. Большое количество предметных областей исследований, проводимых Исследовательской службой, отражаются в смешении в тезаурусе разных типов терминологии - более общей и более конкретной, широко употребляемой и относящейся к более узким предметным областям.

Дескрипторы тезауруса разбиты на 80 тематических областей, называемых top terms (термины верхнего уровня).

1.4.3. Тезаурус ООН UNBIS

Многоязычный Тезаурус UNBIS (UNBIS, 1976), созданный Библиотекой им. Дага Хаммаршельда Департамента общественной информации, содержит терминологию, используемую в качестве дескрипторов при анализе документов и других материалов, относящихся к программам и деятельности ООН. Он используется в качестве списка предметных рубрик Библиографическо-информационной системы ООН (ЮНБИС) и включен в список тематических терминов Системы официальной документации. Будучи многоотраслевым, тезаурус отражает широкий круг вопросов, которыми занимается ООН; термины тезауруса предназначены для их точного и четкого обозначения с учетом специфики предмета.

В настоящее время тезаурус существует на всех официальных языках Организации Объединенных Наций: арабском, китайском, английском, французском, русском и испанском.

1.4.4. Тезаурус по архитектуре и искусству (Art and Architecture Thesaurus)

Тезаурус по искусству и архитектуре (Тезаурус AAT) создается фондом Пола Гетти (www.getty.edu), содержит 34 тысячи понятий (дескрипторов) и 131 тысячу терминов по искусству, архитектуре, архивным материалам и материальной культуре от античности до наших дней.

Дескрипторы тезауруса подразделяются на 7 фасетов: ассоциированные понятия, физические свойства, стили и периоды, АГЕНТЫ (люди и организации), ДЕЯТЕЛЬНОСТЬ, МАТЕРИАЛЫ, ОБЪЕКТЫ (Art and Architecture Thesaurus, 1994).

Каждый фасет подразделяется на иерархии. Всего насчитывается 33 иерархии.

Таким образом, тезаурус AAT отличается, с одной стороны, значительно большей величиной, с другой стороны, более строгой организацией в иерархии. Количество уровней в иерархиях также значительно больше, чем в ранее упомянутых тезаурусах. На наш взгляд, это объясняется тем, что основным назначением тезауруса является индексирование не только документов по культуре и искусству, но и собственно музейных объектов, что требует большого количества конкретных сущностей. При этом такая направленность тезауруса носит более структурированный характер, чем широкие области ранее упомянутых тезаурусов.

Каждому дескриптору может соответствовать несколько терминов (аскрипторов), которые включают термины в различных грамматических числах, термины в инвертированном порядке, варианты написания термина, а также синонимы различного происхождения. Наличие большого количества морфологических форм терминов не менее, чем в два раза завышает число терминов, описанных в тезаурусе.

Дескрипторы тезауруса снабжены стандартными для тезаурусов отношениями ВЫШЕ-НИЖЕ и АССОЦИАЦИЯ.

Основное внимание разработчики уделили установлению для каждого дескриптора отношения ВЫШЕ. Для некоторых дескрипторов описаны два отношения ВЫШЕ, одно из которых считается основным, другое вспомогательным.

При описании отношений НИЖЕ, если имеется несколько оснований классификации, то под каждое основание классификации заводится отдельный дескриптор. Например, =вместилища= делятся по форме (=мешки=, =бочки=, =ящики= и т.п.), =вместилища по функции= (=вместилища для церемоний=, =вместилища. для денег=, =вместилища для тканей= и т.п.), =вместилища по расположению= (=седельные сумки=, =настенные сумки= и др.). Отношения АССОЦИАЦИЯ занимают относительно небольшой процент всего набора отношений тезауруса AAT.

Авторы тезауруса считают, что наиболее полное покрытие тезаурус обеспечивает для искусства Западной Европы и Америки.

Пример словарной статьи:

athletic shoes (спортивные ботинки)
Note: Shoes designed to be worn for sports (ботинки, предназначенные для спорта)
Terms:
athletic shoes
athletic shoe
trainers (athletic shoes)
trainer
shoes, athletic
shoes, sport
shoes, training
sport shoes
training shoes

Hierarchical position (позиция в иерархии):
<shoes by function>, (ботинки по функции)
shoes (footwear), (ботинки)
<footwear by form>, (обувь по форме)
<accessories worn on the legs or feet>, (аксессуары, носимые на ногах)
<costume accessories worn>, (носимые аксессуары костюма)
<costume accessories>, (аксессуары костюма)
costume, (костюм)
Furnishings and equipment,
Objects facet.

В последнее время разработчики тезауруса вместо обобщенного отношения ассоциации стали использовать конкретные виды семантических отношений, например, понятие embroidery (visual works) **(**вышивка (продукт труда)) связано с понятием embroidering (вышивание) отношением «activity/event producing is»:

embroidery (visual works)
activity/event producing is embroidering

Всего предлагается использовать около 40 различных семантических подвидов отношения ассоциации.

1.4.5. Тезаурус в области медицины MeSH

Тезаурус MeSH развивается Национальной медицинской библиотекой США для индексирования и поиска документов в медико-биологической сфере (Medical Subject Headings, 1992). В настоящее время (2009) год тезаурус содержит более 25 тысяч

дескрипторов. Дескрипторы снабжены толкованиями и списком синонимов или близких по смыслу терминов (entry terms).

Отношения между дескрипторами могут быть иерархическими – такие отношения представлены в виде иерархических деревьев – и ассоциативными. Расположение в дереве размечается посредством специальных меток – номеров в дереве, и каждый дескриптор может входить в несколько деревьев, то есть ему сопоставлено несколько таких номеров. На рис. 1.1. показана словарная статья дескриптора ВОСПАЛЕНИЕ ЛЕГКИХ. Данный дескриптор относится к двум иерархическим деревьям – одно дерево ЛЕГОЧНЫЕ ЗАБОЛЕВАНИЯ, второе – РЕСПИРАТОРНЫЕ ИНФЕКЦИИ.

National Library of Medicine - Medical Subject Headings

2007 MeSH

MeSH Descriptor Data

Return to Entry Page

Standard View. Go to Concept View; Go to Expanded Concept View

MeSH Heading	Pneumonia
Tree Number	C08.381.677
Tree Number	C08.730.610
Annotation	GEN or unspecified; prefer specifics
Scope Note	Inflammation of the lungs.
Entry Term	Experimental Lung Inflammation
Entry Term	Lung Inflammation
Entry Term	Pneumonitis
Entry Term	Pulmonary Inflammation
Allowable Qualifiers	BL CF CI CL CN CO DH DI DT EC EH EM EN EP ET GE HI IM ME MI MO NU PA PC PP PS PX RA RH RI RT SU TH UR US VE VI
Date of Entry	19990101
Unique ID	D011014

Рис.1.1. Словарная статья тезауруса MeSH

Номера дескриптора в иерархических деревьях могут изменяться с развитием тезауруса, при этом каждый дескриптор имеет уникальный идентификационный номер, который остается неизменным в течение всего времени существования дескриптора.

На верхнем уровне иерархии тезауруса находится 16 дескрипторов: АНАТОМИЯ, ОРГАНИЗМЫ, БОЛЕЗНИ, ХИМИЧЕСКИЕ ВЕЩЕСТВА И ЛЕКАРСТВА, МЕТОДЫ И ОБОРУДОВАНИЕ, ПСИХИАТРИЯ И ПСИХОПАТОЛОГИЯ ... ПЕРСОНЫ, ГЕОГРАФИЯ и др.

В тезаурусе также имеется стандартный набор квалификаторов (allowable qualifiers), которые могут быть добавлены к дескриптору для сужения тематики, например, BS – кровоснабжение, PP – физиопатология, MI – Микробиология и др. Такие квалификаторы особенно важны в многотематических документах, в которых упоминается много понятий в разных аспектах, тогда посредством квалификаторов можно точнее определить, какое понятие с какой точки зрения рассматривается.

Таким образом, в тезаурусе MESH, как и тезаурусе AAT достаточно большой объем дескрипторов сочетается со значительной структурированностью и иерархичностью, что, несомненно, также связано с особенностями предметной области тезауруса.

1.5. Правила индексирования документов дескрипторами информационно-поискового тезауруса

Рассмотрим особенности применения информационно-поисковых тезаурусов для ручного индексирования документов экспертами-индексаторами.

Правила индексирования документов регулируются несколькими ГОСТами (ГОСТ 7.66-92; ГОСТ 7.59-2003). Приведем некоторые нормативные положения, регулирующие процесс ручного индексирования

Под индексированием понимается выражение содержания документа и/или смысла информационного запроса на информационно-поисковом языке. Для обеспечения эффективного информационного поиска основное содержание документа (а при необходимости - его форму и назначение) следует представлять с необходимой и достаточной полнотой и точностью в поисковом образе документа (ПОД) в виде терминов индексирования.

Индексирование следует проводить на основе непосредственного анализа документа с учетом характера информационно-поискового массива, элементом которого становится ПОД, характера информационных потребностей пользователей данной информационно-поисковой системы (ИПС), в соответствии с общими принципами индексирования и особенностями их применения в конкретной организации.

Одним из основных методов индексирования является так называемое координатное индексирование, то есть индексирование, цель которого состоит во всестороннем отражении содержания документа или запроса путем включения в поисковый образ всех необходимых для этого терминов индексирования.

Метод координатного индексирования базируется на представлении о том, что основное смысловое содержание документа может быть с достаточной степенью точности и полноты выражено набором ключевых слов, содержащихся в индексируемом тексте.

Координатное индексирование может быть свободным или нормализованным (контролируемым). Свободное координатное индексирование означает индексирование ключевыми словами, выбранными непосредственно из полного текста документа и представленными в ПОДе в терминологии автора без нормализации, с минимальным контролем над лексикой и без учета того, какие ключевые слова уже использовались ранее для индексирования таких же или близких по смыслу документов.

При нормализованном индексировании поисковый образ документов составляется из дескрипторов нормализованного списка – тезауруса.

Процесс нормализованного индексирования включает следующие этапы, которые осуществляют в указанной ниже последовательности:

- анализ и определение содержания документа, как объекта индексирования;
- выбор понятий, характеризующих основное содержание документа;
- выбор терминов индексирования для обозначения понятий;
- формирование поискового образа документа из терминов индексирования.

Число характеристик и понятий, отраженных в ПОД, определяет его полноту и является важнейшим показателем качества индексирования. В ПОДе необходимо отразить все понятия, которые могут иметь ценность для пользователей системы. В документе может быть выявлено более одной темы из сферы интересов пользователей.

Число терминов индексирования, приписываемых одному документу, определяется количеством сведений, содержащихся в документе. Ограничение числа терминов должно быть основано на содержательном отборе наиболее важных понятий.

Полнота индексирования, принятая в каждой информационно-поисковой системе, определяется ее функциональным назначением. Объем документа также сильно влияет на полноту индексирования. Необходимо учитывать указанные факторы и на их основе производить экспертный отбор понятий из документа, не стремясь включить в ПОД все упомянутые в нем понятия.

Поскольку понятия, упоминаемые в документе, могут быть разной значимости относительно его основного содержания, в ГОСТе 7.66-92 обсуждаются возможности проставления весов для дескрипторов индексатором: информационный вес термина индексирования отражает в ПОД важность данного понятия для данного документа. Число градаций информационного веса определяется потребностями конкретной поисковой системы. Важными категориями дескрипторов в документе, которые следует различать, являются:

- понятия, выражающие главную тему документа;
- понятия, выражающие побочные темы документа;
- понятия, использованные в документе как вспомогательные для изложения его содержания.

Допускается использовать указатель отрицательного веса, которым помечают термины индексирования для указания на то, что данное понятие не рассматривается в документе.

В качестве примера инструкции, регулирующей индексирование по конкретному информационно-поисковому тезаурусу, рассмотрим положения, принятые в информационной службе ООН, в которой для индексирования используется тезаурус UNBIS (см. раздел. 1.4.3) (UNBIS Guidelines, 2009).

В документах службы отмечается, что для определения основного содержания документа не является достаточным просматривание только заголовков документов. Нужно дополнительно обращать внимание на заголовки подразделов, на рефераты, содержание, названия глав, введение и заключение, приложения.

Индексаторы должны выбирать понятия, которые наилучшим образом выражают основное содержание текста. В дополнение к определению основных тем документа, процесс индексирования включает определение подтем, которые могут быть полезны в поиске специальной информации или для уточнения каких-либо аспектов основных тем документа. При этом индексатор должен учитывать интересы потенциальных клиентов и запросы, которые они могут задать. Индексаторы должны задавать следующие вопросы:

- Какие понятия документа могут быть интересны пользователям информационной системы?
- Какие термины индексирования и их комбинации лучше всего отвечают основным направлениям поиска?

Индексаторы должны учитывать, что в тексте могут быть просто упоминаемые сущности или примеры и не индексировать такие сущности. Для этого полезно задавать себе следующие вопросы:

- Является ли эта сущностей темой документа, или это простое упоминание?
- Найдет ли пользователь, ищущий по этой теме достаточно информации в тексте, чтобы оправдать выбор этого понятия как темы текста?

При переводе сформулированной темы на язык дескрипторов тезауруса индексаторы должны выбрать наиболее соответствующий и наиболее специфичный дескриптор тезауруса. При этом индексатор должен осознавать, что слова документа могут отличаться от терминов тезауруса. Например, документ может обсуждать проблемы коренных народов Америки (*indigenous peoples of the Americas*), но наиболее подходящий термин тезауруса будет AMERINDIANS, а не INDIGENOUS PEOPLES. Документ может относиться к статистическим данным, но дескриптор STATISTICAL DATA тезауруса UNBIS используется только, когда документ действительно использует статистические данные, иначе используется дескриптор STATISTICS. Документ, обсуждающий нефть OIL, может использовать термин OIL INDUSTRY, но индексатор должен знать, что в тезаурусе UNBIS дескриптор OIL INDUSTRIES относится только к промышленности по извлечению масла из растений, а для переработки нефти нужно использовать дескриптор PETROLEUM INDUSTRY. Индексаторы должны проверять комментарии к дескрипторам, чтобы удостовериться, что они проиндексировали текст правильно.

Документ информационной службы ООН обращает внимание на сложность индексирования больших документов, в которых главная тема документа развивается большим количеством более специфичных тем, которые также хотелось бы отразить при индексировании документа. С 1999 года индексаторы могут применять ранжированное индексирование, присваивая ранг 1 понятиям основной темы документа и величину 2 вторичным сущностям. Дескрипторов 1-го уровня обычно не более 5.

Дескрипторы 2-го уровня коррелируют с фактором полноты индексирования, обычно повышают полноту поиска, отражают большую специфичность и показывают несколько аспектов основной темы.

Таким образом, мы видим, что ручное индексирование документов по информационно-поисковому тезаурусу является сложной процедурой, требующей очень хорошего знания структуры и состава тезауруса.

Серьезной проблемой ручного индексирования является также субъективность, непоследовательность индексирования: один индексатор может поставить в соответствие тексту дескриптор более низкого уровня, другой - дескриптор более высокого уровня.

Кроме того, определенную сложность представляет собой последовательный учет тематической структуры связного текста: один индексатор может счесть обсуждаемое в каком-то фрагменте текста важным и отразить в приписываемых ключевых словах или дескрипторах тезауруса, другой индексатор для того же или похожего текста посчитает эту «локальную тему» неважной и не отразит ее в терминах индексирования.

В результате исследований, проходивших в рамках известного Крэнфилдского эксперимента в начале 60-х годов, было показано, что значимый процент неудач поиска связан с неправильным индексированием документа, что до трети неудач поиска можно было бы избежать, если бы индексаторы индексировали последовательным образом. Точнее, индексаторы допускали ошибку в каждых пяти документах их ста, и эта ошибка обычно состояла в неуказании релевантного понятия (Tonta, 1992), то есть полнота индексирования была недостаточной.

Кроме того, данные других экспериментов по анализу неудач в информационном поиске в 60-70-х годах обнаружили, что у неподготовленного пользователя имеются проблемы с использованием нормализованных словарей (тезаурусов) и языков запросов, что приводит к большому количеству неудач поиска. Большинство пользователей не знали роли нормализованных словарей в информационных системах, не понимали структуру нормализованного индексирования и языков индексирования. Пользователи пытались выразить свои запросы собственными словами, которые не совпадали с приписанными документу дескрипторами, что и вызывало неудачи поиска.

Взаимодействие всех этих факторов приводит к тому, что серьезные усилия по разработке и ведению информационно-поисковых тезаурусов, обеспечению качественного ручного индексирования не привели к лучшим показателям информационного поиска по сравнению с поиском по словам (Salton, 1986; Sparck Jones, 1981). Вместе с тем, как мы увидим в разделе 1.6.3., использование комбинированных технологий, сочетающих пословный поиск и поисковые образы документов, содержащих дескрипторы тезауруса, приводит к значительному улучшению качества поиска.

1.6. Информационно-поисковые тезаурусы в приложениях автоматической обработки документов

1.6.1. Автоматическое индексирование по информационно-поисковым тезаурусам.

Поскольку основными элементами информационно-поискового тезауруса являются термины предметной области, описанные как дескрипторы и аскрипторы, то может показаться, что сопоставление информационно-поискового тезауруса и документа

осуществить достаточно просто путем непосредственного сопоставления единиц тезауруса с документами.

Однако для большинства документов такое автоматическое сопоставление не сможет отразить основное содержание документа:

- важные термины документа могут быть не найдены в тезаурусе, поскольку выражены в нем несколько иначе,
- менее значимые термины найдут прямое отражение в тезаурусе и выйдут на первый план и т.п.

В работе **(**Pouliquen и др., 2003) приводятся данные, полученные на основе 587 документов, проиндексированных вручную дескрипторами тезауруса EUROVOC. Только 31 процент документов явно содержит в тексте дескрипторы, приписанные документу индексаторами. При этом в 9 из 10 случаев дескрипторы, найденные в тексте документа, не приписаны индексаторами.

Поэтому исследуются более сложные методы автоматизации индексирования по информационно-поисковым тезаурусам.

В работе (Hlava, Heinebach, 1996) излагается подход к автоматическому индексированию по тезаурусу EUROVOC, основанному на правилах. Правила могут быть простыми и сложными. Простые правила не содержат условий. Сложные правила содержат такие условия как Близость (на расстоянии трех слов по тексту, в одном предложении, в том же самом поле, например, поле реферата), Местонахождение (в заголовке, в тексте реферата или документа, начало предложения, конец предложения), Формат (с большой буквы, все большими буквами).

Примеры сложных правил:

IF (near "Technology" AND with "Development")
USE Community programme
USE development aid
ENDIF

IF (near "Technology" AND with "Regional Innovation" AND with "Development")
USE Community programme
USE common regional policy
USE technology transfer
ENDIF

Основная процедура создания сложных правил заключается в следующем:

- создается множество простых правил, заключающихся в представлении дескрипторов и синонимов тезауруса EUROVOC в виде текстовых строк,
- на основе простых правил обрабатываются документы Европейского парламента и автоматически полученные дескрипторы, сравниваются с наборами дескрипторов EUROVOC, проставленных в ручной работе индексаторами.
- простые правила, производящие слишком много шума, то есть проставляющие дескрипторы автоматически значительно чаще, чем ставят люди, преобразуются в сложные правила, путем снабжения их дополнительными условиями.

Всего было создано около 40 тысяч правил.

При обработке текста отбираются 20 наиболее частотных дескрипторов, порожденных по документу, они и рассматриваются как автоматически приписанные дескрипторы. Для оценки качества работы описанной системы автоматического индексирования для разных типов документов проводилось сравнение с наборами дескрипторов, приписанных вручную. Приводятся данные, что было показано 42% полноты автоматического индексирования.

Архивы поисковых образов документов могут быть использованы для реализации статистических методов автоматического индексирования по информационно-поисковым тезаурусам.

В работе (Steinberger и др., 2000) автоматическое приписывание дескрипторов тезауруса EUROVOC полнотекстовым документам основывается на предварительном нахождении соответствия между словами документа и дескрипторами тезауруса на основе статистических мер (chi-square или log-likelihood) (Manning, Shutze, 1999). Вес соответствия отдельного слова ключевому слову тем выше, чем выше совместная частотность использования данного слова и данного ключевого слова относительно частотности во всей коллекции.

Например, дескриптору тезауруса FISHERY MANAGEMENT (управление рыболовством) соответствуют следующие слова (в порядке убывания веса): *fishery, fish, stock, fishing, conservation, management, vessel, и т.д.*

На второй стадии (собственно, индексирование) для каждого слова документа проверяется, каким дескрипторам тезауруса оно соответствует. Если такие дескрипторы имеются, то слово добавляет к весу дескриптора для данного текста натуральный логарифм веса, полученного на первом этапе. После обработки всех слов текущего текста получается суммированный вес дескрипторов тезауруса.

Например, для Резолюции по правам языковых и культурных меньшинств в Европейском союзе были получены следующие дескрипторы (в порядке убывания веса). *Community programme, Young person, cultural policy, CEEC, European Union и т.д.*

В статье **(**Pouliquen и др., 2003) для автоматического индексирования по тезаурусу EUROVOC процедура автоматического индексирования рассматривается как процедура определения сходства векторов, один из которых вектор слова текста, а второй вектор слов, ассоциированных с дескрипторами тезауруса, по одной из статистических мер совместной встречаемости в документе и его поисковом образе (частотность, нормализованная частотность, log-likelihood). Для процедуры сопоставления векторов использовались такие меры, как формула косинусов (Salton, 1989), формула OKAPI (Robertson и др., 1994), скалярного произведения (формула косинусов без нормализации), линейные комбинации этих формул.

При сравнении результатов с дескрипторами, приписанными людьми для 6 дескрипторов, получивших наиболее высокий вес, были получены следующие результаты: 46, 2 точность, 49,9 полнота, 48.0 F-мера (см. главу 13).

Также в рамках этой работы был выполнен эксперимент по вторичному индексированию человеком. Было получено, что согласие между индексаторами находилось в пределах 74-84 процентов для английских и испанских текстов.

В работе (Montejo-Raez и др., 2004) задача приписывания документам дескрипторов информационно-поискового тезауруса рассматривается как задача автоматической рубрикации, в которой рубрикатором является набор дескрипторов тезауруса. Предлагается использовать подходы машинного обучения, при которых в качестве положительных примеров приписывания конкретного дескриптора рассматриваются документы, к которому индексаторы приписали этот дескриптор, и как отрицательные примеры, документы, к которым данный дескриптор не приписан.

Эксперименты проводились на коллекции рефератов по ядерной физике, использовался тезаурус DESY (http://www-library.desy.de/schlagw2.html).

1.6.2. Проблема вариантности терминов и автоматическое индексирование

Часть проблем с сопоставлением терминов тезауруса и текста связана с тем, что в тексте они употреблены в несколько иной форме (термин разбит дополнительным словом, употреблена однородная конструкция и т.п.), поэтому многими исследователями делаются усилия найти наиболее эффективные способы автоматического сопоставления тезауруса и документа (Большакова, Васильева, 2008; Bolshakova, 2004).

Авторы статьи (Nenadic и др., 2004) классифицируют вариантность терминов на следующие 5 групп:

- орфографические варианты – использование пробелов или дефисов, орфографические варианты (*tumour – tumor*), разная (латинская или греческая) транскрипция (*oestrogen vs. Estrogen*),
- морфологические и словообразовательные варианты: *cellular gene - cell gene*;
- лексические варианты - включают лексические синонимы - *carcinoma – cancer*;
- структурные варианты – посессивное использование существительных или использование существительных с предлогом (*clones in human – human clones)*) варианты предлогов *(cell in blood - cell from blood*), использование сочинительных конструкций (*adrenal glands and gonads*);
- аббревиатуры (DNA – deoxyribonucleic acid).

В статье (Jacquemin, Tzoukermann, 1999) описывается система Fastr, которая позволяет находить в тексте варианты терминов информационно-поискового тезауруса.

Система содержит набор правил, описывающих совокупность трансформаций, которые могут происходить с терминами ИПТ в реальных текстах.

Эти трансформации делятся на два класса: синтаксические вариации и морфосинтаксические вариации.

Синтаксические трансформации включают следующие виды:

- слабые синтаксические вариации, при которых происходит замена предлога внутри термина (*drying by vacuum, drying under vacuum*) или включение определителей: артиклей или указательных местоимений (*milk from cows – milk from these cows*);
- вариация включения, когда прилагательное или наречие помещаются внутри термина. Также допускается вставка в термин более сложной последовательности слов: *Legislation in production – legislation in certain areas of production*;
- вариации координации, при которой внутрь термина вставляется фрагмент сочинительной конструкции (*transfer of energy – transfer of mass and energy*).

Морфо-синтаксические вариации включают случаи, когда хотя бы одно слово термина перешло в другую часть речи, и одновременно возможно произошла синтаксическая вариация. Различаются четыре вида таких вариаций:

- переход прилагательного в существительное,
- переход существительного в прилагательное,
- переход существительного в однокоренное существительное,
- переход существительного в глагол.

Проведенные эксперименты показали 78% точности распознавания исходных терминов в случае синтаксических вариаций. Оценка морфосинтаксических вариаций показала, что их точность значительно меньше и составляет 54.7 % точности. Таким образом, естественной платой за более гибкое сопоставление дескрипторов тезауруса с документами является снижение точности распознавания единиц тезауруса.

Авторы работы (Nenadic и др., 2004) пишут об улучшении качества извлечения терминов на базе учета орфографических и морфологических вариантов, аббревиатур, и указывают на проблемы работы с предложными и сочинительными конструкциями, поскольку среди предложных конструкций имеется множество нетерминологических конструкций, что увеличивает шум при извлечении терминов, а нормализация сочинительных конструкций порождает множество лишних вариантов.

1.6.3. Сочетание свободных запросов и запросов на основе информационно-поисковых тезаурусов

В настоящее время в мире существует достаточно много информационных систем, предоставляющих пользователям как возможности поиска информации по свободному запросу на естественном языке, так и с помощью дескрипторов информационно-поисковых тезаурусов, сопоставленных документам профессиональным индексаторами.

Одним из направлений использования поисковых образов документов является привлечение этой информации при обработке свободных запросов пользователей, сформулированных на естественном языке. Первым шагом на таком пути может быть нахождение корреляций между словами документов и дескрипторами тезауруса или рубриками рубрикатора, подобно описанным в разделе 1.6.1. (Plaunt, Norgard, 1998).

Появление таких корреляций дает возможность при обработке свободного запроса пользователя определить наиболее соответствующие этому запросу рубрики и/или дескрипторы и предложить их пользователю, который может тем или иным образом включить их в запрос. Например, можно сложить веса дескрипторов (рубрик), соответствующих каждому слову запроса и получить упорядоченный список наиболее релевантных запросу дескрипторов (рубрик) (French и др., 2002).

Так, если пользователь ищет по запросу струйные принтеры в информационной системе, в которой документы прорубрицированы по американскому рубрикатору «Стандартная промышленная классификация», то такая обработка запроса позволит показать наиболее соответствующие запросу рубрики такие как,

3579 Офисная техника и детали

3825 Инструменты для измерения и тестирования электричества и сигналов

2893 Чернила для принтеров.

Тезаурусные поисковые образы документов могут быть использованы и для автоматического расширения свободного запроса пользователя дескрипторами тезауруса (Petras 2004; Petras 2005).

Описанные в работе эксперименты проводились на двуязычной коллекции немецких и английских документов по общественным наукам. База содержит более 150 тысяч немецких документов и 26 тысяч английских документов. Документы реферативного характера содержат заголовок публикации, реферат и дескрипторы Тезауруса по общественным наукам (Schott, 2000), приписанных индексаторами. Эксперименты выполнялись в рамках предметно-ориентированного задания форума по многоязыковым информационным системам CLEF (Kluck, 2003).

Для каждого слова запроса выявлялись два наиболее коррелирующих с этим словом дескриптора тезауруса и добавлялись в запрос. Было получено, что в этом случае исходные показатели эффективности поиска для 25 запросов (средняя точность – см. п. 11.2) возросла с 0.4547 до 0.5144, то есть более чем на 13 процентов для немецкого языка, и с 0.4513 до 0.4818 для английского языка.

1.7. Почему традиционный информационно-поисковый тезаурус сложно использовать как ресурс для автоматической обработки текстов в задачах информационного поиска

Основной целью разработки традиционных информационно-поисковых тезаурусов является использование их единиц (дескрипторов) для описания основных тем документов в процессе ручного индексирования. При этом сам процесс индексирования по такому тезаурусу базируется на лингвистических, грамматических знаниях, а также знаниях о предметной области, которые имеются у профессиональных индексаторов текстов. Индексатор сначала должен прочитать текст, понять его и затем изложить

содержание текста, пользуясь дескрипторами, указанными в информационно-поисковом тезаурусе. Индексатор должен хорошо понимать всю терминологию, использованную в тексте, - для описания основной темы текста ему понадобится значительно меньшее количество терминов.

При автоматической обработке текстов человека-посредника между текстом и описанием его содержания в виде дескрипторов нет. Есть только автоматический процесс и Тезаурус, который должен содержать и те знания, которые содержатся в традиционных информационно-поисковых тезаурусах, и те знания (насколько это возможно), которые использует индексатор для определения основной темы текста.

Таким образом, информационно-поисковый тезаурус, предназначенный для автоматической обработки текстов, должен содержать значительно больше информации о языке предметной области. Кроме того, отношения между терминами, указанные в тезаурусе, должны быть значительно более формализованы для использования их в автоматических режимах

В следующих разделах мы рассмотрим эти проблемы подробнее.

Наибольшая часть примеров, приводимых нами в следующих разделах, будет основываться на тезаурусе EUROVOC. Мы рассматриваем этот тезаурус как типичный пример информационно-поискового тезауруса, при разработке которого многие решения обусловлены направленностью на ручное индексирование документов и удобством для человека-индексатора, и, по большей мере, наш выбор этого тезауруса как источника примеров обусловлен следующими обстоятельствами:

- тезаурус EUROVOC – это рабочий инструмент информационных служб парламентов европейских государств;
- имеется русскоязычный перевод тезауруса, что позволяет использовать русскоязычные эквиваленты дескрипторов как примеры;
- тезаурус EUROVOC – это один из немногих тезаурусов, который реально используется для ручного индексирования документов в настоящее время в России.

1.7.1. Нехватка информации о языке предметной области

Нехватка информации о языке предметной области в информационно-поисковых тезаурусах проявляется несколькими разными способами**.**

Во-первых, как мы указывали в разделах 1.1.1 и 1.5, некоторые дескрипторы снабжены подробными правилами их использования, которые предназначаются для индексаторов и наличие этих правил говорит о том, что в текстах предметной области те же термины употребляются по-другому. Так, в разделе 1.5. указывалось, что документ может относиться к статистическим данным, но дескриптор STATISTICAL DATA тезауруса UNBIS используется только, когда документ действительно использует статистические данные, а не просто упоминает их.

Во-вторых, как указывалось в разделе 1.3., разработчики тезаурусов предпочитают не включать в синонимичные ряды дескрипторов синонимы, которые являются очевидными для человека, однако для компьютера эти варианты должны быть обозначены.

Так, например, дескриптор ОХРАНА ОКРУЖАЮЩЕЙ СРЕДЫ помимо указанных в тезаурусе EUROVOC вариантов и синонимов может быть выражен также следующими словами и терминами, не описанными в тезаурусе, но встречающимися в текстах российских правовых актов: *защита природы, природозащитный, природоохранный, природоохранительный (меры, деятельность, процесс*); дескриптор ОХРАНА ЛЕСОВ - *защита лесов, защита лесного фонда, лесозащитный (деятельность, мероприятия), лесоохрана, лесоохранный;* дескриптор СУДЕБНЫЕ РАСХОДЫ – *судебные издержки***,** дескриптор РАСХОДЫ НА ОБОРОНУ – *оборонные расходы, военные расходы, военный бюджет, оборонный бюджет* и еще сотни примеров.

В третьих, как также указывалось в разделе 1.3., разработчики тезаурусов в своем изложении иерархии понятий стараются остановиться на достаточно высоком уровне иерархии и не включать более конкретные термины.

Так, в тезаурусе EUROVOC отсутствуют такие конкретные термины как *минтай, солдаты, пшеница*. Между тем, например, среди законодательных документов широко представлены такие документы, в которых обсуждается *минтай*, но нет слова *рыба*, обсуждаются *солдаты,* но нет слова *военнослужащий*, обсуждается *пшеница*, но нет слова *зерно* и многие другие подобные примеры. Такие тексты не могут быть проиндексированы правильно из-за нехватки информации в тезаурусе.

Наконец, в традиционном информационно-поисковом тезаурусе не указана неоднозначность некоторых терминов, описанных в тезаурусе только в одном из значений, что несущественно для человека-индексатора, но необходимо для автоматической обработки.

Примеры неоднозначных терминов тезауруса, включенных в русскую версию EUROVOC в одном значении, таковы: *кожа* (как кожевенная продукция и кожа человека), *печать* (как СМИ, как штамп, как процесс печатания), *питание* (еда и электрическое питание), *корма* (питание животных и часть корабля), *образование* (как обучение и как создание чего-либо). Средства описания и работы с многозначностью лексики необходимы для любого ресурса, использующегося для автоматической обработки текстов

Для преодоления различий между реальными текстами и информационно-поисковыми тезаурусами при автоматическом индексировании необходимо применять алгоритмы, подобные описанным в разделе 1.6.1. Однако нужно отметить, что такая процедура автоматического индексирования является по сути процедурой автоматической рубрикации по сверхбольшому рубрикатору, качественная реализация которой чрезвычайно сложна (см. главу 13).

1.7.2. Использование отношений между дескрипторами в автоматическом режиме

Автоматическое индексирование предполагает и автоматизацию поиска, то есть поиск с автоматическим расширением запроса. Рассмотрим проблемы автоматического применения отношений между дескрипторами тезауруса на примере тезауруса EUROVOC.

Традиционно исследователи (Tudhope, Taylor, 1997; Chen и др., 1993) указывают на проблемы использования отношения ассоциации при автоматическом расширении запросов. Действительно, и в тезаурусе EUROVOC можно найти многочисленные примеры ассоциативных отношений, на которые невозможно уверенно опереться при автоматическом расширении запроса:

ОХРАНА ДЕТСТВА
АСЦ *ПРОСТИТУЦИЯ*

Ищем тексты о детях, получаем тексты о проституции, из которых лишь некоторые о детях. В обратную сторону тоже не лучше: ищем тексты о проституции, получаем тексты о детях.

МОНОГРАФИИ
АСЦ *ТИПОГРАФИИ*

Ищем тексты о монографиях, получаем тексты о типографиях и наоборот.

Неудачным с точки зрения автоматического расширения запроса представляется требование создателей EUROVOC, чтобы каждый дескриптор имел максимум один вышестоящий дескриптор. Из-за этого отношения ВЫШЕ-НИЖЕ, которые часто могут использоваться для расширения запроса, переводятся в симметричные ассоциации,

использование которых, по крайней мере, в одну сторону, не кажется оправданным, например,

ПРОМЫШЛЕННЫЕ АВАРИИ
ВЫШЕ *ПРОМЫШЛЕННАЯ БЕЗОПАСНОСТЬ*
АСЦ *АВАРИИ.*

Ищем по *АВАРИИ*, получаем тексты о *ПРОМЫШЛЕННЫХ АВАРИЯХ*. В обратную сторону ищем тексты о *ПРОМЫШЛЕННЫХ АВАРИЯХ*, получаем тексты о любых *АВАРИЯХ.*

Отношения ассоциации, рассмотренные выше, могут быть сочтены фактором ошибки или субъективизма. Ведь в любом словаре могут встретиться ошибки и неточности. Рассмотрим поподробнее и подвергнем автоматизированной процедуре проверки в информационно-поисковой системе отношения ассоциации, которые, на первый взгляд, не вызывают сомнений в их полезности.

Пусть два дескриптора тезауруса С1 и С2 связаны ассоциативным отношением. Использование отношения для расширения запроса заключается в том, что если запрос содержит С1, то можно расширить запрос, включив в него и С2, и получить дополнительно количество релевантных запросу документов.

Запросы в информационной системе могут состоять из различного числа терминов и слов. С точки зрения тезауруса простейшим запросом является запрос, ссылающийся на один дескриптор тезауруса. Все другие запросы, ссылающиеся на два или более понятий, должны обрабатываться как функция от элементарного запроса.

Мы предполагаем, что потенциальное качество расширения запроса на базе отношений информационно-поискового тезауруса может изучаться на простых запросах. Если поисковые характеристики расширения элементарных запросов являются низкими, то качество расширения сложных поисковых запросов не может быть лучше. Если тезаурусные отношения дают возможность эффективного расширения запроса для простых случаев, то это является важным шагом для изучения способов расширения сложных запросов.

Смысл такого рода элементарных запросов таков: «найти все о *С*», и мы будем обозначать его как SQ(*C*).

Рассмотрим два понятия *C1* и *C2*, между которыми установлено отношение *R*. Выполняя простой запрос SQ(*C1*), мы хотим узнать, может ли отношение *R* с понятием *C2* быть использовано для расширения этого простого запроса. То есть, можно ли в выдачу по запросу SQ(*C1*) с некоторыми весами добавить документы, содержащие только *C2*. Следовательно, чтобы проверить полезность такого расширения для запроса SQ(*C1*), не нужно выполнять реальное вычисление запроса с расширением, а нужно рассмотреть документы, содержащие *C2*, и выяснить, какой процент документов релевантен SQ(*C1*).

Рассмотрим пример тезаурусной статьи из тезауруса EUROVOC для понятия ЗЕМЕЛЬНЫЙ КАДАСТР.

По определению российского законодательства, земельный кадастр имеет следующее определение:

> *Земельный кадастр – это систематизированный свод документированных сведений, получаемых в результате проведения государственного кадастрового учета земельных участков, о местоположении, целевом назначении и правовом положении земель Российской Федерации и сведений о территориальных зонах и наличии расположенных на земельных участках и прочно связанных с этими земельными участками объектов.*

Дескриптор *ЗЕМЕЛЬНЫЙ КАДАСТР* в Тезаурусе EUROVOC имеет ассоциативные связи с такими дескрипторами:

- ГРАДОСТРОИТЕЛЬНОЕ ЗАКОНОДАТЕЛЬСТВО,

- МЕСТНЫЕ НАЛОГИ;
- НАЛОГ НА НЕДВИЖИМОСТЬ;
- РАЗРЕШЕНИЕ НА СТРОИТЕЛЬСТВО.

Выполним поиск по запросу *земельный кадастр* в коллекции стенограмм заседаний Государственной Думы Федерального Собрания Российской Федерации в Университетской системе Россия (www.cir.ru), которая соответствует области применения тезауруса EUROVOC, и проанализируем содержание первых десяти документов в выдаче (стенограммы заседаний Государственной Думы ФС РФ 25.10.2000 – 14.06.2002):

При поиске по стенограммам мы имеем

- только один фрагмент обсуждения как-либо касается проблемы разрешений на строительство в следующей фразе: «Наконец, кадастровая оценка земли. Посмотрите, что делается вокруг Москвы. Вокруг Москвы - леса первой группы. Эти леса нещадно вырубаются, люди строят дачи. Каким-то хитрым постановлением леса первой группы переводятся в земли общего пользования впрямую, а потом там продаются земли» (выступление Немцова Б.Е. на заседании Государственной Думы ФС РФ от 15 июня 2001 года)
- только один фрагмент обсуждения касается законов о строительстве, предлагая рассматривать незавершенный строительный объект как «нормальную недвижимость (стенограмма от 14 июня 2002 года)
- только один документ обсуждает земельный кадастр как источник информации для налоговых органов, но обсуждается проблема налогов на доходы: «базовая доходность с единицы площади одного рабочего места» (стенограмма от 6 июня 2002 года).

Проанализировав первые 50 документов выдачи УИС РОССИЯ по словам *земельный кадастр,* получаем, что 41 документ был релевантен понятию *ЗЕМЕЛЬНЫЙ КАДАСТР* (остальные 9 документов обсуждали назначения в профильном комитете Государственной думы). Из них

- 11 документов были релеванты запросу «Налог на недвижимость»;
- 9 документов – запросу «Местные налоги»;
- 9 документов – запросу «Градостроительное законодательство»;
- 3 документа – запросу «Разрешение на строительство».

Если мы на том же множестве документов рассмотрим документы, выданные на запрос «Налог на недвижимость», то среди первых 50 документов мы обнаружим лишь 5 документов, релевантных запросу «Земельный кадастр».

Таким образом, мы видим, что если при поиске по каждому из четырех вышеперечисленных понятий, будут автоматически добавлены документы, обсуждающие земельный кадастр, то точность поиска «катастрофически» упадет.

Рассмотрим, что же происходит, чему посвящены другие тексты выдачи. В стенограммах обсуждались такие вопросы как составление Земельного кадастра, регистрация прав на недвижимость, кадастровая стоимость земельного участка, купля-продажа земли и другие вопросы.

Таким образом, мы видим, что с земельным кадастром связан ряд ситуаций. Только в относительно небольшой части из них земельный кадастр сильно связан с перечисленными выше четырьмя понятиями, а в других связь с этими понятиями отсутствует, тексты же могут обсуждать любую из этих ситуаций, поэтому плохие поисковые характеристики вышеперечисленных ассоциативных связей закономерны.

На наш взгляд, установление таких ассоциативных связей нарушает правило, которое пытается ввести стандарт Z39.19 «отношение стоит устанавливать между двумя дескрипторами, если при употреблении одного термина другой термин как бы подразумевается». В приведенном примере использование каждого дескриптора из пары не подразумевает другого дескриптора этой же пары. Например, для разрешения на

строительство необходимо множество документов, а не только выписка из земельного кадастра, а сведения из земельного кадастра могут понадобиться для принятия многих других решений.

При этом, безусловно, правило, устанавливаемое стандартом, абсолютно неформализовано, сформулировано очень нечетко, и его практически невозможно последовательно применять на практике. Как можно более четко сформулировать это правило, мы рассмотрим в разделе 17.4.3.

1.8. Тезаурусы и рубрикаторы в информационно-поисковых системах

В настоящее время в информационно-поисковых системах значительно более широко, чем информационно-поисковые тезаурусы, используются рубрикаторы – классификационные системы.

ГОСТ 7.74-96 определяет классификационную систему следующим образом:

> ***Рубрикатор (классификационная система)*** - *это средство формализованного представления содержания документов, данных и информационных запросов посредством кодов или описаний классов логически упорядоченного множества понятий. Информационные классификационные системы являются одним из типов информационно-поисковых языков.*

Рубрикаторы могут быть иерархическими и фасетными.

Иерархический рубрикатор – это классификационная структура, основанная на отношениях подчинения.

Иерархическими являются библиотечные рубрикаторы такие, как УДК (Универсальная десятичная классификация), ББК (Библиотечно-библиографическая классификация), ГРНТИ (Государственный рубрикатор научно-технической информации).

Фасетный рубрикатор - это классификационная структура, основанная на делении классифицируемого множества по нескольким классификационным признакам одновременно. Так, новостное сообщение может классифицироваться как по основной теме, так и по региону, в котором произошло событие данной новости.

Используются и смешанные формы рубрикаторов.

Может возникнуть вопрос, в чем заключается отличие между рубрикаторами и тезаурусами

Имеется главное теоретическое отличие терминов тезауруса от рубрик рубрикатора. Термины тезауруса являются фундаментально языковыми, в то время как рубрики соответствуют концептуальным категориям (Bates, 1988).

Цель разработки информационно-поискового тезауруса – это, используя реально существующие термины предметной области, найти хорошие, компактные слова и фразы для описания основных тем документов, сведя синонимы и квазисинонимы к дескрипторам тезауруса.

Цель создания рубрикаторов, которая не всегда достигается, но всегда ставится, - это разработать совершенно отдельные концептуальные категории, которые взаимно не пересекаются. Идеально не должно быть пересечений между рубриками и не должно быть промежутков, то есть ни одна подобласть не должна остаться вне рубрик рубрикатора.

Для того, чтобы определить рубрики достаточно строго и исключить пересечение значений, часто необходимо называть рубрики длинными и «неуклюжими» именами, например, «Тропические и субтропические фрукты и орехи; полевые культуры». Такое словосочетание не встретить в тезаурусе, его назначение - четко определить отдельную концептуальную категорию. Поскольку работать с такими сложно сформулированными сущностями достаточно тяжело, им обычно присваивается некоторая система классификационных кодов.

Таким образом, рубрикатор создается сверху, разделением предметной области на подобласти, а тезаурус – снизу, начиная от терминологии конкретных документов.

Процесс присваивания рубрик документам – рубрицирование – в современных информационных системах может осуществляться вручную, автоматическом или автоматизированном режимах. Подробнее различные способы рубрицирования и их особенности будут рассмотрены в главе 13.

Заключение к главе 1

Информационно-поисковые тезаурусы, создаваемые в том виде, как это закреплено международными и национальными стандартами, предназначены для использования их в ручном режиме индексирования. По своей сути такой тезаурус является искусственным языком описания, построенным на основе естественного языка, имеется значительная дистанция между лексическим составом документов предметной области и словарным составом информационно-поискового тезауруса в этой предметной области.

Именно поэтому традиционные информационно-поисковые тезаурусы, разработанные для ручного индексирования, сложно использовать при автоматическом индексировании документов, применять в других приложениях информационного поиска, хотя такие тезаурусы содержат в себе много полезной информации о предметной области.

Не случайно большое место в исследованиях по применению тезаурусов в информационном поиске занимают тезаурусы другого типа – тезаурусы типа WordNet, словарный состав которых является значительно более подробным, значительно более близок лексике документов. Структура и принципы создания тезауруса WordNet будут рассмотрены в следующей главе.

Глава 2. Тезаурус английского языка WordNet

Одним из наиболее известных лексических ресурсов в сфере компьютерной лингвистики и автоматической обработки текстов является компьютерный тезаурус WordNet. Большое количество экспериментов выполнено с этим тезаурусом и в рамках различных приложений информационного поиска.

WordNet версии 3.0 включает приблизительно 155 тысяч различных лексем и словосочетаний, организованных в 117 тысяч понятий, или совокупностей синонимов (synset), общее число пар лексема – значение составляет более 200 тысяч.

Разработка тезауруса была начата в 1984 году в Принстонском университете США под руководством известного психолингвиста Джорджа Миллера. В 1995 году WordNet появился в Интернет в свободном доступе и вызвал всплеск исследований по его использованию в различных компьютерных приложениях автоматической обработки текстов. Результаты применения WordNet в автоматической обработке текстов оказались не столь однозначно положительными, но WordNet открыл новую эпоху разработки сверх больших структурированных лингвистических ресурсов, вызвал появление большого числа последователей в разных странах, создающих такие ворднеты для своих языков, а также стал базой для многоплановых дискуссий и исследований того, на основе каких принципов должны строится большие лингвистические ресурсы, пригодные для разнообразных приложений в области компьютерной лингвистики.

Первоначально WordNet создавался как модель человеческой памяти. Многие решения по представлению описаний слов в WordNet мотивируются психолингвистическими ээкспериментами. Однако, по мнению самих авторов ресурса, WordNet вызвал значительно больший интерес у компьютерных лингвистов, чем у психолингвистов (Fellbaum, 1998; Поляков, 2002).

В данной главе мы рассмотрим основные принципы создания тезауруса WordNet, способы представления лексической информации, а также рассмотрим основные направления критики, которым подвергался данный ресурс. Все это является важным для последующего обсуждения результатов использования WordNet в приложениях информационного поиска.

2.1. WordNet: основные принципы

Основоположник WordNet Джордж Миллер формулирует основные гипотезы, лежащие в основе разработки WordNet, следующим образом (Miller, 1998):

- гипотеза отделимости: описание лексического компонента естественного языка может быть отделено от других уровней (морфологического, синтаксического);
- гипотеза «образца» (patterning hypothesis): существует такое формальное описание слов, которое может быть применено к большинству слов языка;
- гипотеза о покрытии (comprehensiveness hypothesis): для эффективного использования компьютерного словаря в приложениях автоматической обработки текстов, такие словари должны быть очень большой величины.

Основным отношением в WordNet является отношение синонимии. Наборы синонимов – синсеты – являются основными структурными элементами WordNet.

Понятие синонимии, используемое разработчиками WordNet, базируется на критерии, что два выражения являются синонимичными, если замена одного из них на другое в предложении не меняет значения истинности этого высказывания.

При этом не требуется заменяемости синонимов во всех контекстах – по такому критерию в естественном языке было бы слишком мало синонимов. Используется значительно более слабое утверждение, что синонимы WordNet должны быть взаимозаменимы хотя бы в некотором множестве контекстов. Например, замена *plank* (доска, планка) для слова *board (доска)* редко меняет значение истинности в контексте

плотницкого дела, но существуют контексты, где такая замена не может считаться приемлемой.

Именно определение синонимии в терминах заменимости делает необходимым разделение WordNet на отдельные подструктуры по частям речи. Лексемы различных частей речи (существительные, прилагательные, глаголы, наречия) хранятся отдельно и описания, соответствующие каждой части речи, имеют различную структуру.

Синсет может рассматриваться как представление лексикализованного понятия (концепта) английского языка. Авторы ресурса считают, что синсет существительных представляет понятия существительных, глаголы выражают глагольные концепты, прилагательные – концепты прилагательных и т.п. Кроме того, предполагается, что такое разделение соответствует психолингвистическим экспериментам, которые показывают, что представление информации о прилагательных, существительных, глаголах и наречиях устроено в человеческой памяти по-разному.

Большинство синсетов снабжены толкованиями, подобными толкованиям в традиционных словарях, - это толкование рассматривается как одно и то же для всех синонимов синсета. Если слово имеет несколько значений, то оно входит в несколько различных синсетов.

Для установления отношений между синсетами используется метод лингвистических тестов (Cruse, 1986). При таком методе каждому потенциальному лексическому отношению между словами X и Y сопоставляются высказывания, сформулированные на естественном языке и содержащие в качестве компонентов X и Y. Если составленное диагностическое высказывание для слов X и Y истинно, то соответствующее лексическое отношение между этими словами может быть установлено.

В следующих разделах будут подробно рассмотрены принципы описания в WordNet существительных, прилагательных, глаголов.

2.2. Существительные в WordNet

Основными отношениями, установленными в WordNet между существительными, являются родовидовое отношение, отношение часть-целое и отношение антонимии.

Самым многочисленным отношением между синсетами существительных является родовидовое отношение, при этом видовой синсет называется гипонимом, а родовой гиперонимом. Это транзитивное иерархическое отношение, подобное ISA-отношению в исследованиях по искусственному интеллекту.

Синсет X называется гипонимом синсета Y, если носители английского языка считают нормальными предложения типа «*An X is a (kind of) Y" ("X – это (вид) Y").*

Авторы тезауруса подчеркивают, что на практике различие между синонимией и гипонимией не всегда очевидно. Кроме того, если традиционные словари могут в качестве различных значений одного и того же слова включить и более широкое, и более специализированное значение, например, *board (доска)* в широком смысле, и в более специализированном как *surfboard (доска для серфинга),* при разработке WordNet предпочтение отдавалось решениям, в которых одно и то же слово не представлено и в синсете гипонима, и в синсете гиперонима.

Отношения между синсетами образуют иерархическую структуру (рис. 2.1.). При построении иерархических систем на базе родовидовых отношений обычно предполагается, что свойства вышестоящих понятий наследуются на нижестоящие – так называемое свойство наследования. Таким образом, существительные в WordNet организованы в виде иерархической системы с наследованием. Разработчиками были сделаны систематические усилия, чтобы для каждого синсета найти его родовое понятие, его гипероним.

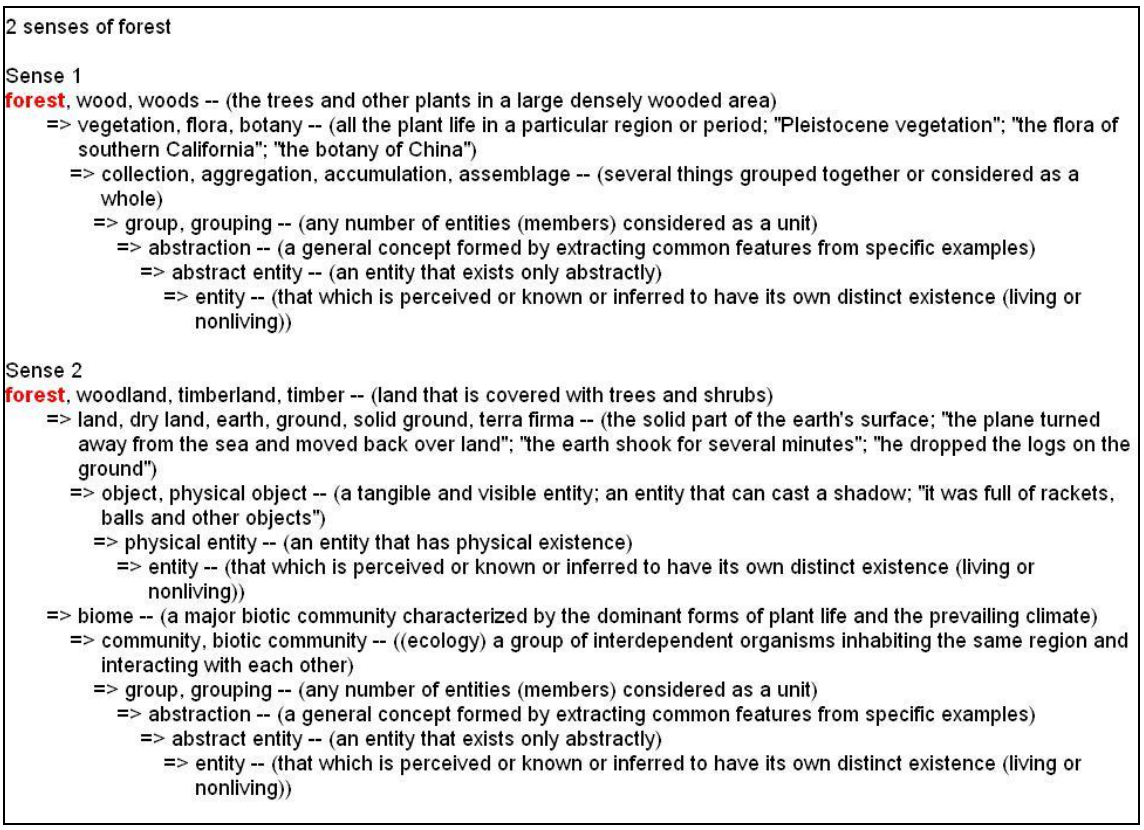

Рис.2.1. Гиперонимы для двух значений существительного forest (лес): лес как совокупность деревьев и лес как территория, на которой растут деревья.

WordNet разделяет существительные на несколько иерархий, каждая со своим начальным понятием. Всего для существительных имеется 25 синсетов верхнего уровня, такие как *{act,activity} (деятельность), {animal, fauna} (животное), {artifact} (продукт труда), {food}(пища), {process} (процесс), {quantity, amount} (количество)* и др.,

Между существительными устанавливаются также отношения ЧАСТЬ-ЦЕЛОЕ, называемое отношением меронимии, синсет-часть называется меронимом, а синсет-целое холонимом. Для установления этого отношения применяется следующий лингвистический тест:

X является частью Y, если можно сказать, что X – это часть Y (An x is a part of Y) или Y имеет X как часть (A y has an x as a part).

Внутри отношения меронимии дополнительно выделяются отношения *быть_элементом* и *быть_сделанным_из*, например,

собственно часть: цветок как орган растения является часть цветкового растения
flower, bloom, blossom -- (reproductive organ of angiosperm plants esp. one having showy or colorful parts)
PART OF: angiosperm, flowering plant -- (plants having seeds in a closed ovary)

элемент: человек является элементом человечества
homo, man, human being, human -- (any living or extinct member of the family Hominidae)
MEMBER OF: genus Homo -- (type genus of the family Hominidae)

вещество: стекло является материалом для стеклянной посуды, стекляных тарелок и др.
glass -- (a brittle transparent solid with irregular atomic structure)
SUBSTANCE OF: glassware, glasswork -- (articles made of glass)
SUBSTANCE OF: plate glass, sheet of glass -- (glass formed into a thin sheet)

Для частей характерно, что у многих разных сущностей части могут называться одинаково, например, *point* (острие) может быть у *стрелы, ножа, иголки, карандаша, булавки* и т.п. В таких случаях описываются все такие холонимы, например,

point -- (sharp end; "he stuck the point of the knife into a tree"; "he broke the point of his pencil")

PART OF: awl -- (a pointed tool for marking surfaces or for punching small holes)

PART OF: icepick, ice pick -- (pick consisting of a steel rod with a sharp point; used for breaking up blocks of ice)

PART OF: knife -- (edge tool used as a cutting instrument; has a pointed blade with a sharp edge and a handle)

PART OF: needle -- (a sharp pointed implement (usually steel))

PART OF: pencil -- (a thin cylindrical pointed writing implement; a rod of marking substance encased in wood)

PART OF: pin -- (a small slender (often pointed) piece of wood or metal used to support or fasten or attach things)

Считается, что меронимы могут наследоваться гипонимами, например, если крыло и клюв описаны как части птицы, то все виды птиц наследуют эти части.

Авторы подчеркивают, что одной из проблем описания отношений меронимии является то, что части описываются несколько выше, чем это необходимо. Например, часто утверждается, что колесо – это часть транспортного средства, но тогда сани не являются транспортным средством. Однако часто такая ситуация является следствием того, что понятие необходимого уровня не лексикализовано в языке. Для данного конкретного примера WordNet вводит специальное дополнительное понятие {wheeled vehicle} – колесное транспортное средство.

Еще одним отношением, установленным для существительных, является отношение антонимии. Отношение антонимии является отношением между конкретными словами, не между синсетами. Кроме того, отношение антонимии не наследуется на синсеты-гипонимы. Предполагается, что отношение антонимии должно быть явным образом описано. Примерами отношений антонимии в WordNet являются следующие: *победа – поражение, счастье – несчастье, мужчина – женщина.*

2.3. Описание прилагательных в WordNet

Прилагательные в WordNet делятся на качественные прилагательные и относительные (Miller K., 1998).

Семантическое описание качественных прилагательных значительно отличается от описания других основных категорий слов и базируется не на отношении гипонимии, а на отношении антонимии. Авторы считают, что важность этого отношения для качественных прилагательных проявляется в психолингвистических тестах: когда человека просят назвать ассоциацию на качественное прилагательное, он чаще всего называет его антоним. Например, самая частая ассоциация на слово *good (хороший)* – это слово *bad (плохой)* и наоборот.

Важность антонимии для организации качественных прилагательных становится понятной, если учесть, что функцией этих прилагательных является выражение величин

атрибутов, и эти атрибуты обычно являются биполярными. Антонимичные прилагательные обычно выражают противоположные полюса атрибута.

Авторы WordNet подчеркивают, что не всегда противоположные по смыслу слова рассматривается носителями языка как антонимы. Так, например, два близких по смыслу слова *heavy (тяжелый), weighty (тяжеловесный)* имеют два разных антонима *light (легкий), weightless (невесомый)*. Кроме того, некоторые слова не имеют непосредственных антонимов, например, слово *ponderous (громоздкий)*. Возможный антоним - прилагательное *light* - является безусловным антонимом прилагательного *heavy.*

Если отношение антонимии возникает между выделенными парами прилагательных, то возникает вопрос, куда отнести такое прилагательное как *ponderous.* Было принято решение такие прилагательные описывать через отношение сходства с одним из тех прилагательных, которые имеют антонимы.

Таким образом, качественные прилагательные в WordNet представлены как биполярные кластеры: центральным является отношение антонимии, для каждого из двух антонимов определены близкие по смыслу прилагательные. Например, описание пары значений прилагательных *slow (медленный, медленно двигающийся*) и *fast (быстрый, быстро двигающийся)* выглядит следующим образом: указана пара антонимов *slow (vs. fast),* для которых описываются так называемые головные синсеты, приводятся толкования и примеры. Для каждого головного синсета описываются «сателлитные» синсеты, которые представляют синсеты, семантически близкие соответствующему головному синсету (и часто являющиеся его специализациями) такие, как *lazy (ленивый, медленно двигающийся), slow-moving (медленно двигающийся)* и др.

slow (vs. fast) -- (not moving quickly; taking a comparatively long time; "a slow walker"; "the slow lane of traffic"; "her steps were slow"; "he was slow in reacting to the news"; "slow but steady growth")

=> bumper-to-bumper -- (used of traffic; "bumper-to-bumper traffic")

=> dilatory, laggard, poky, pokey -- (wasting time)

=> dragging -- (passing painfully or tediously slowly; "the dragging minutes")

=> drawn-out -- ((used of speech) uttered slowly with prolonged vowels)

=> lazy -- (moving slowly and gently; "up a lazy river"; "lazy white clouds"; "at a lazy pace")

=> long-play, long-playing -- ((used of records) playing at a slower speed and for a longer time than earlier records)

=> slow-moving -- (moving slowly; "slow-moving cars")

=> sluggish, sulky -- (with little movement; very slow; "a sluggish stream")

Отдельную группу составляют прилагательные цвета. Все оттенки цветов представляются как синсеты-сателлиты к прилагательному *colored (цветной*), которому противопоставлено как антоним прилагательное *uncolored (нецветной*). Оттенки от белого к черному представлены как семантически близкие синсеты к прилагательному *achromatic (бесцветный).*

Значение относительных прилагательных представлено как отсылка к соответствующему синсету существительных. Для некоторых прилагательных одно из значений представлено как качественное прилагательное через антонимическую пару, а второе как относительное прилагательное. Например, для прилагательного *agricultural (сельский, сельскохозяйственный)* первое значение отсылает к синсету слова *agriculture (сельское хозяйство),* а второе значение представлено как элемент пары антонимов *сельский – городской.*

Sense 1

agricultural -- (relating to or used in or promoting agriculture or farming; "agricultural engineering"; "modern agricultural (or farming) methods"; "agricultural (or farm) equipment"; "an agricultural college")

Pertains to noun agriculture (Sense 2)

=>farming, agriculture, husbandry -- (the practice of cultivating the land or raising stock)

=> cultivation -- ((agriculture) production of food by preparing the land to grow crops)

Sense 2

agrarian, agricultural, farming(prenominal) -- (relating to rural matters; "an agrarian (or agricultural) society"; "farming communities")

=> rural (vs. urban) -- (living in or characteristic of farming or country life; "rural people"; "large rural households"; "unpaved rural roads"; "an economy that is basically rural")

2.4. Описание глаголов в WordNet

Для описания глаголы были разделены на семантические поля. На первом этапе были отделены глаголы, обозначающие действия и события, от глаголов, обозначающих состояния. Первая группа глаголов была разделена на 14 семантических полей: глаголы движения, восприятия, контакта, коммуникации, соревнования, изменения, познания, потребления, создания, эмоций, обладания, ухода за телом, и глаголы, относящиеся к социальному поведению.

Авторы WordNet подчеркивают, что границы между группами являются достаточно расплывчатыми. Например, многие глаголы не могут однозначно классифицированы как глаголы познания или коммуникации (например, *подтверждать, судить* и др.). Также, например, глагол *whistle* (*свистеть*) в предложении «*The bullet whistled past him*» (*Пуля просвистела мимо него*) может классифицироваться и как глагол издания звука, и как глагол движения. Если такие глаголы представлять как однозначные, они должны относиться к более чем одному семантическому полю. В WordNet глаголы чаще описывались как полисемичные, если обнаруживалось, что они могут быть отнесены одновременно к разным семантическим полям.

Между глаголами в WordNet устанавливается свой набор отношений.

Отношение следования (Entailment). Отношение устанавливается между синсетами глаголов V1 и V2, если из предложения «*Someone V1*» логически следует «*Someone V2*». Например, из того, что «*Человек идет*», следует, что «*Человек делает шаг*», поэтому:

Sense 1

walk, go on foot, foot, leg it, hoof, hoof it -- (use one's feet to advance; advance by steps)

=> step, take a step

Другой пример: из «храпеть» следует «спать», так как из предложения «Он храпит», следует «Он спит».

Отношение тропонимии. Лингвистический тест, который использовался для определения гипонимии между существительными, «*An x is an y*» не подходит для глаголов, поскольку требует, чтобы x и y были существительными. Поэтому потребовалось предложить другой лингвистический текст для установления

иерархических отношений между глаголами а также было введено специальное название отношения – тропонимия. Используемый лингвистический тест таков: «*To V1 is to V2 in some particular manner*» (*Делать V1 означает делать V2 особым способом*), например, «*Mumbling is talking is some particular manner*» («*Бормотать – это говорить особым способом*»).

Отношение тропонимии представляет собой специальный вид отношения следования.

Глагольные иерархии, образованные отношением тропонимии, образуют более узкую, но более кустистую структуру, чем существительные. В подавляющем большинстве случаев число уровней иерархии не превышает четырех.

Отношение причины. Отношение причины связывает два глагольных синсета, один из которых может быть назван каузатив, например, *давать*, а второй результатив (*иметь*). Кроме того, WordNet устанавливает отношение причины от каузативных транзитивных глаголов к соответствующим инхоативным нетранзитивным значениям тех же слов. В качестве примеров можно рассмотреть такие глаголы как *blacken (чернить – чернеть), develop (развивать – развиваться), break (сломать – сломаться), shrink (сократить – сократиться).*

Отношение причины систематически представлено среди глаголов перемещения, соединяя каузативное и некаузативное употребление *blow (выдуть – выдуться).*

Отношение причины также может быть рассмотрено как специальный случай следования. Если V1 необходимо вызывает V2, значит, из V1 также следует V2.

2.5. Исследования конкретных проблем представления лексической информации в WordNet и последующие модификации тезауруса

Появление WordNet и возможность его свободного использования вызвали большое число исследований по применению этого тезауруса в самых различных приложениях автоматической обработки текстов. Некоторые из этих приложений и результаты применения в них WordNet будут рассмотрены в третьей части книги.

Большое количество экспериментов привело к массовому выявлению и обсуждению проблем и недостатков WordNet, препятствующих его эффективному применению.

В данном разделе мы рассмотрим некоторые из таких проблем, возникшие дискуссии, а также изменения в структуре новых версий WordNet, которые были сделаны в результате этих обсуждений.

2.5.1. Отсутствие отношений между частями речи

При разработке WordNet был выдвинут принцип раздельного описания разных частей речи. Между различными частями речи, имеющими одинаковое значение, не было установлено никаких отношений

Так, например, такие синсеты как *adorn1 (украшать) и adornment2 (процесс украшения)* не были никак связаны между собой:

Adorn1 -- (make more attractive by adding ornament, colour, etc.)

Adornment2 -- (the action of decorating yourself with something colorful and interesting)

Это вызывало серьезные проблемы в приложениях, поскольку одно и та же мысль могла быть выражена разными частями речи (Climent и др., 1996).

Кроме того, в различных языках для выражения одной и той же идеи могут использоваться лексемы разных частей речи. Поэтому иерархии синсетов, построенные на основе конкретных частей речи, становятся в большой мере зависимыми от естественного

языка разработки, поскольку в некотором естественном языке может не оказаться возможности выразить некоторое понятие той или иной частью речи.

Начиная с версии WordNet 2.0, в ресурс были введены отношения между однокоренными синсетами, относящимися к разным частям речи и связанными между собой по смыслу. Такие отношения обозначаются RELATED_TO (Miller, Fellbaum, 2003):

Adorn#v1 -- RELATED TO -> adornment#n2

Abandon#v1 -- RELATED TO -> abandonment#n3

Rule#v6 -- RELATED TO -> ruler#n1

Catch#v4 ---- RELATED TO -> catcher#n1

Всего было размечено 21.5 тысячи пар синсет существительного – синсет глагола.

В настоящее время выполнена автоматизированная семантическая разметка отношений между синсетами разных частей речи (Clark и др., 2008), которая указывает специфическое семантическое отношение между существительным и глаголом:

abandonment#n3	*EVENT*	*of*	*abandon#v1*
ruler#n1	*INSTRUMENT*	*of*	*rule#v6*
catcher#n1	*AGENT*	*of*	*catch#v4.*

2.5.2. Слишком много значений в WordNet

Серьезное обсуждение возникло по поводу описания значений многозначных слов в WordNet. Во многих работах признается, что различия значений в WordNet слишком тонки для таких компьютерных приложений как машинный перевод, информационный поиск, классификация текстов, вопросно-ответные системы и др. В (Chugur и др., 2002) было показано, что среднее количество значений в WordNet больше, чем в традиционных лексикографических словарях.

Особенно большое количество значений имеют глаголы и прилагательные. Так, глагол *give* имеет 44 значения, а прилагательное *good* – 21 значение.

Некоторые из описанных значений плохо отделимы друг от друга. Например, значения глагола *give: give5* и *give 21*. В обоих случаях в определении присутствует одно и то же слово *bestow* в том же значении:

Give5: **give**, *pay -- (convey, as of a compliment, regards, attention, etc.; bestow;* ``Don't pay him any mind"; "***give*** the orders"; "Give him my best regards"; "pay attention" *)*

Give21: **give**, *render -- (bestow;* ``**give** homage"; "render thanks" *)*

Часть выделенных значений сочетается только с узким набором слов, например, значение give19:

*Give19: give - (***give** *(as medicine); ``I gave him the drug") – дать (как лекарство)*

Как известно, число значений тех или иных лексических единиц может значительно различаться в различных лексических ресурсах, словарях. Но большое количество значений в WordNet препятствует его применению в приложениях автоматической обработки текстов. Кроме того, проблема лексической многозначности для компьютерных приложений усугубляется тем, что синсеты WordNet, соответствующие близким по смыслу значениям многозначных слов, в большинстве случаев, не имеют между собой никаких отношений.

Эти проблемы привели к постановке вопроса о том, каким образом и какие типы значений многозначного слова могут быть объединены («кластеризованы») (Chugur и др., 2000, Peters и др., 2000; Agirre, Lacalle, 1996; McCarthy, 2006) для целей работы в приложениях автоматической обработки текстов, когда для значений многозначного

слова из кластера можно не делать различий, и это не приведет к снижению качества работы этого приложения.

Для рассмотрения предложенных подходов кратко остановимся на выделяемых типах отношений между значениями отдельного слова.

2.5.2.1. Отношения между значениями одного и того же слова

Лингвисты выделяют несколько типов отношений между парами значений одного и того же слова (Апресян, 1995; Кобозева, 2000; Зализняк 2006).

Омонимией называется случайное внешнее совпадение двух разных слов, что проявляется в том, что между значениями нет общих элементов смысла, например,

Лук (оружие) – лук (растение)

Брак (изъян) – брак (женитьба)

Значения слова называются **полисемичными**, если между ними существуют общий элемент смысла, например, значения слова *клапан* в словосочетаниях:

Клапан мотора – клапан фагота – сердечный клапан – клапан кармана

имеют общий элемент - *«часть предмета, закрывающая отверстие в нем».*

Содержательно отношения между значениями многозначного слова делятся на следующие основные типы (Кобозева, 2000).

Метафорическое отношение между значениями характеризуется как основанное на сходстве, подобии обозначемых явлений. Так, язык пламени, язык колокола и язык во рту человека похожи по форме.

Метонимические отношения между значениями основаны на смежности обозначаемых объектов и явлений. Так, работой называется не только определенный вид деятельности, но и место, где эта деятельность происходит, а также ее результат.

Метонимические переносы достаточно часто бывают регулярными – **регулярная полисемия** (Апресян, 1995; Кронгауз, 2001). Среди наиболее частых метонимических переносов отмечаются следующие:

- действие – результат действия: сочинение, остановка, украшение,
- вместилище – содержимое: стакан, кастрюля,
- населенный пункт – люди, живущие в населенном пункте: город, деревня, поселок,
- наука – предмет науки: семантика, синтаксис и др. (Апресян, 1995).

Использование данной классификации отношений между значениями слова являлось одним из важных направлений в попытках исследователей найти полезную для приложений кластеризацию значений WordNet.

2.5.2.2. Подходы к кластеризации значений WordNet

Одной из первых идей по объединению значений WordNet для компьютерных приложений было исследование, насколько явление регулярной многозначности может служить основой для такой процедуры.

Работа (Buitellar, 1998) была посвящена изучению масштабов регулярной полисемии в WordNet. Для этого все многозначные существительные были расклассифицированы по их основным семантическим типам, задаваемым наиболее высоким гиперонимом, к которым они относятся, таким как *артефакт, каузальный агент, форма, действие* и др. Далее все такие существительные были разбиты на группы в соответствии с наборами семантических типов, в которые попадают их значения. Так, например, существительное *банан*, которое имеет значение пищевого продукта и растения, попадает в ту же группу, что и такие слова как *кориандр, грейпфрут* и др.

Всего было выделено 126 семантических типов, которые охватывают 39937 существительных в 317 регулярных классах полисемии.

В работе (Peters и др., 2000) рассматриваются возможные направления кластеризации значений в WordNet, основанные на различных типах лексической многозначности. Рассматривается три возможных типа многозначности, которые могут быть использованы для кластеризации значений.

Первое направление – обобщение, которое заключается в том, что если различные значения одного и того же слова имеют один и тот же гипероним, то можно попытаться найти общее между всеми этими значениями, которое и рассматривать как кластеризованное значение. Такие значения могут располагаться в сети WordNet друг под другом (одно является гиперонимом для другого) – явление, называемое **автогипонимией.**

В таких случаях более высокое по иерархии значение может представлять значение кластера.

Также обобщение может быть сделано для значений, являющихся так называемыми «сестрами», т.е. значениями, являющимися гипонимами одного и того же гиперонима. Например, в WordNet значения слова *table (table2 и table3)* имеют один и тот же гипероним *piece of furniture – предмет мебели*:

Table2 – a piece of furniture having a smooth flat top supported by one or more vertical legs "it was a sturdy table"

Table3 – a piece of furniture with tableware for a meal laid out on it: "I reserved a table at my favourite restaurant".

Наконец, еще одной возможностью формального обнаружения обобщающего значения являются так называемые синсеты-близнецы (twins) - т.е. те синсеты, в которых по крайней мере три синонима совпадают.

Например, близецами являются такие синсеты как:

violate, go against, break *-- (fail to agree with; be in violation of; as of rules or patterns; "This sentence violates the rules of syntax")*

и

*transgress**, offend, infract**, violate, go against**, breach**, break** -- (act in disregard of laws, rules, contracts, or promises; "offend all laws of humanity"; "violate the basic laws or human civilization"; "break a law"; "break a promise").*

Второй тип возможного кластерного значения – это кластер, основанный на метонимии. Этот тип покрывает случаи так называемой регулярной полисемии: *организация - здание, дерево - древесина, материал - продукт, вместилище – содержимое вместилища* и др. В данном исследовании такие пары задавались вручную

Третий тип семантической кластеризации основан на явлении диатезы - вариативности в управлении глаголов, во многих случаях различия между транзитивным (нетранзитивным), каузативным (инхоативным) использованием нужны лишь для выражения некоторых сторон предиката, в то время как базисное значение остается одним и тем же.

В (Chugur и др., 2000) исследуется вопрос, какая группировка значений была бы полезной для задач информационного поиска. Предполагается, что некоторые значения могут быть кластеризованы для разных приложений, в то же время существуют примеры пар значений, кластеризация которых была бы полезна в информационно-поисковых приложений, при этом в других приложениях было бы полезно их различать. Примером такой пары значений являются следующие синсеты:

Bet

1. The act of gambling (ставить - вкладывать в банк в азартных играх)

2. The money risked on a gamble (ставка в азартных играх).

Отмечается, что исследования регулярной многозначности не приводят к выделению полезных кластеров для информационно-поисковых задач, так как, как представляется авторам данной работы, некоторые образцы регулярной полисемии хорошо бы не различать для задач информационного поиска, в то время как другие хорошо бы сохранить отдельно. Так, например, полезно было бы кластеризовать такие пары регулярной полисемии как *container/quantity (вместилище – объем вместилища) и music/dance (музыка – танец)*. Однако такие образцы как *animal/food (животное-пища), plant/food (растение- пища), animal/skin (животное-шкура), language/people (язык-народ)* хорошо бы различать, поскольку, как представляется они употребляются в разных типах текстов.

Поэтому нужны дополнительные исследования критериев кластеризации значений для информационно-поисковых задач.

В работе сравниваются два дополнительных критерия группировки значений. Первый критерий заключается в том, чтобы группировать значения, которые встречаются в одних и тех же текстах. Для этого используется семантически размеченный значениями WordNet корпус Semcor (Landes и др., 1998). Второй критерий группирует значения, которые получают одни и те же переводы в нескольких языках. Пересечение кластеров, построенных на основе этих двух критериев, составляет 55-60 процентов, что показывает некоторую корреляцию между кластерами, но оставляет сомнения в полезности каждого из критериев.

В заключении авторы работы (Chugur и др., 2000) рассматривают основные типы отношений между различными значениями (см. п.2.5.2.1), которые могут привести к полезным кластерам значений для информационного поиска.

Рассматривается следующие четыре типа отношений между значениями:

- обобщение/спецификация – автогипонимия;
- метонимия;
- метафора;
- омонимия.

В таблице 2.1. приводится корреляция между типами отношений между значениями и полезностью кластера для информационного поиска:

	Метонимия	Обобщение	Метафора	Омонимия
Кластеры, полезные для инф. поиска	4 (27%)	11 (73%)	0	0
Значения, которые нужно различать	5 (45%)	0	3 (27%)	3 (27%)

Таблица 2.1. Корреляция между типами отношений между значениями многозначного слова и возможными кластерами значений для информационного поиска из работы (Chugur и др., 2000).

Проведенный анализ типов отношений между значениями слова показал, что:

- тип обобщение/спецификация образует полезный кластер для информационного поиска;
- типы метафора и омонимия не приводят к полезным кластерам для информационного поиска;
- отношение метонимии ведет себя двояко, что требует дополнительных исследований.

В работе (Gonzalo, 2004) подчеркивается, что проведенные эксперименты по кластеризации значений привели к выводу, что типология отношений между разными значениями многозначных слов является более полезной, чем формирование кластеров значений, поскольку «прикладная» близость значений зависит от приложения.

Например, указание, что одно из значений является метафорой исходного значения, является важным различием для приложений информационного поиска и вопросно-ответных систем, поскольку такие значения относятся к разным тематическим полям. Однако для приложений машинного перевода это различие может быть несущественно, поскольку метафорический перенос может быть сходным в разных языках.

В работе (Fellbaum, Miller, 2006) подводится итог всем исследованиям по «прикладному» объединению значений, введению недоопределенности значений. Подчеркивается, что кластеризация значений может проводиться на основе различных взаимоисключающих критериев (семантических, синтаксических, предметно-ориентированных), что, видимо, подтверждает мысль работ (Chugur и др., 2000; Gonzalo, 2004) о разной значимости разных подразделений значений для конкретных приложений автоматической обработки текстов.

По причине упомянутой позиции авторов ресурса никаких значительных изменений в структуре значений WordNet не производилось.

Проблема автоматического выбора значений WordNet в практических приложениях может быть смягчена за счет использования информации из семантически размеченного по значениям WordNet корпуса текстов SemCor (Landes и др., 1998).

Корпус SemCor представляет собой подмножество известного Брауновского корпуса и включает 352 текста. В 186 текстах все знаменательные слова (существительные, прилагательные, глаголы, наречия) размечены следующей информацией: часть речи, лемма, значение по WordNet. В остальных текстах размечены только глаголы. Всего размечено около 200 тысяч слов.

В последних версиях WordNet значения упорядочены по мере встречаемости в этом корпусе (первое значение соответствует самому частотному значению).

В экспериментах по автоматическому разрешению многозначности слов часто используется информация о самом частотном значении слова в корпусе SemCor, которое выбирается в сложных случаях (подробнее см. раздел 10.2).

2.5.3 Проблемы описания отношений между синсетами существительных

Многие исследователи использовали для своих экспериментов, прежде всего, синсеты существительных из WordNet. Поэтому особое внимание и обсуждение исследователей было посвящено системе отношений между этими синсетами. В данном разделе мы рассмотрим наиболее активно обсуждавшиеся вопросы установления отношений между синсетами.

Во-первых, это так называемая «теннисная» проблема - проблема нехватки отношений между синсетами, относящимися к одной и той же тематической области (Miller, 1998). Во-вторых, мы рассмотрим дискуссию по поводу принципов установления отношений гипонимии-гиперонимии.

2.5.3.1. «Теннисная проблема»

Одной из серьезных проблем WordNet, препятствующей его использованию в приложениях, является так называемая «теннисная проблема»: принадлежащие одной предметной области, сфере деятельности, ситуации синсеты оказываются очень далеко друг от друга в структуре WordNet.

Дж. Миллер (Miller, 1998) пишет, что, если кто-либо захочет обратиться к WordNet, чтобы узнать о специализированном словаре теннисной области, то выяснится, что в WordNet очень много слов из этой сферы, но они совершенно разделены, будучи включенными каждый в свою классификацию: синсет *теннисный инвентарь* включен в иерархию артефактов, синсет *теннисный корт* включен в иерархию местоположений, различные синсеты теннисных ударов в иерархию действий. Получается, что

существительные, которые часто употребляются в одних и тех же текстах, в WordNet не имеют между собой никаких общих отношений. Такая же проблема возникает, естественно, с тематической лексикой из других областей деятельности.

Отсутствие такого рода отношений оказывает серьезное негативное воздействие на использование WordNet в автоматических процедурах разрешения лексической многозначности, вызывает проблемы в информационном поиске.

В ряде исследований было предложено решать данную проблему введением в WordNet информации о принадлежности синсетов определенным тематическим доменам. Домены, такие, как «теннис», «политика» или «образование», группируют синсеты в сценарии или схемы. Так, домен «теннис» включает такие синсеты, как «гейм», «теннисный мяч», «теннисная ракетка», «тай-брейк» и т.д.

Работа (Magnini, 2000) описывает процесс создания иерархической системы таких доменов и процедуру автоматизированной приписки доменов синсетам WordNet.

Разработка иерархической системы доменов началась с 250 рубрик, собранных по различным словарям и затем была дополнена и уточнена на базе Десятичной классификации Дьюи. Была получена иерархия из 115 доменов, организованных по 4 уровням иерархии, включающая, например, такие домены как, например, «сельское хозяйство», «археология», «астрология», «биология», «ветеринария» и др..

Кроме того, была выделена специальная область, в которую входят синсеты WordNet, не принадлежащие никаким тематическим доменам, поскольку они могут употребляться в текстах многих предметных областей. Такая специальная предметная область получила название Factotum.

Область Factotum включает два типа синсетов:

- «общие» синсеты, которые трудно отнести к какой-либо предметной области, например, *человек, мужчина, день.* Эти синсеты располагаются обычно высоко в иерархии WordNet и содержат очень многозначные слова:

  Man 1 – an adult male person (мужчина)

  Man 3 – the generic use of the word to refer to any human being (человек)

  Date 1 – day of the month (день месяца)

  Date3 – appointment, engagement (назначение);

- синсеты, которые можно рассматривать как стоп-синсеты: числа, дни недели, цвета. Такие синсеты могут встретиться в самых разнообразных контекстах, но обычно их вклад в содержание текста невелик.

Всего область Factotum включает 6450 синсетов, включая 2780 стоп-синсетов и 3670 «общих» синсетов.

Для того чтобы разметить все множество синсетов WordNet, была реализована автоматизированная процедура, состоящая из следующих шагов:

1) Вручную размечается относительно небольшое количество синсетов верхнего уровня,

2) Автоматически по связям (гипонимия, тропонимия, меронимия, антонимия) пометки распространяются на другие синсеты,

3) Можно задать исключения, например, для синсета кресло парикмахера (barber_chair), которое является частью парикмахерской (barbershop) и поэтому получает домен КОММЕРЦИЯ (COMMERCE).

Процедура была выполнена только для существительных.В работе приводятся данные о количестве приписанных в результате автоматизированной приписки синсетов для некоторых доменов:

Сельское хозяйство	*248*
Археология	*47*
Питание	*2563*
Астрология	*16*
Биология	*20266*
Медицина	*2660*
Ветеринария	*36 и др.*

В настоящее время разметку последних версий WordNet по тематическим областям можно получить с сайта http://wndomains.itc.it/wordnetdomains.html

Вместе с тем остаются вопросы по отношению к введению в систему, построенную на основе одних единиц, набора других единиц с неопределенным относительно исходных единиц статусом среди которых:

- вариативность возможного набора доменов;
- небольшая наполненность некоторых доменов, и большое количество синсетов в других доменах;
- необходимость разных систем доменов для разных задач;
- отсутствие полностью выверенной разметки синсетов набором доменов (выверить вручную очень трудоемко, если выверять в процессе решения различных задач, то далеко не все проблемы (неточности, ошибки)) приписки удастся быстро обнаружить.

2.5.3.2. Проблемы родовидовых отношений WordNet.

Как уже указывалось в разделе 2.1, основным принципом установления отношений в Wordnet было применение так называемых диагностических высказываний. В частности, для установления отношений гипоним-гипероним использовалось проверочное высказывание: *An X is a (kind of) Y (X - это Y*). Однако в процессе экспериментов и дискуссий выяснилось, что такому высказыванию могут удовлетворять несколько совершенно разных отношений между синсетами (Miller, 1998).

Одной из серьезных проблем, приводящих к неправильным путям иерархии и, следовательно, препятствующих применению в приложениях автоматической обработки текстов, является проблема установления таких отношений, когда вышестоящее понятие частично характеризует нижестоящее. Часто это связано с проблемой смешения понятий-типов и понятий-ролей (подробнее см. главу 7).

Указывая на смешение типов и ролей в Wordnet, Н. Гуарино (Guarino, 1998) привел следующие примеры описания из WordNet:

Человек – это живое существо и каузальный агент.

Яблоко – это фрукт и еда.

Н.Гуарино указывает, что каждое из этой пары отношений отличается от другого: Человек всегда живое существо, но он (она) начинает играть роль каузального агента только в некоторых ситуациях. Та же проблема возникает для яблока, которое всегда плод растения и в некоторых ситуациях может быть пищей: «Проблема в том , что человек и яблоко – это типы сущностей, в то время как каузальный агент и пища – это роли.»

Один из аргументов в пользу различения типов и ролей в лингвистических онтологиях – это то, что они различаются в способах наследования свойств. WordNet не различает эти два типа понятий и помещает их в одни и те же иерархии.

В соответствии с онтологическими подходами (см. главу 7) понятия-типы не должны находиться в иерархиях ниже понятий-ролей. Более радикальный подход заключается в том, чтобы разделить иерархии типов и ролей.

Одна из авторов WordNet К. Фелбаум (Fellbaum, 2002), отвечая на эту критику Н. Гуарино, заявляет, что в таких ресурсах, как WordNet, неоднородные классификации имеют право на существование, поскольку такие ресурсы рассматриваются в настоящее время, прежде всего, как инструменты для компьютерной обработки текстов, а не только как совершенные онтологии, которые должны соответствовать строгим онтологическим принципам.

Вместе с тем важно подчеркнуть, что установление связей между синсетами, которые выполняются не в любых контекстах, а лишь при некоторых условиях, приводит к ложному срабатыванию этих связей, к неправильному выводу как раз при автоматической обработке текстов.

Используемые диагностические высказывания для установления отношений между гипонимами и гиперонимами привели к смешению и других отношений.

Дело в том, что в первых версиях WordNet не делалось различий между синсетами-категориями классов как множествами сущностей, имеющих между собой общие свойства, например, как синсет *state, nation* – (государство), и примерами классов, то есть конкретных сущностей, например, синсет *United States, United States of America - США* (подробнее о смешении такого рода см. раздел 6.4).

По этой причине отношения между классами и отношения «пример-класс» обозначались одинаково. Такое неразличение стало предметом критики со стороны разработчиков онтологий (Gangemi и др., 2001a; Oltramari и др., 2002).

Первоначально авторы WordNet не предполагали менять структуру WordNet (Miller, Hristea, 2006), поскольку считали, что WordNet – это лексический, а не онтологический ресурс. Однако со временем рост значимости онтологических исследований, а также сходство иерархии существительных из WordNet с онтологией стали очевидными.

В результате были предприняты усилия по разметке синсетов существительных как примеров и как классов, а также различению таксономических отношений между классами, и отношениями «пример-класс» (Miller, Hristea, 2006).

Для автоматизации проведения уточненной разметки было выдвинуто предположение, что синсеты-примеры должны обладать следующими тремя свойствами:

- это должны быть синсеты существительных,
- синсеты должны содержать слова с прописной буквы,
- будучи уже конкретными сущностями, синсеты-примеры не должны иметь гипонимов.

Таких синсетов оказалось 24073, причем выяснилось, что есть достаточное количество синсетов, удовлетворяющих этим требованиям, но при этом являющихся обозначением классов понятий. Поэтому авторы рассмотрели все выделенные синсеты и вручную разметили их как классы или примеры классов. В частности, выявились интересные случаи классов и экземпляров, подтребовавшие отдельного рассмотрения.

Основным критерием разметки было существование единственного референта для синсета. Так, Бетховен как композитор – это пример класса, а Бетховен как музыка («*играть Бетховена*») – это класс, поскольку относится к классу музыкальных произведений. Если слово имеет конкретное число денотатов (два, три и т.д., что означает многозначность слова), то все соответствующие синсеты размечаются как примеры, как, например, Bethlehem на Ближнем востоке и Bethlehem в Пенсильвании.

Одной из проблем разметки была разметка синсетов, соответствующих естественным языкам. В частности, возник вопрос, являются ли конкретные диалекты языка примерами класса. Было решено, что с онтологической точки зрения языки – это не примеры классов, примерами являются конкретные речевые акты.

Сложным случаем оказалась также разметка синсетов, соответствующих священным текстам, таким, как Библия, Коран и другие. Для данного случая было решено, что сами синсеты священных текстов рассматриваются как классы, а их конкретные версии – являются их примерами.

Названия конкретных денежных единиц были размечены как классы, например, синсет *гонконгский доллар* не является примером синсета *доллар*.

В итоге всего 7671 синсетов были признаны синсетами-примерами. Все выявленные отношения пример-класс были размечены специальным образом. Результаты разметки стали доступны пользователям в версии WordNet 2.1.

Заключение к главе 2

Тезаурус WordNet как общедоступный лингвистический ресурс большой величины вызвал огромный интерес во всем мире.

Часть исследователей видит проблемы WordNet в чрезмерной простоте его структуры. Однако эта простота позволила обеспечить большой объем тезауруса, что, в свою очередь, позволило организовать многочисленные эксперименты по применению этого ресурса в реальных приложениях автоматической обработки текстов.

Результаты экспериментов позволили исследователям увидеть проблемы WordNet с точки зрения практических приложений, описать те подводные камни, которые могут поджидать разработчиков новых больших лингвистических ресурсов, предназначенных для автоматической обработки текстов.

Именно поэтому всем исследователям, которые разрабатывают или собираются разрабатывать, новые лингвистические ресурсы для практических приложений в настоящее время, очень важно хорошо владеть сведениями о принципах устройства WordNet, о возникших проблемах, об экспериментах, направленных на изучение и преодоление этих проблем. Кроме того, WordNet продолжает свое развитие, поскольку его разработчики реагируют на критику, на результаты экспериментов и вводят новые типы информации в свой ресурс, уточняют имеющиеся описания.

Ценность WordNet состоит еще и в том, что формализованные отношения между значениями слов позволяют исследователям быстро составлять свои собственные словари, списки слов и выражений для решения частных задач.

Ресурсы типа WordNet разрабатываются в настоящее время для многих языков мира. При этом разработчики стараются учесть выявленные проблемы, предложить новые решения. Принципы реализации новых ресурсов типа WordNet мы рассмотрим в следующей главе.

Глава 3. EuroWordNet и тезаурусы типа WordNet для разных языков

Идея создания тезаурусов типа WordNet (далее будем называть их ворднетами) для своих языков показалась привлекательной исследователям во многих странах.

Разработчиков новых ворднетов можно разделить на две категории. Часть разработчиков считает, что важным делом является точное воспроизведение структуры и состава англоязычного WordNet (обычно называемого Принстонский WordNet по месту работы его авторов), поскольку предполагается, что таким образом обеспечивается более тесная связь с англоязычным ресурсом и лексической системой английского языка.

При этом подходе синсеты нового ворднета создаются на основе синсетов Принстонского WordNet, отношения между синсетами копируются. Такая разработка рассматривается как более быстрая, легкая, порождает структуру, совместимую с англоязычным ворднетом. Часто значительная часть работы производится автоматизированными методами на основе двуязычных электронных словарей (Farreres и др., 1998) Но одновременно такой ворднет может унаследовать недостатки исходного ворднета, неточности могут усилиться, могут быть перенесены чуждые создаваемому ворднету отношения. По такой модели создавались такие ворднеты как испанский ворднет, баскский ворднет, один из ворднетов итальянского языка MultiWordNet.

Другие разработчики полагают, что для создания качественного ресурса собственного языка необходимо учесть специфику его лексической системы, а также учесть критику и проблемы Принстонского WordNet. При таком подходе разработчики развивают собственную структуру синсетов, руководствуясь общими принципами построения ворднетов. Такой метод использовался при создании таких ворднетов как голландский, немецкий и датский ворднеты, тезаурус русского языка RussNet (Азарова и др., 2003; Азарова и др., 2004).

Для некоторых языков появляется два ресурса типа тезаурус WordNet, созданных на основе упомянутых подходов. Например, для итальянского языка один тезаурус ItalWordNet (Roventini и др., 2000) сделан в рамках проекта EuroWordNet, в котором было введено значительное количество нововведений (см. следующий раздел), а другой MultiWordNet (Pianta и др., 2002) копирует структуру англоязычного WordNet.

Также две разные программы действий провозглашают разработчики русских ворднетов (Сухоногов, Яблонский 2005; Азарова и др., 2003).

В этой главе будут рассматриваться в основном те проекты, которые пытаются творчески развить структуру создаваемых тезаурусов, обычно ставя своей целью улучшение их применимости в приложениях автоматической обработки текстов.

3.1. Общие принципы организации EuroWordNet

Первым проектом, который провозгласил цель построения ворднетов для нескольких европейских языков и в котором были сделаны попытки внести улучшения в структуру такого рода лингвистических ресурсов, был проект EuroWordNet, который включал в себя два этапа. На первом этапе (1996-1999) ворднеты создавались для голландского, испанского и итальянского языков. На втором этапе – для французского, чешского, немецкого и эстонского языков (Vossen, 1998; Vossen, 2003; Climent и др., 1996).

Поскольку проект EuroWordNet был многоязычным, то перед разработчиками стоял серьезный выбор, нужно ли стремиться к разработке языково-независимой структуры, с которой необходимо сопоставить единицы каждого языка, или, может быть, нужно иметь единую систему синсетов – новая единица в иерархической сети может быть включена, если хотя бы один язык из рассматриваемых имеет лексему или устойчивый оборот с таким значением.

По принятому в проекте решению каждый ворднет должен сохранять специфику своего языка. При этом каждый ворднет должен содержать отсылки на значения Принстонского WordNet, что позволяет сравнивать ворднеты, обнаруживать непоследовательности в построении ворднетов и видеть различия в устройстве разных языковых систем (рис. 3.1).

Одновременно в рамках проекта была создана небольшая классификация верхнего уровня, к которой должен был приписан каждый создаваемый ворднет**.**

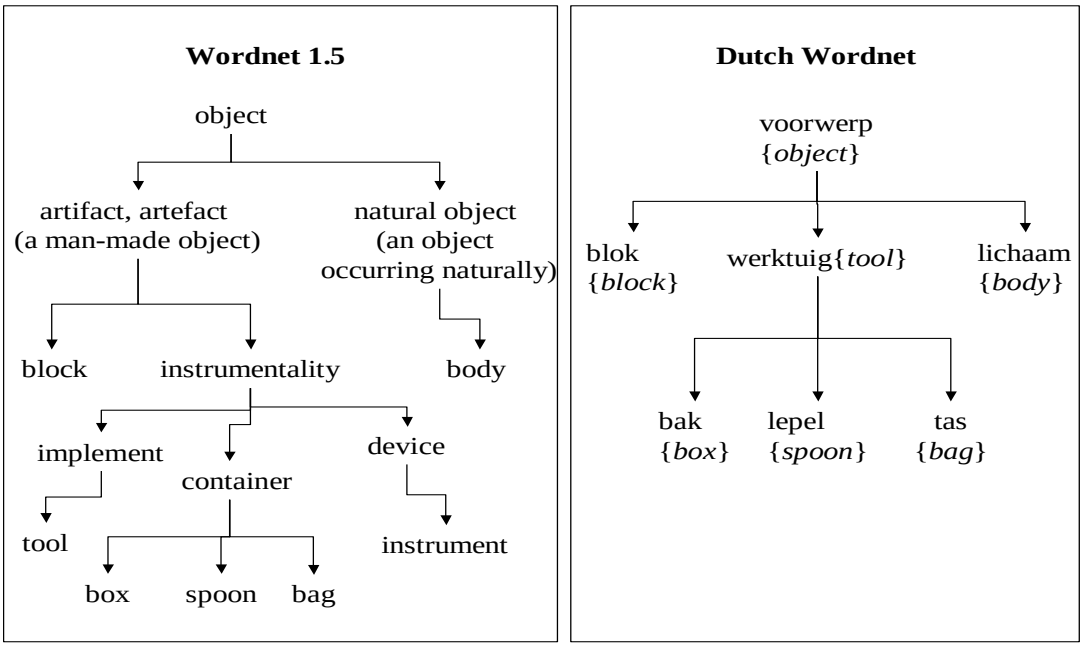

Рис.3.1. Различия в классификации объектов в англоязычном и голландском ворднетах (Vossen, 2003)

Основные предполагаемые применения ворднетов – это предсказание той или иной возможной замены лексических единиц в тексте для целей информационного поиска, генерации текстов, машинного перевода, разрешения лексической многозначности.

Отношения между лексемами должны выявляться в процессе применения классических лингвистических тестов (см. раздел 2.1. и Cruse, 1986).

Учитывая сложности, которые возникали при применении тезауруса WordNet в компьютерных приложениях, разработчики европейских ворднетов предложили ряд существенных нововведений в структуре создаваемых ворднетов.

Большой класс изменений касается описания отношений между синсетами, которые можно разделить на следующие группы:

- приписывание дополнительных атрибутов отношениям,
- введение отношений между частями речи,
- введение дополнительных отношений.

3.2. Отношения в EuroWordNet

3.2.1. Атрибуты дизъюнктивности/конъюктивности

Приписанные синсету отношения могут выполняться одновременно (возможна конъюнкция отношений) или выборочно (отношения дизъюнктивны). Так, обычно отношения «часть» конъюктивны – все части автомобиля одновременно составляют автомобиль. Гипонимы синсета обычно дизъюнктивны. Таким образом, обычно имплицитно предполагается конъюнктивность меронимов и гиперонимов, и дизъюнктивность гипонимов.

Вместе с тем могут возникать ситуации, когда явно полезно указать дизъюнктивность или конъюнктивность какой-либо совокупности отношений. Например, полезно иметь возможность отражения дизъюнктивность таких частей как пропеллер и реактивный двигатель у самолетов.

Для возможности отражения таких отношений между отношениями введены атрибуты отношений: ci - для отражения конъюнктивности, di - для отражения дизъюнктивности.

Тогда, фрагмент описания частей самолета можно выглядеть таким образом:

{ самолет

HAS PART: c1 *дверь*

HAS PART: c2d1 *реактивный двигатель*

HAS PART: c2d2 *пропеллер}*

Информация о том, что собака является и животным, и домашним питомцем записывается так:

{ собака

HYPERONYM: c1 *млекопитающее*

HYPERONYM: c1 *домашний питомец}*

Возможность нахождения дверей в разных объектах можно отразить так:

{ дверь

PART OF: d1 *автомобиль*

PART OF: d2 *помещение*

PART OF: d3 *вход}*

А то, что альбинос может быть животным или растением:

{альбинос

HYPERONYM: d1 *растение*

HYPERONYM: d1 *животное}*

Авторы EuroWordNet считают, что такое описание отношений позволит в некоторых случаях уменьшить число различных значений. Кроме того, такая возможность полезна для описания валентностей глаголов, то есть сочетаемости глаголов с другими словами в предложении.

3.2.2. Отношения между разными частями речи

Как уже указывалось, первоначально в Принстонском WordNet не были установлены отношения между различными частями речи.

Поскольку это вызывало серьезные проблемы в приложениях, в проекте EuroWordNet были введены дополнительные отношения между частями речи:

- xpos-synonymy – частеречная синонимия,
- xpos-antonymy – частеречная антонимия,
- xpos-hyponymy - частеречная гипонимия.

Таким образом, упомянутые в разделе 2.5.1 отношения между синсетами *adornment2 (процесс украшения*) и *adorn1 (украсить*) могли быть описаны отношением частеречной синонимии:

{adorn V} XPOS_SYNONYM *{adornment N}*

3.2.3. Новые отношения

Существенным дополнением в описание отношений между синсетами стало введение семантических отношений (ролей) таких как *агент, инструмент, объект, место* и обратные к ним отношения (Табл. 3.1). Подобные отношения в настоящее время вводятся и в Принстонский WordNet 3.0. (Clark, 2007).

{*hammer-молоток*}	ROLE_INSTRUMENT	{*to hammer – прибивать молотком*}
{*to hammer – прибивать молотком*}	INVOLVED_INSTRUMENT	*{hammer - молоток}*
{school - школа}	ROLE_LOCATION	*{to teach - учить}*
{to teach - учить}	INVOLVED_LOCATION	{*school - школа}*

Таблица 3.1. Примеры семантических ролей между синсетами в EuroWordNet

Кроме того, были введены отношения типа Co-role relations, которые выражают использование лексем из синсетов при описании ролей в одних и тех же ситуациях (Табл. 3.2.).

гитарист	HAS_HYPERONYM CO_AGENT_INSTRUMENT	*исполнитель* *гитара*
игрок	HAS_HYPERONYM ROLE_AGENT CO_AGENT_INSTRUMENT	*человек* *играть музыку* *музыкальный инструмент*
играть музыку	HAS_HYPERONYM ROLE_INSTRUMENT	*to make (создавать)* *музыкальный инструмент*

Таблица 3.2. Примеры отношений между синсетами, которые участвуют в одних и тех же ситуациях

3.2.4. Описание предметных областей (domains)

EuroWordNet включает в свою структуру также описание предметных областей – доменов. Это нововведение призвано преодолеть проблему WordNet, описываемую как теннисная проблема, когда принадлежащие одной предметной области, сфере деятельности, ситуации синсеты, оказываются далеко друг от друга в структуре WordNet (см.п. 2.5.3.1.).

Именно в рамках проекта EuroWordNet было предложено упоминавшееся решение, сгруппировать синсеты в домены. Предполагалось, что введение доменов должно быть особенно полезно для информационно-поисковых задач. Домены представляют собой отдельные объекты и могут быть организованы между собой в иерархии.

Эксперименты с доменами в ворднетах были продолжены и в следующем европейском проекте, связанном с ворднетами, Meaning (Atserias и др., 2004; Castillo и др., 2004).

3.2.5 Межъязыковой индекс ILI

Для того, чтобы установить связи между различными языками в проекте EuroWordNet, синсеты каждого ворднета имеют отсылку на так называемый межязыковой индекс (interlingual index - ILI), в качестве которого выбираются синсеты Принстонского WordNet. Индекс представляет собой неупорядоченный список синсетов с толкованиями.

Для наиболее точного описания соответствия конкретных синсетов каждого языка и преодоления лексических пропусков, которые могут возникнуть в том или ином языке, предоставляется возможность использования нескольких разных отношений эквивалентности от синсетов конкретного языка к индексу ILI:

- EQ_SYNONYM: имеется прямое соответствие между синсетом языка и синсетом индекса;
- EQ_NEAR_SYNONYM: синсету соответствует несколько синсетов индекса,
- HAS_EQ_HYPERONYM: синсет является более специфичным, чем имеющиеся синсеты индекса,
- HAS_EQ_HYPONYM: синсет может быть связан только с более специфичными синсетами индекса.

Так, испанское слово *dedo,* соответствующее русскогому слову *палец*, находится в отношении HAS_EQ_HYPONYM таким английским синсетам из индекса ILI как *toe (палец ноги) и finger (палец руки).*

3.3. Ворднеты для других языков

В данном разделе будут рассмотрены особенности представления лексической информации, предлагаемые разработчиками разных ворднетов.

3.3.1. Немецкий ворднет GermaNet

GermaNet является ресурсом, созданным по принципам WordNet, а не просто немецким вариантом синсетов Принстонского WordNet (Kunze, Wagner, 1999).

Характеристики GermaNet на дату апрель 2010 года: 61659 синсетов, 84586 лексических единиц, 76709 разных лексических единиц, отношений между синсетами - 73686 (http://www.sfs.uni-tuebingen.de/GermaNet/).

Особенностью описания существительных в GermaNet является ввод искусственных синсетов со специальной пометкой для объединения в отдельные классы гипонимов, разделяемых по одному и тому же признаку. Например, такими синсетами являются синсет *?Abstammender Mensch – Люди по происхождению*, или *?ausgebildeter_Mensch – Обученные люди*. Для того, чтобы включить понятие дилетанта – вводится еще одно понятие *?ausgebildeter_Mensch?*, что означает (Человек_по_образованию), которое, таким образом, разделяется на три гипонима: *учащиеся, обученные люди, необученные люди.*

Рассматривая примеры регулярной полисемии существительных (такие как вместилище – его содержимое, процесс - результат, место - жители) (см. раздел 2.5.2.1), разработчики ресурса указывают, что используют два метода ее описания в GermaNet:

- создание отдельных синсетов для каждого такого значения, что приводит к дополнительным значениям, которые необходимо автоматически разрешать при обработке текста,
- установление нескольких отношений гипоним – гипероним, но в таких случаях необходимо, чтобы все нижестоящие гипонимы имели такую же полисемию.

Для описания глаголов в GermaNet добавлено отношение каузации между глаголом и прилагательным, отражающим состояние, к которому приводит обозначаемое глаголом действие, например, *zerschleißen (изнашивать) – zerstört (изношенный).*

В отличие от WordNet в GermaNet используется множественная классификация глаголов (например, глаголы движения) классифицируются по субъекту движения, одновременно по свойству транзитивности, а также по направлениям движения, что делает сеть классификаций более плотной.

Кластерный подход описания прилагательных, предложенный в WordNet, изменен на иерархическую структуру описания прилагательных, подобно существительным и глаголам.

3.3.2. Датский ворднет DanNet

Разработка датского ворднета началась в 2005 году. В период до 2007 года планировалось разработать ворднет величиной 40 тысяч понятий, 30 тысяч понятий из которых соответствуют существительным (Pedersen, Sorensen 2006; Pedersen и др. 2006). Разработка DanNet базируется на толковом словаре современного датского языка DDO и семантическом лексиконе датского языка SIMPLE (Lenci и др., 2000, McShane и др., 2004).

Разработчики датского ворднета особое внимание обращают на построение правильной структуры таксономий, поскольку, как мы уже упоминали в разделе 2.5.3.2., одной из проблем Принстонского WordNet'а является смешение нескольких разных отношений под одним и тем же названием гипоним-гипероним.

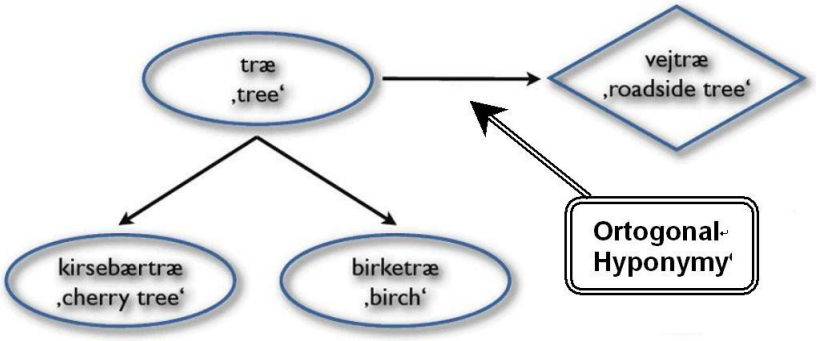

Рис.3.2. Отделение синсета «дерево у обочины» от основной иерархии в DanNet (Asmussen и др., 2007).

Для улучшения качества таксономии в DanNet авторы ресурса предполагают строже использовать диагностические высказывания для установления отношений гипоним-гипероним, а именно, устанавливать это отношение, если, действительно, можно сказать «*X – это тип Y*».

Авторы проекта обращают внимание на то, что потенциальные гипонимы, для которых не выполняется высказывание «*X – это тип Y*» (например, плохо звучит, что *roadside tree* (*дерево у обочины*) является типом дерева), коррелируют с введенными в (Cruse, 2002) понятием номинального типа.

Лексемы номинального типа (Cruse, 2002) в отличие от лексем естественного типа характеризуются тем, что это лексемы «одного свойства», то есть они характеризуются одним простым свойством, например, *"rattletrap" – колымага (об автомобиле), "getaway car" (автомобиль, на котором преступник скрылся с места преступления), "roadside tree" (дерево у обочины), tanning agent (дубильное вещество)*. Поэтому например, *дерево у обочины* не может рассматриваться как лексема естественного типа, несмотря на то, что относится к природным объектам. Среди гипонимов лексемы *человек* много единиц номинального типа, такие как *пассажир, читатель, идиот* и др.

Авторы подчеркивают, что включение таких единиц в таксономию делают ее запутанной, поэтому все такие единицы в датском ворднете описываются как единицы, ортогональные таксономии (см. рис. 3.2.). Считается, что «номинальные типы находятся

на том же уровне иерархии как и их гиперонимы, относительно таксонимов». Это позволяет отразить тот факт, что такие номинальные типы совместимы с таксонимами, например, самые разные типы автомобилей могут оказаться колымагами или использоваться для бегства с места преступления.

Для отличения номинальных типов предлагается использовать различные лингвистические тесты. Один из тестов, отражающий способность номинального типа быть совместимым с разными таксонимами, звучит следующий образом: «X – это любой Y, который …». Также предлагается использовать отрицательный тест «являться видом», так, например, странно звучит утверждение, что *пассажир является видом человека*.

Помимо лексем естественного типа разработчики предполагают описывать таксономические отношения между лексемами функционального типа, включая, например, такие, как названия профессий (*хирург – врач*).

Таким образом, в данном ресурсе сделаны серьезные усилия, чтобы выделить в группах гипонимов подгруппы несовместимых между собой таксонимов. Однако, представляется, что проблемой такого подхода является существование достаточно большого числа промежуточных случаев (например, лесное дерево может рассматриваться как номинальный тип, но не любое дерево может расти в лесу), сложность определения, что такое одно свойство, наличием гипонимов и таксонимов у номинальных типов (например, *пассажир* имеет такие гипонимы как *транзитный пассажир*, и таксонимы(?) *авипассажир, пассажир метро*). Возникает вопрос, оправдаются ли усилия, вложенные в различение всех этих типов, лучшей эффективностью обработки текстов на основе созданного ресурса.

Еще одной характеристикой таксономии в DanNet, на которую обращают внимание разработчики ресурса, является принадлежность синсетов и отношений между ними к общеупотребительной лексике. Разработчики DanNet поясняют свое решение на примере классификации домашней мебели: *стул > мебель_для_ сидения> мебель>объект*. При этом в области страхования для обозначения предметов домашнего обихода имеется термин *household effects (домашнее имущество)*, которое потенциально могло бы быть вставлено в цепочку таксономических отношений. Однако авторы ресурса считают, что этого делать не нужно, поскольку в классификациях необходимо придерживаться «интуитивной позиции неспециального использования языка» (Asmussen и др., 2007).

Для сравнения в Принстонском WordNet’е для класссификации, например, животных и растений, используется биологическая терминология из биологической систематики (Asmussen и др., 2007).

3.3.3 Компьютерный тезаурус русского языка RussNet

Компьютерный тезаурус RussNet, который разрабатывается на кафедре математической лингвистики Санкт-Петербургского государственного университета, строится на основании принципов, общих для wordnet-словарей (Fellbaum, 1997). Словарь RussNet является оригинальным ресурсом (Азарова и др., 2003) в том смысле, что он не переводится с Принстонского WordNet, а создается как отдельный ресурс.

В методологическом плане стандартная процедура построения RussNet включает следующие положения: (Азарова и др., 2005):

1) словарь опирается на корпус современных текстов 1985-2004 годов общим объемом около 21 млн. словоупотреблений, основу которого составляют газетные и журнальные статьи на темы повседневной жизни, экономики, политики, науки, культуры и спорта;
2) ядерная структура тезауруса задается примерно двумя тысячами наиболее частотных слов (существительных, глаголов, прилагательных, наречий), которые встречаются более 100 раз на миллион словоупотреблений в корпусе современных текстов;

3) разные значения некоторого слова, представленные в тезаурусе, упорядочены в соответствии с частотностью их употребления в корпусе текстов;
4) в RussNet представлена, как правило, общая, нетерминологическая лексика, хотя предполагается, что расширения базовой структуры будут включать терминологические элементы, которые тесно связаны с определенными тематическими областями;
5) синсеты национального тезауруса соотносятся с Межъязыковым лингвистическим индексом (ILI), предложенным в рамках проекта EuroWordNet – см раздел 3.2.5.

В структуру RussNet были внесены следующие нововведения по сравнению с другими ворднетами:

1) среди синонимов синсета выделяется доминантный синоним, представляющий собой наиболее нейтраальный и частотный способ выражения соответствующего лексического значения;
2) основным инструментом при разграничении значений слова в RussNet является контекстный анализ. При принятии решений о том, сколько и какие значения должны быть описаны у многозначного слова, выделяются статистически значимые маркеры, в качестве которых может выступать и определенная грамматическая форма, и принадлежность к некоторому семантическому дереву родовидовой иерархии RussNet, или оба этих показателя вместе. Эти признаки должны проявляться устойчиво: более чем в 33% контекстов для рассматриваемого значения в корпусе;
3) значения слов, частотность появления которых в корпусе составляет менее 1% контекстов для слова, считаются окказиональными (неустойчивыми) и не включаются в тезаурусное описание.

Для задания частотного упорядочения значений многозначного слова используется разметка выборочной совокупности контекстов корпуса. Эта процедура производится вручную, что приводит к большим затратам времени.

В проекте уделяется отдельное внимание вопросу включения в RussNet словосочетаний. Авторы ресурса считают, что хотя при включении в толковые словари словосочетаний обычно во внимание принимается несколько критериев (лексическая ограниченность, воспроизводимость некоторой конструкции в неизменном виде и др.), граница между свободными и устойчивыми словосочетаниями устанавливается довольно субъективно (Азарова и др., 2005).

При разработке RussNet разработчики опираются, прежде всего, на данные, полученные при статистической обработке корпуса текстов. Используется несколько показателей таких, как абсолютная частота сочетания слов, относительная частота сочетания слов (в частности, используются коэффициенты типа тест Стьюдента и MI-коэффициент (коэффициент взаимной информации)) на основе меняющегося контекстного диапазона - «окна» (Manning, Shutze, 1999).

3.3.4. Ворднет итальянского языка MultiWordNet

Ворднет итальянского языка MultiWordNet (http://multiwordnet.itc.it, Pianta и др.2002) в 2005 году включал 58 тысяч лексических значений итальянского языка, 41500 разных лемм, 32700 синсетов, для которых установлены соответствия с англоязычными синсетами Принстонского ворднета. MultiWordNet также включает разметку пометами предметных областей (см. п.2.5.3.1).

Основной моделью построения MultiWordNet является разработка его синсетов в максимально полном соответствии с синсетами Принстонского ворднета, то есть, итальянские синсеты вводятся на основе существующих английских синсетов, отношения между итальянскими синсетами импортируются из принстонского ворднета.

В тех случаях, когда для очередного англоязычного синсета в итальянском языке нет переводного эквивалента, или имеется приблизительный (более специфический или более общий эквивалент), то вводятся специальные «пустые» синсеты.

В связи с принятой технологией разработки отношения MultiWordNet практически полностью повторяют отношения принстонского WordNet, добавлено только отношение NEAREST соединяющее в случае лексической лакуны итальянский синсет с ближайшим по смыслу англоязычным синсетом (или синсетами).

Разработка ресурса выполнялась автоматизированным методом с помощью двуязычного словаря с последующей ручной проверкой.

MultiWordNet является одним из двух ворднетов итальянского языка.

В рамках проекта EuroWordNet создавался другой ворднет итальянского языка другой группой разработчиков под названием ItalWordNet (Roventini и др., 2000).

Основное различие между проектами разработчики MultiWordNet видят в более тесной связи своего ресурса с англоязычным ворднетом, когда для каждого итальянского синсета сразу устанавливается отношение с англоязычным синсетом.

Впрочем, разработчики понимают, что возможно такая модель разработки могла привести к заимствованию чуждых для итальянского языка лексических и семантических отношений, которые по необходимости будут устраняться.

3.3.5 Проект Meaning

Европейский проект Meaning является продолжением проекта EuroWordNet (Atserias и др., 2004), (Castillo и др., 2004).

Авторы проекта Meaning мотивируют необходимость продолжения работ в данном направлении тем, что десятки человек-лет были затрачены для создания ворднетов для разных языков, но этих усилий недостаточно, чтобы обеспечить качество многоязычных приложений компьютерной обработки текстов.

Прогресс в этой области связан с решением двух промежуточных задач: автоматическое разрешение лексической многозначности и масштабное обогащение лексических баз знаний.

Проблема, однако, заключается в том, что существуют взаимозависимые факторы:

1) для того чтобы достичь качественного разрешения лексической многозначности, необходимо значительно больше лингвистического и семантического знания, чем имеется в текущих лексических базах знаний (к примеру, в ворднетах),
2) для того чтобы обогатить существующие лексические базы знаний необходимо получать информацию из корпусов с качественной семантической разметкой.

В проекте планировалось выполнить три последовательных цикла масштабного разрешения лексической многозначности и извлечения знаний для пяти европейских языков, включая баскский, испанский, итальянский, голландский и английский языки.

Последовательные циклы работ должны были состоять из следующих этапов (Bentivogli и др., 2003; Atserias и др., 2004):

- разработка и обучение высокоточных автоматических систем разрешения лексической многозначности (см. главу 10) и разметка с помощью этих систем сверх больших корпусов,
- использование частично размеченных данных и лингвистических процессоров для пополнения знаний в ворднетах,
- дополнительное обучение систем разрешения лексической многозначности.

Данные работы должны привести к пополнению лингвистической информации на основе обработанного корпуса, а также к многоуровневой лингвистической аннотации самого корпуса. Накопленные знания должны храниться в Многоязычном Центральном Репозитории.

3.3.6. Словосочетания в WordNet и ворднетах других языков

Многие исследователи подчеркивают, что возможность ввода словосочетаний в систему синсетов очень важна и для отражения соответствий между синсетами разных языков, и для различных приложений автоматической обработки текстов.

Как уже указывалось, в качестве синсетов Принстонского WordNet'а включаются лексикализованные понятия, которые соответствуют значениям отдельных знаменательных слов и некоторых словосочетаний. Однако, как подчеркивают разработчики новых ворднетов, границы лексикализации многословных выражений очень трудно определить (Agirre и др., 2006). Указывается, что, с одной стороны, в синсеты Принстонского WordNet'а наряду с фразеологическими единицами регулярно включаются свободные словосочетания (Азарова и др., 2005б). С другой стороны, в ворднетах, создававшихся в рамках Европейских проектов EuroWordNet, BalkaNet, Meaning, введение синсетов на базе значений словосочетаний серьезно ограничивается (Bentivogli, Pianta 2003; Alonge и др., 1998, Agirre и др., 2006).

При введении в лингвистические ресурсы словосочетаний необходимо решить два вопроса.

Во-первых, нужно определить критерии ввода словосочетаний, поскольку невозможно описать и эффективно использовать все словосочетания, которые могут упоминаться в естественных текстах.

Во-вторых, нужно определиться, какого рода информацию, отношения с другими синсетами, необходимо приписывать синсетам, соответствующим словосочетаниям.

Для ввода словосочетаний обсуждаются такие критерии как (Calzolari и др., 2002; Bentivogli, Pianta, 2004; Sag и др., 2002; Азарова и др., 2005б):

- высокая частотность,
- высокая степень взаимной ассоциации,
- синонимичность словосочетания отдельной лексеме,
- существование переводных эквивалентов - отдельных слов в других языках,
- значительная многозначность слов-компонентов.

При выполнении для словосочетаний одного и более такого рода критериев решение остается за разработчиком ресурса, его лингвистической интуицией (Bentivogli, Pianta, 2003). Решение лингвиста обычно связано с тем, насколько большой объем знаний о мире, не выводимый из компонентов словосочетаний, ассоциируется с этим словосочетанием.

Чтобы дать возможность описывать в ворднетах необходимые словосочетания, в работе (Bentivogli, Pianta, 2004) предлагается вводить специальную структуру для представления свободных словосочетаний, которые авторы работы называют фразовым синсетом (phraseset) и которая может объединять множество синонимичных словосочетаний.

До введения таких структур в итальянском ворднете MultiWordNet при обнаружении лексических пропусков в итальянском языке по отношению к английскому языку заводился пустой синсет, снабженный комментарием, фразовые синсеты могут дать дополнительную важную информацию для работы с такими лексическими пропусками. Так, например, в итальянском ворднете MultiWordNet при установлении соответствия англоязычному синсету *toilet_roll (рулон туалетной бумаги, туалетный рулон)* создается пустой синсет, а также создается фразовый синсет. А для англоязычного синсета *dishcloth (полотенце для посуды)* в MultiWordNet имеются как синсет, так и фразовый синсет:

Примеры:

1) Eng_synset {*toilet_roll*}
   Ita_synset {*GAP*}
   Ita_phraseset {*rotolo_di_carta_igienica*}

2) Eng_synset {*dishcloth*}

Ita_synset {*canovaccio*}

Ita_phraset {*strofinaccio_dei_piatti, strofinaccio_da_cucina*}

Для описания внутренней структуры словосочетания разработчики MultiWordNet предлагают описывать отношение *composed-of* **(***состоять_из),* которое соединяет фразовый синсет со словами-компонентами.

Разработчики баскского ворднета (Agirre и др., 2006) вводят в свой ресурс пока только фразеологические словосочетания, которые зафиксированы в толковых словарях, и помечают введенные синсеты специальной отметкой. Для описания отношений синсета-словосочетания разработчики баскского ворднета предлагают использовать набор отношений INVOLVED, взятый из номенклатуры отношений EuroWordNet и определяемых следующим образом: отношение INVOLVED должно использоваться для описания аргументов сущностей 2 порядка (процессов, действий), например, как отношения involved_theme, involved_instrument и др (см. раздел 3.2.3.).

Текущая версия баскского ворднета включает 356 синтагматических синсетов. Итальянский ворднет MultiwordNet включает 1216 фразовых синсетов.

Таким образом, можно констатировать, что пока некоторого единого решения, как правильно поступать с включением словосочетаний в ворднеты, не выработано.

3.3.7. Общеупотребительная лексика и терминология предметных областей в тезаурусах типа WordNet

Разрабатываемые ворднеты естественных языков имеют своей целью описание общеупотребительного национального языка. Поэтому считается, что они должны содержать преимущественно общую лексику, и не должны включать термины отдельных предметных областей.

Однако в Принстонском WordNet можно обнаружить достаточно большое количество терминов из разных сфер деятельности. Ресурс содержит большое количество названий из биологической систематики (см. раздел 2.5.3.1), термины (инструменты, оборудование) из технической области, термины лингвистики и психолингвистики.

Это связано с тем, что разработчики Принстонского WordNet во многом пользовались уже готовыми классификациями и не контролировали содержания вводимых синсетов по текстовым корпусам.

При разработке следующих ворднетов большое внимание уделяется обоснованию выбора лексики, значений на основе корпусов своего языка.

Предполагается, что для применения созданного ресурса типа ворднет в конкретной предметной исходный ворднет должен быть расширен терминами предметной области, соответствующие синсеты должны быть встроены в иерархии ворднета. Причем высказывается предположение, что добавленные синсеты будут встраиваться на нижних уровнях построенных иерархий, как бы продолжая их (Magnini, Speranza, 2002).

Было создано несколько ворднетов в конкретных предметных областях: области архитектуры (Bentivogli и др., 2004), морского судоходства (Roventini, Marinelli, 2004; Marinelli, Tiberi, Bindi 2008), в юридической области (Sagri и др., 2004), в области медицины (Buitellar, Sacalenau, 2001), экономики (Magnini, Speranza, 2002).

3.4. Сравнение модели представления знаний в информационно-поисковых тезаурусах и тезаурусах типа WordNet

Рассмотрев основные принципы устройства информационно-поисковых тезаурусов и тезаурусов типа WordNet, можно сделать некоторые выводы о сходстве и различии используемых моделей представления знаний в этих тезаурусах.

Наиболее бросающееся в глаза различие состоит в том, что информационно-поисковые тезаурусы описывают определенную предметную область, а WordNet содержит информацию о значениях общей лексики языка. Однако это различие не является принципиальным, поскольку, как указывалось в предыдущем разделе, можно строить тезаурусы типа WordNet и для конкретных предметных областей.

Более значимые различия имеются в выборе единиц тезаурусов.

Как мы видели в главе 1, в информационно-поисковых тезаурусах имеется множество ограничений на включение в тезаурус языковых единиц: дескрипторы должны быть четко отделены по смыслу друг от друга, многозначность языковых единиц практически не представлена, ограничивается глубина иерархий и т.д. Это приводит к возникновению существенного расхождения между единицами тезауруса и языковыми единицами, упоминаемыми в текстах предметной области. В тезаурусах типа Wordnet такой разницы нет: если существует слово или выражение с определенными значениями, то оно включается в тезаурус в соответствующем количестве значений.

Существенно различным является подход к включению в эти два типа тезаурусов словосочетаний. Как мы указывали в разделе 1.1.2, в информационно-поисковых тезаурусах имеется достаточно подробный перечень правил, которыми должен руководствоваться разработчик тезауруса при вводе в тезаурус многословных дескрипторов. Разработчики WordNet заявляют о необходимости того, чтобы словосочетание было «лексикализовано» без уточнения критериев, а это, в свою очередь, приводит к тому, что ввод новых словосочетаний в WordNet, а особенно в тезаурусы типа Wordnet, создаваемые для других языков, серьезно ограничивается.

Если сравнивать систему отношений в стандартных информационно-поисковых тезаурусах и тезаурусах типа WordNet, то, прежде всего, нужно брать для сравнения отношения между синсетами существительных WordNet, поскольку дескрипторы информационно-поисковых тезаурусов – это обычно существительные и группы существительного.

Здесь мы видим сходство в небольшой величине набора отношений стандартного информационно-поискового тезауруса и Принстонского WordNet, что несомненно объясняется разнообразием описываемых сущностей. При этом однако в наборе отношений информационно-поискового тезауруса имеется отношение ассоциации, которое при всей высказанной по поводу его критике позволяет лучше описать отношения между сущностями предметной области, чем отношение «часть-целое» и «антонимии».

В последнее время в ряде работ отмечается, что и разработчики информационно-поисковых тезаурусов и разработчики ворднетов включают в свои тезаурусы более разнообразные наборы отношений между единицами (Soergel и др., 2004, Clark и др., 2008).

Заключение к главе 3

Задача разработчиков новых ворднетов для своих языков может показаться более легкой, чем задача разработчиков первого тезауруса WordNet, поскольку модель ресурса уже известна.

Однако в разработке новых ресурсов необходимо учесть критику Принстонского WordNet, удачи и неудачи в прикладных экспериментах. Поскольку было высказано много критических замечаний, каждый разработчик должен выбрать для себя наиболее необходимые изменения в структуре и составе своего создаваемого ворднета, что является непростой задачей.

Можно заметить, что по величине ворднеты других языков значительно меньше, чем Принстонский WordNet. Частично это объясняется тем, что Принстонский WordNet включает достаточно много специальной терминологии, особенно в области биологии (что можно видеть по количеству синсетов в домене биологии – более 20 тысяч- см.

раздел 2.5.3.1.), а также значительный блок синсетов именованных объектов – более 7.5 тысяч (см. раздел 2.5.3.2.).

Разработчики новых ворднетов включают лексику именно общеупотребительного языка, минимизируют включение синсетов, соответствующих именованным сущностям. Также во вновь создаваемых ворднетах значительно более ограничен ввод синсетов, базирующихся на значениях словосочетаний, чем в исходном Принстонском WordNet.

ЧАСТЬ 2. ФОРМАЛЬНЫЕ И ЛИНГВИСТИЧЕСКИЕ ОНТОЛОГИИ

Тезаурусы и рубрикаторы как формализованные информационные ресурсы известны достаточно давно. В последние 15 лет стал активно обсуждаться такой тип информационных ресурсов как онтологии. Часто можно слышать такие вопросы как «Чем тезаурусы и рубрикаторы отличаются от онтологий» или «Являются ли тезаурусы и рубрикаторы онтологиями». Читая статьи о таком ресурсе как WordNet, можно встретить ссылку на него как на тезаурус или как на онтологию. В данной части книги мы рассмотрим наиболее базовые вопросы, связанные с определением и созданием онтологий. Также будут рассмотрены соотношения между терминами *онтология, тезаурус, рубрикатор.*

Глава 4. Онтологии как ресурсы для представления знаний о мире

4.1. Определения онтологии

Слово «онтология» имеет два значения**:**

- **Онтология 1.** – Философская дисциплина, которая изучает наиболее общие характеристики бытия и сущностей;
- **Онтология 2.** – Артефакт, структура, описывающая значения элементов некоторой системы.

В данной книге мы будем использовать слово *онтология* во втором значении как некоторый компьютерный ресурс, представляющий собой некоторое описание взгляда на мир применительно к конкретной области интересов.

На формальном уровне, онтология - это система, состоящая из набора понятий и набора утверждений об этих понятиях, на основе которых можно строить классы, объекты, отношения, функции и теории.

Одно из самых известных определений онтологии, сформулированное Т. Грубером таково (Gruber, 1993):

Онтология – это точная спецификация концептуализации.

Концептуализация – это структура реальности, рассматриваемая независимо от словаря предметной области и конкретной ситуации. Например, если мы рассматриваем простую предметную область, описывающую кубики на столе, то концептуализацией является набор возможных положений кубиков, а не конкретное их расположение в текущий момент времени.

Более поздней модификацией определения Грубера является такое определение (Gomez-Perez и др., 2004):

Онтология – это формальная спецификация согласованной концептуализации.

Под согласованной концептуализацией подразумевается, что данная концептуализация не является частным мнением, а является общей для некоторой общности людей.

Сформулировано еще достаточно много разных определений онтологии (Клещев, Шалфеева; 2005). В работе (Guarino, Giaretta, 1995) было проанализировано семь различных определений онтологии и предложили следующее определение:

Ontology is a logical theory which gives an explicit, partial account of a conceptualization (Онтология – это формальная теория, ограничивающая возможные концептуализации).

При всем различии к определению онтологии многие авторы соглашаются в наборе основных компонентов онтологии.

Основными компонентами онтологии являются:

- классы или понятия;
- атрибуты;
- отношения;
- аксиомы;
- экземпляры.

Часто используется очень широкая трактовка **классов (понятий**) онтологии. При широкой трактовке утверждается, что классы (понятия онтологии) могут быть абстрактными и конкретными, элементарными и составными, реально существующими и воображаемыми. Другими словами, классом (понятием) может быть любая сущность, о которой может быть дана какая-либо информация (Corcho, Gomes-Perez; 2000).

Экземпляры (индивиды) представляют собой единичные сущности, принадлежащие классам онтологии.

Единицы онтологии (классы и экземпляры) могут иметь свойства - **атрибуты**. Каждый атрибут обычно имеет имя и значение, и используется для хранения информации, которая специфична для данной единицы.

Отношения представляют тип взаимодействия между понятиями области. Они формально определяются как подмножество произведения n множеств: R: C1 x C2…x Cn. Пример бинарного отношения – отношение часть-целое. Различие между отношениями и атрибутами заключается в том, что отношения связывают между собой два класса, а атрибут описывает внутренние свойства объектов посредством конкретных значений.

Наиболее важным среди отношений в онтологиях является так называемое таксономическое отношение (также известное как отношение класс-подкласс, родовидовое отношение, is-a отношение).

Аксиомы (правила вывода) используются, чтобы записать высказывания, которые всегда истинны. Они могут быть включены в онтологию для разных целей, например, для определения комплексных ограничений на значения атрибутов, аргументы отношений, для проверки корректности информации, описанной в онтологии, или для вывода новой информации.

Видно, что термину «онтология» удовлетворяет широкий спектр структур, представляющих знания о той или иной предметной области. В качестве в разной степени формализованных онтологий разными авторами рассматривается множество различных компьютерных ресурсов (Хорошевский, 2008; Welty и др., 1999; Клещев, Шалфеева, 2005; Obrst, 2003), в том числе и известных задолго до начала исследований по онтологиям таких как словари, рубрикаторы, тезаурусы.

4.2. Виды онтологий

Рассмотрим некоторые из типов онтологии в порядке от менее формализованных ресурсов к более формализованным ресурсам (Lassilla, McGuinness, 2001).

Уже **словарь** с определениями, глоссарий может рассматриваться как онтология с пустым множеством отношений (Гаврилова, Хорошевский, 2000; Хорошевский, 2002).

Простейшая модель онтологии с отношениями строится обычно на основе отношений класс-подкласс. Такие модели часто называются **таксономиями.**

Возможно построение онтологии и на других типах отношений, например, на основе отношения Часть-целое. В таком случае такая онтология называется **партономией.**

Рубрикаторы представляют собой иерархически организованные онтологии. При этом отношения между рубриками не сводятся к одному и тому же типу отношений, смысл отношений между разными рубриками может различаться.

Информационно-поисковые тезаурусы также рассматриваются как онтологические ресурсы. Такие тезаурусы имеют обычно таксономические отношения, а также ряд дополнительных отношений. Как мы уже указывали, часто в тезаурусах происходит совмещение под одним именем отношения ВЫШЕ-НИЖЕ разного рода отношений, то есть отношения устанавливаются не всегда формальным образом.

Тезаурусы типа WordNet, особенно классификация существительных, также рассматриваются как ресурсы онтологического типа. Как мы уже указывали, структура Принстонского WordNet достаточно интенсивно обсуждалась с формальных онтологических позиций. Некоторые изменения, вносимые в следующие версии этого ресурса, вызваны именно такого рода обсуждением, как, например, выделение из отношений гипонимии-гиперонимии отношений класс-экземпляр.

Часто возникает вопрос, можно ли кратко сформулировать основные особенности тезаурусов как вида онтологических ресурсов. Рассмотренные виды тезаурусов (тезаурус Роже, информационно-поисковые тезаурусы, тезаурусы типа WordNet) позволяют выделить следующие отличительные особенности этого вида онтологических ресурсов:

- единицы тезаурусов имеют тесную связь с естественным языком, обычно снабжаются вариантами их выражения на естественном языке;
- тезаурусы не имеют внутренней структуры понятий, то есть представления свойств и атрибутов в виде фреймов. Знания о мире, предметной области представлены в виде отношений между понятиями;
- аксиомы (правила вывода) сводятся к свойствам транзитивности и наследования.

Следующий тип онтологических моделей - это **модели с некоторым широким набором отношений**. Такие модели могут иметь или не иметь представление свойств и атрибутов понятий в виде фреймов. Для разных видов *отношений* может указываться кардинальность (соотношение количеств экземпляров связываемых сущностей) и модальность (возможность/ обязательность) связей.

Большей выразительностью обладают **онтологии, включающие ограничения на область значений свойств**. Значения свойств берутся из некоторого предопределенного множества (целые числа, символы алфавита) или из подмножества концептов онтологии (множество экземпляров данного класса, множество классов). Можно ввести дополнительные ограничения на то, что может заполнять свойство.

В целом, с необходимостью выразить больше информации, выразительные средства онтологии (и ее структура) усложняются. Например, может потребоваться заполнить значение какого-либо свойства экземпляра, используя математическое выражение основанное на значениях других свойств и даже других экземплярах. Многие онтологии позволяют объявлять два и более классов дизъюнктивными (непересекающимися). Это означает, что у данных классов не существует общих экземпляров.

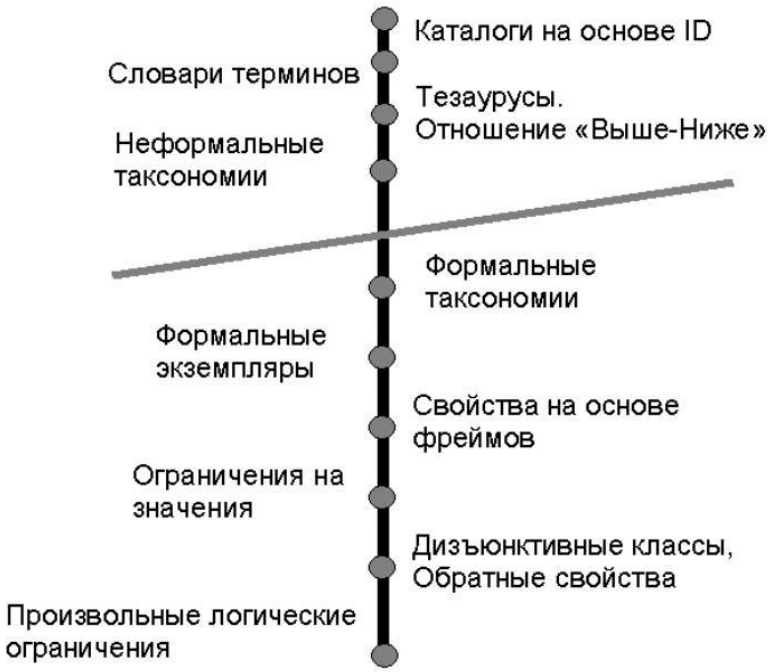

Рис. 4.1. Классификация онтологий в (Lassilla, McGuinness, 2001).
Косая черта разделяет системы, представляющие
«машино-понятные» и «человеко-понятные» описания

Наиболее формализованные онтологии представляют собой **логические теории**, построенные на произвольных логических утверждениях о понятиях – аксиомах. Для описания таких формальных онтологий применяются различные логики (дескриптивные логики, модальные логики, логика предикатов первого порядка) и различные языки описания онтологий DAML+OIL, OWL, CycL, Ontolingua.

Онтологии такие как тезаурусы, рубрикаторы, понятия которых не определяются полностью в терминах формальных свойств и аксиом, иногда называются **легкими онтологиями** (lightweight ontologies) (Gomez-Perez и др., 2001). Дж. Сова (http://www.jfsowa.com/ontology/ontoshar.htm) называет такие онтологии **терминологическими онтологиями.**

Приверженцы формальных подходов считают такие легкие онтологии не настоящими онтологиями, а ресурсами онтологического типа.

Для отражения спектра онтологий по степени формальности представления, использованию тех или иных формальных элементов часто используется диаграмма типа изображенной на рис. 4.1. Каждая точка соответствует наличию некоторых ключевых структур в онтологии, отличающих ее от других точек на спектре. Косая черта условно отделяет онтологии от других ресурсов, имеющих онтологический характер.

4.3. Два основных подхода к построению онтологий

В проектировании онтологий условно можно выделить два направления. Первое связано с представлением онтологии как формальной системы, основанной на математически точных аксиомах. Этот подход тесно связан с различными логическими формализмами (предикатов первого порядка, дескриптивной, модальной логики и т.п.). Это направление онтологических исследований является продолжением работ в рамках классического искусственного интеллекта, изучающих способы представления знаний.

Второе направление связано с разработкой онтологий для компьютерной обработки текстов. Онтологии дают возможность использовать знания о мире, которые необходимы для выполнения многих этапов анализа текста. При этом, с одной стороны, формальность описания в таких онтологиях значительно ниже, чем в онтологиях, создаваемых в рамках первого подхода. С другой стороны, формальный логический вывод на основе онтологий при анализе текста часто является необходимым, поскольку в связном тексте значительный объем информации не указывается явно (Леонтьева, 1981; Леонтьева, 2006; Chavez и др., 2009).

При всей кажущейся важности развития онтологий в рамках первого подхода, именно в рамках второго подхода создаются сверхбольшие ресурсы, используемые в широких предметных областях, в то время как в рамках первого подхода создаются относительно небольшие ресурсы (ресурсы с относительно небольшим числом понятий – экземпляров может быть достаточно много). Так, большое количество широкоизвестных медицинских онтологических ресурсов представляет собой тезаурусы, не обладающие высокой степенью формализации своей структуры (Gene ontology, 2009).

Так, в работах (Hepp, 2007; Hovy, 2005) указывается, что исследователи написали очень много работ о потенциальных преимуществах использования формальных онтологий, о необходимости использования онтологий в качестве центральных блоков семантической сети и других семантических систем. Однако количество и качество «реальных», «неигрушечных» онтологий, имеющихся на сегодняшний день, чрезвычайно мало, то есть не построено практически полезных онтологий для большого количества предметных областей.

Здесь часто можно встретиться с мнением, что отсутствие формальных онтологий большой величины происходит из-за того, что это «недалекие» бизнесмены не хотят понять, какие преимущества несет с собой использование формальных онтологий.

Однако, на самом деле, на пути создания масштабных формальных онтологий существуют реальные технические и социальные проблемы.

В работе (Tsujii, Ananiadou, 2005) указывается, что тогда как небольшие онтологии могут быть построены методом сверху-вниз, разработка подробных онтологий для реальных приложений – нетривиальная задача. Более того, во многих предметных областях, знание, нужное для распространения и интеграции, содержится в основном в текстах. Из-за внутренних свойств человеческого языка, непростой задачей является

связать знания, содержащиеся в текстах, с онтологиями, даже если бы они были построены для данной предметной области. То есть предполагается, что такие однозначные и последовательные концептуальные модели играют менее значительную роль в распространении знаний, чем предполагают сторонники формального онтологического подхода.

В работе (Hepp, 2007) описываются следующие существенные проблемы на пути развития формальных онтологий.

Во-первых, подавляющее число предметных областей продолжает развиваться, пополняться новыми понятиями, отношения между некоторыми понятиями меняются. Создаваемые онтологии будут всегда отставать от существующего понятийного аппарата предметной области. Чем более подробной является онтология (а для практической применимости она должна быть подробной), тем больше динамика ее изменений. Отсутствие новых понятий в онтологии не позволяет использовать семантические технологии для поиска по запросам, включающим новые понятия, или аннотирования документов.

Во-вторых, создание онтологий требует серьезных ресурсов. Для того, чтобы затраты были оправданы, требуется применимость созданных онтологий пользователями. Должны возникнуть реальные пользователи, которые должны поверить в полезность онтологии и начать ее применять, что достаточно сложно на первых этапах появления онтологии.

Третьей проблемой является проблема понятности онтологии для пользователей так, чтобы она могла правильно применяться и интерпретироваться (Fox, Gruningen, 1997). На основе спецификаций и документации онтологии пользователи должны правильно интерпретировать семантику всех ее элементов. Кроме того, как показывает практика, далеко не всякий специалист в предметной области может хорошо разбираться в формальных онтологических спецификациях. Чем больше степень формализованности онтологии, тем труднее ее понять пользователю.

Также и Джон Сова в онтологическом форуме (*ontolog.cim3.net/forum/ontolog-forum/2008-12/msg00015.html)* высказывается по поводу предполагаемой в проекте Семантическая сеть (Semantic Web) разметки сайтов семантическими тегами, для обеспечения более качественного поиска информации в Интернет и обращает внимание на следующий вопрос: если теги формально определены, как можно быть уверенным, что люди, которые используют эти теги, реально прочитали и поняли формальные определения?

Если пользователи будут проставлять теги несколько различным образом, то в условиях примения процедур формального вывода это может привести к противоречиям: «Если от 5.5 до 33% данных может оказаться неправильными, то утверждения о необходимости формальной точности в аксиомах и процедурах доказывания оказываются под вопросом» (там же).

В работе (Hepp, 2007) приводится следующий рисунок (см. рис. 4.2.), который показывает, что чем больше формальная выразительность онтологии, тем меньше потенциальный круг ее пользователей, поскольку пользователям трудно понять описания онтологии для того, чтобы применить ее в своей деятельности.

Таким образом, вопрос о создании и качественном применении больших строго формальзованных онтологий является достаточно сложным, что связано как со сложностью создания таких ресурсов, так и со сложностью их понимания, применения, описания с их помощью реальных материалов. Вышесказанное не означает, что можно пренебречь любой степенью формализации, поскольку неформализованный ресурс сложно использовать в автоматических режимах работы компьютерных приложений, а непоследовательность описаний сущностей ведет к нарушению процедур логического вывода. Осознавая описанные проблемы, каждый разработчик онтологий должен иметь в виду, что существует ряд противоречивых требований к онтологии (формальная

строгость, практическая применимость, величина, понятность пользователям), и осознанно делать свой выбор.

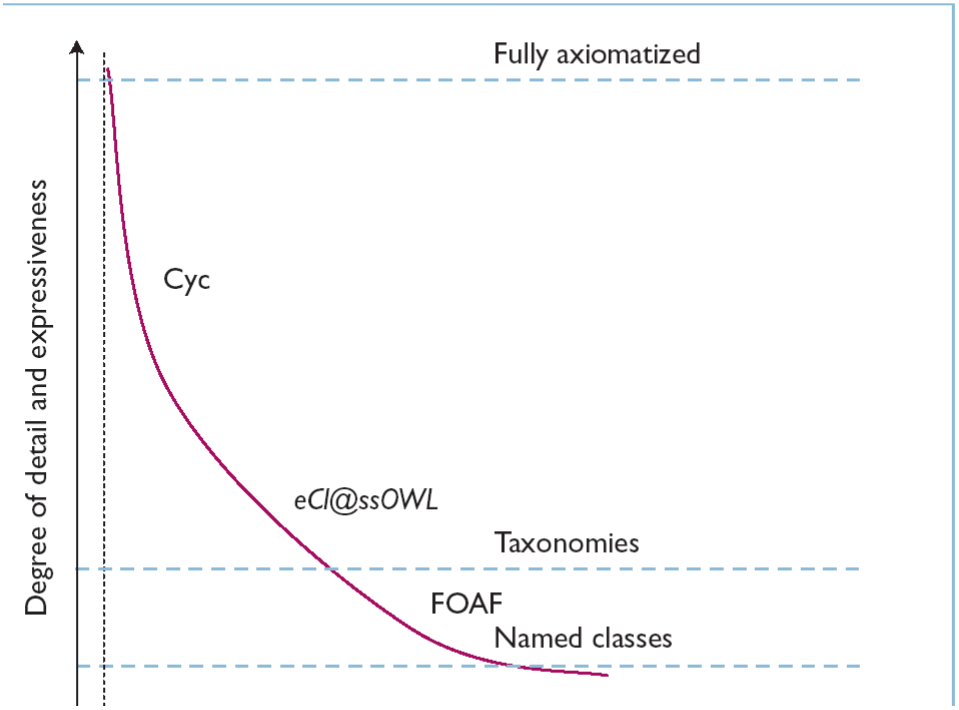

Рис.4.2. Соотношение между степенью формальности онтологии и величиной сообщества, которая может ее использовать (Hepp, 2007).

4.4. Принцип независимости онтологии от естественного языка. Лингвистические онтологии

Одним из важных вопросов формальной организации онтологий, является вопрос о связи единиц онтологии со значениями естественного языка. Часто заявляется, что формальные онтологии должны быть независимы от естественного языка.

Авторы работы (Mahesh, Nirenburg, 1995) считают, что онтология должна быть независима от конкретного естественного языка в двух аспектах:

- она не содержит единиц, специфичных для того или иного языка, хотя названия понятий для удобства могут быть даны на естественном языке;
- понятия онтологии не имеют взаимнооднозначного соответствия со значениями слов конкретных языков. Многие понятия онтологии не могут быть сопоставлены ни одному конкретному слову в языке, другие понятия могут соответствовать многим словам в языке и наоборот.

Doug Lenat (Lenat и др., 1995), руководитель известного проекта в области представления знаний Cyc, в рамках которого предполагалось формализовать знания здравого смысла (common sense) и использовать их, в частности, для обработки текстов на естественном языке, считает, что учет значений слов может только запутать ("words are often red herrings"), что значения слов делят мир неоднозначно, а линии деления происходят из самых различных причин: исторических, физиологических и т.п.

Однако в настоящее время некоторые исследователи (Brewster и др., 2005) указывают на следующий факт: «несмотря на то, что все авторы статей по онтологиям подчеркивают, что понятия являются кирпичиками любой онтологии, мы манипулируем понятиями посредством слов. Во всех онтологиях, которые известны, слова используются, чтобы представлять понятия. Следовательно, то множество явлений в мире, которые не вербализованы, не могут быть смоделированы. Мы можем описать это явление как

Онтологическая гипотеза Сепира-Уорфа, то есть то, что не описывается словами, не может быть отражено в онтологии…».

Как было уже упомянуто, утверждая независимость онтологии от языка, разработчики для называния понятий и отношений онтологии могут использовать слова естественного языка. Это стандартная практика, которая использовалась еще в системах представления знаний, создаваемых в рамках работ в сфере искусственного интеллекта (McCarthy, Hayes, 1969). Несмотря на мнение Дага Лената о значениях слов, в CYC также названия многих сущностей носят явно языковой характер, например, понятие концепт #$Skin (кожа), #$FemaleAnimal (самка), #$mother (предикат *мать*) и многие другие.

Y.Wilks (Wilks, 2002) указывает, что предикаты в языках представления, которые выглядят как слова естественного языка, и есть слова естественного языка и странно, что этот факт так яростно отрицается. Таким образом, подчеркивает Y.Wilks, символы в языках представления фундаментально базируются на естественном языке, что язык представления - это средство человеческой коммуникации с присущим ему динамизмом, многозначностью и возможностью расширенного толкования. Так, ключевые предикаты длительного проекта CYC за годы развития проекта изменили свои значения (Wilks, 2008). Этот факт не учитывается в сообществах разрабочиков онтологий таких как Семантическая сеть, которые напрасно верят, что их онтологии достигнут точности в значениях терминов (понятий и т.п.) (Nirenburg, Wilks, 2001).

С учетом вышеприведенных мнений уже не таким парадоксальным оказывается понятие так называемой лингвистической онтологии, то есть онтологии, понятия которой в значительной мере связаны со значениями языковых единиц, терминов предметной области (Gomez-Perez и др., 2001; Magnini, Speranza, 2002).

Лингвистические онтологии охватывают большинство слов языка или предметной области, и одновременно имеют онтологическую структуру, проявляющуюся в отношениях между понятиями. Поэтому лингвистические онтологии могут рассматриваться как особый вид лексической базы данных и особый тип онтологии. При этом лингвистические онтологии являются относительно слабо формализованными, то есть являются «терминологическими» онтологиями по определению Дж. Совы.

Примерами лингвистических онтологий являются Принстонский WordNet и ворднеты других языков (см. главы 2, 3). Также примерами лингвистических онтологий являются информационно-поисковые тезаурусы, поскольку их единицы – дескрипторы - в подавляющем большинстве основываются на реальных терминах предметной области

4.5. Онтологии и автоматическая обработка текстов

Как уже указывалось, для того, чтобы сделать автоматическую обработку текстов более качественной и надежной, необходимо использовать знания и о языке, и об окружающем мире. Знания о мире могут быть представлены с помощью онтологий - систем понятий, для которых описаны отношения и заданы правила вывода (Нариньяни, 2001; Рубашкин, Лахути, 1998; Рубашкин, Лахути, 1999).

Чтобы применить онтологию для автоматической обработки текстов, в частности для решения задач информационного поиска, необходимо понятиям онтологии сопоставить набор языковых выражений (слов и словосочетаний), которыми понятия могут выражаться в тексте.

Процедура сопоставления понятий онтологий и языковых выражений может быть осуществлена различными способами:

Во-первых, онтология может быть сделана заранее, путем логической классификации, а затем к ее единицам могут быть приписаны языковые единицы (Gruber, 1993). При этом предлагается создавать онтологию путем логического анализа, «сверху-вниз». Имена вводимых понятий (желательно) должны отражать те признаки, которые заложены в основу деления. В результате получаются имена понятий достаточно

громоздкие, неестественные, с ними трудно оперировать как разработчикам, так и возможным пользователям.

Другой проблемой такого подхода является то, что при приписывании языковых выражений к логически обоснованной системе понятий получается, что одно и то же слово может соответствовать слишком большому количеству таких «правильных» понятий в зависимости от контекста, возникает излишняя многозначность лексической единицы.

Поскольку в настоящее время существуют тезаурусы типа ворднет, содержащие большой объем лексической информации, то активно обсуждаются методы автоматического приписывания некоторой формальной онтологии языковых единиц из этих тезаурусов (Reed, Lenat, 2002; Pazienza, Stellato, 2006, Peter и др., 2006; Prevot и др., 2006).

Лингвистические онтологии отличаются от формальных онтологий по степени формализации. Поэтому второй путь предполагает, что разработчики такого рода ресурсов разрабатывают иерархию лексических значений естественного языка, а для более строгого описания знаний о мире необходимо снабдить эти ресурсы отношениями из формальных онтологий.

Так, содержанием одного из проектов является установление отношений между WordNet, с одной стороны, и формальной онтологией верхнего уровня SUMO – Standartized Upper Merged Ontology, с другой стороны (Niles, Pease, 2003). Проект состоит в том, чтобы установить соответствие между синсетами WordNet и понятиями онтологии, при котором каждый синсет WordNet либо напрямую сопоставляется с понятием онтологии, либо является гипонимом для некоторого понятия, либо примером понятия онтологии.

Участники другого проекта OntoWordNet (Gangemi и др., 2003) считают, что недостаточно провести формальную склейку ресурса типа WordNet и формальной онтологии, необходима значительная реструктуризация исходного лексического ресурса.

Третий путь – попытаться разработать единый ресурс, в котором были бы сбалансированы обе части: система понятий – и система лексических значений, что заключается в разумном разделении этих единиц в создаваемом ресурсе и аккуратном описании их взаимосвязей (Mahesh, Nirenburg, 1996, Nirenburg, Raskin, 2004, Hirst, 2003). При создании такого сбалансированного ресурса ввод понятий в онтологию требует непременного учета существующих лексических значений, то есть необходимо создавать сбалансированный ресурс, который должен являться лингвистической онтологией.

Таким образом, мы видим, что все обсуждаемые в настоящее время основные пути адаптации созданных формальных онтологий к приложениям автоматической обработки текстов включают в себя сопоставление этих онтологий с лингвистическими онтологиями.

В следующих разделах мы опишем лингвистические ресурсы MicroKosmos и FrameNet, которые также могут рассматриваться как лингвистические онтологии и которые понадобятся нам в дальнейшем рассмотрении.

4.5.1. Онтология Microkosmos

Онтология МикроКомос (более позднее название OntoSem) является одним из известнейших онтологических ресурсов. Эта онтология разрабатываются в рамках подхода, называемого «онтологическая семантика» (Nirenburg, Raskin, 2004). Онтология предназначается для использования в приложениях автоматической обработки текста и построению семантического, языково-независимого представения содержания предложений текста. Для поступающего текста производится предобработка, морфологический анализ, синтаксический анализ, семантический анализ, результаты которого представляются как Представление текст-смысл (Text-Meaning Representation - TMR).

Все сущности в онтологии Микрокосмос разделены на объекты, события и свойства. Объекты, события и свойства являются концептами (понятиями) онтологии, которые описываются фреймами. Фреймы – это наборы слотов с одним или более фасетов. Слоты в совокупности описывают понятия, определяя, как данное понятие соотносится с другими понятиями онтологии (посредством отношений) и буквенным и числовым ограничениям (посредством атрибутов). Лексикон системы описывает значения слов и словосочетаний, устанавливая ссылки от них на понятия онтологии.

Каждый слот – атрибут или отношение – определен как понятие в своей собственной иерархии.

Основными особенностями онтологии являются**:**

- независимость от конкретного естественного языка;
- независимость мотивации. Добавление понятий в онтологию не диктуется лексиконом языка. Развитие онтологии и пополнение лексикона системы - два равноправных взаимодействующих процесса, которые помогают друг другу и в то же время ограничивают друг друга;
- хорошая структурированность;
- последовательность и сочетаемость с лексиконом, семантическим анализатором и т.п.;
- понятность и простота. Онтологию должно быть легко обходить и представлять. Например, онтология не использует And-Or деревья с дизъюнктивным наследованием, поскольку такое наследование достаточно трудно воспринимать.

Имена в онтологии могут выглядеть как английские слова или фразы, но их семантика отличается и выражается набором четко определенных отношений между понятиями.

Понятие языковой зависимости (независимости) значения демонстрируется на примере существования в немецком языке слова s*chimmel – белая лошадь*. Авторы онтологии подчеркивают, что нет необходимости вводить отдельное понятие для отражения значения данного слова, для описания значения этого слова правильнее ввести словарную статью с ссылкой на понятие ЛОШАДЬ и с описанием значения свойства ЦВЕТ - «белый».

Словарная статья языкового значения в онтологии может иметь простую структуру, представляя собой ссылку на понятие онтологии, и достаточно сложную структуру, содержащую и ссылку на понятие онтологии и особенности конкретной лексической единицы (Nirenburg и др., 2004: Nirenburg, Raskin, 2004).

Например, все глаголы изменения в онтологии приписаны одному и тому же понятию Change-event. Особенности слов описываются в словарной статье, например, для глагола *увеличить (increase*) указывается, что в семантической роли ТЕМА этого глагола должна выступать СКАЛЯРНАЯ_ВЕЛИЧИНА (например, цена или высота) и указывается, что значение этой величины меняется на большее.

Значение слова *сионист* представлено в словаре как POLITICAL ROLE (политическая роль), которая является агентом (AGENT_OF) a SUPPORT_EVENT, темой которого является Израиль. Значение слова *асфальтировать* описывается как COVER_EVENT (событие покрывания), инструментом которого является понятие *АСФАЛЬТ.*

Авторы указывают, что нет необходимости иметь отдельные понятия для описания значений слов s*ibling (родные брат или сестра), brother (родной брат), sister (родная сестра).* Вводится одно понятие SIBLING, и с помощью значений атрибута *gender (мужской или женский пол)* в словаре системы могут быть описаны значения слов *sister и brother.*

Поскольку авторами сделаны значительные усилия по ограничению величины онтологии, то размер онтологии МикроКосмос (OntoSem) составлет порядка 6 тысяч

понятий, каждое из которых описывается в среднем 16 свойствами. Лексикон системы составляет несколько десятков тысяч слов и выражений.

Основные этапы разработки онтологии, по мнению разработчиков, должны состоять в следующем:

1) установление того, является ли значение слова достаточным основанием для введения нового понятия,
2) нахождение места понятия в онтологии, определение того, какие существующие понятия онтологии могут служить наилучшими родовыми понятиями для нового понятия;
3) описание свойств нового понятия, которые должны отличаться от свойств родовых понятий, видовых понятий, не только заполнением слотов, но и более содержательным образом, наличием других свойств.

Таким образом, провозглашаемая языковая независимость не должна вводить в заблуждение. По своей сути онтологии OntoSem и MikroKosmos являются, несомненно, лингвистическими онтологиями, поскольку основным принципом, обосновывающим введение новых понятий, является существование слов с таким значением в большом количестве языков.

При этом принцип языковой независимости этих онтологий подчеркивает, что при построении лингвистической онтологии необязательно жесткое следование системе значений конкретного языка. Лингвистическая онтология может учитывать систему значений конкретного языка или совокупности языков, и при этом следовать принципам введения понятий, провозглашаемых в формальных онтологиях (см. главу 5).

4.5.2. FrameNet как лингвистическая онтология

Одним из известных в настоящее время проектов в области описания лексической семантики является лингвистический ресурс FrameNet, который создавался под руководством известного лингвиста Чарльза Филмора (Fillmore, Atkins, 2000; Fillmore и др., 2003) в рамках концепции фреймовой семантики. Цель проекта – создать онлайновый лексический ресурс, основанный на фреймовой семантике, и обеспечить его базой в виде текстового корпуса. Проект направлен на описание семантической и синтаксической сочетаемости слов – валентностей – для каждого слова в каждом известном смысле.

В 2009 году ресурс содержал 960 иерархически организованных фреймов, с которыми ассоциировано более 11 тысяч лексических единиц.

Например, фрейм APPLY_HEAT (НАГРЕВАНИЕ ЕДЫ) описывает ситуацию, в состав которой входят такие слоты, как ПОВАР, ЕДА, НАГРЕВАТЕЛЬНОЕ ОБОРУДОВАНИЕ. Данный фрейм вызывается такими словами как *bake, blanch, boil, broil, brown, simmer, steam*, etc. Слоты фрейма называются фреймовыми элементами FE, а вызывающие фрейм слова – лексическими единицами (LU) этого фрейма. В качестве корпусных данных для этих описаний размечено более 135 тысяч предложений

По сути FrameNet представляет собой онтологию ситуаций, представленных в виде фреймов и связанных между собой иерархическими отношениями. FrameNet – это, несомненно, лингвистическая онтология, поскольку для описания нового фрейма необходимым условием является существование лексических единиц, которые вызывают этот фрейм.

Основными иерархическими отношениями между фреймами являются следующие:

Отношение **Is_A** устанавливается в тех случаях, когда каждый фреймовый элемент родительского фрейма связан с соответствующим элементом нижестоящего фрейма. Например, фрейм МЕСТЬ (REVENGE) является нижестоящим для фрейма REWARDS_AND_PUNISHMENTS (НАГРАДЫ И НАКАЗАНИЯ).

Отношение **Using** указывается, если нижестоящий фрейм предполагает родительский фрейм как бэкграунд, например, фрейм СКОРОСТЬ предполагает фрейм

ДВИЖЕНИЯ, однако не все фреймовые элементы родительского фрейма должны быть связаны с фреймовыми элементами нижестоящего фрейма.

Отношение **Subframe** описывает нижестоящий фрейм как подсобытие вышестоящего события, например, фрейм КРИМИНАЛЬНЫЙ ПРОЦЕСС имеет подфреймы АРЕСТ, СУД, и ПРИГОВОР.

Отношение **Perspective on** показывает, что нижестоящий фрейм описывает точку зрения вышестоящего, не ориентированного на определенные точки зрения фрейма. Например, фреймы НАНЯТЬ_НА_РАБОТУ и ПОЛУЧИТЬ_РАБОТУ являются такими подфреймами для фрейма ТРУДОУСТРОЙСТВО (EMPLOYMENT_START) с точки зрения нанимателя и работника соответственно.

Также используются отношения предшествования **Precedes**, отношение причины **Causative_of.**

4.5.3. От информационно-поисковых тезаурусов к формальным онтологиям

Рассмотрим, какое влияние оказали современные онтологические исследования на концепцию разработки информационно-поисковых тезаурусов. Критика информационно-поисковых тезаурусов с позиции формальных онтологий в связи с тем, что в тезаурусах недостаточно хорошо структурированы отношения и отсутствует последовательность в их установлении, привели к постановке вопроса о возможности преобразования информационно-поискового тезауруса в более формальзованный онтологический ресурс (Wielinga и др., 2001).

Задача преобразования информационно-поисковых тезаурусов в формальные онтологии была поставлена разработчиками достаточно известных тезаурусов таких, как тезауруса в области сельского хозяйства AGROVOC и тезауруса в области образования ERIC (Soergel и др., 2004). Предполагается, что при таком преобразовании могут быть улучшены разнообразные функции использования информационно-поисковых тезаурусов, включая:

- более качественное взаимодействие с пользователями, помощь в формулировании запросов;
- интеллектуальное расширение запросов;
- автоматизированная помощь индексаторам и база для систем автоматического индексирования и рубрицирования текстов,
- поддержка для приложений, создаваемых в рамках искусственного интеллекта, и исследований в области Семантической сети.

Предлагается, прежде всего, преобразовать систему отношений тезауруса в более формализованный набор предикатов и описать правила вывода (аксиомы).

Так, например, в работе (Soergel и др., 2004) в качестве примеров модификации информационно-поискового тезауруса по сельскому хозяйству AGROVOC приводятся следующие словарные статьи (AGROVOC, 1999):

Исходные статьи тезаруса (NT – отношение НИЖЕ, BT - отношение ВЫШЕ):

milk
    NT cow milk
    NT milk fat

cow
    NT cow milk

Cheddar cheese
    BT cow milk

Указанные статьи действительно показывают смешение разных отношений, что не противоречило существующим стандартам в области разработки информационно-поисковых тезаурусов. Как мы видим, отношения между понятиями *МОЛОКО-КОРОВЬЕ-МОЛОКО, МОЛОКО-МОЛОЧНЫЙ ЖИР, КОРОВА – КОРОВЬЕ МОЛОКО, МОЛОКО – СЫР ЧЕДДЕР,* выражено одним и тем же отношением BT-NT.

Преобразованные словарные статьи должны более четко различать конкретное семантическое отношение, и, таким образом, один тип тезаурусных отношений преобразуется в четыре разных отношения:

milk
<*includesSpecific*> cow milk
<*containsSubstance*> milk fat

cow
<*hasComponent*> cow milk

Cheddar cheese
<*madeFrom*> cow milk

Тезаурусы обычно не содержат понятий с атрибутами, в проекте преобразования тезауруса AGROVOC в онтологию предполагается использовать атрибутную структуру для описания некоторых понятий, например, описывать, что понятие МАТЬ это РОДИТЕЛЬ с атрибутом *женский.*

На построенной системе отношений предполагалось ввести правила вывода, например:

Правило 1:
Part_X <*mayContainSubstance*> Substance_Y:
IF Animal_W <*hasComponent*> Part_X
AND Animal_W <*ingests*> Substance_Y

(Если животное W имеет в качестве компонента часть X, и животное W съедает вещество Y, то часть X может содержать вещество Y).

Правило 2:
Food_Z <*containsSubstance*> Substance_Y**:**
IF Food_Z <*madeFrom*> Part_X
AND Part_X <*containsSubstance*> Substance_Y

(Если пища Z делается из части X, и часть X содержит вещество Y, то пища Z содержит вещество Y).

Предполагается, что система, имея такие правила вывода, может автоматически получить, что сыр-чеддер содержит (*containsSubstance**)*** молочный жир, и, что если коровы на ферме съели корма, зараженные ртутью, то, сыр, сделанный из этого молока, также, возможно, будет заражен ртутью (*Cheddar cheese <mayContainSubstance>mercury).*

Другой пример преобразования информационно-поискового тезауруса в формализованную онтологию – это нововведения, предлагаемые для тезауруса в области образования ERIC (Thesaurus of ERIC, 1990).

Исходные статьи тезауруса:

Reading instruction
BT *Instruction*
RT *Reading*
RT *Learning standards*

Reading ability
BT *Ability*
RT *Reading*
RT *Perception*

Предполагается преобразовать в следующий набор отношений:

Reading instruction	isa	*Instruction*
Reading instruction	has domain	*Reading*
Reading instruction	governed by	*Learning standards*
Reading ability	isa	*Ability*
Reading ability	has domain	*Reading*
Reading ability	supported by	*Perception*

А также предполагается установить следующие правила вывода:Правило 1

If X isa (type of) instruction and X has domain Z
and Y isa ability and Y has domain Z
Then X should consider Y

(Если X - это инструкция, и X имеет область Z, и Y – это способность, и Y имеет область Z, тогда инструкция X должна учитывать Y.

Правило 2:

If X should consider Y and Y is supported by W
Then X should consider W.

(Если X должно учитывать Y, и Y поддерживается W, то X должно учитывать W).

Проект преобразования тезауруса AGROVOC в онтологию действительно стал реализовываться (Liang и др., 2006). Речь идет об автоматизированном преобразовании исходного набора тезаурсных отношений в онтологические отношения. Всего предложено более 70 отношений между понятиями тезауруса (http://www.fao.org/aims/cs_relationships.htm).

Взаимосвязь между некоторыми отношениями вызывает вопросы и, например, отношение таксономии taxonomic relationships и отношение класс-подкласс hasSubclass указаны в списке как отдельные отношения. Как известно, большое количество отношений между сущностями, тем более плохо определенных, влечет дополнительные проблемы с последовательностью и субъективностью их установления. Про реализацию описания правил ввода и использование их в автоматических процедурах обработки текста пока ничего не известно.

Кроме того, на пути применения таких «информационно-поисковых онтологий» в реальных приложениях информационного поиска и автоматической обработки текстов в широких, плохо структурированных предметных областях (какими, собственно, и являются предметные области «Сельское хозяйство» и «Образование») имеются определенные трудности.

Действительно, чтобы правила логического вывода действительно работали, помимо изменений в описании понятий и терминов предметной области, нужно иметь автоматические средства обработки естественно-языковых текстов, позволяющие в неограниченном связном тексте точно и полно извлекать последовательности фактов, уметь прослеживать кореферентность, следить за временем извлекаемых фактов: в корма попала ртуть, эти корма принадлежат данной ферме, коровы этой фермы съели именно эти корма, изготовление сыра чеддер этой фермой произведено в период времени сразу после того, как эти коровы съели эти корма и т.п.

Кроме того, в тексте слова *корма* и *ртуть* могут оказаться в разных частях длинного предложения, или в разных предложениях текста, например, из-за использования эллиптической конструкции или местоимения и т.п., что значительно усложнит выявление этого факта.

Понятно, что в настоящее (и ближайшее) время ни одна из существующих систем автоматической обработки текстов, извлечения знаний из текстов не может обеспечить такой уровень точности и полноты получения информации из текстов, на которых надежно можно было обосновывать работу таких правил вывода.

Таким образом, по нашему мнению, значительные трудозатраты на такого рода формализацию информационно-поисковых тезаурусов могут и не привести к улучшению качества автоматической обработки текстов и созданию ресурсов, лучше приспособленных к автоматическим режимам работы, чем существующие информационно-поисковые тезаурусы.

Заключение к главе 4

Таким образом, современные исследования в области онтологий развиваются в нескольких направлениях, изучая как аксиоматические способы представления знаний о мире, так и менее формализованные методы.

Создание онтологий на строгих формальных принципах в настоящее время связано с проблемами масштабируемости описания, с проблемами понимания пользователями, с существованием других формальных точек зрения на ту же сферу понятий.

Создание массово используемых понятийных ресурсов связано пока с относительно слабой формальзованностью описаний понятий, с основанием понятий онтологии на существующих языковых значениях. Нестрогость таких онтологий может естественно привести к проблемам в логическом выводе, который считается важным следствием создания онтологических ресурсов.

Таким образом, решая конкретные прикладные задачи особенно в широких предметных областях, необходимо делать осознанный выбор уровня сложности формализма представления знаний о предметной области.

Глава 5. Единицы онтологии: понятия

5.1. Понятия как единицы мышления и понятия в онтологиях

В литературе по компьютерным онтологиям трудно найти хорошее определение понятия как единицы онтологии. Б. Смит (Smith, 2004) указывает, что во многих случаях термин «понятие» используется вместо «слова», когда нужно абстрагироваться от конкретного естественного языка, специфических синтаксических особенностей. Иногда понятие – это идея, разделяемая людьми, использующими соответствующие слова или термины (Smith, 2004).

Тем не менее можно отметить, что понятия современных онтологических ресурсов имеют прямые аналогии с философской классической теорией понятия (Margolis E., Laurence S., 2006), в которой понятие определяется как единица системы с уникальным набором свойств и отношений. В качестве примера такого «классического» определения понятия можно привести определение (Степанов, 1990):

> *Понятие – мысль, отражающая в обобщенной форме предметы и явления действительности посредством фиксации их свойств и отношений; последние (свойства и отношения) выступают в понятии как общие и специфические признаки, соотнесенные с классами предметов и явлений.*

Таким образом, при определении понятия:
- устанавливаются его существенные признаки (характеристики),
- выявляются его связи с другими понятиями,
- определяется его место в системе понятий данной области знания.

Кроме того, значимым фактором рассмотрения понятий как единиц онтологии является их понимание как единиц, фиксирующих существующие знания о внешнем мире, предметной области (Smith, 2004; Gangemi и др., 2001b).

Такое понимание отражается в практических рекомендациях по введению понятий (класссов) в компьютерных онтологиях. Так, во многих руководствах по разработке онтологий указывается, что важно различать класс (понятие онтологии) и его имя:

- классы представляют понятия предметной области, а не те слова, которые обозначают эти понятия;
- синонимы одного и того же понятия не представляют разные классы, синонимы – всего лишь разные имена понятия (Noy, McGuinness, 2001).

Каждое понятие по определению должно быть элементом системы понятий и в то же время должно быть отделимо по своим свойствам от близких по смыслу понятий. В проектировании онтологий это положение раскрывается следующими рекомендациями по структуризации онтологии (Bouaud и др., 1995):

*1) Принцип сходства**:***

Принцип сходства контролирует сходство понятия по отношению к его родовому понятию. Нижестоящее понятие (понятие-потомок) должно разделять тип своего родового понятия (понятия-родителя). Таким образом, все понятия-потомки одного и того же понятия-родителя имеют между собой нечто общее. Отнесенность к типу понятия-родителя является необходимым условием описания понятия-потомка как видового по отношению к данному понятию-родителю. Потомок должен наследовать свойства родителя.

2) Принцип специфичности

Понятие-потомок должно отчетливо отличаться от понятия-родителя, что является необходимым и достаточным условием для понятия потомок. Это отличие может

выражаться в дополнительном свойстве, которое присуще потомку, или наличием дополнительных семантических ролей, например, при описании действий.

3) Принцип оппозиции

Понятие должно отчетливо отличаться от понятий одного уровня и должно быть представлено различие между каждой парой понятий этого уровня.

Следствием из этих принципов является правило, что понятие-родитель должно иметь более одного понятия-потомка. Если понятие имеет только одно непосредственное понятие-потомок, то, возможно, при моделировании допущена ошибка или онтология неполная.

Рассмотрим, как на основе изложенных принципов анализируется конкретная онтология. В работе (Bodenreider и др., 2004) исследуется соответствие медицинского терминологического ресурса SNOWMED CT (Табл. 5.1.), следующим формальным онтологическим принципам:

- каждое понятие должно иметь хотя бы одного родителя,
- понятия, имеющие потомков, должны иметь по крайней мере двух потомков,
- понятиям-потомкам следует иметь одного родителя,
- описание каждого понятия-потомка должно отличаться от описания понятия-родителя,
- все роли понятия-родителя должны наследоваться понятием-потомком или уточняться,
- отличие понятия-потомка от понятия-родителя должны проявляться либо в уточнении заполнителя роли или введении новой роли.

Роль	Значение
Каузативный агент	Вирус
Onset	sudden onset; Gradual onset
Severity	Severities
Episodicity	Episodicities
Course	Courses
Associated Morphology	Inflammation
Finding site	meninges stricture

Таблица 5.1. Пример описания понятия ВИРУСНЫЙ МЕНИНГИТ из онтологии SNOWMED CT

Исследование показало, что ресурс содержит 269864 понятия. 196237 понятия не имеют понятий-потомков – понятия-листья. Из 73267 понятий с понятиями-потомками, 23 174 понятия (31,5%) имеют только одно понятие-потомок. 8034 понятия имеют более 10 непосредственных потомков (11%), и 150 понятий имеют более 99 потомков, что, видимо, связано с недостаточной проработанностью классификации.

Каждое понятия, за исключением корня, имеет хотя бы одного родителя. Число родителей понятия может быть от 1 до 13.

Из рассмотренных 377681 пар потомок-родитель, 51% не проявили никакого различия между описаниями понятия-потомка и понятия-родителя.

В 7226 случаях некоторые роли, присутствующие у понятия-родителя, не наследовались и не уточнялись в понятии-потомке. В 21799 случаях, хотя у родителя и у потомка присутствует одна и та же роль, значения этой роли не являются ни идентичными, не состоят ни в таксономическом отношении, ни в отношении часть-целое. Обычно эта проблема возникает у понятий с множественными родителями: роль, которая не соответствует роли одного родителя, обычно находит свое соответствие с ролью другого родителя.

Качественный анализ понятий с одним понятием-потомком показал, что это явление может быть связано с тремя разными ситуациями:

- неполнота описания;
- единственное нижестоящее понятие представляет собой гибрид между двумя родительскими понятиями;
- понятие-потомок и понятие-родитель не демонстрируют никаких отличий в описании, и, таким образом, скорее всего, нижестоящий класс является излишним.

5.2. Критерии для ввода нового понятия

При разработке онтологии достаточно сложным вопросом являются критерии ввода новых понятий. На практике такие критерии связаны обычно с проверкой того, добавляет ли ввод нового понятия полезную и важную информацию для работы предполагаемого приложения.

Так, в работе (Noy, McGuinness, 2001) указывается, что необходимость введения нового класса для онтологии может возникнуть, если

1) у предполагаемого класса есть слот, которого нет у других классов, например, красные вина характеризуются свойством «уровень танина»,
2) у предполагаемого класса есть ограничения на слот, отличные от ограничений других классов, например, у класса ДЕСЕРТНЫЕ ВИНА – значение слота СОДЕРЖАНИЕ _САХАРА – «сладкий»,
3) у предполагаемого класса есть специфические отношения.

Если в разрабатываемой онтологии для понятия могут быть определены атрибуты посредством описания слота в фрейме понятия, то может возникнуть вопрос, в каких случаях нужно определить новый класс (понятие онтологии), а в каких случаях можно ввести лишь различные значения атрибута.

На примере онтологии вин (Noy, McGuinness, 2001) приводят следующий пример анализа ситуации в случае предполагаемого ввода нового понятия БЕЛОЕ ВИНО в противовес вводу различных значений слота ЦВЕТ для понятия ВИНО.

Авторы работы предлагают проанализировать следующие факторы:

- насколько важно БЕЛОЕ ВИНО для предметной области;
- если понятия с разными значениями слота становятся ограничениями для различных слотов в других классах, то для разделения следует создать новый класс. В противном случае разделение представляется в значении слота;
- какова изменчивость свойства, то есть как часто экземпляр класса меняет значения этого свойства. Если у экземпляра значение свойства постоянно – это является дополнительным критерием для введения отдельного класса с таким значением свойства.

Авторы онтологии MikroKosmos (Nirenburg, Raskin, 2004) излагают сходные принципы введения нового понятия в лингвистическую онтологию. Они обращают внимание на следующие положения:

- желаемый уровень подробности. Если предполагается, что в данной предметной области не понадобится то или иное знание об объектах или ситуациях предметной области, то не нужно вводить соответствующие единицы в онтологию;
- понимание того, является ли значение общим для многих языков или является свойственной данному изменению отклонению от языково-независимого значения;
- понимание того, в каких процедурах и процессах работы системы могут возникнуть проблемы, если не будет добавлено данное понятие.

Хорошими основаниями для ввода нового понятия также являются:

- отличие понятия-потомка от понятия-родителя в наборе отношений, не считая отношения к видовым понятиям;
- отличие в более чем одном атрибуте;
- если ограничение на заполнения какого-то слота в свойстве понятия-родителя содержат сложную дизъюнктивную формулу, а вводимое понятие имеет значительно более строгие ограничения, то такое значение имеет хорошие основания для ввода в онтологию в качестве понятия.

5.3. Понятие и значение в лингвистических онтологиях

В процессе разработки лингвистической онтологии, то есть онтологии, которая разрабатывается для обработки текстов на естественном языке и/или ставит своей целью адекватный учет существующих языковых значений, возникают практические вопросы соотношения понятий онтологии и представленных в языке значений.

Представляется, что при вводе понятий лингвистической онтологии также важно обеспечить выполнение требований к понятиям онтологии, перечисленных в разделе 5.1., связанных с четкой отличимостью понятия от соседних по иерархии понятий. Поскольку именно эти требования создают основу для создания последовательного описания отношений и иерархии онтологии, а также снижают зависимость структуры онтологии от естественного языка, носителями которого создается эта онтология.

При этом возникают три основные проблемы, для преодоления которых необходимо иметь четко сформулированные принципы:

- проблема отличения понятия от его имени, поскольку непросто отличить понятие и его название, если работа ведется с языковыми значениями,
- проблема разбиения на понятия совокупности значений близких по смыслу слов – квазисинонимов,
- проблема выделения разных понятий для отражения близких значений одного и того же многозначного слова.

В следующих подразделах мы подробно рассмотрим эти проблемы.

5.3.1. Разбиение на понятия совокупности значений квазисинонимов

Как известно, в любом языке существуют совокупности близких по смыслу слов – квазисинонимов (Апресян, 1995). Несмотря на свою смысловую близость, квазисинонимы могут различаться по понятийному содержанию, сфере употребления, оценочному содержанию, сочетаемости и др. Неслучайно существует специальный жанр синонимических словарей (НОСС, 2003), которые подробно разъясняют особенности употребления таких синонимов. При этом значения многих квазисинонимов различаются не по одному параметру, а по нескольким, видоизменяются в зависимости от контекста.

Для многих таких совокупностей квазисинонимов чрезвычайно трудно установить однозначное соответствие на других языках, поскольку, чаще всего, на другом языке данной совокупности квазисинонимов соответствует другая совокупность близких по смыслу слов, которая характеризуется своей системой параметрических различий и, соответственно, своими особенностями.

Если при описании соотношений между значениями руководствоваться принципами возможности синонимичной подстановки в одни и те же предложения, как было принято при создании Принстонского WordNet (см. главу 2), то это означает, что производится попытка построить классификацию сразу по нескольким основаниям, поскольку синонимическая замена слова должна учитывать и понятийный, и стилевой, и оценочный и другие компоненты значения.

Понятно, что такое построение иерархии сразу по нескольким основаниям осуществить часто невозможно, все построение становится очень изменчивым при

переходе от языка к языку. Именно в таких ситуациях проявляется серьезная зависимость построенной классификации от языка.

Главным компонентом классификации естественно является понятийный компонент значения, который часто очень сложно отделить от других компонентов. Так, например, сколько понятий онтологии оптимально (и на основе каких принципов) следует сопоставить следующему ряду слов со значением ОШИБКА: *ошибка, погрешность, недосмотр, просмотр, ляп, промах, оплошность, осечка, прокол, упущение, недочет,* а также *ослышка, описка, опечатка, оговорка.*

В качестве другого примера может быть рассмотрена группа слов, относящаяся к ситуации «Драка» (система значений дана по Большому толковому словарю русского языка - 1998). Здесь также возникает вопрос, как эта россыпь значений должна быть отражена в системе понятий онтологии, описывающей ситуацию драки:

драться *1. бить друг друга, устраивать драку*

подраться *1. сов. к Драться*
2. Драться некоторое время

передраться *1. Поссориться, подраться друг с другом (о двух лицах)*
2. Подраться (обо всех, о многих)
3. Поочередно подраться со всеми, со многими.

свалка 5. всеобщая драка, потасовка

потасовка - ссора с дракой

побоище - ожесточенная кровавая драка

мордобой - битье по лицу

поножовщина - драка с применением ножей.

Для снижения зависимости лингвистической онтологии от конкретного языка в работах (Edmonds, Hirst, 2000; Hirst, 2003) предлагается для описания близких по смыслу слов в лингвистических онтологиях ввести еще один уровень представления - понятийно-семантический уровень.

Понятийно-семантический уровень задает относительно грубую понятийную иерархическую систему, которая основывается на денотативных, независимых от контекста, свойствах значений слов. Каждому такому понятию поставлен в соответствие набор синонимов, а их особенности (стилистические, отношение говорящего, коннотации и т.п.) описываются в дополнительных, внутрипонятийных структурах.

Авторы работы (Edmonds, Hirst, 2000) подчеркивают, что часто может оказаться, что определить, какие близкие по смыслу слова лучше описать в рамках внутренней структуры понятия, а какие разнести в разные понятия, очень непросто. С одной стороны, можно надеяться на интуицию лингвиста. С другой стороны, возможно, взгляд на понятийную структуру с точки зрения другого языка может действительно лучше проявить границы понятий.

Таким образом, лингвистическая онтология, которая хоть и учитывает существующие лексические значения, все же должна оставаться онтологией. По общим принципам организации онтологической иерархии (см.раздел 5.1) ее основные элементы - понятия должны иметь четкие, независимые от контекста отличия от соседних понятий. Чем четче эти различия между понятиями, тем более независимой от конкретного естественного языка становится онтология, несмотря на то, что источником для введения того или иного понятия могло быть значение слова или выражения в конкретном естественном языке.

5.3.2. Выделение разных понятий для отражения близких значений одного и того же слова

Сложным случаем при разработке лингвистических онтологий является наличие у слова нескольких близких по смыслу значений. Поскольку общеизвестно, насколько тяжело системе автоматической обработки текстов бывает разобраться с близкими значениями слова, то также важно выработать принципы для описания таких совокупностей близких значений отдельногослова.

Как мы видели в разделе 2.5., при применении WordNet были выявлены серьезные проблемы приложений в связи со слишком большим количеством описанных значений, после чего разработчиками было проведено значительное количество экспериментов, с целью кластеризации значений, выявлению групп близких значений, позволяющих улучшать качество применения WordNet в автоматической обработке текстов. Однако было предложено слишком много разных принципов группировки значений и непонятно, какие принципы нужно предпочесть {Fellbaum, 2002).

В проекте OntoNotes (Hovy и др., 2006) предлагается способ отражения набора лексических значений многозначного слова совокупностью понятий на основе рассмотрения конкретных примеров употребления из корпуса. Сопоставляя примеры употребления и системы значений слова, нужно разделять значения на наиболее далекие друг от друга группы, создавать точку ветвления на дереве, затем для каждой такой точки повторять процесс

Рассматривая глагол *drive,* для которого WordNet выделяет 22 отдельных значения, авторы проекта предлагают формировать наиболее очевидные группы значений, которые для глагола drive таковы, и которые и являются предлагаемыми понятиями:

1) drive mad – Cause-mental- instability – привести в бешенство
2) группы смыслов физического движения – Cause-movement-in–Desired-Direction, (Вести или путешествовать на транспортном средстве)
3) группа смыслов нефизического характера, – Cause–State-Change–toward Desired-Value (Изменение-состояния-к-желаемой-величине).

Далее можно продолжать онтологизацию значений слова в зависимости от объяснительной необходимости или потребности приложения. Каждый шаг онтологизации требует введения нового понятий в растущую онтологию. В результате нескольких шагов два независимых эксперта выделили 7 наиболее важных групп смыслов глагола *drive.*

По мнению авторов работы (Hovy и др., 2006), хорошим принципом для остановки процесса онтологизации является ситуация, когда не находится очевидного разбиения оставшейся группы смыслов на подгруппы, или возможно одинаково обоснованные разбиения на подгруппы по разным основаниям. Также подчеркивается полезность многоязычного рассмотрения для наиболее адекватного разделения «пространства смыслов» и «пространства понятий».

В работе приводятся примеры объединения значений глагола drive. К значению Cause-movement-in–Desired- Direction («Вести или путешествовать на транспортном средстве») относятся 7 значений из WordNet:

WN1: Can you ***drive*** *a truck (водить)?*

WN2: ***drive*** *to school (ехать),*

WN3: ***drive*** *her to school (везти),*

WN12: this truck ***drives*** *well (едет),*

WN13: He ***drives*** *taxi (водит),*

WN14: The car ***drove*** *around the corner (повернул),*

*WN16: **Drive** the turnpike to work.*

Отметим, что с точки зрения носителя русского языка эта «транспортная» группа значений глагола *drive* не так очевидна, поскольку соответствует значениям нескольких разных слов: *водить, ехать, везти, повернуть*, и, значит, зависимость системы понятий от исходного языка разработки сохраняется в серьезной степени.

В противовес тенденции ряда исследований к сокращению числа значений языковых единиц, представленных в лингвистических онтологиях, высказываются мнения, предупреждающие против чрезмерной кластеризации разных значений даже в благих целях облегчения автоматической обработки текстов. Так, Н. Гуарино (1998) критикует несколько существующих онтологий за многозначность онтологических узлов, например, за трактовку понятия ОКНО одновременно и как артефакта, и как отверстия.

Проблема возникает из-за того, что слово *окно* в различных контекстах может обозначать =изделие= (как во фразе «разбить окно») или =отверстие= (как во фразе «выглянуть в окно»), и разработчики лингвистических онтологий стремились описать оба типа употреблений посредством одного понятия онтологии

Эта критика связана с тем, что по мнению Гуарино многозначность в онтологических узлах не должна быть разрешена ни в какой форме. Чтобы соответствовать принципу отсутствия многозначности узлов, онтология должна иметь различные узлы в различных местах онтологии для таких понятий как ОКНО- ИЗДЕЛИЕ и ОКНО-ОТВЕРСТИЕ, при этом эти сущности ОКНО-ОТВЕРСТИЕ и ОКНО-ИЗДЕЛИЕ очень тесно связаны между собой.

Мы продолжим обсуждение этой проблемы в разделе 5.6.2., где приведем возражения авторов критикуемой онтологии.

Как видно, проблема близких значений многословных слов, которая сложна и для составителей толковых словарей, многократно усложняется при представлении таких значений в словарном ресурсе предназначенном для автоматической обработки текстов.

5.4. Смешение понятия и его имени в Принстонском WordNet и других ворднетах

Проблемы со слишком большим количеством значений в Принстонском WordNet были рассмотрены в разделе 2.5. В данном разделе будет рассмотрена проблема описания близких по смыслу слов в Принстонском WordNet и других ворднетах.

Первоначально авторы WordNet считали, что WordNet – это лексический, а не онтологический ресурс. Однако, со временем рост значимости онтологических исследований, а также сходство иерархии существительных из WordNet с онтологией стали очевидными (Miller, Hristea, 2006). Поэтому на основе WordNet правомерно рассматривать проблемы, возникающие при создании лингвистических онтологий.

В WordNet можно найти многочисленные примеры смешения понятия и его названия. Это связано с тем, что основным отношением в WordNet является отношение синонимии. Наборы синонимов – синсеты – являются основными структурными элементами WordNet. Авторы WordNet считали два выражения синонимичными, если замена одного из них на другое в предложении не меняет значения истинности этого высказывания.

Этот основной принцип устройства WordNet приводит к тому, что не выполняется один из важнейших принципов разработки онтологий – это различение собственно понятия и способов его называния, то есть вводятся разные синсеты для разных способов наименования одной и той же сущности.

Имеется несколько типов смешений понятий и их названий в ресурсах типа WordNet.

Во-первых, смешение понятий и их названий проявляется в поддержке разных иерархий для разных частей речи. Действительно, с помощью какой бы части речи в

тексте не было бы упомянуто понятие ПРИВАТИЗАЦИЯ (*приватизировать, приватизационный, приватизация*) – это всегда ссылка на одно и то же понятие разными лексическими средствами, от изменения части речи не должны меняться отношения этого понятия с другими понятиями.

Кроме того, различие в описаниях отношений разных частей речи, имеющих между собой прямое смысловое соответствие, увеличивает долю непоследовательно выполненных описаний. Например, в синсете WordNet

engagement, participation, involvement, involution -- (the act of sharing in the activities of a group; "the teacher tried to increase his students' engagement in class activities")

как синонимы указываются существительные engagement и participation. А в соответствующем глагольном синсете глагол participate упоминается только в толковании.

prosecute, engage, pursue -- (carry out or participate in an activity; be involved in; "She pursued many activities"; "They engaged in a discussion")

Если части речи конкретных слов существенны для проводимой обработки текстов, они могут быть извлечены из морфологического словаря, или конкретные текстовые входы, сопоставленные понятию, могут иметь соответствующие пометы частей речи и (или) морфологических классов.

Авторы проекты EuroWordNet (см. главу 3) рассматривали возможность соединения всех частей речи-дериватов к одному синсету, поскольку такое разделение противоречит принципам разработки онтологических ресурсов (Climent.и др., 1996). Однако, в конце концов, решение о соединении частей речи принято не было.

Вторым типом проявления смешения понятия и его названия является использование разных синсетов для описания старых и новых названий, названий понятия в разных диалектах языка, в разных текстовых жанрах и т.п.

В принстонском WordNet можно найти многочисленные примеры того, что особенность употребления слов приводит к введению нового синсета.

Например, для отражения способов разговорного упоминания человеческого носа заведен специальный синсет

beak, honker, hooter, nozzle, snoot, snout, schnozzle, schnoz -- (informal terms for the nose – разговорные варианты слова «нос»),

который является гипонимом синсета для слова *нос*

nose, olfactory organ -- (the organ of smell and entrance to the respiratory tract; the prominent part of the face of man or other mammals; "he has a cold in the nose").

Разговорная лексика, имеющая отношение к деньгам, также собрана в отдельный синсет:

boodle, bread, cabbage, clams, dinero, dough, gelt, kale, lettuce, lolly, lucre, loot, moolah, pelf, scratch, shekels, simoleons, sugar, wampum -- (informal terms for money)

Некоторые синсеты отражают специфику диалектов английского языка, как например, название домашнего осла в британском английском:

Moke 1 -- (British informal)

=> domestic ass, donkey, Equus asinus -- (domestic beast of burden descended from the African wild ass; patient but stubborn)

Разработчики русского WordNet – RusNet специально рассматривают вопросы синонимии, и ее описании в синсетах. Они разделяют синонимию на 5 подвидов: абсолютную синонимию, дупликатную синонимию, стилистическую синонимию, экспрессивную синонимию и деривационную синонимию (*дом: домик, домина*). Такие

виды синонимов как стилистические и экспрессивные синонимы описываются в том же синсете, что и нейтральные слова, но снабжаются дополнительными пометами.

Для деривационной синонимии предлагается заводить отдельные синсеты и особые виды отношений: деривационный гипоним и деривационный гипероним. Авторы ресурса считают, что ввод словообразовательной компоненты не дает считать такие единицы как *домик* и *домина* просто экспрессивными синонимами и отражать их в едином синонимическом ряду.

Однако, с точки зрения разработки онтологий, такое понятие, как деривационный синсет, не имеет четких признаков отличия от своего вышестоящего понятия, поскольку дом любой величины в разных контекстах может быть назван домиком или доминой.

Еще одним проявлением различий синсетов и понятий как единиц представления является описание денежных единиц, используемых в различных странах под одними и теми же названиями, например, как франк или сантим. С точки зрения языка, могут быть введены соответствующие синсеты, как в WordNet:

franc -- (the basic monetary unit in many countries; equal to 100 centimes)

centime -- (a fractional monetary unit of several countries: France and Algeria and Belgium and Burkina Faso and Burundi and Cameroon and Chad and the Congo and Gabon and Haiti and the Ivory Coast and Luxembourg and Mali and Morocco and Niger and Rwanda and Senegal and Switzerland and Togo)

Однако с точки зрения представления на понятийном уровне такие единицы невозможны:

- все эти франки и сантимы имеют разную ценность, соответствие между собой,
- общее между ними только название;
- в любой момент соответствующее государство может ввести другое название своих единиц, не меняя их относительной стоимости.

Таким образом, если мы считаем своей единицей представления понятие, то должна быть введена отдельная понятийная единица для денежной единицы каждой страны, например, *швейцарский франк, американский доллар, канадский доллар и т.п.*

5.5. Квазисинонимы в Принстонском WordNet

Если при описании соотношения между значениями руководствоваться принципами возможности синонимичной подстановки в одни и те же предложения, как было принято при создании Принстонского WordNet, то это означает, что квазисинонимы необходимо классифицировать сразу по нескольким основаниям, поскольку синонимическая замена слова должна учитывать и понятийный, и стилевой, и оценочный и другие компоненты значения.

Следствием принципа синонимичной подстановки является то, что WordNet имеет значительное количество синсетов, которые трудно отличимы друг от друга, что также нарушает онтологические принципы описания понятий.

Так, например, имеется четыре различных синсета, обозначающие *сходство, подобие*, каждый следующий из которых является гипонимом для предыдущего и при этом является практически не отличимым от своего гиперонима:

sameness -- (the quality of being alike; "sameness of purpose kept them together")

similarity -- *(the quality of being similar) - сходство*

*likeness, alikeness, similitude -- (*similarity *in appearance or character or nature between persons or things; ``man created God in his own likeness'') – сходство по внешности, характеру или природе между людьми или объектами).*

resemblance -- (similarity in appearance or external or superficial details) – сходство во внешности или во внешних или поверхностных деталях.

5.6. Понятие и значение в онтологии MicroKosmos

5.6.1 Отражение значений квазисинонимов

В онтологии МикроКосмос проблема квазисинонимов решается за счет объединения квазисинонимов к одному и тому же понятию онтологии, и описания особенностей конкретных лексем в словарных статьях словаря.

Авторы онтологии приводят пример, что все глаголы изменения в онтологии приписаны одному и тому же понятию Change-event (Nirenburg, McShane, 2004). Особенности слов описываются в словарной статье, например, для глагола увеличить (increase) указывается, что в семантической роли ТЕМА этого глагола должна выступать СКАЛЯРНАЯ_ВЕЛИЧИНА (например, цена или высота) и указывается, что значение этой величины меняется на большее.

Если мы обратимся к сайту ресурса, то мы увидим, что ситуация с реализацией изложенных принципов достаточно сложная. Так, понятию CHANGE_EVENT сопоставлен в лексиконе большой список слов, которые, по мнению авторов онтологии, соответствуют этому понятию, например: *acclimatization* (акклиматизация – приспособление к другому климату*), commerzialization* (коммерциализация), contamination (загрязнение), *damage* (повреждать), *deteriorate* (ухудшать*), improve* (улучшать) и многие другие – для этих слов не было заведено отдельных понятий.

В то же время среди нижестоящих по иерархии понятий можно увидеть следующие: ADJUST (адаптировать, приспособить), CORRECT-EVENT (исправление, коррекция), DIVIDE (делить), INTEGRATE (интегрировать), RESTRUCTURE (реструктуризация) и др.

Непонятно, почему для одних значений слов были заведены отдельные понятия, а для других нет. Почему значение слова *acclimatization* не заслуживает отдельного понятия, хотя есть важное отношение к климату, биологическим процессам, а значение слова *adjust* такой концепт получило?

Помимо вопросов последовательности/непоследовательности описания имеются и явные последствия для процедур автоматической обработки текстов.

Так, сложной становится процедура установления, какие все-таки слова из большего списка словарных входов к понятию CHANGE-EVENT, могут рассматриваться как синонимы, какие соотношения между этими словами. Невозможно указать отношение между *асфальтированием* и, например, *дорожными работами.*

Кроме того, относительно небольшая величина онтологии приводит к тому, что при работе в конкретном приложении и конкретной предметной области многое придется доделывать и вводить дополнительные понятия даже для слов, которые уже учтены в онтологии.

Таким образом, на наш взгляд, в приведенных примерах из онтологии MikroKosmos проблема квазисинонимов решается путем чрезмерного переобобщения, что может привести к проблемам в реальных предметных областях. Необходимо выделить дополнительный уровень понятий, который поможет более четко разделить слова, не сваливая их в единый большой мешок.

5.6.2. Описание близких значений многозначных слов в онтологии MikroKosmos

Основным правилом, провозглашаемым при работе с близкими значениями многозначных слов в онтологии MikroKsomos, является правило редукции полисемии, которое заключается в том, что нужно решить, сколько значений словаря может представлять данное словарное значение, и объединить столько значений, сколько возможно так, чтобы осталось как можно меньше разных значений.

Принципы для различения значений таковы:

- значение-кандидат должно быть отчетливо отличимым от уже описанных значений;
- необходимо проверять, требует ли значение дополнительного разъяснения, если использовано в коротком предложении. Если необходим дополнительный контекст, чтобы разобраться, какое значение используется, то значение не должно вводиться, а должно быть отнесено к одному из существующих значений;
- необходимо проверять, имеется ли свойство при описании данного значений, которое заполнено слишком малым числом заполнителей. Если да, то также либо значение должно быть отнесено к одному из более общих значений, либо описано в рамках описания многословной конструкции.

Казалось бы такая процедура должна снижать проблемы разрешения многозначности, однако, с другой стороны, эта процедура может привести к нарушению структуры онтологии. Именно эту онтологию Н. Гуарино (Guarino, 1998) критикует за трактовку понятия ОКНО одновременно и как артефакта, и как места (см.раздел 5.3.2.).

Также Гуарино критикует эту же онтологию за представление понятия КОММУНИКАТИВНОЕ СОБЫТИЕ одновременно как видового понятия для понятий SOCIAL_EVENT (социальное событие) и MENTAL_EVENT (ментальное событие).

Как мы уже упоминали, Н. Гуарино считает, что многозначность в онтологических узлах не должна быть разрешена. Чтобы соответствовать принципу отсутствия многозначности узлов, онтология должна иметь различные узлы в различных местах онтологии для таких понятий как WINDOW-ARTIFACT и WINDOW-PLACE, MENTAL-COMMUNICATION-EVENT и SOCIAL -COMMUNICATION-EVENT.

Отвечая Н.Гуарино, авторы онтологии (Nirenburg, Raskin, 2004: Стр. 129) указывают, что факт того, что английское слово *window* имеет два значения не имеет решающего значения при построении онтологии, поскольку не считается, что отношение между значениями естественного языка (или точнее значениями все известных языков) и понятиями онтологии должны иметь однозначное соответствие.

В качестве обоснования своей позиции авторы приводят аргумент, что они не знают такого естественного языка, в котором имеющееся слово для понятие WINDOW не реализовывало оба значения: значение отверстия и значение артефакта. Эта семантическая универсалия является сильнейшим аргументом в пользу того, что люди могут совмещать эти два понятия

Авторы онтологии также подчеркивают, что «попытки расщепить онтологические понятия на все меньшие однозначные единицы ведет к резкому увеличению многозначности, и поэтому делает разрешение многозначности более сложным. Если онтология делается менее многозначной, то при работе в приложении, где-то возрастает нагрузка на обработку многозначности (Там же)».

Вместе с тем нужно согласиться с Н. Гуарино в том, что нарушение онтологической структуры, введение взаимоисключающих отношений понятий, также представляет собой проблему, поскольку, если предполагается использовать описанные отношения между понятиями для логического вывода, необходимо будет сначала определить, применимо ли это отношение для данного контекста, а это означает, что проблема выбора значения многозначного слова просто сместилась на другой этап.

Кроме того, при рассмотрении разбиения значений на отдельные понятия онтологии нужно учитывать не только существующие в языке однословные единицы, но и словосочетания, которые также нужно автоматически обрабатывать при анализе текста, такие как *оконная рама, оконное стекло, отверстие окна* (см. раздел 16.4.3.).

5.7. Понятия и значения в ресурсе FrameNet

Авторы ресурса FrameNet уделяют большое внимание рассмотрению взаимотношения понятий-фреймов и языковых единиц Описывается достаточно большой

набор принципов, который должен регулировать, в каком случае два близких по смыслу слова могут быть отнесены к одному и тому же фрейму, а в каких случаях эти два слова должны быть разнесены в разные фреймы.

Так, все лексические единицы, относящиеся к одному и тому же фрейму, должны иметь одинаковое число фреймовых элементов в имплицитных и эксплицитных контекстах. Если число существенных, синтаксически значимых элементов варьируется в предложениях, то фрейм должен быть расщеплен, чтобы отразить это варьирование. Так, систематически по разным фреймам разделяются каузативы (*уменьшать*) и так называемые инхоативы (*уменьшаться*), например, инхоативный фрейм называется «ИЗМЕНЕНИЕ ПОЗИЦИИ НА ШКАЛЕ», а каузативный фрейм «ВЫЗВАТЬ ИЗМЕНЕНИЕ ПОЗИЦИИ НА ШКАЛЕ» (между ними установлено отношение каузации).

Авторы ресурса подчеркивают, что здесь может быть законное возражение, что присутствие агента или причины относится слишком к тонкой лингвистической интуиции, и что игнорируется факт, что всякое изменение в мире чем-нибудь каузируется. Предлагаемое обоснование заключается в том, что существуют лексические единицы, которые проявляют либо то, либо другое поведение. Например, глагол *gain* имеет только инхоативное употребление по отношению к ситуации изменения на шкале, а глагол *lower* позволяет только каузативное использование. Во-вторых, межязыковое сопоставление показывает, что многие языки различают каузативы и инхоативы с помощью словообразования.

Вместе в одни и те же фреймы группируются лексические единицы, которые связаны с различными языковыми реализациями:

- пассивные залоги;
- видо-временные конструкции;
- композиции с экстра-тематическими фреймовыми элементами;
- антонимы. Например, прилагательные *high* and *low* помещены в фрейм POSITION_ON_A_SCALE (ПОЗИЦИЯ НА ШКАЛЕ). Подобно этому, глаголы *love* and *hate* относятся к фрейму EXPERIENCER_SUBJ. Однако, так называемые конверсивы (Cruse, 1986, Апресян, 1995), такие как *buy* (*купить*) and *sell (продать),* которые отражают противоположные точки зрения на то же самое событие помещаются в разные фреймы, поскольку они имеют разные наборы участников. Также реверсивные пары *привязать, отвязать* описываются в разных фреймах, поскольку соответствуют разным типам действий.

Также объединяются любые различия, возникающие вследствие речевого контекста:

- дейксис, то есть зависимость от точки зрения говорящего (come vs. go);
- диалекты языка (*lorry* vs. *truck*; *fixture* vs. *regular season game*);
- оценки (*criticize* vs. *praise*; *genius* vs. *moron*).

Представляется, что отнесение лексических единиц к разным фреймам в рамках проекта FrameNet также является своего рода разбиением на понятийный единицы. Авторы подчеркивают, что основой отнесения к одному и тому же фрейму является не только сходный набор фреймовых элементов, лексические единицы одного и того же фрейма должны иметь один и тот же набор пресуппозиций, ожиданий.

С этой точки зрения, кажется странным объединение в один фрейм FORMING_RELATIONSHIPS (Установление отношений) таких слов как *befriend (подружиться), divorce (развестись), marry (жениться)*, так как каждое из этих слов имеет разный набор пресуппозиций и ожиданий. Также в один фрейм объединены лексемы *купить* и а*рендовать*. Основой объединения служит, видимо, введение фреймового элемента "длительность" (duration). Однако в случае аренды по умолчанию необходимо вернуть собственность тому же владельцу, то есть ожидание отличается от ситуации покупки. К фрейму KILLING (Убийство) относятся достаточно разные

лексические единицы: *задушить, самоубийство, морить голодом, погром, застрелить, геноцид, холокост, детоубийство, убийство матери*.

Таким образом, авторы FrameNet делают значительные усилия, чтобы описывать взаимоотношения между понятиями-фреймами и лексическими единицами последовательно, минимизируя субъективность описания, что еще раз подтверждает сложность этой проблемы.

5.8. Понятия и значения в информационно-поисковых тезаурусах

Мы уже указывали, что информационно-поисковые тезаурусы можно рассматривать как лингвистические онтологии, поскольку их единицы – дескрипторы – обычно вводятся на основе реально существующих в предметной области терминов.

Поскольку многие решения в области построения информационно-поисковых тезаурусов связаны со спецификой их применения в ручном индексировании, с удобством человека-индексатора, это находит непосредственное отражение в представлении в тезаурусе квазисинонимов и многозначных слов.

Так, многие близкие по смыслу термины могут быть представлены в тезаурусе одним термином-дескриптором, а остальные не включаются в тезаурус совсем, поскольку их включение как дескрипторов увеличивает субъективность индексирования, а включение как аскрипторов может затруднить восприятие индексатора (см. разделы 1.3., 1.7.1).

Включение различных значений слов и выражений минимизируется, представительство возможных значений не является необходимым, поскольку в процессе использования тезауруса имеется человек-посредник. Однако в результате возникает серьезная разница между языком документов предметной области и единицами тезауруса, что затрудняет автоматическое применение тезауруса при обработке текста (см. раздел 1.7).

Заключение к главе 5

Одним из основных принципов построения формальной онтологии заключается в том, что создаваемая онтология должна быть независима от естественного языка. Вместе с тем разработчикам онтологий очень трудно избежать влияния языковых значений, языковой многозначности, поскольку в онтологиях имена понятий и отношения носят мнемонические имена, знания о понятиях во многих предметных областях хранятся в виде текстов.

В так называемых лингвистических онтологиях понятийные единицы создаются на основе реально существующих языковых значений. Однако, как ни парадоксально, лингвистическая онтология может быть значительно более независимой от исходного языка разработки, если ее понятия будут иметь четкие отличия от близких по смыслу понятий в понятийной системе.

При формировании отличимых понятий в лингвистической онтологии возникают существенные проблемы, а именно:

- проблема различения понятия и его имени,
- проблема представления близких значений многозначных слов,
- проблема разбиения на понятия совокупности близких значений квазисинонимов.

Существующие лингвистистические онтологии используют разные принципы формирования своих понятий. Таблица 5.2 обобщает основные характеристики решений, принятых авторами различных ресурсов/

Проблема формирования понятия	WordNet	MikroKosmos	Информационно-поисковые тезаурусы	FrameNet
Смешение понятия и его имени	Понятие и имя часто смешиваются	Понятие и имя не смешиваются за счет разделения онтологии и лексикона	Понятие и его имя редко смешиваются	Не смешиваются. Понятие имеет однозначное название
Представление близких по смыслу значений многозначных слов	Детальное описание различных значений	Описание близких по смыслу значений слов минимизируется	Близкие значения многозначных слов не включаются	Близкие значения многозначных слов описываются, если они соответствуют разным фреймам
Отношения между значениями многозначных слов	Нет отношений между значениями многозначных слов	Близкие значения обобщаются к одному понятию онтологии	-	Отношения межу фреймами
Квазисинонимы	Совокупности квазисинонимов могут быть произвольно расщепляться на понятия	Квазисинонимы обобщаются к одному понятию онтологии	Квазисинонимы исключаются или описываются как аскрипторы к дескриптору	Квазисинонимы приписываются к одному и тому же фрейму

Таблица 5.2 Особенности описания отношений между понятиями и языковыми значениями в различных лингвистических онтологиях

Глава 6. Установление отношений в онтологиях. Отношение класс-подкласс

Установление отношений между понятиями онтологии в широких гетерогенных предметных областях является непростым видом деятельности.

В последующих главах мы опишем рекомендуемые принципы и проблемы установления отношений в онтологиях. Будет рассмотрены отношения, которые могут быть применимы в подавляющем большинстве предметных областей, а именно: отношение *класс-подкласс* (родовидовые отношения), отношение роли, отношение *часть-целое,* а также применяемое по большей мере в онтологиях верхнего уровня отношение *онтологической зависимости.*

Начнем рассмотрение онтологических отношений с основного отношения онтологических и многих других компьютерных ресурсов отношения «класс-подкласс».

6.1. Проблемы установления отношения «класс-подкласс»

Отношение между классами и подклассами понятий может носить разное название в зависимости от терминологических традиций в области использования ресурса: таксономическое отношение, родовидовое отношение, IS-a отношение, отношение гипонимии и гиперонимии (в лексических ресурсах). Далее в тексте мы будем ссылаться на это отношение как «родовидовое отношение».

Родовидовые отношения обладают такими важными свойствами как транзитивность и наследование, на которых основывается логический вывод во многих компьютерных системах (Осипов, 1997).

Пусть T(X,Y) – родовидовое отношение между понятиями X и Y, X является видом (подклассом) Y, R (X, Z) – это произвольное отношение между понятиями X и Z.

Тогда свойства родовидового отношения могут быть записаны следующим образом:

$T(X, Y) \wedge T(Y, Z) \rightarrow T(X, Z)$ - транзитивность родовидового отношения,

$T(X, Y) \wedge R(Y, Z) \rightarrow R(X,Z)$ – свойство наследования по родовидому отношению.

Наиболее исторически ранними принципами установления родовидовых отношений, используемых и в работах по искусственному интеллекту, и компьютерной лингвистике, было использование ставших классическими диагностических высказываний (Cruse, 1986). Например, если понятие X является видом понятия Y, то можно сказать, что «X – это Y», «X, Z и другие Y», «к числу Y относятся X» (см. также п. 2.2.)

Однако позже выяснилось, что одни и те же выражения естественного языка (и в частности, применяемые диагностические тесты) могут с онтологической точки зрения соответствовать значительно различающимся отношениям между сущностями внешнего мира, в том числе обладающими совсем другими свойствами (Guarino, 1998). Поэтому многие методические руководства по разработке понятийных ресурсов рекомендуют осуществлять дополнительные проверки для устанавливаемого родовидового отношения.

Наиболее распространенной рекомендацией для проверки правильности установления родовидовых отношений является ответ на вопрос, если объект является экземпляром одного класса, будет ли он обязательно (т.е. по определению) экземпляром некоторого другого класса (см. также п.1.2.1.1):

> *Если класс A – надкласс класса B, то каждый экземпляр класса B также является экземпляром A (Noy, McGuinness, 2001; Z39.19; Gomez-Perez и др., 2004).*

Однако ситуации, в которых происходит смешение родовидовых отношений с отношениями других типов, значительно более разнообразны и при разработке онтологий, других понятийных ресурсов необходимо учитывать такого рода проблемы.

Особенно серьезно эти проблемы стоят перед разработчиками понятийных ресурсов для автоматической обработки текстов, информационно-поисковых приложений в широких предметных областях. В таких приложениях ресурсы, с одной стороны, должны в значительной мере учитывать существующую понятийную систему языка (группы языков). С другой стороны, для сохранения необходимых свойств моделируемых отношений эти отношения должны устанавливаться на основе понятийного, онтологического анализа, а не только с использованием языковых диагностических высказываний.

В то же время нужно подчеркнуть, что онтологии, создаваемые и для другого рода компьютерных приложений, не связанных с обработкой текстов, достаточно трудно отделить полностью от естественного языка. Единицы онтологий часто носят языковые или мнемонические названия, тем самым дополнительно «провоцируя» применение неоднозначных языковых тестов (см. п.4.4.).

В любом случае, на наш взгляд, в «языковую ловушку» может попасть разработчик понятийных ресурсов в самых различных областях и для различных компьютерных приложений. Поэтому важно описать наиболее частые случаи проблемного установления родовидовых отношений, а также возможные способы выявления таких неточностей в момент описания. Кроме того, при использовании транзитивности родовидовых отношений локальная неточность может перерасти в серьезное искажение в процессе многошагового логического вывода.

Далее мы рассмотрим типы проблемного установления родовидовых отношений, а также возможные критерии для проверки правильности установления этих отношений.

6.2. Возможные критерии проверки правильности установления родовидовых отношений

Критерии проверки правильности установления родовидовых отношений связаны с проверкой выполнения свойств транзитивности и наследования.

На проверке транзитивности родовидового отношения основано следующее правило:

> *Нижестоящее понятие и вышестоящее понятие должны относиться к одному и тому же наиболее общему семантическому классу, такому как =действие=, =свойство=, =объект= и т.п.*

Так, стандарты и методические руководства по разработке информационно-поисковых тезаурусов рекомендуют использовать такой принцип для описания иерархических отношений в тезаурусах.

В качестве реальной ситуации, при которой неправильный семантический класс помог выявить неточно установленное родовидовое отношение, приведем следующий пример. При установлении отношений в тезаурусе РуТез (см. часть 4) первоначально была установлена следующая цепочка родовидовых отношений:

> *РЕКА* – выше – *ВОДОЕМ* – выше – *ВОДНЫЙ ОБЪЕКТ* – выше – *ВОДА* – выше – *ВЕЩЕСТВО,*

в результате чего получилось, что все конкретные реки относятся к семантическому классу *ВЕЩЕСТВО*, что неправильно.

В этой цепочке наиболее проблематичным является отношение *ВОДНЫЙ ОБЪЕКТ* – выше – *ВОДА,* изменение которого на другой тип отношения устранит возникшую проблему (подробнее см. п. 6.5).

Второй тип критериев проверки правильности установления родовидовых отношений связан с проверкой свойства наследования.

Проверка может носить частный характер, быть связанной именно с конкретной парой понятий. Например, в словарях изюм определяется как «сушеные ягоды винограда». Следует ли из этого определения, что нужно установить родовидовое отношение между понятиями *ИЗЮМ* и *ЯГОДА ВИНОГРАДА?* С точки зрения наследования свойств ответ на этот вопрос должен быть отрицательным, поскольку изюм не несет многих свойств ягод как плодов некоторого растения: он не растет, не зреет, его не собирают.

Проверка свойств наследования может производиться и на основе общезначимых формальных свойств понятий. Так, для анализа правильности родовидовых отношений Н. Гуарино и К. Велти (Guarino, Welty, 2002) предлагают проверять наследование на видовые понятия такого свойства вышестоящего понятия как «критерий идентичности» .

Суть критерия идентичности некоторого понятия заключаются в том, чтобы определить, что означает, что две сущности, представляющие примеры одного и того же понятия, являются одним и тем же, как может сущность меняться, сохраняя свою идентичность, какие свойства существенны для сохранения своей идентичности и др., Можно говорить о достаточных условиях идентичности, то есть какие условия используются, чтобы определить идентичность и о необходимых условиях идентичности, то есть, что следует из того, что два объекта идентичны.

Например, два человека должны быть признаны одним и тем же лицом, если они находились в одном и том же месте в одно и то же время. Таким образом, условием идентичности физических лиц является физическое совпадение нахождения по месту и времени. Если предполагаемое родовое и видовое понятие имеют разные условия идентичности, то это означает, что между ними не может быть установлено родовидовое отношение.

В дальнейших разделах будут рассмотрены конкретные типы ошибочного описания родовидовых отношений и показано, какие именно критерии могут помочь не допустить такого рода ошибки.

6.3. Смешение типов и ролей

Одной из частых проблем, встречающихся при описании родовидовых отношений, является смешение так наззываемых типов и ролей в одной иерархии.

Например, отношения «тип-тип» (*береза – это дерево*) и отношения «тип-роль» (*яблоко – это пища*) в равной степени могут быть выражено всеми диагностическими тестами, применяемыми для установления родовидовых отношений. Различие заключается в том, что береза остается деревом в каждый момент своего существования, а яблоко может быть использовано в пищу, может быть использовано для других целей, может вообще никак не использоваться.

Достаточно распространенной ошибкой при описании предметной области является размещение понятий-ролей как родовых понятий над понятиями-типами. Например, поскольку работодателем может быть человек или организация, то понятие *РАБОТОДАТЕЛЬ* представляется как вышестоящее, родовое понятие, а понятия *ЧЕЛОВЕК* и *ОРГАНИЗАЦИЯ* представляются как нижестоящие, видовые понятия (Steinmann, 2000). Однако такое представление неточно описывает свойства сущностей, поскольку не каждый человек является работодателем.

Во многих случаях анализ отношения может выявить нарушение основного принципа установления родовидовых отношений о принадлежности всех примеров нижестоящего понятия к классу вышестоящих понятий (см. п. 6.1.), как это происходит при неправильном установлении отношения *ЧЕЛОВЕК* – выше – *РАБОТОДАТЕЛЬ.* Для работы системы логического вывода такая неточность приведет к тому, что система для

каждого экземпляра понятия *ЧЕЛОВЕК* будет делать вывод, что это экземпляр понятия *РАБОТОДАТЕЛЬ,* что в общем случае неверно.

В других случаях проблема не столь очевидна. Например, при установлении отношения *ЯБЛОКО – ПИЩА* разработчик онтологии может учитывать особенности моделируемой предметной области, в которой все или подавляющее большинство яблок могут рассматриваться как пища.

В главе 7 мы подробно рассмотрим, как можно определить ролевые понятия, как можно описать знание о основных ролях того или иного понятия, оставаясь в рамках простых моделей представления знаний и не нарушая принципов установления родовидовых отношений.

Несмотря на то, что размещение ролей как родовых понятий для типов не подчиняется одному из наиболее известных принципов описания родовидовых связей, который заключается в том, что все примеры видового понятия должны всегда быть примерами родового понятия, но эта проблема остается серьезной, поскольку «провоцируется» многими текстовыми источниками.

Например, следующий фрагмент (http://www.giord.ru/0705211117391.php):

> *наиболее используемыми консервантами являются: поваренная соль, этиловый спирт, уксусная, сернистая, сорбиновая, бензойная кислоты и некоторые их соли*

может показаться хорошим источником информации для того, чтобы описать виды консервантов: поваренная соль, этиловый спирт и т.п.

Определение электролита:

> *Электролит - проводник второго рода; вещество, обладающие ионной проводимостью. Электролитами являются:*
>
> - *расплавы солей, оксидов или гидроксидов;*
> - *растворы солей, кислот или оснований в полярных растворителях; а также + твердые электролиты.*

может показаться основанием, например, для установления отношения, что соль является видом электролита.

Однако в таких случаях нужно помнить, что *консервант и электролит* являются ролями веществ - вещество становится консервантом или электролитом только, если попадает в некоторые условия. А поваренная соль и соль как химическое соединение являются типами веществ.

Устанавливая родовидовую связь от типа к роли, мы сообщаем системе некорректное знание, состоящее, например, в том, что любое вещество, относящееся к классу солей, в любой момент времени своего существования в любой ситуации, является электролитом, что далеко не так.

6.4. Смешение отношений класс-подкласс и класс-экземпляр

Современное онтологическое моделирование (Cyc Ontology Guide; Guarino, 1998; Noy, McGuinness, 2001) достаточно четко отличает отношения экземпляр-класс от родовидовых отношений. Это отношение связывает индивидуальные сущности, например, такие как конкретный город – Москва и классы сущностей как *ГОРОД.* Отношение экземпляр-класс характеризуется тем, что в отличие от родовидовых отношений, не является транзитивным отношением.

Многие руководства указывают, что экземпляры – это самые конкретные понятийные единицы, представленные в базе знаний. Так, в (Noy, McGuinness, 2001) приводится пример, что, если в моделируемой предметной области необходимо описать только подбор сочетаний вина и еды, то нас не будут интересовать конкретные материальные бутылки вина. Поэтому такие термины как *Sterling Vineyards Merlot*,

вероятно, будут самыми конкретными используемыми понятийными единицами. Следовательно, *Sterling Vineyards Merlot* будет экземпляром в базе знаний и между этим вином и классом вин должно быть установлено отношение экземпляр-класс.

Сложность, приводящая к смешению этих двух видов отношений, заключается в том, что вопреки сложившемуся мнению отношение экземпляр-класс может встретиться на любом иерархическом уровне понятийной системы, а не только на самых нижних уровнях.

Так, понятие *СПАНИЕЛЬ* связано родовидовым отношением с понятием *СОБАКА* и отношением экземпляр-класс с понятием *ПОРОДА СОБАК,* понятие *ШКОЛЬНЫЙ УЧИТЕЛЬ* связано родовидовым отношением с понятием *ПЕДАГОГИЧЕСКИЙ РАБОТНИК,* и отношением экземпляр-класс с понятием *ПРОФЕССИЯ.* В таких случаях различать эти отношения не всегда просто.

Для различения родовидовых отношений и отношений экземпляр-класс можно воспользоваться принципом идентичности (см. п.6.2.), который утверждает, что у родового понятия и видового понятия должны быть одни и те же критерии идентичности.

Если мы выполним анализ критерия идентичности, например, для понятий *СПАНИЕЛЬ* и *ПОРОДА СОБАК*, то увидим, что критерии идентичности для спаниелей и породы животных различаются. Породы собак идентифицируются с их позицией в некоторой классификации собак. С другой стороны, примеры спаниелей могут, в простейшем случае, идентифицироваться через расположение в пространстве/ времени их тел – два спаниеля различны, если они находились в одно и то же время в разных местах. Поэтому *ПОРОДА СОБАК* не может являться родовым понятием для понятия *СПАНИЕЛЬ.* Спаниель является не подвидом понятия ПОРОДА СОБАК, а его примером.

Точно также конкретный учитель идентифицируется своим физическим местоположением, а профессии - некоторым набором характеристик: полученного образования, опыта работы, необходимых умений. Поэтому понятие *ШКОЛЬНЫЙ УЧИТЕЛЬ* – это экземпляр понятия *ПРОФЕССИЯ*, а не вид.

6.5. Смешение родовидовых отношений и отношений часть-целое

Приведенный во разделе 6.2.пример ошибочной цепочки отношений:

РЕКА – выше – *ВОДОЕМ* – выше – *ВОДНЫЙ ОБЪЕКТ* – выше – *ВОДА* – выше – *ВЕЩЕСТВО*

также соответствует одному из распространенных типов проблем, возникающих при описании родовидовых отношений.

Суть проблемы заключается в том, что некоторая сущность имеет существенную часть, которая занимает значительную долю объема этой сущности, и тогда возникает желание перенести на объемлющую сущность родовидовые отношения этой части.

Эта ошибка не распознается диагностическими тестами. Так, высказывания «река – это вода», «река и другая вода» звучат нормально.

Кроме того, ошибка «провоцируется» толкованиями в словарях: «Река – естественный значительный непрерывный водный поток…» (БТС, 1998). Значит, можно из определения сделать вывод, что река – это вода.

Определение понятия *ВОДНЫЙ ОБЪЕКТ* по Водному кодексу Российской Федерации (Федеральный закон РФ № 167-ФЗ от 16.11.1995) таково:

Водный объект - сосредоточение вод на поверхности суши в формах ее рельефа либо в недрах, имеющее границы, объем и черты водного режима. (ст. 1)

Из такого определения можно сделать вывод, что водный объект – это вода.

В таких случаях обычно помогает анализ наиболее абстрактных семантических классов для видового понятия и для родового понятия – обычно происходит изменение такого семантического класса.

Кроме того, неправильное отношение распознается анализом идентичности для видового и родового понятия: разрушение реки не приводит к разрушению воды – вода просто уходит в другое место.

Таким образом, более аккуратное описание отношений между понятиями РЕКА и ВОДА может выглядеть следующим образом:

РЕКА
ЧАСТЬ *РЕЧНАЯ ВОДА*

РЕЧНАЯ ВОДА
ЦЕЛОЕ *РЕКА*
ВЫШЕ *ВОДА*

Другим примером той же проблемы является, например, отношение между понятиями *КОМПАНИЯ* и *ГРУППА ЛЮДЕЙ* (Guarino, Welty, 2002).

6.6. Смешение родовидовых отношений и отношений происхождения

Еще один вид смешения отношений, уже упоминавшийся во разделе 6.2, связан ошибочным описанием отношения происхождения как родовидового отношения: *ИЗЮМ – ВИНОГРАД*. Как и в предыдущем случае, такая ошибка часто основывается на словарных определениях. Так, янтарь определяется в словарях как «ископаемая смола хвойных деревьев» (БСЭ), но неправильно описывать, что понятие *ЯНТАРЬ* – это вид понятия *СМОЛА*, янтарь происходит от смолы.

Такую ошибку можно распознать за счет анализа свойств и отношений видового и родового понятия. Видовое понятие, полученное смешением отношения происхождения, не наследует многих свойств и отношений родового понятия, а также не наследует принадлежность к классам понятий верхнего уровня (*ЖИВОЕ – НЕЖИВОЕ*) (см. п.6.2).

6.7. Смешение описания сущности и знака

Еще одна проблема неправильного описания отношений класс-подкласс связана с проблемой смешения описания сущности с описанием названия этой сущности. Как уже упоминалось в разделе 5.1., при разработке онтологий важным принципом является различение сущности и ее имени или совокупности имен.

Смешение сущности и ее знака может проявиться и при установлении родовидовых отношений. Так, при описании в некоторой онтологии понятия *ДАРВИНИЗМ* может быть сделана попытка ввести такие понятия как *БИОЛОГИЧЕСКИЙ ТЕРМИН* или *ПОНЯТИЕ БИОЛОГИИ* и описать *ДАРВИНИЗМ* как видовое понятие для этих понятий.

Однако такое описание относится к классификации знака - термина *дарвинизм*. *ДАРВИНИЗМ* как понятие может быть, например, отнесено к классу *БИОЛОГИЧЕСКОЕ УЧЕНИЕ*.

Заключение к главе 6

Таким образом, существует ряд типовых проблем, возникающие при установлении родовидовых отношений в онтологических ресурсах. В нашей работе мы видели, что сотрудники, только начинающие работать с классификациями в рамках разработки тезаурусов, онтологий, непременно совершают практически все из вышеперечисленных ошибок.

Частично это связано с тем, что неявно предполагается, что в определениях терминов, которые даются в различных пособиях, энциклопедиях, словарях, после слова «это» непременно находится родовое понятие, хотя на самом деле спектр отношений, через которые определяется термин или толкуется слово достаточно широк (Шелов, 2003).

Кроме того, часто мысленно употребляемые лингвистические тесты типа «X – это Y» также могут привести к установлению неправильного отношения.

Таким образом, при установлении родовидовых отношений разработчик онтологического ресурса должен использовать знать о возможных типах смешений отношений и в сложных случаях применять набор формальных критериев.

Впрочем, как мы увидим в разделе 18.2.2. особенности реализации реальной прикладной задачи могут приводить к необходимости нарушений тех или иных формальных критериев при моделировании отношений. Однако важно, чтобы такое нарушение критериев было осознанным выбором.

Глава 7. Описание ролей в компьютерных ресурсах

Как было описано в разделе 6.1.2, наиболее часто родовидовые отношения могут быть спутаны с отношением «тип-роль». В этом разделе мы подробно остановимся на особенностях рассмотрения понятий-ролей в онтологическом моделировании

7.1. Концепция роли в онтологических исследованиях

Онтологи обычно различают сущности (то, что есть) и события (то, что случается). Роли в этом разделении занимают «промежуточную» позицию: роли – это то, что есть, но только в контексте того, что случается. В течение многих лет понятие роли активно обсуждается в таких областях как концептуальное моделирование и представление знаний. Наиболее часто роль рассматривается посредством двух дополнительных понятий: игрок и контекст. Например, для роли *студент* игроком является человек, а контекст определяется отношением к высшему учебному заведению.

При рассмотрении ролей существенным фактором является то, что роль сообщает сущности некоторые внешние характеристики.

Источник проблемы описания ролевых отношений лежит в различии между внутренними и внешними характеристиками. Внутренние характеристики, такие как размер или форма, описывают сущность в изоляции. В противоположность, внешние характеристики описывают сущность относительно других сущностей и событий. Например, свойство «обычно использоваться для забивания гвоздей» является внешней характеристикой молотка, поскольку устанавливает отношение между молотком и гвоздями.

Внешние и внутренние характеристики сущности обладают разными свойствами. Например, такие внутренние характеристики как возраст могут изменяться в течение времени, но они всегда приложимы к сущности. Напротив, внешние характеристики, такие как зарплата работника, могут оказаться совершенно неприменимыми в некоторых ситуациях, например, когда работник участвует в спортивном соревновании. Более того, в отличие от внутренних характеристик, внешние характеристики могут быть противоречивы, например, как величина зарплаты человека на разных работах.

Steimann (Steimann, 2000) выделяет 15 характеристик ролей, из которых наиболее существенными являются следующие пять:

1) Роли создаются и исчезают динамически.

Из-за того, что роль представляет внешние характеристики сущности, связанные с его участием в некотором событии, роль создается, когда начинается событие. Если сущность прекращает участие в событии, то роль может прекратить существовать и ее свойства более не будут действительными.

2) Роль может передаваться между сущностями.

Например, роль менеджера может переходить от одного человека к другому. Заметим, что многие из характеристик ролей могут передаваться без изменения, тогда как другие могут «перевычисляться» в зависимости от новой сущности, играющей роль. Например, работник получает премию 20 процентов за то, что он менеджер, а заработная плата может пересчитываться в зависимости от конкретного лица.

3) Сущность может играть разные роли одновременно.

Например, одно и то же лицо может быть работодателем и наемным работником.

4) Сущности разных, не связанных между собой типов, могут выступать в одинаковых ролях.

Например, крекер (для человека) и муха (для лягушки) могут выступать в роли пищи.

5) Роли могут играть роли.

Это отражает ситуацию, что работник может быть, например, руководителем проекта, что, в свою очередь, является ролью работника.

7.2. Критерии распознавания ролей

Дж.Сова (Sowa, 1988) определяет понятие-роль следующим образом: «Подтипы сущности могут быть двух видов: натуральные типы и ролевые типы, которые являются подтипами натуральных типов в конкретных образцах отношений (particular pattern of relationships). Человек, например, является натуральным типом, а учитель – это подтип человека в ситуации обучения». Сова предлагает простой тест для определения, является ли понятие ролью:

> *r – является ролевым типом, если сущность может быть охарактеризована как r только при рассмотрении другой сущности, действия или состояния.*

В соответствии со взглядом Дж. Совы роли ассоциируются с отношениями, но при этом они сущности, а не отношения.

В работе (Guarino, 1998) Н. Гуарино отмечает, что тест Совы для различения типов и ролей недостаточен: например, нечто может быть охарактеризовано как автомобиль, только если оно имеет, по крайней мере, колеса и мотор, но автомобиль является типом, а не ролью.

В работе (Guarino, Welty, 2002) условие, сформулированное Совой, заменяется на условие так называемой внешней онтологической зависимости:

> *Понятие C1 называется внешне зависимым от понятия C2, если для всех примеров C1 должен существовать пример C2, который не является частью или материалом примера C1.*

Авторы работы (Masolo и др., 2004) замечают, что точнее это условие можно сформулировать так:

> *пример C2 не должен быть внутренним для примера C1, то есть не должен быть частью, или материалом, или качеством (цвет).*

Например, *сын* является внешне зависимым, поскольку существует только в рамках семьи по отношению к своим родителям. С другой стороны, автомобиль не является внешне зависимым, поскольку требует существования мотора, который является частью автомобиля.

Таким образом, данное условие формализует определение ролей, данной Совой.

Вместе с тем, на основе такого определения в класс ролей попадают еще дополнительные сущности такие, как качества: =цвет=, =вес=, =скорость=. Например, если синий – это цвет, то обязательно существует хотя бы один объект, цвет которого синий, при чем этот объект не является частью цвета. Понятие цвета поэтому является зависимым, но не кажется подходящим к понятию роли.

Поэтому в работе (Guarino, Welty, 2002) вводится еще одно условие, которое вместе с условием внешней зависимости дает лучшее определение понятию «роль».

> *Понятие C является семантически жестким (rigid), если любой пример понятия C остается примером C в течение всего своего существования.*

Например, щенок перестает быть щенком, все еще оставаясь собакой, поэтому собака и животное – это жесткие сущности, а щенок не является жестким понятием.

Таким образом, понятие C называется ролью, если оно является внешне зависимым и не является семантически жестким (Guarino, Welty, 2002).

В соответствии с этим определением качества не могут быть ролями, поскольку они являются семантически жесткими: если цвет прекратит быть цветом, то он станет чем-то еще, потеряет свою идентичность.

Таким образом, оба условия вносят вклад в определение роли. Первое условие описывает, что сущность-роль должна рассматриваться в рамках чего-либо объемлющего, что в определении Дж. Совы называлось *particular pattern of relationships*, второе условие помогает формализовать условие конкретности, особенности, упоминаемое в определении Дж.Совы (Sowa 1998, 2000).

Введенные понятия онтологической зависимости и семантической жесткости помогают формализовать понятие натурального типа.

Понятие C называется натуральным типом, если оно существенно независимо и семантически жестко.

Таким образом, собака – это пример натурального типа. Цвет не является ни ролью, ни натуральным типом: цвет - семантически жесткий, но является онтологически зависимым. Щенок или хромая собака также не являются ни натуральным типом, ни ролью, поскольку они являются независимыми от внешних сущностей и являются семантически жесткими.

7.3. Типы понятий-ролей

По типу контекста, определяющего роль, можно выделить три вида понятий-ролей (Loebe, 2005; Mizoguchi и др., 2007):

- реляционные роли (способы участия в отношении), например, роль «друг» в отношении «дружба», роли муж, жена в отношении «супружество»,
- процессуальные роли (способы участия в событии), например, роли сырья, продукции в процессе производства,
- социальные роли (способы участия в некотором сообществе), например, «профессор», «премьер-министр».

Каждый вид понятий-ролей имеет свои особенности. Так, для социальных ролей характерно, что такие роли не ограничиваются временем участия в конкретном событии: музыкант остается музыкантом, даже когда спит. Кроме того, в отличие от реляционных и процессуальных ролей социальные роли онтологически зависят от социальных институтов: организаций, нормативных систем, предприятий.

Реляционные и процессуальные роли могут быть объединены понятием так называемой абстрактной роли, которое контрастирует с понятием социальной роли. Абстрактные роли могут единообразно характеризоваться через рассмотрение одной сущности – игрока в определенной контексте. Игроки абстрактных ролей рассматриваются с точки зрения их внешних свойств в противоположность рассмотрения их как внутренних свойств, например, рассматривая части и качества.

Это общее понимание абстрактных ролей противопоставляется социальным ролям, которые имеют свои собственные свойства, отношения, процессы, в которых они участвуют. Рассматривая пациента как социальную роль, в которой пациент имеет собственный идентификационный номер, другие особенности в форме прав, норм и обязанностей, трудно найти дополнительные роли как в случае отношения «пациент-врач». Действительно, социальные роли скорее агрегируют различные процессуальные и реляционные роли. При рассмотрении социальных ролей контекст становится имплицитным, фокус смещается к особенностям собственно роли, так же как их отношения к игрокам (Loebe, 2005).

В работе (Mizoguchi и др., 2007) отдельно выделяется класс составных ролей, то есть ролей, который играют другие роли. Например, премьер-министры многих стран должны быть одновременно гражданами своих стран. Таким образом, в классификации ролей выделяется класс примитивных ролей, которые зависят от одного контекста, и составных ролей, которые зависят от совокупности контекстов.

7.4. Роли как части контекста

В работе (Loebe, 2005) реляционные роли рассматриваются как части произвольного отношения r, а процессуальные роли как части процесса p.

Роль в процессе обычно рассматривается как то, что описывает, как сущность участвует в процессе. Обычно роли в процессе не рассматриваются как собственно процессы. Однако процессуальная роль может быть определена как часть процесса, относящаяся к одному участнику. Процессуальная роль – это «фрагмент» процесса, в котором один участник остается неизменным. Процесс может быть разложен на свои процессуальные роли.

Работа (Loebe, 2005) поясняет идею процессуальных ролей следующим образом.

Представим сцену фильма, включающую несколько действующих лиц. Предположим, что сцена была подправлена так, что одно из действующих лиц было удалено. Но по длительности сцена не меняется. Другим иллюстрирующим примером может служить представление мима. В строгом варианте мимы не используют предметов, тем не менее они действуют так, как будто предметы находятся вокруг них. Это означает, что они выполняют процессуальные роли, в то время как дополняющие эти роли другие процессуальные роли отсутствуют.

Утверждается, что и для социальной роли всегда можно найти социальное сообщество, частью которого является эта социальная роль. Такими социальным сообществом может быть юридически оформленная организация (профессор в ВУЗе), группа лиц (предводитель банды), социальная система (президент государства).

Контекст роли может пониматься как наиболее исчерпывающее ЦЕЛОЕ, в котором роль интерпретируется как часть. Рассматривая наиболее общее понимание отношения «часть-целое», можно рассмотреть отношение между ролью и контекстом (role_of) как специализацию отношения «часть-целое».

Рассмотрение реляционных ролей как специализации отношения часть-целое может показаться неочевидным. Однако, реляционные роли используются в компьютерном моделировании и представлении знаний, и здесь они проявляют особенности частей по отношению к отношениям как их целому.

7.5. Представление ролей в компьютерных ресурсах

Существует три традиционных, относительно несложных подхода к представлению ролей (Fan и др., 2001). Первый подход (1) рассматривает роль только как метку, приписанную к участнику ситуации. Например, роль нанимателя отмечает роль агента в событии найма на работу. Такой подход сходен с описанием семантических ролей в семантической модели управления глагола (Апресян, 1995).

Этот подход прост, но он не представляет роли отлично от сущностей (комбинируя внешние и внутренние характеристики понятий в единое представление сущности). Как следствие, становится невозможным описать собственные свойства ролей. Например, пятое свойство ролей из раздела 7.1 говорит о том, что роли могут играть роли, то есть у ролей есть своя собственная классификация, которую при данном способе представления невозможно отразить.

Второй подход представляет роли и сущности отдельными понятийными единицами, однако комбинирует эта два типа понятий в рамках одной иерархии. В одной иерархии роли и типы могут быть объединены двумя способами.

Во-первых (2.1), роли могут описываться как вышестоящие понятия для типов, которые могут их занимать.

Например, поскольку работодателем может быть человек или организация, то понятие *РАБОТОДАТЕЛЬ* (Рис.7.1) представляется как вышестоящее, родовое понятие, а понятия *ЧЕЛОВЕК* и *ОРГАНИЗАЦИЯ* представляются как нижестоящие, видовые понятия. Против такого представления выступают многие онтологи (Guarino, 1998; Sowa, 2000; Steinmann, 2000). Действительно, такое представление неточно описывает свойства сущностей, поскольку не каждый человек является работодателем. Нарушается основной принцип установления родовидовых отношений (см. п. 6.1.)

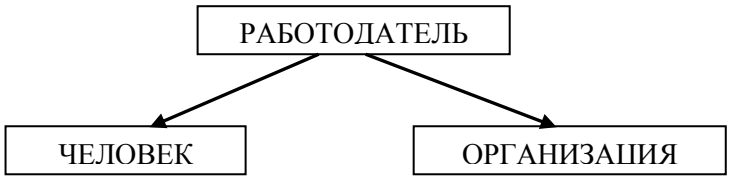

Рис.7.1. Расположение роли над типами сущностей нарушает основной принцип установления родовидовых отношений

Во-вторых (2.2), роли можно описывать как нижестоящие понятия для сущностей, которые могут их занимать.

Тогда понятие *РАБОТОДАТЕЛЬ* может быть представлено, например, как нижестоящее понятие для понятия *ОРГАНИЗАЦИЯ* (Рис.7.2). Однако, если нужно отразить знание, что работодателем может быть и человек, то ситуация несколько усложняется. Если теперь представить понятие РАБОТОДАТЕЛЬ как подтип понятий *ЧЕЛОВЕК* и ОРГАНИЗАЦИЯ, то получится, что работодатель одновременно и человек, и организация.

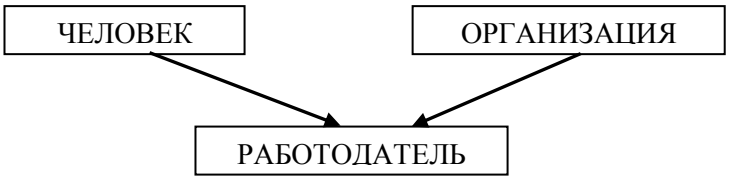

Рис. 7.2. Расположение роли под двумя возможными типами может привести к неправильному логическому выводу

Чтобы описать, что работодатель может быть человеком или организацией, может быть введено дополнительное понятие, например, с названием *СУБЪЕКТ ДЕЯТЕЛЬНОСТИ* (Рис.7.3), подтипами которого являются понятия *ЧЕЛОВЕК* и *ОРГАНИЗАЦИЯ*.

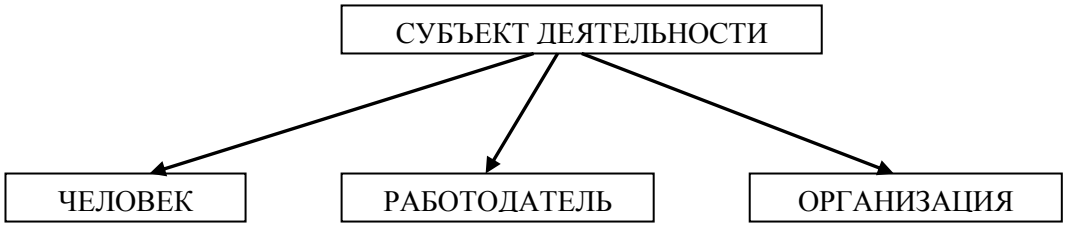

Рис. 7.3. Введение дополнительного понятия для отражения сложных взаимоотношений между типами и ролями

Далее устанавливается отношение между понятием *РАБОТОДАТЕЛЬ* и понятием *СУБЪЕКТ ДЕЯТЕЛЬНОСТИ.* Таким образом, понятие *РАБОТОДАТЕЛЬ* оказывается на одном уровне иерархии с понятиями *ЧЕЛОВЕК* и *ОРГАНИЗАЦИЯ,* что, с одной стороны, не описывает первоначального утверждения, что работодатель является либо человеком, либо организацией, а с другой стороны все-таки может использоваться как модель для представления ролевых понятий.

В работе (Gangemi и др., 2001b) авторы для уточнения возможностей совмещения представления ролей и типов в одних и тех же иерархиях разделяют роли на два подтипа: *материальные роли и формальные роли*.

Как указывают авторы, формальные роли не несут идентичности, то есть не относятся ни к какому конкретному типу, могут быть применены к любому типу. В качестве формальных ролей можно привести пример таких ролей как *часть* и *целое, инструмен*т и т.п.

В качестве материальных ролей рассматриваются такие роли как *студент* (должен относиться к типу человек) или *еда* (является физической сущностью).

Авторы считают, что формальные роли должны представляться только в таксономиях ролей, материальные роли могут быть представлены как видовые понятия для классов и ролей, а сами могут подчинять как видовые понятия только материальные роли.

Наконец, в третьем подходе (3) предполагается, что иерархия ролей должна быть представлена отдельно от иерархии типов. Иерархия ролей подчиняется самому верхнему узлу иерархии. В таком представлении понятия-роли описываются независимо от типов, но каждый пример роли существует только как пример типа, то есть пример роли не может существовать независимо от примера типа, которые может занимать эту роль.

Существуют и значительно более сложные представления ролей, назначения которых предоставить формализм, в рамках которого можно описать все особенности ролей (см. например, (Masolo и др., 2004; Mizoguchi и др., 2007)).

Однако, при представлении ролей с помощью усложненных формализмов, возникает ряд специфических проблем. Авторы работы (Mizoguchi и др., 2007) в качестве примера трудностей приводят роли врача и медсестры в онтологии больницы. Авторы работы указывают, что люди предпочитают рассматривать эти сущности как базисные понятия, типы при построении таких онтологий, поскольку предполагается, что в онтологии больницы врач всегда врач, а медсестра –всегда медсестра. Необходимость в усложненном рассмотрении этих сущностей как ролей возникает, когда врач заболевает и приходит в больницу на прием в качестве пациента.

7.6. Роли в тезаурусах

Проблемы описания понятий-ролей, отношений тип-роль находят свое отражение и в процессе построения и использования тезаурусов разных типов.

Так, одной из проблем WordNet, на которую обращали внимание многие исследователи, является смешение нескольких разных отношений под именем отношения гипонимии-гиперонимии (см. п. 2.5.3.2.)

Указывая на смешение типов и ролей в Wordnet, Н. Гуарино (Guarino, 1998) привел следующие примеры описания из WordNet:

Человек – это живое существо и каузальный агент.

Яблоко – это фрукт и еда.

Человек всегда живое существо, но он (она) начинает играть роль каузального агента только в некоторых ситуациях. Та же проблема возникает для яблока, которое всегда плод растения, и в некоторых ситуациях может быть пищей. Проблема в том, что

человек и яблоко – это типы сущностей, в то время как каузальный агент и пища – это роли.

Одна из авторов WordNet К. Феллбаум, отвечая на эту критику Н. Гуарино, заявляет (Fellbaum, 2002), что в таких ресурсах, как WordNet неоднородные классификации имеют право на существование, поскольку такие ресурсы рассматриваются в настоящее время, прежде всего, как инструменты для компьютерной обработки текстов, а не только как совершенные онтологии, которые должны соответствовать строгим онтологическим принципам.

Во-первых, указывает К. Феллбаум, если уничтожить «неправильные» отношения, то теряется важная информация. В некоторых случаях семантическая информация об отношениях между словами, не отвечающая строгим принципам, может быть более полезна, чем более обоснованное семантическое отношение.

Во-вторых, считает К. Феллбаум, для лексических ресурсов, которые используются для компьютерной обработки текстов, полезно иметь подробную сеть отношений. Это важно для разрешения многозначности, разрешения референции, методов выявления лексической связности текстов.

Более того, К. Феллбаум (Fellbaum, 2002) предлагает расширить множество отношений, подобных отношением тип-роль. Предлагается ввести и использовать другой тип отношений, который назван в (Cruse, 1986) парагипонимией, а в качестве лингвистического теста установления такого рода отношения применять следующие пары предложений:

X's and other Y's & It's an X, but it's not a Y. (X и другие Y, & Это X, но не Y)

Однако проблемой такого предложения для описания отношений «тип-роль» является то, что под вышеуказанные тесты подходят многие экзотические роли, например, в некоторых ситуациях мухи могут стать едой (например, в голодное время), а бутылки музыкальным инструментом (Trautwein, Grenon 2004). Лингвистические тесты не препятствуют установлению такого рода отношений, например:

Мухи и другие виды еды & Это муха, а не еда.

Отметим, что высказанное предложение не было реализовано в WordNet.

Вместе с тем, понятно, что, если использовать отношения вне правильного контекста, что часто возникает в связи с динамичностью ролей, то это может привести к серьезным ошибкам обработки текстов.

Если обратиться к информационно-поисковым тезаурусам, то стоит отметить, что современные руководства и стандарты по информационно-поисковым тезаурусам (Z 39.19; Will, 2004) рекомендуют придерживаться строгих принципов в представлении ролей. Как было описано в п. 1.2.1.1., рекомендуется устанавливать иерархические отношения в информационно-поисковых тезаурусах в тех случаях, когда отношения истинны независимо от контекста, - только в таких случаях дескрипторы информационно-поискового тезауруса могут быть организованы в иерархии.

В упомянутом примере (см. п.1.2.1.1), обсуждающем правильность установления родовидового отношения от дескриптора МЫШИ к дескриптору ВРЕДИТЕЛИ, последний как раз и является ролью. Авторы руководств и стандартов считают, что такое отношение представлять как родовидовое неправильно, поскольку имеются лабораторные мыши и домашние мыши, которые не являются вредителями.

Заключение к главе 7.

В данной главе были рассмотрены подходы к представлению ролей в онтологических и лингвистических ресурсах. Мы показали, что различие между онтологическими характеристиками понятий-типов и понятий-ролей имеет существенное

значение для представления знаний о предметной области. В различных исследованиях ведется достаточно интенсивная дискуссия о принципах определениях и описания ролей.

На наш взгляд, описанные проблемы представления ролей обязательно нужно учитывать при разработке онтологических ресурсов, предназначенных для автоматической обработки текстов.

Глава 8. Отношения часть-целое

Отношение ЧАСТЬ-ЦЕЛОЕ играет существенную роль во многих предметных областях. Необходимость описания этого отношения возникает при создании таких разных ресурсов как информационно-поисковые тезаурусы, лингвистические ресурсы для компьютерной обработки текстов, онтологии, в объектно-ориентированном программировании.

Для компьютерных приложений особое значение представляет такое свойство отношения ЧАСТЬ-ЦЕЛОЕ как транзитивность, на основе которой может строиться многошаговый логический вывод. Для некоторых типов отношений ЧАСТЬ-ЦЕЛОЕ может выполняться наследование свойств и операций от части к целому и от целого к части, что также может стать базой для логического вывода.

Специфической особенностью этого отношения является его разнообразие: отношение может быть установлено между сущностями различных семантических типов: физическими объектами, процессами и действиями, географическими регионами, свойствами и состояниями, коллекциями и множествами, абстрактными сущностями такими как числа. Изучающая это отношение отрасль философии - мереология - не накладывает никаких ограничений на типы части и целого, лишь постулируя три основных аксиомы (Varzi, 2006; Gangemi и др., 2001b) (см.п.8.1).

Другой стороной явного отсутствия ограничений на участников отношения ЧАСТЬ-ЦЕЛОЕ является то, что его описание осложнено наличием достаточно разнообразных подвидов этого отношения, а также существованием совокупности отношений, которые могут быть ошибочно отождествлены с этим отношением. Разработчики компьютерных ресурсов в конкретных предметных областях сталкиваются с достаточно сложной ситуацией такого описания отношений ЧАСТЬ-ЦЕЛОЕ, которое сохранило бы его важнейшие свойства.

В данном разделе мы рассмотрим основные проблемы описания отношения ЧАСТЬ-ЦЕЛОЕ, его определения в рамках философии и лингвистики, а также примеры его моделирования в конкретных компьютерных ресурсах.

8.1. Определение отношения ЧАСТЬ-ЦЕЛОЕ в философии и лингвистике

В классической мереологии обычно постулируются три аксиомы (Varzi 1996, Varzi 1998):

1) Транзитивность. Части частей являются частями целого:

P(A, B) ∧ P(B,C)->P(A,C) :
Если A – часть B и B - часть C, то A – это часть C.

2) Рефлексивность. Все является частью самого себя:

P(A, A) :
A – это часть A.

3) Несимметричность: ничто не является частью своих частей:

P(A, B) ∧ ¬EQ (A, B) -> ¬P(B, A) :
Если A - часть B и A не совпадает с B, то B – это не часть A (Varzi, 2006, Gangemi и др., 2001b)

Поскольку отношение обладает свойством рефлексивности, выделяются еще собственно-части, то есть части, не равные своему целому.

В лингвистике для определения отношения ЧАСТЬ-ЦЕЛОЕ широко используются лингвистические тесты, то есть некоторые заданные предложения, в которые подставляются анализируемые сущности. При этом *часть* обычно называется меронимом, а *целое* – холонимом.

Естественным тестом для определения меронимии является предложение *X – это часть Y,* которое должно звучать нормально для X и Y, интерпретируемых как родовые понятия*: палец – это часть руки, страница – это часть книги* (Cruse, 1986).

Другими возможными лингвистическими тестами для отношения ЧАСТЬ-ЦЕЛОЕ являются следующие:

Части У включают X, Z и т.д.: Части слова включают корень, приставку, суффикс

X и другие части Y.: Суффикс и другие части слова

Однако, многие авторы отмечают, что если применять лингвистические тесты, то возникают серьезные проблемы с транзитивностью отношения ЧАСТЬ-ЦЕЛОЕ. Так, например, рассмотрим следующую совокупность утверждений: *рука – это часть дирижера, дирижер – это часть оркестра,* но странно, если сказать, что *рука – это часть оркестра.*

8.2. Разнообразие отношений ЧАСТЬ-ЦЕЛОЕ

Многие исследователи указывают, что отношение ЧАСТЬ-ЦЕЛОЕ представляет собой скорее совокупность несколько отличающихся отношений, чем четкое отделяемое отношение (см. также п. 8.3.):

- наиболее центральный тип этого отношения представляют физические объекты
- сущности, длящиеся во времени, могут иметь части, которые называются стадиями, фазами или этапами.
- сущности такие как группы (*племя, команда*), классы (*пролетариат, буржуазия*) и коллекции (*куча*) состоят в отношении меронимии со своими элементами.
- если и ЦЕЛОЕ и ЧАСТЬ являются неисчислимыми, то можно говорить об отношении ингредиентов, например, *заварной крем и молоко.*
- если ЦЕЛОЕ – исчислимое, а часть неисчислимое, то это так называемое отношение объект-материал: *бокал - стекло*.
- если ЦЕЛОЕ неисчислимое, а ЧАСТЬ – исчислимое, то говорят об отношении вещество-частица: *песок – песчинка, снег – снежинка, дождь – капля*.(Cruse, 1986).
- не изменяющиеся во времени состояния и свойства могут быть описаны как имеющие части: *Самоконтроль – это часть зрелости* (Cruse, 1986).
- другими периферийными меронимами могут быть названы характеристики событий, состояний. На них возможна ссылка как на части. Однако более нормально их называть признаками. *Рождественский пудинг – это часть (признак) Рождества*.

На сходство в некоторых случаях признаков и частей указывается в работе (Никитина, 1987). А.И. Уемов (Уемов, 1963) отмечает, что в логике, между вещью, свойством и отношением нет четкой границы, и все зависит от способа рассмотрения. Дж. Сова (Sowa, 2000) считает, что части в отличие от свойств и атрибутов могут существовать отдельно от своего целого.

F. Loebe (Loebe, 2007) рассматривает роли как части отношений и процессов (см. п. 7.4.) Например, процесс может быть разделен не на временные части, на части по отношению к конкретным участникам: «Когда Джон двигает ручкой, он и ручка формируют участников процесса, и процессуальная роль, которую играет Джон, включает

то, что Джон делает в течение этого процесса. Мим, который двигает воображаемой ручкой, может быть хорошей иллюстрацией понятия процессуальной роли.»

8.3. Классификация отношений ЧАСТЬ-ЦЕЛОЕ

Предложены различные классификации отношений ЧАСТЬ-ЦЕЛОЕ.

Основная идея авторов работы (Winston и др., 1987) состоит в том, чтобы классифицировать различные способы, которыми части соотносятся к своему целому, введением шести различных типов отношений меронимии, различаемых на основе признаков функциональности, гомеомерности и отделимости:

1) *Функциональные части* ограничены своей функцией, в их пространственном или временном положении. К примеру, ручка чашки может быть расположена в ограниченном числе позиций, если она должна функционировать как ручка.
2) *Гомеомерные части* представляют собой объект того же типа, что и их целые, к примеру, кусок – пирог, тогда как негомеомерные части отличны от своих целых, например, дерево-лес.
3) *Отделимые части* могут, в принципе, быть отделенными от целого, к примеру, ручка – чашка, тогда как неотделимые не могут, например, сталь – велосипед.

На основе комбинации этих трех признаков можно выделить следующие типы отношений ЧАСТЬ-ЦЕЛОЕ:

1) *Компонент/Интегральный объект:* Интегральные объекты характеризуются своей структурой, тогда как их компоненты отделимы и имеют специальную функциональность. Например, *Колеса – это части автомобиля, фонология – это часть лингвистики.*
2) *Член/Коллекция:* Члены не играют никакой функциональной роли по отношению к своему целому, они отделимы: *дерево – часть леса*.
3) *Порция/Масса:* Целое рассматривается как гомогенный агрегат и его порции подобны целом: *Кусок – это часть пирога*.
4) *Материал/Объект:* Материал, из которого сделан объект, не отделим от этого объекта, не имеет функциональной роли и негомеомерный: *Сталь – часть велосипеда.*
5) Фаза/Деятельность: Фаза, как компонент, имеет функциональную роль, но не отделима: ложка – часть процесса еды, глотание – часть процесса еды.
6) *Место/местность*: часть гомеомерна, так как каждая часть пространственного региона подобна целому региону, но не отделима: *Оазис – это часть пустыни*.

Классификация отношений часть-целое на основе других принципов приводится в работе (Gerstl , Pribennow, 1996).

Прежде всего, части подразделяются на структурные части (structural dependent parts), то есть те, которые зависят от структуры целого, и временные части (temporarily-constructed parts).

Структурные части – это части, на которые интегральный объект может быть отделен или те части, которые существовали как отдельные объекты в момент создания интегрального объекта. Структурные части разделяются на компоненты, элементы и количества (quantites).

Временные части делятся на сегменты, куски и порции.

Сегменты – это части объекта, выделяемые на основе некоторых внешних критериев, то есть это такие топологические понятия как низ, верх, отверстие, дыра. В отличие от компонентов сегменты не демонстрируют четких границ между собой. Так, например, мраморная статуя может не иметь внутреннего разделения на голову и шею.

Кусок – это часть объекта, получившаяся в результате произвольного механического разделения этого объекта.

Порции - это части, определяемые на основе внутренних критериев, как деревянная часть дома.

При всем разнообразии подвидов отношения ЧАСТЬ-ЦЕЛОЕ, существуют классы отношений, которые не рекомендуется путать с отношениями ЧАСТЬ-ЦЕЛОЕ.

Например, к таким отношениям относится отношение «местонахождение внутри»: тот факт, что некто находится в комнате, не означает, что этот некто является частью комнаты(Noy, Wallace, 2005).

8.4. Проблема транзитивности отношения ЧАСТЬ-ЦЕЛОЕ

В философии постулируется транзитивность отношения ЧАСТЬ-ЦЕЛОЕ. Это свойство отношения является существенным для многих компьютерных приложений. Однако разные авторы приводят многочисленные примеры нарушения транзитивности этого отношения (Cruse, 1986; Winston и др., 1987; Motschnig-Pitrik, Kaasboll, 1999; Johansson, 2004), например:

Несомненно, что
Ручка двери является частью двери, дверь является частью дома,
но странно сказать, что
ручка двери является частью дома;

Несомненно, что
Клеточное ядро является частью клетки, клетка является частью органа,
но странно сказать, что
клеточное ядро является частью органа;

Несомненно, что
Ветка является частью дерева, дерево является частью леса,
но странно сказать, что
ветка является частью леса;

Несомненно, что
Ручка ложки часть ложки, ложка – это часть процесса поедания супа,
но странно сказать, что
ручка ложки есть часть процесса поедания супа.

Рассматривая разные виды отношения ЧАСТЬ-ЦЕЛОЕ, авторы обычно подчеркивают, что проблемы с транзитивностью связаны со смешением разных видов отношений ЧАСТЬ-ЦЕЛОЕ.

В работе (Winston и др., 1987) проблемы с транзитивностью объясняются следующим образом: пока используется один тип отношения, то часть-целое всегда транзитивно. Однако, когда смешиваются различные отношения меронимии, то возникает проблема с транзитивностью.

В работе (Cruse, 1986) подчеркивается, что правильно сформированная иерархия состоит из элементов одного и того же типа. Так, если элемент меронимии физический объект, то и все другие элементы меронимии должны быть такими же (например, вес тела не должен фигурировать среди его частей). Если один элемент – является географической областью, то и другие должны быть такими же (так Вестминстерское аббатство не является частью Лондона), если один элемент - абстрактное существительное, то и другие должны быть такими же.

Д. Круз (Cruse, 1986) приводит в пример одно из известных видов меронимии - человеческое тело. «Почему меронимия не идет дальше? К семье, населению, биосфере? Что квалифицирует тело как ЦЕЛОЕ? Частичный ответ заключается в том, что переход от тела к семье означает переход от связного физического объекта к сущности, не имеющей физической связности. Семьи имеют части, конечно, но это люди, а не тела.»

В работе (Motschnig-Pitrik, Kaasboll, 1999) предлагается выделить те отношения часть-целое, которые, комбинируясь, дают приемлемые результаты транзитивности, и отделить те отношения часть-целое, которые могут привести к ошибочным транзитивным заключениям. Если моделировать такие отношения как *член/коллекция, материал/объект* отношениями, отличными от отношений ЧАСТЬ-ЦЕЛОЕ, то авторы утверждают, что оставшиеся типы отношений демонстрируют транзитивное поведение, даже если комбинируются произвольным образом. Таким образом, группа отношений компонент/объект, порция/масса, фаза/деятельность, место/местность может быть названа базовыми отношениями ЧАСТЬ-ЦЕЛОЕ. В рамках любой комбинации базовых отношений ЧАСТЬ-ЦЕЛОЕ действует правило транзитивности, независимо от комбинации конкретных видов отношений.

Другое мнение высказывается в философской работе (Varzi, 2006). Автор работы утверждает, что проблемы с транзитивностью отношения ЧАСТЬ-ЦЕЛОЕ и приводимые контрпримеры связаны с неявным сужением понятия «часть» в обыденной речи. То, что ручка двери, являясь функциональной частью двери, может не рассматриваться как функциональная часть дома, не означает, что ручка не является вообще частью дома. Напротив, ручка двери проявляет все обычные свойства частей: масса ручки является частью массы дома; она занимает часть пространства, занятого домом; она будет уничтожена, если уничтожить дом; если уничтожить ручку двери, то и дом будет поврежден.

Если рассмотреть пример: *рука дирижера – дирижер – оркестр,* то также можно видеть, что масса руки является частью массы оркестра, рука дирижера занимает часть пространства занимаемого оркестром, если будет повреждена рука дирижера, это может вызвать и (может быть даже серьезные) проблемы с функционированием оркестра.

Сужение понятия «часть» заключается в том, что на интерпретацию понятия «часть» накладываются дополнительные условия (то есть дополнительное требование, что часть должна быть функциональной и т.п.) и при этом, действительно, свойство транзитивности может не выполняться. В более общем виде, если x – ф-часть (то есть часть с дополнительным условием ф) от y и y – ф-часть от z, x не обязательно является ф-частью от z. Модификатор отношения ф – может не быть транзитивным, но эта ситуация говорит лишь об отсутствии транзитивности у отношения ф-часть, а не у обобщенного отношения в целом.

Отметим, что если проанализировать упомянутые в п. 8.3 отношения «местонахождение внутри», то можно заметить, что многие повреждения и даже уничтожение находящегося внутри комнаты, могут не приводить ни к каким повреждениям самой комнаты.

8.5. Вертикальные» отношения между частью и целым

Помимо свойств транзитивности существенную роль в компьютерных системах могут иметь так называемые вертикальные отношения между частью и целым. Вертикальные отношения между частью и целым описывают зависимость свойств части от свойств целого и свойств целого от свойств части (нескольких частей) (Artale и др., 1996).

Части могут быть входить в состав различных целых, так, например, ручки, шестеренки, ножки могут входить в состав множества разных предметов. Поэтому отношение часть-целое можно рассматривать с позиции факультативности-обязательности отношения относительно части и относительно целого.

Так, пальцы обязательны для руки, в то время как ручка факультативна для двери. При этом здесь не идет речь о логической необходимости. На руке может и не быть пальца (пальцев), но тогда рука воспринимается как имеющая физический недостатки (Cruse 1986).

Для описания вертикальных отношений между частью и целым также используется понятие *зависимости существования (экзистенциальной зависимости)* (Artale и др., 1996; Gangemi и др., 2001b), то есть рассматривается, насколько целое может существовать, если не существуют его части, и могут ли существовать части, если не существует целое. Здесь для анализа отношений *часть-целое* фактически используются отношения онтологической зависимости (см. далее главу 9).

Если целое не может существовать, пока не существуют его некоторые части – эти части называются существенными частями. Если часть не может существовать, пока не существует целое, то такая часть называется зависимой частью. Если существует максимум одно целое, содержащее данную часть, то такая часть – эксклюзивна для целого.

Таким образом, можно сказать:

Грибы и водоросли являются существенными частями лишайника,

Деревья являются существенными частями леса.

Этаж является эксклюзивной и зависимой частью здания.

Передовица является эксклюзивной и зависимой частью газеты и др.

В число вертикальных отношений между частями и их целыми относится также и возможность наследования некоторых свойств как от целого к части, так и от части к целому (Artale и др., 1996). Так, примером свойства, которое целое наследует от своих частей является его дефектность: во многих случаях целое дефектно, если дефектны его части.

Свойства, которые части наследуют от своего целого, включают, например, свойство расположения (быть под столом), которое может выполняться как для целого, так и для его частей.

Могут возникать отношения между свойствами частей и свойствами целых: например, пространство, которое обычно занимает часть находится внутри пространства, занимаемого целым, или вес одной части меньше веса целого.

Примером операции, которая наследуется от целого к части, является операция форматирования документа (Motschnig-Pitrik, Kaasboll, 1999): если форматируется целый документ, то, значит, форматируются все его части.

8.6. Отношение ЧАСТЬ-ЦЕЛОЕ в компьютерных ресурсах и подходах

При моделировании отношения ЧАСТЬ-ЦЕЛОЕ в различных компьютерных ресурсах разработчики стараются учесть особенности этого отношения так, чтобы его можно было использовать для тех или иных операций в соответствующих программных приложениях. Рассмотрим некоторые подходы к описанию отношения ЧАСТЬ-ЦЕЛОЕ.

8.6.1. Отношение ЧАСТЬ-ЦЕЛОЕ в объектно-ориентированных моделях

В 90-х годах в научных публикациях, посвященных объектно-ориентированным подходам в программировании, развернулась дискуссия о необходимости выделения отношений ЧАСТЬ-ЦЕЛОЕ из общего набора атрибутов, приписываемых классам (Motschnig-Pitrik, Kaasboll, 1999; Artale и др., 1996), в связи с особыми свойствами этого отношения.

Результаты этой дискуссии нашли свое воплощение, например, в языке для моделирования объектно-ориентированных систем UML (Леоненков, 2001). В этом языке среди всех отношений ассоциации выделяются отношения агрегации, которые служат для

описания отношений между агрегатом (целое) и его составной частью. Дополнительно среди отношений агрегации выделяются отношения композиции, которые устанавливаются в тех случаях, когда части целого имеют такое же время жизни, что и само целое. Эти части уничтожаются вместе с уничтожением целого.

Отношение *композиции* - частный случай отношения *агрегации*. Это отношение служит для спецификации более сильной формы отношения "часть-целое", при которой составляющие части тесно взаимосвязаны с целым. Особенность этой взаимосвязи заключается в том, что части не могут выступать в отрыве от целого, т.е. с уничтожением целого уничтожаются и все его составные части.

В качестве примера отношения композиции обычно приводится пример окна компьютерной программы, которое может включать подзаголовок, главное меню и др.

Таким образом, мы видим, что в языке UML отдельно моделируется отношение ЧАСТЬ-ЦЕЛОЕ, и, кроме того, отдельно выделяется подвид этого отношения, характеризующийся жесткой зависимостью части от целого.

8.6.2. Отношения ЧАСТЬ-ЦЕЛОЕ в информационно-поисковых тезаурусах и WordNet

Как было указано в пп. 1.2.1, 1.2.1.2 в рамках информационно-поисковых тезаурусов отношения ЧАСТЬ-ЦЕЛОЕ могут входить в состав иерархических отношений. Иерархические отношения обычно рассматриваются как несимметричные и транзитивные. При установлении иерархических отношений важна независимость от контекста.

В частности, в тех случаях, когда имеется множественная принадлежность части к целому, то между такими терминами не должно устанавливаться иерархическое отношение. Между такими дескрипторами может быть установлено отношение ассоциации. Например, карбюраторы являются частями не только автомобилей. Поэтому дескрипторы *КАРБЮРАТОР* и *АВТОМОБИЛЬ* не должны быть связаны отношением ЧАСТЬ-ЦЕЛОЕ в информационно-поисковом тезаурусе (Will, 2004).

Таким образом, с точки зрения разработки информационно-поисковых тезаурусов не рекомендуется описывать как отношения ЧАСТЬ-ЦЕЛОЕ такие отношения, упомянутые в п. 8.3. качестве примеров этого отношения, как:

- *Сталь – велосипед*, поскольку сталь может быть в разных артефактах, не только в велосипеде,
- *Рука – музыкант*, поскольку руки не только у музыкантов,
- *Кусок – пирог*, поскольку многие другие вещи можно разделить на куски,
- *Дерево – лес,* поскольку деревья растут не только в лесу

Подход к отношениям часть-целое в тезаурусе WordNet принципиально другой. Как мы уже указывали в разделе 2.2. отношения часть-целое, устанавливаются в WordNet на основе лингвистического теста

> *X является частью Y, если можно сказать, что X – это часть Y (An x is a part of Y) или Y имеет X как часть (A y has an x as a part).*

Внутри отношения ЧАСТЬ-ЦЕЛОЕ дополнительно выделяются отношения *быть_элементом* (человек - часть человечества)и *быть_сделанным_из (стекло – часть стеклянного изделия).* Синсет-часть может быть сопоставлен большому количеству синсетов-целое, как, например, *point* (острие) может быть у *стрелы, ножа, иголки, карандаша, булавки* и т.п.

Приведем еще примеры различных отношений *часть-целое* из WordNet (табл. 8.1. - цифры во втором столбце таблицы означают номера значений слов в WordNet):

Синсет-часть	Синсет-целое	Трактовка WordNet
Air 1	*Wind_1*	Воздух входит в состав ветра
Artillery 1	*Battery_1*	Артиллерийское орудие входит в состав артиллерийской батареи
Bucharest	*Romania*	Бухарест входит в состав Румынии
Cellulose	*Paper_1*	Целлюлоза входит в состав бумаги
Iron 1	*Steel_1*	Железо входит в состав стали
Chew 2	*Eating*	Жевание часть процесса еды
Computer 1	*Computer network*	Компьютер входит в состав компьютерной сети
Wing 1	*Angel_1*	У ангела есть крыло
Wing 1	*Bird_1*	У птицы есть крыло
Snow 1	*Snowball*	Снег входит в состав снежка
Palm	*Hand*	Ладонь – это часть руки

Табл.8.1. Примеры отношений часть-целое в WordNet

Отметим, что к каждой паре приведенных примеров применимы лингвистические тесты, которые используются для диагностики отношения *часть-целое* (см. п. 2.2). При этом очевидно, что многие из приведенных примеров отношений часть-целое в WordNet не могли бы быть установлены, в соответствии с рекомендациями, принятыми для информационно-поисковых тезаурусов, говорящими о том, что отношение ЧАСТЬ-ЦЕЛОЕ в информационно-поисковых тезаурусах должно устанавливаться в тех случаях, когда одно понятие включено в другое понятие независимо от контекста.

Например, этому правилу не соответствуют такие пары как:

Air – wind – *не всякий воздух обязательно ветер,*

Computer – computer network – *не всякий компьютер входит в состав сети,*

Snow – Snowball – *не всякий снег входит в состав снежка,*

Iron – Steel – *не всякое железо входит в состав стали и др.,*

Таким образом, мы видим, что методологии построения тезаурусов разного типа включают в себя существенно разные принципы установления отношений часть-целое.

8.6.3. Отношение ЧАСТЬ-ЦЕЛОЕ в онтологиях верхнего уровня

Рассмотрим, какие решения по моделированию отношения ЧАСТЬ-ЦЕЛОЕ принимаются в онтологиях верхнего уровня.

В онтологии SUMO (Niles, Pease, 2001) отношения ЧАСТЬ-ЦЕЛОЕ определены только над осязаемыми (tangible) пространственными сущностями – объектами. Такое ограничение не является типичным для общей мереологии.

В этой онтологии отношение ЧАСТЬ-ЦЕЛОЕ подразделяется на следующие подвиды: член, компонент, кусок (piece), собственно часть, поверхностная часть. Поверхностные части делятся на поверхность, верх, низ и бок.

В онтологии OpenCYC (Cyc Ontology Guide) отношение ЧАСТЬ-ЦЕЛОЕ определяется в очень общем смысле. Единственное ограничение на аргументы отношения заключается в том, что они должны быть конкретными сущностями. Отношение ЧАСТЬ-ЦЕЛОЕ включает такие подвиды как пространственные части, временные части, «концептуальные» части (например, содержать_информацию) , члены группы и т.п.

Физические части в онтологии OpenCYC включают следующие подвиды:

- стенки полости;
- внешние части;
- внутренние части;
- визуальные отметки.

В онтологии DOLCE (Masolo и др., 2003) отношение «объект – материал этого объекта» (ваза – глина) рассматривается как отдельное отношение «составляет» (constitute), не являющееся отношением *часть-целое*:

> *X составляет Y тогда и только тогда, когда X может быть субстратом после разрушения Y.*

Такое решение связано с тем, что объект (ваза) и материал, из которого сделан объект, считаются различными сущностями. Если предположить, что между глиной и вазой существует отношение *часть-целое*, то глина должна совпасть с вазой, поскольку у глины и вазы совпадают части, а значит, и по аксиомам мереологии глина и ваза совпадают.

Это решение является программным для авторов онтологии, поскольку в DOLCE принят так называемый мультипликативный подход: считается, что различные сущности могут быть совмещены по пространству и времени. Причина, по которой это возможно, заключается в том, что такие сущности могут иметь несовместимые существенные свойства. Классический пример таких сущностей: ваза и глина, из которой сделана эта ваза: ваза не переживет радикальное изменение формы или топологии, в то время как кусок глины останется тем же независимо от этих изменений. Таким образом, эти две сущности различны, хотя и совмещены по времени и пространству. Считается поэтому, что кусок глины составляет вазу (*constituted*), но что сама ваза не является куском глины.

Еще одна черта DOLCE – явное разделение на «постоянные» и «происходящие» сущности. Различие между ними состоит в том, что «постоянные» сущности имеются в наличии целиком и неизменно в некотором фиксированном промежутке времени (например, стол, дом в течение периода своего существования).

«Происходящие сущности» разворачиваются во времени и в каждый момент в некотором временном интервале они могут быть различными, по-разному себя проявлять, иметь разный состав, (например: ураган, жизненный цикл), однако при этом их идентичность сохраняется.

Другой способ разделения «постоянных» сущностей и «происходящих» сущностей заключается в следующем: сущность является «постоянной», если она существует больше, чем в один момент времени, и утверждения о частях должны быть сделаны относительно временной шкалы. Другими словами различие между категориями базируется на фундаментальном различии отношения части для двух категорий: «постоянные» сущности нуждаются в описании отношения *часть-целое* с добавлением индекса времени, а «происходящие» сущности – нет.

Поэтому в DOLCE рассматриваются два вида отношений *часть-целое*: постоянное отношение *часть-целое* и отношение *часть-целое* в момент времени *t*. Постоянное отношение *часть-целое* устанавливается между «происходящими» сущностями, а временное отношение *часть-целое* устанавливается между «постоянными» сущностями.

Для описания отношений между «происходящими» сущностями и «постоянными» сущностями вводится отношение *участия (participation):*

> *"Обычное представление об участии состоит в том, что «постоянные» сущности вовлечены в «происходящие» сущности. В онтологии, базирующейся на строгом разделении между «происходящими» сущностями и «постоянными» сущностями, участие не может быть просто частью. Участвующие «постоянные» сущности не являются частями «происходящих» сущностей, только «происходящие» сущности могут быть частями «происходящих» сущностей. Отношения участия имеют индекс по времени, для того, чтобы учитывать вариации участия во времени (постоянное участие, временное участие)".*

Постоянные сущности «живут» во времени, участвуя в тех или иных «происходящих» сущностях. Например, человек, который является «постоянной» сущностью, участвует в дискуссии, которая является «происходящей сущностью. Человеческая жизнь также является «происходящей» сущностью, в которой человек участвует всю свою жизнь.

Отметим, что в такой онтологии верхнего уровня как OCRE на счет отношения *участия* принято полностью противоположное решение: отношение *участия* рассматривается как частный случай отношения *часть-целое* (Masolo и др., 2003).

Заключение к главе 8

Мы показали, что в настоящее время существует достаточно широкий спектр подходов к отношениям *часть-целое*. Это отношение исследуется довольно давно, однако полного согласия между исследователями по их классификации, границам, пока не достигнуто.

Как представляется, определении того набора отношений, которые будут представляться в создаваемой онтологии как отношения *часть-целое*, полезно применять не неоднозначные лингвистические тесты, а проводить анализ, подобный рассуждениям в приведенным п. 8.4., работе (Varzi, 2006), онтологическом анализе Дж.Совы (Sowa, 2000), когда рассматривается, как изменение или уничтожение предполагаемых частей и целого влияет друг на друга.

При определении отношений *часть-целое* в создаваемом онтологическом ресурсе важно прислушаться к рекомендациям стандартов и руководств по разработке информационно-поисковых тезаурусов и, прежде всего, стремиться описывать такие части, которые жестко «привязаны» к своему целому, что связано с онтологической зависимостью части от целого.

Этот вывод связан с тем, что все приводимые примеры кажущейся нетранзитивности отношения *часть-целое* упоминают такие части, которые нежестко связаны с приводимыми в примерах целыми, и именно такие отношения часть-целое не рекомендуют устанавливать стандарты и руководства по разработке информационно-поисковых тезаурусов.

Глава 9. Отношения онтологической зависимости

Отношение *онтологической зависимости* между сущностями A и B состоит в установлении факта зависимости существования A от существования B (Lowe, 2005). Это отношение известно со времен Аристотеля, который заметил, что невещественные сущности такие как качества и количества зависят от вещественных сущностей.

Известный философ Гуссерль считал отношение зависимости центральным отношением онтологии. Никола Гуарино (Guarino, 1998b) называет теорию зависимости одним из основных инструментов анализа сущностей в рамках Формальной Онтологии.

В настоящее время это отношение стало активно использоваться при построении онтологий верхнего уровня, однако, на наш взгляд, важность этого отношения для онтологического моделирования конкретных предметных областей пока еще недостаточно изучена.

9.1. Определение и свойства отношения онтологической зависимости

Зависимость сущностей может быть разного содержания. П. Симонс (Simons, 1987) рассматривает несколько разных видов зависимости между сущностями:

- физиологическая зависимость (человек A зависит от лекарства B так, что A не может жить, если не будет принимать это лекарство);
- причинная зависимость (детонация мины A зависит от предварительного воспламенения B так, что детонация не произойдет, если не будет предварительного воспламенения);
- логическая зависимость (пропозиция p зависит от пропозиции q, тогда и только тогда, если p не может быть истинным, пока q не является истинным);
- функциональная зависимость (давление P фиксированной массы идеального газа зависит от температуры T и объема V тогда и только тогда, когда P не может изменяться, пока не изменяется по крайней мере один из параметров T или V);
- практическая зависимость (умение A зависит от умения Б тогда и только тогда, когда умение A не может быть достигнуто, пока не достигнуто умение Б).

Онтологическая зависимость отличается от всех вышеперечисленных зависимостей. Для выявления онтологической зависимости нужно ответить на следующий вопрос: может ли сущность (X) существовать сама по себе, или подразумевает существование чего-либо еще (Y). Так, свойство белизны зависит от вещества, например, от куска бумаги тогда и только тогда, когда это свойство не может существовать без этого куска бумаги.

> ***Определение.*** *X онтологически зависит от Y тогда и только тогда, когда X существует только, если Y существует.*
>
> *D(X, Y) = def (существует (X) -> существует (Y))*

Для отношений онтологической зависимости можно сформулировать пары предложений подобно тому, как определяются семантические отношения между лексическими единицами (диагностические тесты лексического отношения):

> *X онтологически зависит от Y, если из утверждения «X существует» следует утверждение «Y существует».*

В связи с отношениями зависимости наиболее распространенными используемыми аксиомами является аксиомы рефлексивности и транзитивности (Gangemi и др., 2001b):

D (x, x) – рефлексивность отношения зависимости;

$D(x, y) \wedge D(y, z) \rightarrow D(x, z)$ - *транзитивность отношения зависимости.*

9.2. Виды отношения онтологической зависимости

Существует много форм онтологической зависимости.

Во-первых, можно рассмотреть онтологическую зависимость существования конкретной сущности (*экзистенциальная зависимость*), или онтологическую зависимость существования свойства сущности (*зависимость свойств*), в том числе свойство принадлежности к некоторому классу сущностей - *концептуальная зависимость* (Masolo и др., 2003; Gangemi и др., 2003).

Например, человек зависит от своего мозга – экзистенциальная зависимость. Свойство сущности «быть_гаражом» зависит от существования автомобиля – концептуальная зависимость. Если автомобили в некотором гипотетическом мире исчезнут, то постройка останется, но ее свойство «быть гаражом» исчезнет.

Во-вторых, можно выделить *строгую зависимость (rigid),* то есть зависимость от существования конкретной сущности, или *родовую зависимость (generic),* то есть зависимость от существования класса сущностей. Так, человек зависит от своего мозга строгой зависимостью – мозг не может быть заменен, а от своего сердца – родовой зависимостью – сердце может быть заменено.

В-третьих, можно рассмотреть течение времени и выделить *константную зависимость*, то есть зависимость от существования некоторой сущности в текущий момент времени, или *историческую зависимость*, то есть зависимость от существования некоторой сущности в предшествующий отрезок времени.

$CD(x,y)=def\ EX(x,t) \rightarrow EX(y,t)$ – *константная зависимость*

$HD(x,y)=def\ EX(x,t) \rightarrow (EX(y,t') \wedge t'=<t)$ – *историческая зависимость*

Так, ребенок зависит от своей матери исторической зависимостью, если бы мать не существовала в предшествующий промежуток времени, ребенок не мог бы родиться.

В-четвертых, могут быть выделены *внутренняя онтологическая зависимость*, то есть зависимость от внутренних свойств или частей сущности, и *внешняя онтологическая зависимость*, то есть онтологическая зависимость от существования некоторой отдельной сущности.

Отношения внутренней и внешней зависимости могут быть определены следующим образом:

внутренняя зависимость:

$$ID(x, y) = def\ D(x, y) \wedge P(y, x);$$

внешняя зависимость:

$$ED(x, y) = def\ D(x, y) \wedge \neg P(y, x);$$

здесь $P(x,y)$ обозначает отношение часть-целое.

Н. Гуарино указывает на важность выделения отношений внешней зависимости из всего спектра отношений зависимостей. В работе (Guarino, Welty, 2000) подчеркивается, что, вообще говоря, необходимо более строгое определение внешней зависимости, которое должно исключить из рассмотрения и другие внутренние характеристики сущности как материал, из которого сделана сущность, его свойства (например, цвет), а, кроме того, те сущности, которые необходимо существуют, например, вселенная.

Наконец, зависимость может быть *односторонней* или *двусторонней*.

Рассмотрим в качестве примера отношения между матерью и ее ребенком. Мы видим здесь двустороннюю онтологическую зависимость.

Зависимость ребенка от матери является следующей:

- экзистенциальная,
- историческая,
- строгая,
- внешняя,

то есть существование ребенка как конкретной сущности зависит от существования конкретной матери (другой сущности) в предшествующий период времени.

Зависимость матери от ребенка является следующей:

- концептуальная,
- константная,
- строгая,
- внешняя,

то есть свойство женщины быть матерью в текущий момент времени зависит от существования в текущий момент времени ее ребенка как конкретной сущности.

Рассмотрим отношения между автомобилем с одной стороны и гаражом и двигателем с другой стороны.

Гараж односторонне зависит от автомобиля:

- концептуальной,
- константной,
- родовой,
- внешней зависимостью,

то есть некоторая постройка не перестанет существовать, если в мире исчезнут все автомобили, но ее свойство «быть_гаражом» зависит от существования класса сущностей «автомобили».

Автомобиль зависит от своего двигателя:

- экзистенциальной,
- константной,
- родовой,
- внутренней зависимостью,

то есть конкретный автомобиль зависит от существования класса конкретных двигателей (двигатель может быть заменен); и эта зависимость является внутренней, поскольку двигатель является частью автомобиля.

Сложным вопросом является вопрос о зависимости между такими сущностями как, например, *Сократ* и *Жизнь_Сократа*. С одной стороны, представляется, что, имеется взаимозависимость, поскольку Сократ зависит от своей жизни. С другой стороны, интуитивно, представляется, что *Жизнь_Сократа* должна быть более зависимой от *Сократа*.

В работе (Masolo и др., 2004) уточняется понятие внешней онтологической зависимости. Авторы указывают на необходимость аккуратного описания отношений зависимости в следующем тривиальном случае: собственно Socrate и {Socrate} как множество из одного элемента. Поскольку, когда существует X, то существует {X}, то получается, что X внешне зависит от {X}, что противоречит интуиции.

Поэтому авторы этой работы обсуждают полезность такого понятия как зависимость по определению (Fine, 1995):

> *сказать, что сущность X зависит от Y, это означает сказать, что Y необходимо (eliminably) должно быть использовано в любом определении X.*

В результате авторы уточняют отношение внешней родовой зависимости следующим образом: понятие X является внешне зависимым от понятия Y, если выполняются два условия:

- определение понятия X необходимо включает понятие Y
- если для любой сущности x, которая классифицируется как X, найдется сущность y, которая классифицируется как Y, являющаяся внешней для сущности x, то есть не являющейся частью, материалом или свойством сущности x.

Таким образом определенная внешняя зависимость показывает зависимость сущности *Жизнь_Сократа* от *Сократа*, поскольку определение сущности *Жизнь_Сократа* необходимо требует существования сущности Сократ (Masolo и др., 2004).

9.3. Онтологическая зависимость в онтологиях верхнего уровня

Для Дж. Совы (Sowa, 2000) в построении онтологии верхнего уровня одним из существенных параметров является зависимость понятий друг от друга (prehension). Зависимость может быть внешняя (extrinsic) и внутренняя (intrinsic) (рис.9.1). Если исчезновение одной из сущности меняет структуру или существование другой сущности, то это отношение внутреннее.

Для создаваемых в настоящее время онтологий верхнего уровня понятие зависимости, зависимых сущностей является существенным. Рассмотрим некоторые из таких онтологий.

Онтология BFO (Basic Formal Ontology) (Masolo и др., 2003; Grenon, 2003) является онтологией универсалий, целью которой является структурирование онтологий конкретных научных областей.

В этой онтологии подразделение второго уровня непосредственно связано с понятием зависимости.

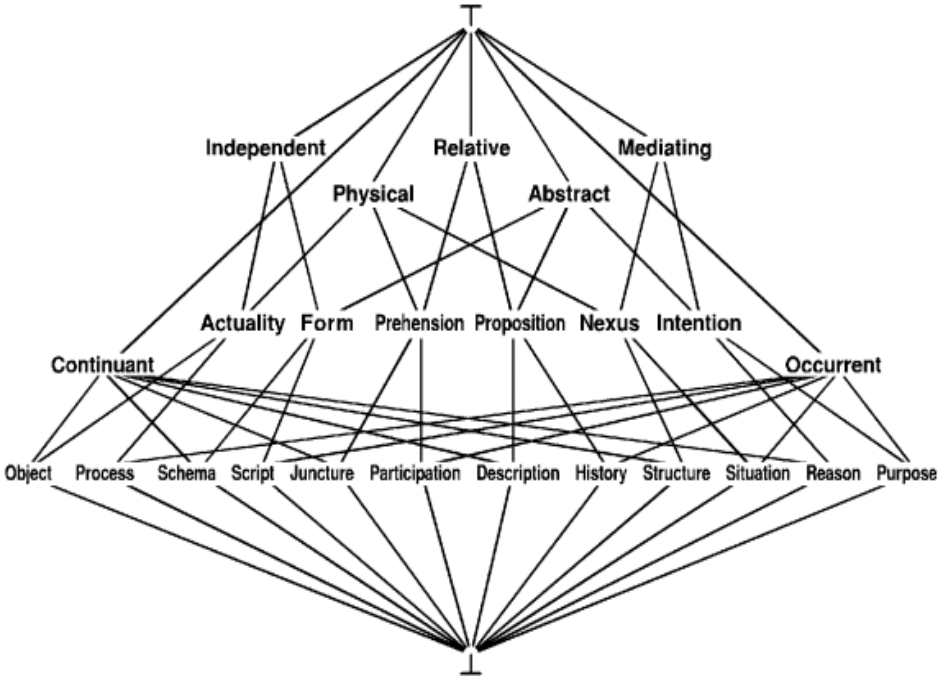

Рис.9.1 Онтология верхнего уровня Дж.Совы (Sowa, 2000) (т.н. «бриллиант» Дж.Совы). Понятие зависимости (prehension) входит в состав основных категорий.

На первом уровне все сущности делятся на сущности, существующие во времени, и сущности, происходящие во времени. Временные сущности делятся на пространственные

регионы (пространство, плоскость, прямая, точка), независимые сущности, то есть те которые могут существовать отдельно от других сущностей, и зависимые сущности (Dependent entities), существование которых всегда связано с существованием других сущностей – к таким сущностям относятся функции, роли и качества (см. рис. 9.2).

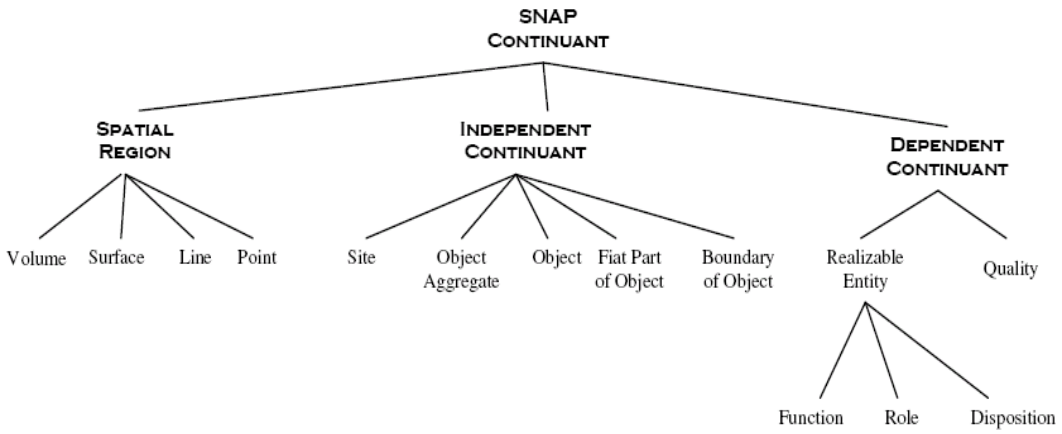

Рис.9.2. Классификация сущностей второго уровня в онтологии BFO

Соответственно, одним из основных отношений онтологии является отношение между зависимыми и независимыми сущностями - *inhere:* C1 *inheres* in C2 at t, что означает, например, что некоторое качество (например, краснота) зависит от некоторого объекта (например, яблоко или красный жакет), все то время, когда это качество существует, и оно зависит все время своего существования от одного и того же объекта.

В онтологии Dolce (Masolo и др., 2003) отношение зависимости также входит в состав основных отношений онтологии.

Разработчики Dolce рассматривают разнообразный набор отношений зависимости:

- специфическая зависимость – зависимость от существования конкретного примера сущности (соответствует строгой зависимости п.9.2);
- родовая зависимость – зависимость от существования класса сущностей;
- отношения односторонней и двусторонней зависимости;
- отношения пространственной зависимости.

На важность отношений зависимости указывают и разработчики и исследователи предметных онтологий.

В работе (Kumar, Smith, 2004) приводятся примеры онтологической зависимости в области биологии: не может быть клеточного движения без клеток, биологические процессы зависят от органов, клеток и молекул – такая зависимость является строгой.

В руководстве по использованию онтологии генов GO, которая включает три независимых таксономии: молекулярные функции, биологические процессы и клеточные компоненты, указывается, что некоторые термины онтологии GO, предполагают присутствие в онтологии других терминов, например, если имеется термин «*регулирование X*», то «*процесс X*» также должен существовать и быть описанным в онтологии (http://www.geneontology.org/GO.usage.htm), что соответствует существованию отношения онтологической зависимости между сущностями *РЕГУЛИРОВАНИЕ X* и *X*.

Рассматривая индексирование генных продуктов терминами онтологии GO, авторы работы (Burgun и др., 2004) отмечают, что если продукт проиндексирован термином Ti, и имеется термин Ti0, от которого зависит термин Ti, то продукт должен быть проиндексирован и термином Ti0.

9.4. Нетаксономические отношения информационно-поискового тезауруса и отношение онтологической зависимости.

На наш взгляд, попытки сформулировать принципы установления отношений в информационно-поисковых тезаурусах имеют явные аналогии с проблемой установления отношений онтологической зависимости.

Во-первых, требование при описании отношения ЧАСТЬ-ЦЕЛОЕ того, чтобы это отношение выполнялось вне зависимости от контекста, сильно связано с зависимостью существования части от существования целого, что и показывают примеры, приведенные в разделе 8.6.2.

Во-вторых, наиболее общее правило установления ассоциативного отношения между дескрипторами, рекомендуемое американским стандартом (Z39.19 , см. также п.1.2.2), говорящее о том, что при установлении этого отношения один термин как бы подразумевает другой – вообще говоря, во многих случаях это требование означает случаи концептуальной зависимости и зависимости по определению между понятиями, когда введение одного понятия, термина невозможно пока не введено другое понятие, термин.

Рекомендации по использованию определений терминов для введения ассоциативных отношений также, как представляется, являются попытками нащупать наиболее существенные зависимости между терминами предметной области

В-третьих, если мы рассмотрим примеры рекомендуемых отношений ассоциации, мы увидим, что для большинства терминов, связанных ассоциациями в таких примерах, соответствующие понятия связаны односторонними или двусторонними отношениями онтологической зависимости. Например, в примерах из п. 1.2.2 имеем:

1) научная дисциплина – объект изучения или специалист в этой дисциплине:

математика - математик

неврология - нервная система

Специалист по дисциплине является концептуально зависимым от дисциплины, а область изучения является концептуально зависимой от своего объекта изучения.

2) операции или процессы и их агент или инструмент

контроль температуры – термостат

охотник – охота

Понятно, что приспособления (*термостат)* концептуально зависят от своего назначения, а роли (охотник) зависят от той ситуации, в которой они задействованы.

3) объекты или процессы и их контрагенты

растения – гербициды

Если контрагент создан специально для противодействия некоторому объекту или процессу, это означает, что он является онтологически зависимым от этого объекта (процесса).

Таким образом, разработчики информационно-поисковых тезаурусов интуитивно включают в правила установления тезаурусных отношений те или иные виды онтологической зависимости.

В следующем разделах мы подробнее рассмотрим взаимосвязи между ассоциативными отношениями информационно-поисковых тезаурусов, отношениями онтологической зависимости и использованием тезаурусных отношений для автоматического расширения запроса при информационном поиске.

9.5. Анализ отношения АССОЦИАЦИЯ в традиционных информационно-поисковых тезаурусах: тезаурус EUROVOC

Отношение ассоциации является одним из наиболее часто используемых отношений в информационно-поисковых тезаурусах. Как мы уже отмечали в разделе 1.2.2, несмотря на попытки экспликации описания отношений в стандартах и руководствах, установление отношения ассоциации является достаточно субъективной процедурой. Кроме того, как мы показали в разделе 1.7.2 на примере тезауруса EUROVOC, применение ассоциативных отношений при расширении запроса приводит к серьезному снижению точности поиска.

В литературе обычно обсуждается несколько аспектов, связанных с отношением АССОЦИАЦИЯ. Во-первых, в ряде работ предлагается приписывать различные веса ассоциативным отношениям тезауруса (Tudhope, Taylor, 1997; Chen и др., 1993), другая часто работ изучает вопросы необходимости дополнительной семантической классификации отношений АССОЦИАЦИЯ (Tudhope и др., 2001; Rada et al. 1991, Леонтьева и др., 1978), а также исследуются вопросы зависимости весов ассоциативных отношений при расширении запросов от их семантического типа (Jones, 1993).

Вернемся к рассмотренным в разделе 1.7.2 примерам ассоциативных отношений из тезауруса EUROVOC:

ЗЕМЕЛЬНЫЙ КАДАСТР

АСЦ *ГРАДОСТРОИТЕЛЬНОЕ ЗАКОНОДАТЕЛЬСТВО,*

МЕСТНЫЕ НАЛОГИ;

НАЛОГ НА НЕДВИЖИМОСТЬ;

РАЗРЕШЕНИЕ НА СТРОИТЕЛЬСТВО

Анализируя выдачу поисковой системы по коллекции стенограмм Государственной Думы Российской Федерации по запросу *земельный кадастр,* мы показали, что если документы этой выдачи использовать при поиске по запросам *градостроительное законодательство, местные налоги, налог на недвижимость* или *разрешение на строительство*, то точность выдачи по этим запросам значительно снизится.

Рассмотрим, почему же это происходит, чему посвящены другие тексты выдачи, ведь сами по себе представленные отношения не кажутся ошибочными. В стенограммах, полученных по запросу *земельный кадастр,* обсуждались такие вопросы как составление Земельного кадастра, регистрация прав на недвижимость, кадастровая стоимость земельного участка, купля-продажа земли и другие вопросы.

Таким образом, мы видим, что с земельным кадастром связан ряд разнообразных ситуаций. Только в относительно небольшой части из них земельный кадастр сильно связан с перечисленными выше четырьмя понятиями, а в других связь с этими понятиями отсутствует, тексты же могут обсуждать любую из этих ситуаций, поэтому плохие поисковые характеристики вышеперечисленных ассоциативных связей закономерны.

Получается, чтобы сделать ассоциативную связь полезной при автоматическом расширении запроса, необходимо устанавливать такие связи, чтобы они действовали, не пропадали в подавляющем числе ситуаций, в которых участвуют понятие или его конкретные экземпляры.

На наш взгляд, именно отношение онтологической зависимости проявляет такую устойчивость, обеспечивает возможность надежной опоры в разнообразных ситуациях, которые могут обсуждаться в связи с той или иной сущностью.

Так, нетрудно видеть, что при строгой зависимости зависимое понятие не может быть оторвано от конкретного экземпляра главного понятия, поэтому если возникает,

существует, обсуждается конкретный пример такого строго зависимого понятия, то существует и обсуждается пример главного понятия.

В случае родовой зависимости конкретный пример зависимого понятия может быть оторван от главного понятия, с ним может происходить что-то не связанное с главным понятием, но обычно недолго и в относительно небольшой доле примеров зависимого понятия.

При исторической зависимости пример зависимого понятия может достаточно долго существовать без главного понятия и участвовать в самых разных ситуациях, например, *сельскохозяйственная продукция* создается в процессе *сельскохозяйственного производства*, затем продукция значимое время живет «своей жизнью»: перевозится, продается, хранится. Однако многие свойства результата определяются порождающим его процессом.

Таким образом, если для каждого понятия в тезаурусе выявлять понятия, находящиеся с ним в отношении онтологической зависимости, отмечать их, например, направленной ассоциацией, то эти отношения можно было бы использовать для автоматического расширения запроса, поскольку они определяют подавляющее количество ситуаций, которые могут случиться с конкретными экземплярами зависимого понятия.

Так, например, понятие *ЗЕМЕЛЬНЫЙ КАДАСТР* является зависимым понятием от понятия *ЗЕМЕЛЬНЫЙ УЧАСТОК* (родовая зависимость), поскольку понятие *ЗЕМЕЛЬНЫЙ КАДАСТР* не может возникнуть, если не существует этого понятия

Если мы опять вернемся к документам, выданным по запросу *земельный кадастр*, то мы можем видеть, что все эти документы релевантны запросу земельный участок.

Другие упомянутые дескрипторы также имеют отношения зависимости:

- понятие *ГРАДОСТРОИТЕЛЬНОЕ ЗАКОНОДАТЕЛЬСТВО* зависит от понятия *ГРАДОСТРОИТЕЛЬСТВО*;
- понятие *МЕСТНЫЕ НАЛОГИ* зависит от понятия *МЕСТНОЕ САМОУПРАВЛЕНИЕ*;
- понятие *РАЗРЕШЕНИЕ НА СТРОИТЕЛЬСТВО* зависит от понятия *СТРОИТЕЛЬСТВО*;
- понятие *НАЛОГИ НА НЕДВИЖИМОСТЬ* зависит от понятия *НЕДВИЖИМОСТЬ*.

Возникает вопрос, как отношения онтологической зависимости между дескрипторами тезауруса связаны с семантическими отношениями (*часть, результат, причина, содержание* и др.), посредством которых часто предполагается улучшить качество описания ассоциативных отношений в информационно-поисковых тезаурусах (см. также раздел 4.5.2).

В Таблице 9.1. перечислим примеры ассоциативных отношений из тезауруса EUROVOC, которые представляют собой отношения онтологической зависимости. Каждое отношение охарактеризуем также с семантической точки зрения – припишем ему название семантического отношения от главного понятия к зависимому (упорядочено по главному понятию):

Мы видим, как разнообразные семантические отношения могут соответствовать одному и тому же отношению онтологической зависимости. Онтологическая характеризация отношений представляет собой другое, отличное от семантической характеризации измерение отношений, и, на наш взгляд, расширением запроса управляют именно онтологические характеристики отношения.

Главное понятие	Зависимое понятие	Семантическое отношение
Депутат	Кандидат в депутаты	Результат
Дети	Многодетные семьи	Часть
Дети	усыновление	Объект действия
Заболевание	Профилактика заболеваний	Контрагент
Зерно	Импорт зерна	Объект
Злоупотребление властью	Иски об отмене решений	Причина
Инвестиционный риск	Диверсификация рисков	Контрагент
Качество продукции	Знак качества	Содержание
Офицер	Офицерское движение	Субъект
Парламент	Обращение парламента	Субъект
Парламентарий	Парламентский иммунитет	Носитель свойства
Президент	Инаугурация	Субъект

Таблица 9.1. Примеры соответствий между отношениями онтологической зависимости и семантическими отношениями

Таким образом, поисковые характеристики в автоматическом режиме любого тезауруса, созданного для ручного индексирования, могут быть улучшены, если его ассоциативные отношения будут проанализированы с точки зрения теории онтологической зависимости:

- ассоциативные отношения, не являющиеся отношениями онтологической зависимости, помечаются как используемые только в ручном режиме;
- ассоциативные отношения представляющие собой отношения онтологической зависимости, получают направление от главного понятия к зависимому понятию;
- отношения онтологической зависимости между дескрипторами тезауруса, не представленные в виде ассоциативных отношений, дополняются.
- в некоторых случаях, когда ассоциации соединяют близкие по смыслу понятия, а также в некоторых других, которые мы обсудим ниже, ассоциация действительно является симметричной и может быть использована для автоматического расширения запроса в обе стороны.

При использовании тезауруса в автоматическом режиме используются только отношения 2) и 3) в направлении от главного понятия к зависимому понятию.

Анализ 100 первых ассоциаций тезауруса EUROVOC, рассмотренных по алфавитному порядку расположения дескрипторов показал (Loukachevitch, Dobrov, 2004c), что 33 ассоциации представляют собой отношение ВЫШЕ-НИЖЕ и записаны как ассоциации только потому, что в тезаурусе EUROVOC не разрешено два вышестоящих понятия.

Таким образом, они явно несимметричны и могут быть использованы в информационном поиске после их разметки, например,

авария

а промышленная авария
а радиационная авария
а ядерная авария

27 ассоциаций представляют собой отношения, которые могут быть использованы только при ручном составлении запроса, поскольку два ассоциированных понятия связаны между собой лишь в части ситуаций, которые могут с ними случиться, например, *авария – чрезвычайное положение* (далеко не всякая авария приводит к введению чрезвычайного положения, а чрезвычайное положение далеко не всегда возникает из-за аварии);

41 ассоциация представляют собой отношения зависимости и могут быть использованы в одном из направлений (первое понятие в строчке является зависимым от второго; запрос, содержащий второе понятие, может быть расширен первым понятием):

- *абитуриенты – высшее образование*
- *автомобильная промышленность – автомобиль*
- *агентское соглашение – посредничество*

3 ассоциации («истинные ассоциации») представляют собой очень близкие понятия, поэтому поиск может производиться в любом направлении:

- *автомобильные перевозки – автомобильный транспорт,*
- *аграрный сектор – сельское хозяйство*

Заключение к главе 9

Отношения онтологической зависимости стали вводиться в онтологические ресурсы относительно недавно и еще требуют значительного объема исследований.

Наиболее часто эти отношения используются в онтологиях верхнего уровня. Это отношение используется в определении понятий-ролей, а также в определении важных подвидов отношения ЧАСТЬ-ЦЕЛОЕ.

В этой главе мы также показали, что неявно отношение онтологической зависимости используется при обсуждении рекомендаций по установлению отношений часть-целое и ассоциация в информационно-поисковых тезаурусах.

ЧАСТЬ 3. ПРИМЕНЕНИЕ ТЕЗАУРУСОВ В КОНКРЕТНЫХ ПРИЛОЖЕНИЯХ ИНФОРМАЦИОННОГО ПОИСКА

В этой части мы рассмотрим ряд приложений автоматической обработки текстов и информационного поиска, в которых используются тезаурусы и онтологии. Каждая глава посвящена отдельной задаче или приложению.

Структура каждой главы устроена сходным образом. Глава начинается с ввода общих понятий, относящихся к данному приложению, характеризации основных методов, способов тестирования качества выполнения задачи. Далее рассматриваются результаты экспериментов, в которых применялись тезаурусы и онтологии. Проводится сравнение с результатами, полученными без привлечения онтологических ресурсов.

Таким образом, в данной части мы описываем достигнутый уровень качества методов, включающих применение тезаурусов и онтологий для автоматической обработки текстов в приложениях информационного поиска.

Глава 10. Автоматическое разрешение многозначности

Одной из серьезных проблем, которые необходимо решать в рамках широкого круга систем, включающих автоматическую обработку текстов на естественном языке с использованием лингвистических ресурсов, является проблема автоматического разрешения лексической многозначности, то есть выбора между разными значениями слов и словосочетаний, перечисленных в лингвистическом ресурсе (Кобрицов, 2004; Рахилина и др., 2006).

В последние годы проблема разрешения лексической многозначности стала исследоваться как отдельная задача. С 1998 года для тестирования систем автоматического разрешения лексической многозначности проводится специальная конференция Senseval (www.senseval.org).

Подходы к разрешению лексической многозначности достаточно разнообразны. Для разрешения многозначности могут использоваться некоторые внешние источники информации, например, электронные словари и тезаурусы. В качестве тезауруса обычно используется тезаурус английского языка WordNet (см. главу 2). Кроме того, для разрешения многозначности активно исследуется возможность применения методов машинного обучения, для чего обычно используются семантически размеченные корпуса. Применяются и различные комбинации отдельных методов.

10.1. Тестирование разрешения многозначности на конференции Senseval

Исследования методов автоматического разрешения лексической многозначности как отдельной задачи обычно делятся на два направления: разрешение лексической многозначности некоторой совокупности слов (чаще всего, несколько десятков) (см. п.10.1.1.) и разрешение лексической многозначности всех слов текста (см. п. 10.1.2) (Kilgarriff, Rosenzweig, 2000; Snyder, Palmer, 2004).

Для определения качества разрешения многозначности обычно используются два параметра: точность и полнота.

Полнота – это отношение правильно выбранных значений к общему количеству неоднозначных языковых выражений.

Точность – это отношение правильно выбранных значений к общему количеству слов, рассматриваемых системой.

Максимальное качество, которое может достигнуть система автоматического разрешения многозначности, ограничивается согласием между ручной разметкой, сделанной разными экспертами. В настоящее время, согласие между экспертами достигает 95% и выше для четко различимых значений. Для многозначных слов со значениями, близкими по смыслу, согласие между экспертами может составлять 65 – 70%.

Нижняя граница качества разрешения многозначности определяется на основе случайно выбранного значения (предполагается равновероятность значений) или наиболее частотного значения (предполагается, что вероятность одного значения многократно превышает вероятности других значений).

Также в качестве базового метода для сравнения используется так называемый метод Леска, который основан на сопоставлении словарных толкований слов, упомянутых в анализируемом фрагменте текста (Lesk, 1986).

Основные этапы применения метода таковы. Сначала из толкового словаря извлекаются толкования для всех значений слов текстового фрагмента. Для полученных толкований определяется их пересечение между собой и выбираются те значения многозначных слов, толкования которых пересекаются с толкованиями слов-соседей максимально.

В качестве классической иллюстрации метода обычно приводится английское выражение *pine cone,* компоненты которого имеют следующие толкования:

Pine

1. kinds of evergreen tree with needle-shaped leaves

2. waste away through sorrow or illness

Cone

1. solid body which narrows to a point

2. something of this shape whether solid or hollow

3. fruit of certain evergreen trees

Максимальное пересечение между толкованиями достигается при первом значении слова *Pine* и третьем значении слова Cone (Pine#1 ∩ Cone#3 = 2) – именно эти значения и должны быть выбраны для интерпретации этого выражения.

Для разрешения многозначности слов в конструкции более длинной, чем два слова, используется упрощенный алгоритм Леска, который определяет пересечение толкований значений слов с контекстами этих слов в тексте (Kilgarriff, Rosensweig, 2000). Простота алгоритма делает его важным базовым уровнем для сравнения уровня достижения предлагаемых методов разрешения лексической многозначности. Помимо толкований словаря в этом методе могут дополнительно использоваться размеченные корпуса или примеры употребления тех или иных значений слова.

Для понимания уровня, достигнутого современными системами разрешения многозначности, важно рассмотреть, каковы были лучшие результаты, показанные системами автоматического разрешения лексической многозначности на конференции Senseval-3.

10.1.1. Задание «Набор многозначных слов»

Для того, чтобы сформировать набор многозначных слов для тестирования автоматических систем в рамках конференции Senseval, обычно предпринимается специальная процедура.

Прежде всего, многозначные слова классифицируются по их частотности (в Британском национальном корпусе) и уровню их многозначности (по WordNet) (Kilgarriff, Rosenzweig 2000; Michalcea и др., 2004). Для каждой части речи (существительное, глагол, прилагательное) списки, упорядоченные по частоте и многозначности, были поделены на 4 подгруппы, тем самым получилась решетка 4x4. Далее была установлена величина набора образцов – 40 слов, которые были набраны из ячеек решетки в соответствии с количеством слов в каждой ячейке решетки.

Количество примеров из корпуса для каждого образца также базировалось на полученной решетке. Для простых слов (с низкой частотностью и многозначностью) меньшее количество примеров из корпуса было достаточно. Более частотные и более многозначные слова являются более сложными для процедуры разрешения многозначности, и поэтому такие слова должны были быть обеспечены большим количеством примеров из корпуса.

При ручной разметке примеров лексикограф имеет возможность выбрать одно из возможных значений слова, плюс две дополнительные возможности – «неясно» и «ни одно из вышеперечисленных». Была также возможность выбора двух и более значений в случае необходимости.

Для определения качества работы программ в этом задании было выбрано три уровня гранулярности: подробный, обобщенный и смешанный.

На подробном уровне гранулярности засчитывается только единственная совпадающая метка значения. На обобщенном уровне гранулярности все подзначения (обозначенные как 1.1,1.2) собирались к меткам основных значений (таких как 1, 2) и в эталонном файле и файле автоматических результатов, то есть выбор системой значения 1.1 рассматривается как правильный, если в эталонном файле содержатся отметки значений 1, 1.1, или 1.2. На третьем смешанном уровне гранулярности, засчитывались те ответы систем, которые совпадали или были подвидом значений, указанных в эталонном файле.

Были также определены базовые алгоритмы (то есть простые алгоритмы, с помощью которых можно установить минимальный уровень, который должна достигать программа разрешения многозначности):

- случайный выбор значения;
- выбор наиболее частотного значения по коллекции;
- выбор значения по методу Леска (Lesk) – метод сравнения словарных определений с текстами в трех вариантах (по толкованиям и примерам, только по толкования, по толкованиям, примерам и размеченному корпусу).

Результаты Senseval-3 для задания разрешения многозначности для заданного набора многозначных слов по англоязычной коллекции составили около 72% точности для подробного уровня гранулярности, около 79% – для обобщенного уровня гранулярности. Выбор наиболее частотного значения составил 55.2% точности для подробного уровня, 64,5% для обобщенного уровня гранулярности значений.

Для решения этой задачи используются в основном методы машинного обучения, использующие примеры, предоставленные организаторами, а также корпус SemCor, размеченный по значениям WordNet.

В число, используемых для задания «набор многозначных слов», входят такие методы машинного обучения как метод SVM (Support Vector Machines), Метод ближайших соседей, Деревья решений, Решающие списки, Байесовские классификаторы, Нейронные сети и др. В качестве признаков, на основе которых происходит обучение, используются: совместная встречаемость слов, коллокации (устойчивые выражения), биграммы, части речи, отношения между предикатом и его аргументами (подлежащее, дополнения) и др. Лучшие системы Senseval-3 используют комбинации нескольких классификаторов, что показывает, что схемы голосования результатов, комбинирующие несколько алгоритмов работают лучше, чем отдельные классификаторы (Pedersen, 2000).

10.1.2. Задание «все слова текста»

Для тестирования задачи «все слова текста» на конференции Senseval-3 использовались три текста: две статьи из Wall Street Journal и фрагмент из Брауновского корпуса – общий объем 5000 слов (Kilgarriff, Rosenzweig, 2000; Snyder, Palmer, 2004). Всего для тестирования использовались 2081 слов. Аннотирование проводилось по набору значений тезауруса WordNet. Если в WordNet не было подходящего значения, то проставлялась помета U.

По результатам конференции SENSEVAL-3 для английского языка в задаче разрешения многозначности для всех слов текста точность лучшей системы составляет 65.2% (Snyder, Palmer, 2004).

Все лучшие в SENSEVAL-3 алгоритмы разрешения многозначности используют семантически размеченные корпуса по значениям WordNet. Семантическая разметка корпуса обычно используется двумя основными способами: как основа для обучения программы разрешения многозначности, и как информация о наиболее частотном значении, которое выбирается в тех случаях, когда не удалось выбрать значение с помощью основного алгоритма. По оценкам, порядка 60% слов в тестовых текстах употреблены в наиболее частотном значении, полученному по семантически размеченному корпусу SemCor (Snyder, Palmer, 2004).

Согласие между лексикографами-аннотаторами значений достигало – 72,5. Наибольший процент разногласий по разметке значений был связан с небольшим набором трудных слов, например, *national*.

Для каждой системы было выполнено два вида подсчетов. В первом случае отказ системы определить значение рассматривался как U, таким образом, такой ответ засчитывался как правильный только если разметка также была U, и как неправильный, в противном случае.

Второй вид подсчета не учитывал те ответы, в которых система выдала U. Таким образом, точность не менялась, а полнота при таком подсчете понижалась.

При первой системе подсчетов максимальная точность 0.652, средняя точность по системам – 0.522. При второй системе подсчетов – средняя точность 57.4, полнота – 51.9.

Важно отметить, что иногда в счет «благополучно» разрешенных многозначных единиц попадают также и однозначные термины. По нашей оценке, в одном из тестовых текстов около 10% размеченных слов имеют одно значение в WordNet, например, такие слова как *congressional, constituency, salary, legislator, reelection* и др. (Данные получены с сайта http://www.senseval.org/). Если рассчитать точность разрешения многозначности для лучшей системы, не считая этих однозначных слов, то величина точности разрешения многозначности лучшей системы составит 59.9%.)

10.2. Подходы к разрешению лексической многозначности на основе тезаурусных знаний

Различные алгоритмы разрешения лексической многозначности на основе тезаурусной структуры предлагались и тестировались для тезауруса английского языка WordNet.

Одним из классов предлагаемых методов является оценка семантической близости контекста вхождения того или иного многозначного термина к каждому из возможных значений – синсетов.

Такая оценка близости может рассчитываться на основе сравнения путей между синсетами слов контекста и синсетами рассматриваемого многозначного слова.

В работе (Leacock, Chodorow, 1998) предполагается, что два значения тем семантически ближе, чем короче связывающий их путь. Упор делается на отношения гипонимии-гиперонимии и взвешивается длина пути относительно всей глубины таксономии (*D*):

$$Sim_{LC}(C1, C2) = -\log(PathLen(C1,C2)/2D) \qquad (10.1)$$

В работе (Hirst, St-Onge, 1998) предполагается, что два синсета семантически близки, если соединены достаточно коротким путем, который имеет малое количество перегибов:

$$Sim_{HS}(C1, C2) = c_0 - PathLen - k * d, \qquad (10.2)$$

где d – количество перегибов на протяжении пути; c_0 и k – константы. Если такого пути не существует, то $Sim_{HS}(C1, C2) = 0$.

В экспериментах использовались значения констант $c_0 = 8$, $k = 1$, максимальная длина пути 5 шагов.

В ряде работ концептуальное расстояние между синсетами учитывает большее число параметров. Так, для подсчета концептуального расстояния в работе (Agirre, 1995; Agirre, 1996) вводится понятие *концептуальной плотности* и формула ее вычисления, которая, по мнению авторов, наилучшим способом описывает близость между словами. В формуле учитываются следующие параметры:

- длина самого короткого пути в иерархии;
- глубина в иерархии;
- плотность понятий в иерархии;

- число концептов.

Формула вычисления концептуальной плотности выглядит следующим образом:

- c-корень (вершина);
- nhyp – число гипонимов в вершине;
- h - высоту иерархии;
- m-число слов из контекста, которые попали в иерархию.

Тогда формула, которая вычисляет плотность (10.3).

$$CD(c,m) = \frac{\sum_{i=0}^{m-1} nhyp^i}{\sum_{i=0}^{h-1} nhyp^i} \qquad (10.3)$$

$$descendants_c = \sum_{i=0}^{h-1} nhyp^i \qquad (10.4)$$

nhyp в этой формуле вычисляется по формуле (10.4), где descendants-количество потомков в узле.

Эти формулы автор пытался улучшить опытным путем, вводя параметры, и смотря, при каких значениях формула дает наилучшие результаты. В итоге выбор был остановлен на формуле (10.5).

$$CD(c,m) = \frac{\sum_{i=0}^{m-1} (nhyp + \beta)^{i^\alpha}}{descendants_c} \qquad (10.5)$$

Другим направлением выбора значения многозначного слова на основе близости контекста в тексте и окружения слов в тезаурусе являются подходы, основанные на оценке так называемого информационного содержания.

Ф. Резник (Resnik, 1995) вводит характеристику «информационное содержание» (information content), которая определяется как величина вероятности встретить пример понятия C в большом корпусе *P(C)*. Эта вероятностная функция обладает следующим свойством: если *C1* вид для C2, то *P(C1) =< P(C2).* Значение вероятности для наиболее верхней вершины иерархии равно 1. Следуя обычной аргументации теории информации, информационное содержание понятия *C* может быть представлено как отрицательный логарифм этой вероятности:

$$IC(C) = -\log(P(C)). \qquad (10.6)$$

Чем более абстрактным является понятие, тем меньше величина его информационного содержания.

Для решения задачи разрешения лексической многозначности, вводится понятие наименьшего общего вышестоящего (LCS = Least Common Subsumer). Алгоритм базируется на идее, что нужно выбирать такое значение многозначного слова, наименьшее общее вышестоящее которого наиболее информативно.

$$Sim_{Rz}(C1, C2) = IC(LCS(C1, C2)) \qquad (10.7).$$

Авторы работы (Jiang, Conrath, 1997) развивают формулу (10.7) следующим образом:

$$Sim_{JC}(C1, C2) = 2 * IC(LCS(C1,C2)) - (IC(C1) + IC(C2)), \qquad (10.8)$$

то есть учитывается не только коэффициент информационного содержания пересечения путей от синсетов, то и исходное местоположение самих исходных синсетов.

Подчеркнем, что для вычисления информационного содержания, а, значит, и применения описанных выше подходов необходимо иметь семантически размеченный корпус.

В работе (Patwardhan и др., 2002) описывается тестирование ряда предложенных на базе WordNet метрик на материалах конференции Senseval-2. Для 1723 многозначных существительных коллекции метрики применялись в контексте длиной одно слово. Например, для выражения *Plant with flowers,* по этим мерам вычислялось сходство существительных *plant и flower.*Лучший результат был получен для метрики, предложенной в работе (Jiang, Conrath 1997), и составил 39% точности.

В работе (Vossen и др., 2006) предлагается алгоритм разрешения лексической многозначности на основе разметки предметных областей Wordnet (Magnini, Cavaglia, 2000), при которой большинство синсетов тезауруса Wordnet отнесены к той или иной предметной области, а если подходящей предметной области нет, то к специальной области Factotum (см. п. 2.5.3.1)

Выбор значения многозначного слова основывается на проверке соответствия предметных областей этих значений и слов в локальном контексте (4 именные группы слева и 5 именных групп справа) и во всем тексте. Приводятся данные, что с помощью данной системы разрешения многозначности удалось сократить количество значений на 57-65%. При этом подчеркивается, что большинство сокращений относятся к словам из области Factotum (п.2.5.3.1), то есть к словам, не относящимся к конкретным предметным областям таким как *быть, начинаться, человек*.

Подход к разрешению многозначности на основе содержания целого текста тестируется в работе (Galley, McKeown, 2003).

На первом этапе происходит сопоставление с текстом, и в специальную структуру, называемую disambiguation graph записываются все встретившиеся значения. Устанавливаются связи между узлами: гипонимы (видовые понятия), гиперонимы (родовые понятия) и понятия, имеющие с данным понятием одно и то же родовое понятие, так называемые сестры.

На втором этапе происходит разрешение многозначности в предположении «одно значение на текст».

Для каждого значения насчитывается его вес, который представляется как функция, зависящая от типа отношения и от расстояния в тексте между анализируемым вхождением и близким по смыслу значением в тексте. Так, например, синонимы, родовые и видовые значения добавляют вес к соответствующему значению, независимо от своего местоположения в тексте. Выбирается значение, получившее максимальный вес. Зависимость коэффициента добавления веса от расстояния отражена в следующей таблице:

Семантическое отношение	1 предложение	3 предложения	1 абзац	Другое
Синонимы	1	1	0.5	0.5
Гипоним/ Гипероним	1	0.5	0.5	0.3
Синсеты-сестры	1	0.3	0.2	0

Если выбрать значение на основе полученных весов не удалось, то выбирается первое по порядку значение WordNet, которое является наиболее частотным в коллекции SemCor, семантически размеченной по значениям WordNet.

Точность разрешения многозначности на основе данного алгоритма на 35000 существительных 74 текстов корпуса Semcor оценивается как 62.09%.

Авторы работы (Mihalcea и др., 2004) используют алгоритм PageRank для разрешения многозначности на основе WordNet и целого текста как контекста.

Сначала для каждого значимого слова текста отмечаются все синсеты, в которые входит это слово. Такие синсеты становятся вершинами графа, ребрами графа являются отношения, полученные на основе отношений описанных в WordNet, включая:

- традиционные отношения между синсетами: гипонимия, гиперонимия, меронимия и т.п.;
- отношение номинализации, появившееся в WordNet 2.0, которое устанавливается между глаголом и существительным, являющимися дериватами;
- так называемые координатные отношения – отношения между видовыми синсетами являющиеся подвидами одного и того же родового синсета.

Выбирается значение, получившее максимальный PageRank.

Точность разрешения многозначности данного алгоритма для задачи «все слова текста» на тестовом материале Senseval-3 - 50.89%, с учетом наиболее частотного значения – 63.27%.

Заключение к главе 10.

Достигнутые показатели разрешения многозначности для задачи «все слова текста», которые собственно и является базой для последующей обработки текста, не кажутся достаточно высокими, поскольку не достигают и 70% точности.

С другой стороны, и между экспертами лексикографами могут возникать достаточно серьезные расхождения при разметке значений.

Для того, чтобы понять, насколько качество разрешения многозначности и его достигнутый уровень являются существенными для приложений, начат цикл исследований, в рамках которых разрешение многозначности включается в выполняемую задачу, например, в задачу поиска документов (Agirre и др., 2007). С 2008 года такое тестирование проводится в рамках форума по многоязычному информационному поиску CLEF (www.clef-campaign.org).

Глава 11. Тезаурусы в информационном поиске

Современные модели информационного поиска не используют знаний, описанных в тезаурусах и онтологиях, базируются на моделях текста как набора слов, предлагая изощренные методы учета частотностей встречаемости слов в предложении, тексте, наборе документов, совместной встречаемости слов и т.п.

Вместе с тем, существуют типы запросов к поисковым системам, которые являются сложными для современных технологий информационного поиска и, следовательно, качество поиска по этим запроса достаточно низкое. Исследованию таких запросов был посвящен специальный семинар под названием «Надежный доступ к информации» (Reliable Information Access), проведенный в 2003 году. В рамках этого семинара анализировались результаты поиска нескольких поисковых систем по трудным запросам, выявленным в рамках экспериментов конференции по информационному поиску TREC.

Обобщая результаты этих экспериментов, Д. Харман (Harman, 2005) указывала, что при проведении анализа исполнения трудных запросов посредством шестью разными поисковыми системами было выявлено, что проблемы, возникающие в процессе обработки трудных запросов этими системами, были сходны в значительно большей степени, чем это ожидалось. Часто системы возвращали разные документы одного и того же класса, не сумев найти релевантные документы. Среди потенциальных методов, которые могли бы улучшить выдачу систем по таким запросам, указывались методы расширения запросов, в том числе, и с использованием специальных ресурсов – тезаурусов.

В работе (Shah, Croft, 2004) в качестве одного из существенных факторов сложного запроса для современных информационных систем называлось расхождение между словесной формулировкой запроса и описанием релевантных ситуаций в документах коллекции, что, как показано в экспериментах, можно преодолеть с помощью тезаурусов.

Таким образом, одной из потенциальных возможностей преодоления проблем текущих моделей информационного поиска является встраивание в модели поиска знаний, описанных в онтологических ресурсах.

Целью этой главы является рассмотрение результатов работы методов, в которых для поиска документов в процессе автоматической обработки запроса используются тезаурусы и онтологии. Для такого изложения сначала необходимо кратко описать существующие модели информационного поиска.

11.1. Модели информационного поиска

11.1.1. Булевская модель

Исторически первой моделью информационного поиска является Булевская модель. В этом подходе слова запроса соединяются между собой логическими связками: AND ($\&$), OR($\vee$), NOT($\neg$), которые могут быть сгруппированы при помощи скобок. Таким образом, запрос пользователя представляется логической формулой, в которой атомами могут быть термины или какие-либо дополнительные условия (например, тип коллекции или документа, ограничение на расстояние между словами запроса и т.п.).

Поисковая машина, основанная на булевом поиске, возвращает документы, для которых формула запроса принимает истинные значения. Каждому атому формулы сопоставляется множество документов, для которых значение атома истинно. Если атом является термином, то ему сопоставляется множество документов, в которых термин встречается. Затем над множествами выполняются элементарные операции —

объединения, пересечения и дополнения, соответствующие логическим связкам между атомами.

Современные булевские модели информационного поиска включают также операторы близости элементов запроса, которая измеряется либо в количестве промежуточных слов между элементами запроса в документе, либо задается указанием структурной единицы документа (предложение, абзац), в которой должны упоминаться элементы запроса.

Булевская модель обработки запроса имеет ряд недостатков:

- на заданный запрос поисковая машина может вернуть очень много документов (или даже все документы коллекции). В этом случае пользователь вынужден последовательно добавлять условия в запрос, чтобы уменьшить результирующую выборку. Поиск производится методом проб и ошибок. В результате также часто возникает ситуация, когда условия булевского запроса оказываются противоречивы, и пользователь не получает ни одного документа;
- как правило, полезную выборку обозримого размера можно получить, задав сложную логическую формулу. При этом от пользователя требуется не только знание правил построения формул, но и достаточно хорошее знакомство с «языком» предметной области;
- вследствие того, что существует только два значения релевантности: «релевантен» (true) и «нерелевантен» (false), результирующая выборка не может быть упорядочена по релевантности. Все документы одинаково релевантны;
- все атомы формулы имеют одинаковую важность (вес), хотя некоторые из них могут быть «ключевыми», другие — вспомогательными.

В то же время Булевская модель имеет и положительные достоинства. Результаты ее работы хорошо предсказуемы и понятны. В булевском запросе могут быть объединены значения разных характеристик документов, включая как слова, содержащиеся в документе, так и такие характеристики как автор документа, время создания документа и т.д.

Несмотря на недостатки булевской модели, имеются ситуации, когда булевский поиск является предпочтительным, поэтому такая возможность поиска предоставляется многими поисковыми системами как интернет-поисковиками, так и различными коммерческими службами по поиску документов, библиотечными службами.

11.1.2. Векторная модель информационного поиска

Для упорядочения выдачи поисковой системы по мере соответствия ее запросу, необходимо ввести веса соответствия документов запросу, которые должны вычисляться на основе входящих в запрос слов (Buckley и др., 1993).

Простым способом определения значимости слова запроса в документе является частота употребления слова в документе (tf): чем чаще встречается слово запроса в документе, чем выше его вес. Такой способ вычисления веса слов запроса в документе предполагает, что все слова документа имеют одинаковую значимость. Однако слова документа могут иметь большую или меньшую различительную силу. Так, в базе «Законодательство России» практически каждый документ содержит слова *закон, законодательство, Россия, Российский*, поэтому данные слова в данной коллекции имеют низкую значимость для определения релевантности документов. Таким образом, можно предположить, что чем чаще в коллекции документов употребляется некоторое слово, тем меньше его значимость при нахождении релевантных документов.

Частотность употребления слова в коллекции может быть учтена посредством вычисления количества документов в коллекции, в которых содержится это слово, - df. При возрастании df, вес слова в документе должен снижаться. Это можно учесть, умножая частоту употребления слова в документе tf на обратную величину df – idf. Таким образом,

вес слова в документе может быть вычисляться по формуле tf*idf. Idf часто вычисляется по следующей формуле

$$\text{Idf}_{\text{tj}} = \log\left(\frac{N}{n_j}\right), \tag{11.1}$$

где N — число документов в коллекции, n_j — число документов, в которых встретился t_j.

Таким образом, пусть $D=(d_1,..d_N)$ — множество документов коллекции, $T=(t_1,..t_M)$ — множество слов-элементов запроса. Для каждого фиксированного i документ d_i представляется вектором весов

$$w_{ji} = tf_{ji} \cdot idf_{ji},\ j = 1..M, \tag{11.2}$$

где tf_{ji} — частота встречаемости слова t_j в документе d_i, idf_{ji} — величина, обратная частоте встречаемости слова t_j по всем документам коллекции.

После вычисления весов всех слов в документе документ может быть представлен как вектор, в котором каждый компонент соответствует отдельному слову документа. Представление документов и запросов в виде векторов, входящих в них слов, и составляет суть векторной модели информационного поиска.

Запрос также может быть представлен как вектор весов слов. Для определения сходства между векторами запроса и документа используется так называемая косинусная мера.

$$Sim\ (q, di) = (\sum w_{qti} * w_{dti}) / (\sqrt{\sum w_{qti}^2} * \sqrt{\sum w_{dti}^2}) \tag{11.3}$$

Таким образом, теперь соответствие запроса документу измеряется конкретным числом, и все документы могут быть упорядочены в выдаче поисковой системы по этому числу.

К преимуществам векторной модели информационного поиска относится то, что модель предоставляет простую модель для создания упорядоченной выдачи информационной системы. При этом конкретный способ вычисления весов слов в документе может меняться в зависимости от решаемой задачи и рабочей коллекции. Недостатком подхода является предположение о независимости слов в тексте, что противоречит тому, что в тексте используется множество связанных по смыслу слов.

11.1.3. Вероятностные модели информационного поиска

Одним из эффективных типов моделей информационного поиска являются вероятностные модели.

Вероятностные модели базируются на принципе ранжирования на основе вероятности, провозглашенном ван Ризбергеном в 1979 году, который заключается в следующем (Croft и др., 2009):

> *Если поисковая система в ответ на каждый запрос ранжирует документы в коллекции в соответствии с уменьшающейся вероятностью релевантности документа пользователю, который задал запрос,*
>
> *где вероятности оценены максимально точно на базе тех данных, которые доступны системе для этой цели,*
>
> *то общая эффективность системы по отношению к данному пользователю будет максимальной эффективностью, которая может быть получена на имеющихся данных.*

Мы не будем подробно приводить рассмотрения, которые проводятся в рамках данного класс моделей, поскольку это выходит за рамки данной книги. Мы укажем лишь наиболее известную модель оценки релевантности документа запросу, которая сформировалась в рамках этого подхода, а именно так называемую модель BM25, также называемую OKAPI по названию системы, в которой впервые такая схема взвешивания была применена (Robertson и др., 1994).

$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i,D)\cdot(k_1+1)}{f(q_i,D)+k_1\cdot(1-b+b\cdot\frac{|D|}{\text{avgdl}})},$$

где $f(q_i,D)$ это частотность терма qi в документе D, $|D|$ - это длина документа D в словах, avgdl – средняя длина документов в коллекции, k1 и b – это параметры формулы, обычно принимающие значения k1=2.0, b = 0.75. . IDF(q_i) - это обратная частота по коллекции, которая обычно вычисляется как:

$$\text{IDF}(q_i) = \log\frac{N-n(q_i)+0.5}{n(q_i)+0.5}, \qquad (11.5)$$

где N – это общее число документов в коллекции и $n(q_i)$ – это число документов, содержащих q_i.

11.1.4. Языковые статистические модели (language modelling)

Языковые статистические модели являются моделями информационного поиска, относительно недавно адаптированные к этой задаче из других сфер автоматической обработки текста и речи.

«Языковые статистические модели» - это группа статистических методов, которые оценивают вероятность появления последовательности из m слов $P(w_1,\ldots,w_m)$ посредством вычисления вероятностного распределения. Такие модели используются в самых различных сферах автоматической обработки текстов в таких как распознавание речи, машинный перевод, морфологический и синтаксический анализ текста.

В информационном поиске языковые модели используются для установления отношений между запросом Q и документами коллекции, в том смысле, что упорядочение документов при выдаче ответов на запрос определяется на основе оценки вероятности того, что языковая модель, построенная по документу, породит совокупность слов запроса $P(Q|M_d)$ (Ponte, Croft, 1998; Song, Croft, 1999).

Равенство (11.6) представляет собой основную формулу языковой модели информационного поиска для так называемой униграммной модели, то есть в том случае, если все слова запроса рассматриваются как независимые друг от друга сущности:

$$(*)\ P(q1, q2, \ldots qn \mid d) = \prod P(qi|d) \qquad (11.6)$$

Данная формула означает, что вероятность порождения запроса из документа в униграммной модели оценивается как произведение вероятности порождения отдельного элемента запроса из документа. Наиболее естественным способом оценки P(qi|d) является оценка вероятности встречаемости терма qi в документе d посредством так называемой оценки максимального правдоподобия (maximal likelihood estimate – MLE), то есть

$$P(qi|d) = \text{freq}(qi,d)/\text{length}(d) \qquad (11.7)$$

Оценка вероятности последовательностей слов может оказаться достаточно сложной для текстовых коллекций, поскольку некоторые возможные последовательности слов могли никогда не встречаться в базовой коллекции, и не могли использоваться для

качественной настройки языковой модели (training of language model), то есть возникает - так называемая проблема нехватки данных (data sparceness).

По этой причине важным элементом языковых моделей является процедура сглаживания (smoothing) (Chen, Goodman, 1998). Большинство формул сглаживания предложено в рамках моделей, созданных для распознавания речи.

В сфере языковых моделей для информационного поиска ситуация нехватки данных проявляется в том, что если элемент запроса не содержится в документе, то при выбранном способе оценки вероятности получаем P(qi|d)=0 и, следовательно, P (q1, q2, ….qn | d) =0.

Все процедуры сглаживания основаны на некотором снижении оценки вероятности на основе уже встреченных событий (то есть на основе появления термина в документе) и за счет этого появляется возможность дополнительно оценить вероятность событий, которые в конкретном документе не встретились.

Одной из распространенных техник сглаживания является учет вероятности появления слова в коллекции P(qi|C), и тогда обобщенная формула сглаживания выглядит следующим образом:

$$P(w|d)=\begin{cases} P_s(w|d), & \text{если слово запроса встречалось в документе,} \\ \alpha_d P(w|C), & \text{если слово не встречалось в документе.} \end{cases} \quad (11.8)$$

где $P_s(w|d)$ – это сглаженная вероятность P(w|d),

P(wi|C) – это вероятность появления слова в коллекции,

α_d - - коэффициент учета каждой из моделей, в общем случае может зависеть от документа.

Один из простейших вариантов формулы, так называемое сглаживание Jelinek-Mercer, выглядит следующим образом:

$$P(qi|M)= \lambda P(qi|d)+(1-\lambda)P(qi|C) \quad (11.9)$$

Другим примером формулы сглаживания является так называемая формула абсолютного дисконтирования (absolute discounting). Идея метода заключается в понижении вероятности встреченных слов путем вычитания констант вместо умножения их на коэффициенты λ и (1-λ):

$$P(qi|M)=(\max(c(qi,d)-\delta,0)/\Sigma_w c(w,d)) + \sigma P(qi|C) \quad (11.10)$$

где δ – сглаживающая константа величиной от 0 до 1; $\sigma = \delta |d|_u/|d|$; $|d|_u$ – число уникальных слов в документе; |d| - общее количество слов в документе, то есть

$$|d| = \Sigma_w c(w,d)) \quad (11.11)$$

Учет P(qi|C) в языковых моделях играет роль, сходную с учетом обратной частотности (idf) в векторной модели информационного поиска (Zhai, Lafferty, 2001)

Эксперименты в рамках конференции TREC (Ponte, Croft, 1998; Manning и др., 2008) показали эффективность языковых моделей для информационного поиска, однако существенным для эффективной работы методов является процедура подбора подходящей процедуры сглаживания.

В работе (Zhai, Lafferty, 2001) исследовались различные виды сглаживания. На основе этого исследования авторы делают выводы, что некоторые виды сглаживания в информационном поиске лучше подходят для коротких запросов, а другие для более длинных сложных запросов.

11.2. Оценка качества информационного поиска

Качество работы систем информационного поиска оценивается на основе специально разрабатываемых мер. Основными характеристиками качества

информационного поиска являются полнота и точность (Агеев, Кураленок, 2004; Manning и др., 2008)

Полнота (recall, r) — доля релевантных документов в выдаче поисковой системы по отношению ко всем релевантным документам коллекции.

Точность (precision, p) — доля релевантных документов по отношению ко всем документам в поисковой выдаче.

Пусть N — число документов в коллекции, n — число документов в коллекции, релевантных некоторому запросу, m — число документов в выборке, полученной системой на данном запросе, A — число релевантных документов в выборке. Тогда

$$p = A/m, \qquad r = A/n, \tag{11.12}$$

Этих характеристик достаточно, когда система поиска не производит дополнительного ранжирования документов. Если ранжирование документов производится, то нужно оценивать не только общее число найденных релевантных документов, но и на каких местах в выдаче располагаются релевантные документы.

Для определения качества работы поисковой системы в начале списка результатов поиска используется показатель *Точность на уровне n документов (Precision (n)),* который определяется как количество релевантных документов среди первых n документов, деленное на n. Например, если система выдает не более 10 документов на первой странице, то precision (10) отражает качество результатов системы, получаемых на первой странице.

Для оценки качества полной выдачи поисковой системы применяется показатель *средняя точность (average precision),* которая усредняет точность при выдаче каждого из K релевантных документов.

Точность на уровне i-го релевантного документа prec_rel(i) равна precision (pos(i)), если релевантный документ находится в результатх запроса на позиции pos(i). Если i-й релевантный документ не найден, то prec_rel(i)=0.

Средняя точность для заданного вопроса равна среднему значению величины prec_rel(i) по всем k релевантным документам:

$$AvgPrec=(1/k)\ \Sigma prec\_rel(i) \tag{11.13}$$

Усреднение величины средней точности по всем запросам дает величину MAP – mean average precision – число, которое характеризует работу поисковой системы по совокупности запросов.

При ранжированной выдаче значения точности и полноты при разных K могут быть отражены с помощью так называемой кривой «полнота- точность» (см.рис. 11.1).

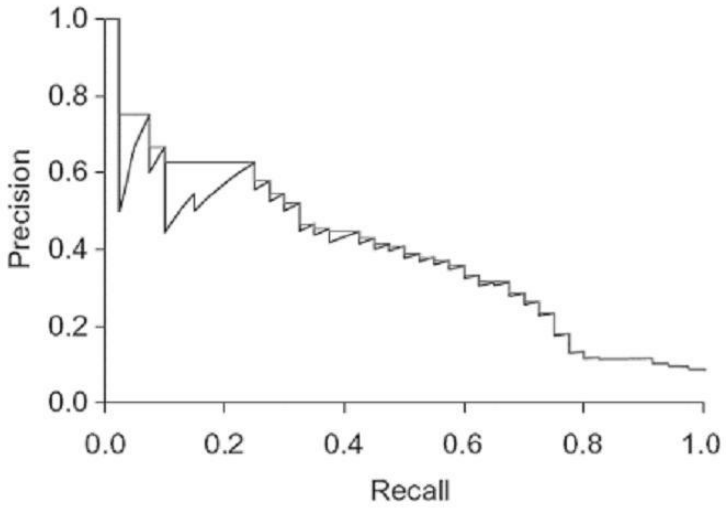

Рис. 11.1. Кривая «полнота-точность»

Получившийся график носит зигзагообразный характер, поскольку если (k+1)-й документ не является релевантным, то полнота выдачи не изменяется, а точность выдачи падает. Если очередной документ является релевантным, то возрастает как полнота, так и точность – кривая отклоняется вверх и вправо.

Для сглаживания этих зигзагов используется понятие интерполированной точности. Интерполированная точность p_{interp} на определенном уровне полноты r определяется как максимальная точность, полученная на уровнях полноты r1, больших чем r: $r1 \geq r$.

$$p_{interp}(r)=\max_{r' \geq r} p(r') \qquad (11.14)$$

Такое приближение убирает «внутренние» зубцы. Интерполированный график показан на рисунке тонкой линией.

Для количественного сравнения работы поисковых систем на разных уровнях полноты используется одинадцатиточечная интерполированная средняя точность (eleven-point interpolated average precision). Для вычисления этой величины по каждому поисковому запросу точность меряется в 11 точках на уровнях полноты 0.0, 0.1, 0.2…0.9, 1.0. Получается список из 11 значений точности, который может усредняться по всем тестируемым поисковым запросам.

Эти 11 значений точности могут быть отражены на графике интерполированной точности (рис. 11.2.). Именно такой график часто показывается при сравнении работы поисковых систем.

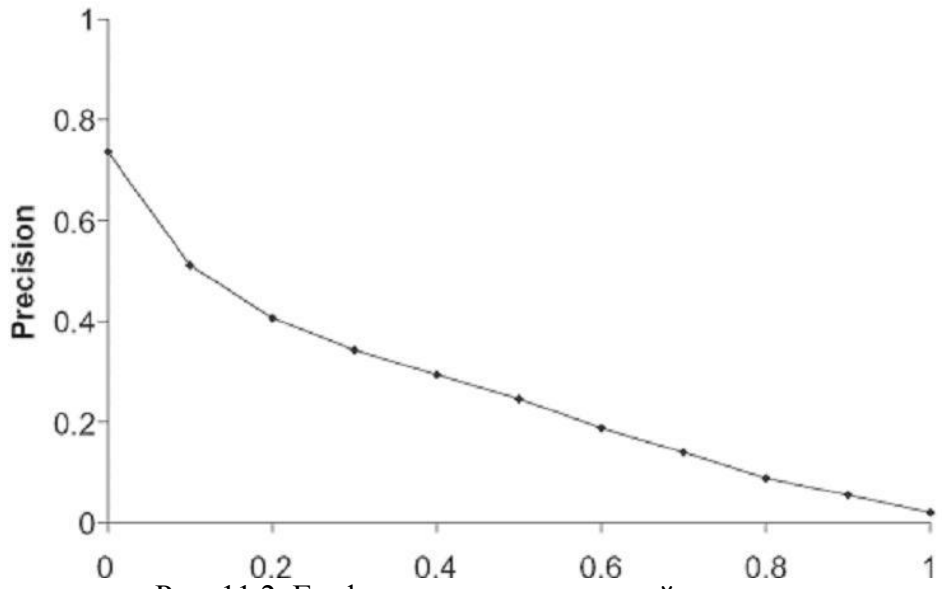

Рис. 11.2. График интерполированной точности

В настоящее время общепринятым является тестировать методы информационного поиска на базе общих коллекций документов в рамках специально проводимых конференций. Первой такой конференцией, впервые организованной в начале 90-х годов 20 века, стала конференция TREC (Text Retrieval Conference). Позже возникли такие конференции как CLEF (Cross Language Evaluation Forum), которая фокусируется на европейских языках и многоязычном поиске, NTCIR (восточно-азитские языки и многоязычный поиск). В России с 2003 года ежегодно собирается семинар по оценке методов информационного поиска – РОМИП (www.romip.ru) (Кураленок и др., 2003).

11.3. Тезаурусы типа WordNet в информационном поиске

Сразу после появления тезауруса WordNet в сети Интернет многие исследователи начали эксперименты по его применению в приложениях информационного поиска, полагая, что появился качественный ресурс, позволяющий резко улучшить качество поиска. Эти надежды были связаны с тем, что WordNet давал возможность использовать важные для задач информационного поиска сведения такие как, сведения о синонимах, значениях слов, лексических отношениях большого количеста слов английского языка.

Одной из первых таких работ была работа (Voorhees, 1994). Однако на основе экспериментов на конференции TREC-5 было показано снижение показателей информационного поиска при использовании WordNet. Автор утверждает, что, с одной стороны, лингвистические технологии должны быть близки к совершенству, чтобы привести к улучшению качества информационного поиска, с другой стороны, что статистические методы частично аппроксимируют лингвистические технологии своими статистическими корреляции. Похожее утверждение было высказано М. Сандерсоном (Sanderson, 1994), который предположил, что подходы, базирующиеся на ворднетах, будут хорошо работать, когда точность разрешения лексической многозначности приблизится к 90 процентам. Поэтому необходимость использования тезаурусов типа WordNet для информационного поиска и классификации документов, в настоящее время, не является общепризнанной.

Однако в последнее время появились работы, в которых учет WordNet при обработке запроса, приводит к значимым улучшениям поиска по сравнению с базовой моделью (см. пп. 11.3.3, 11.3.4.)

В данном разделе мы рассмотрим наиболее известные работы, в рамках которых предлагались различные подходы по интеграции тезауруса WordNet в существующие модели информационного поиска.

11.3.1. Эксперименты по использованию тезауруса WordNet в векторной модели информационного поиска

В работе (Voorhees, 1998) описываются эксперименты по интеграции WordNet в поиск по векторной модели. Целью экспериментов была попытка выполнить поиск документов на основе не отдельных слов, а значений WordNet. Для каждого документа сначала выполняется процедура разрешения многозначности существительных, которая выдает единственный синсет, и в результате которой каждому тексту ставится в соответствие вектор синсетов WordNet. После того, как вектор создан, с ним могут выполняться такие же операции, как и с пословными векторами.

Эффективность использования векторов синсетов сравнивалась с эффективностью информационного поиска на основе стандартного вектора слов. В стандартном прогоне и документы, и запросы представляются как вектора лемм всех значимых слов. В концептуальных прогонах документы и запросы представляются как вектора, состоящие из трех подвекторов:

1) вектор лемм слов, не найденных в WordNet, или тех, многозначность которых не удалось разрешить – например, относящихся к другим частям речи;
2) вектор синсетов для слов с разрешенной многозначностью,
3) леммы для слов с разрешенной многозначностью.

Второй и третий подвектора представляют собой альтернативные представления документа, поскольку одни и те же слова этого документа порождают отдельные элементы каждого вектора.

Для экспериментов было использовано 5 разных коллекций документов (компьютерная область, медицинская область, газетные статьи и др.), и для каждой коллекции было выполнено более 30 различных запросов.

Для каждого запроса стандартный прогон векторной модели сравнивался со следующими комбинациями подвекторов (цифры означают вес каждого из трех подвекторов):

110 – данная комбинация дает одинаковые веса словам, отличным от существительных и синсетам существительных;

211 – данная комбинация учитывает как синсеты, так и леммы существительных, оставшиеся слова поэтому учитываются в двойном размере;

101 – в данной комбинации подвектор синсетов игнорируется, а существительные и другие леммы документа получают одинаковые веса. Однако этот вектор отличается от стандартного прогона, поскольку результат сравнения для системы подвекторов высчитывается как сумма результатов сравнения каждого вектора.

Оценки эффективности информационного поиска на основе показателя средней точности показали серьезное ухудшение эффективности для векторов, включающих синсеты (от 6.2 до 42.3%).

Основная причина такого ухудшения эффективности заключается в том, что процедура разрешения многозначности для слова в запросе может выбрать одно значение, а для того же слова в документе другое значение. Например, при поиске по запросу ”*separation anxiety in infants and preschool children*” (*озабоченность разлукой у детей младшего возраста и дошкольников)* стандартный прогон пословной векторной модели выдает 7 релевантных документов в первых 15 документах, в то время как прогон 110 выдает только один релевантный документ в первых 15 документах. Проблема выдачи по данному запросу состояла в выборе значения слова *separation*, для которого в WordNet описано 8 значений. Процедура разрешения многозначности выбирает такое значение этого слова в запросе, которое не было выбрано ни в одном из релевантных документов.

В другой группе экспериментов по использованию WordNet в информационном поиске исследовалась возможность расширения запроса синонимами или другими словами, связанными со словами запроса отношениями, описанными в WordNet. В таких экспериментах нет необходимости выбора единственного значения слова, что в случае ошибки приводит к серьезному ухудшению результатов поиска.

Для экспериментов были использованы следующие соображения.

Во-первых, расширяться должны только важные для запроса понятия. Важность определяется на основе параметра количество документов, в которых встречается конкретное слово запроса - слова, частотность которых в документах коллекции больше некоторого числа N, не участвуют в расширении запроса. Во-вторых, чтобы смоделировать разрешение многозначности, запрос расширяется только теми словами, которые оказались в окрестностях расширения, по крайней мере, двух слов запроса.

Таким образом, сначала для каждого слова запроса, частотность которых меньше некоторого числа N, и каждого синсета для значений этого слова извлекается список близких по WordNet слов. Те слова, которые встретились по крайней мере в двух таких списках, добавляются к исходному запросу.

Исследовались различные величины N – 10% коллекции и 5% коллекции.

Для расширения запроса использовались синсеты, находящиеся на расстоянии 1 и 2 отношения от исходных синсетов - все виды связей трактовались одинаково.

Добавленные слова могли учитываться с разными величинами весов w=0.3, 0.5, 0.8.

Максимальное улучшение, которое удалось получить – 0.7% средней точности, что не является статистически значимой величиной (N=5%, расстояние – 2, w=0.3).

Авторы подчеркивают, что идея аппроксимации разрешения многозначности путем поиска повторов в списках расширения не является удачной, поскольку чаще всего это решение приводило к добавлению в запрос очень общих слов, таких как *система* и др.

Для того, чтобы исключить из рассмотрения эффект лексической многозначности и исследовать возможности WordNet по расширению поискового запроса, были выполнены эксперименты с ручным выбором значения многозначных слов в запросе.

Для каждого синсета, соответствующего слову запроса, в запрос могут быть добавлены разные слова на основе различных отношений данного синсета, например, синонимы, все слова из нижестоящих синсетов иерархии гипоним- гипероним, все слова, отстоящие на один шаг от текущего синсета.

Чтобы исследовать все такие возможности был образован вектор, состоящий из 11 подвекторов: один для слов исходного запроса, один для синонимов, один для каждого типа отношений существительных в WordNet. Сходство с документами вычислялась как взвешенная сумма результатов сравнений с каждым из подвекторов.

Исследовались четыре варианта векторов:

1) расширение только по синонимам,
2) расширение синонимы + полная иерархия вниз
3) расширение синонимы+ родители+ полная иерархия вниз
4) расширение синонимы+ слова из любых синсетов на один шаг по любому типу отношений.

Тестирование проходило на двух типах вопросов: более длинной и более короткой версии запросов. При поиске по полному запросу ни одной из комбинаций не удалось улучшить результаты поиска более чем на 2 процента.

Короткие вопросы состояли из небольшого списка синсетов, например, {cancer}, {skin_cancer}, {phramaceutical}. Для таких укороченных запросов, используя тип расширения 4), при котором все добавления учитывались с коэффициентом 0.5, было получено 35% улучшение: средняя точность для укороченного запроса без расширения была – 0.1634, с расширением – 0.2205. Средняя точность поиска по полному запросу – 0.3586. Таким образом, при ручном разрешении многозначности удается получить значительное улучшение качества поиска при расширении по тезаурусу WordNet.

Основные выводы автора работы заключались в том, что для успешного применения WordNet в информационном поиске необходимо значительно улучшить эффективность автоматического расширения лексической многозначности.

11.3.2. Эксперименты по семантическому индексированию на базе европейских ворднетов.

В рамках европейского проекта Meaning, который является развитием проекта EuroWordNet, голландская компания Irion Technologies разработала технологию концептуального индексирования TwentyOne, комбинирующую лингвистический и статистический подходы (Vossen и др., 2006). Авторы разработки считают, что неудачи с использованием WordNet в информационно-поисковых приложениях связаны с трудностями встраивания такого рода лингвистических ресурсов в приложения, оптимального использования содержащейся в ворднетах информации.

Основой технологии является статистическая машина поиска, базирующаяся на стандартной векторной модели и обеспечивающая быстрый поиск документов.

Лингвистические технологии используются в двух ролях:

- максимизация полноты выдачи статистической машины за счет синонимии ворднетов;
- максимизация точности выдачи за счет сравнения запросов с конкретными фразами документов, а не с целыми документами.

Фраза представляет собой именную группу (noun phrase). Каждая фраза ассоциируется с отдельными словами, определенной комбинацией слов, а также комбинацией частей слов.

Система TwentyOne использует совокупность факторов для сравнения запроса с фразами текста:

1. Число совпадающих синсетов между запросом и каждой фразой,
2. Степень нечеткого сопоставления между запросом и каждой фразой,
3. Степень деривационного несовпадения, слитного–раздельного написания и т.п.,

4. были ли использованы синонимы,
5. был ли использован тот же язык.

При обработке запроса сначала с помощью векторной модели находятся документы, соответствующие запросу. Затем выданные документы переранжируются так, что сначала выдаются документы, которые имеют наибольшее совпадение по синсетам фраз с запросом. Среди документов, имеющих одинаковое количество сопоставленных синсетов между собственными фразами и запросом, первыми выдаются наиболее похожие по конкретному набору слов. Вес документа по векторной модели используется, если вес по фразам текста получился одинаковым.

Разрешение многозначности в данной системе делается на основе технологии, описанной в (Magnini и др., 2002) и базируется на разметке предметных областей wordnet (см. п. 2.5.3.1.).

Система разрешения лексической многозначности сначала настраивается на наборы слов, относящихся к той или иной предметной области на основе разметки, осуществленной в WordNet. При обработке конкретного документа система сначала присваивает предметную область документу в целом, так называемые микротэги. Затем классифицирует отдельные именные группы внутри контекстного окна 10 именных групп (4 именные группы слева и 5 именных групп справа). В результате этот фрагмент получает один или более тэгов (нанотэги).

При разрешении многозначности конкретного слова сначала выбираются значения, соответствующие нанотэгам. Если нет соответствия с нанотэгами, выбираются значения, соответствующие микротэгам. Если никаких соответствий не обнаружено, выбираются все значения.

Приводятся данные, что с помощью данной системы разрешения многозначности удалось сократить количество значений на основе целого текста: для испанского языка – 48%, для английского языка – 57%. В случае использования контекстных окон сокращения выше: 52% для испанского языка и 65 для английского. При этом подчеркивается, что большинство сокращений относятся к словам из области Factotum (см.п. 2.5.3.1.), то есть словам, не относящимся к конкретным предметным областям, таким как *быть, начинаться, человек*.

В проводимых экспериментах для сравнения были построены четыре индекса:

1) HTM – традиционный пословный индекс;
2) NP - индексы именных групп из запроса, с использованием пословных методов, без использования ворднетов;
3) FULL - полные индексы с использованием ворднетов, но без процедуры разрешения многозначности, что приводит к полному расширению по синонимам и переводам для всех возможных значений слов запроса;
4) WSD - индексы, использующие ворднеты вместе с описанной выше процедурой снижения многозначности на основе предметных областей ворднет.

Полученные индексы тестировались при поиске по документам коллекции Reuter и по коллекции подписей к картинкам в ресурсе Fototeca (Vossen и др., 2005). Базовыми языками для тестирования являются английский и испанский языки. Запросы для тестирования извлекались из самих документов, кроме того, в качестве запросов использовались также запросы, полученные синонимической заменой слов из исходных запросов. В результате тестирования авторы делают вывод о полезности тезаурусов типа WordNet для информационного поиска, однако из-за специфической процедуры формирования тестового набора запросов трудно оценить, насколько этот вывод обоснован в данных экспериментах.

11.3.3. Исследования влияния качества разрешения лексической многозначности на информационный поиск

Вопрос о том, улучшит ли разрешение многозначности слова, поиск по словам в правильном значении, остается дискуссионным. Некоторые авторы (Voorhees, Stevenson) полагают, что если запрос однозначно определяет значение многозначного слова в своем составе, то и в найденных документах, это слово окажется в окружении тех же слов запроса, и тем самым с большой вероятностью будет употребляться в том же значении.

Если же выполняется автоматическая процедура разрешения лексической многозначности, то ошибки в работе этой процедуры могут привести к значительному снижению качества информационного поиска, как это и было показано в экспериментах H.Voorhees (Voorhees, 1994). В работе (Stevenson, 1994) автор вводит в коллекцию искусственную многозначность и тем самым может контролировать процент ее ошибочного разрешения. В исследовании было показано, что при качестве разрешения многозначности хуже 90% эффективность информационного поиска начинает резко снижаться.

В исследовании (Gonzalo и др., 1998) авторы ставят перед собой два вопроса:

1) Абстрагируясь от проблемы разрешения многозначности, какой потенциал несет использование ресурсов типа WordNet для информационного поиска. Такой эксперимент можно выполнить, если сделать вручную разрешение лексической многозначности запросов и документов;
2) Если эффективность использования WordNet для коллекции с разрешенной многозначностью известна, то можно измерить чувствительность качества информационного поиска к ошибкам разрешения многозначности, искусственно внося некоторый процент ошибок в разметку по значениям.

Исследования выполнялись на корпусе SemCor, размеченного значениями WordNet. Были выбраны 171 текстовых фрагментов со средней длиной 1331 словом на документ. Для каждого текста была написана краткая аннотация длиной от 4 до 50 слов, в среднем 22 слова на документ. Эти аннотации использовались как запросы по текстовой коллекции, то есть был ровно 1 релевантный документ на запрос. Аннотации также были размечены по значениям WordNet. На основе стандартного списка стоп-слов английского языка был также автоматически порожден список стоп-синсетов.

В экспериментах использовалась векторная модель в версии информационно-поисковой системы SMART (Salton, 1989) и три типа векторов: исходные слова документа, значения слов, соответствующие словам документа, и синсеты WordNet, соответствующие словам документа (в последнем случае фактически производится дополнение документа синонимами слов).

В процессе эксперимента выяснялось, какой процент документов был возвращен на первом месте в выдаче.

Эксперименты показали, что стандартная векторная модель дает 48% первых релевантных документов, индексирование по значениям слов – 53.2% и индексирование по синсетам – 62%.

Внесение ошибок разрешения многозначности в индексирование по синсетам показало, что 10% ошибок не влияет на качество поиска, что находится в соответствии с работой (Sanderson, 1994). При этом выяснилось, что при уровне 30% ошибок качество поиска превосходит поиск по стандартной модели SMART (54.4%). Таким образом, авторы делают вывод, что если выполнять разрешение многозначности с точностью больше 70%, то это даст преимущество по сравнению с пословными векторными моделями. Важно однако заметить, что за прошедшее время векторные модели значительно усложнились, включая поиск близких по тексту терминов, поиск по абзацам и др.

Для того чтобы изучить, насколько в приложениях информационного поиска можно использовать системы разрешения многозначности с такими показателями, в рамках конференции SemEval-2007 (http://nlp.cs.swarthmore.edu/semeval/), одним из заданий которой является применение алгоритмов разрешения многозначности в рамках задачи информационного поиска (Agirre и др., 2007). Суть задания заключается в следующем: все участники должны выполнять поиск на одной и той же поисковой машине, однако перед поиском необходимо расширить запросы или тексты синонимами или переводами, соответствующими выбранным значениям.

Было предложено три подзадания:

- информационный поиск с автоматическим разрешением многозначности запроса - системы должны автоматически разрешить многозначность слов запроса, расширить запрос синонимами, соответствующими этим значениям и выполнить расширенный поисковый запрос. Документы и запросы на английском языке;
- информационный поиск с автоматическим разрешением многозначности документа – системы должны автоматически разрешить многозначность слов в документах, расширить документы синонимами, соответствующими этим значениям и выполнить поиск на основе исходного поискового запроса;
- двуязычный поиск (с испанского на английский) – для документов автоматически производится разрешение многозначности, документы переводятся в соответствии с полученными результатами разрешения и затем выполняется поиск с использованием исходного поискового запроса.

Результаты систем сравниваются с базисными уровнями: поиск без расширений (noexp), и поиск с полным расширением - запросы расширяются синонимами, соответствующими всем возможным значениям (expall).

В проведенных экспериментах в одноязычном поиске лучший результат был получен при поиске без расширения синонимами noexp - 0.3599 MAP, в двуязычном информационном поиске использованием переводов по всем значениям expall - 0.2617 MAP.

Таким образом, в первом проведенном соревновании с использованием методов автоматического разрешения многозначности системам не удалось получить результаты, превышающие результаты методов, не использующих процедуру автоматического разрешения многозначности.

Организаторы процедуры оценки связывают часть проблем с выбранной базовой системой поиска и намерены продолжать исследования роли автоматического разрешения многозначности в информационном поиске.

11.3.4. Эксперимент по встраиванию тезауруса WordNet в вероятностную модель информационного поиска

В работе (Liu и др., 2004) в качестве базовой модели информационного поиска используется формула OKAPI (Robertson, 1994), к которой добавлен поиск по фразам и используется расширение запроса по отношениям WordNet. После разрешения многозначности слов к запросу добавляются синонимы, гипонимы и слова из определений синсетов. Основное свое внимание авторы концентрируют на коротких запросах (двух или трехсловных запросах).

Значение многозначного слова в запросе выбирается на основе толкований синсетов WordNet. Значение слова в запросе может быть выбрано, если:

- его толкование пересекается с другими словами запроса;
- пересечение его толкования с толкованиями других слов запроса максимально,
- толкование одного из его гипонимов пересекается с другими словами запроса,
- если никакие проверки не привели к выбору значения слова, то берется наиболее частотное значение.

Выбранные значения используются не для того, чтобы построить концептуальный индекс (индекс синсетов), а для того, чтобы найти подходящее расширение запроса.

Учитывая предшествующие неудачи использования WordNet для расширения запросов, авторы вводят дополнительные проверки возможности расширения, а также вес расширения. Важным элементом проверки возможности расширения запросов является предварительная оценка глобальной корреляции между отдельными словами.

Для оценки глобальной корреляции между словами используется следующая формула:

$$Global\_correlation\ (ti,s)=idf(s)*log(dev(ti,s)), \quad (11.15)$$

$$dev\ (ti,s) = (co\text{-}occurrence\ (ti,s)-dfi*sdf/N)/(\ dfi*sdf/N) \quad (11.16)$$

где s – элемент запроса (отдельное слово или словарное выражение), ti – некоторое другое выражение, dfi и sdf – это количество документов, содержащее ti и s соответственно, N – число документов в коллекции, idf (s) – обратная частота встречаемости s, co-occurrence (ti,s)- число документов, в которых встречаются ti и s, dev (ti,s) показывает степень отклонения совместной встречаемости ti и s от независимого употребления.

Рассмотрим, как авторы предлагают расширять запрос, состоящий из двух термов t1 и t2, синонимами.

Терм t11, который является синонимом к терму запроса t1 в синсете S, может быть добавлен в качестве расширения запроса, в одном из двух случаев:

- или S – является доминантным синсетом для терма t11, то есть t11 наиболее часто употребляется в значении, соответствующем синсету S;
- или t2 имеет высокую степень корреляции с t11, и величина корреляции между t2 и t11 больше, чем величина корреляции между t2 и t1.
- При этом расширение производится со следующим весом:

$$w(t11)=f\ (t11,S)/F(t11) \quad (11.17)$$

где f (t11,S) - это частота встречаемости терма t11 в значении S, F(t11) – это сумма всех частот для всех значений t11. Частота значений берется из информации, приписанной синсетам в WordNet, которая, в свою очередь, получена на основе разметки текстового корпуса значениями WordNet. Этот вес интерпретируется как вероятность того, что терм t11 имеет значение S.

Для расширения запроса гипонимами проводятся проверки другого рода.

Пусть U – синсет-гипоним для t1. Синоним из U добавляется к запросу в следующих случаях:

1) U - это единственный гипоним синсета S терма t1. Для каждого терма t11 из U этот терм добавляется к запросу, с весом (11.17), если U – это доминантный синсет t11;
2) U – это не единственный гипоним синсета S терма t1, при этом определение U содержит либо термин t2 или его синонимы. Тогда для каждого терма t11 из U этот терм добавляется к запросу, с весом (11.17), если U – это доминантный синсет t11.

Авторы работы показывают на пяти разных текстовых коллекциях конференции TREC, что применение технологии разрешения многозначности к коротким запросам и на этой основе расширение запроса приводит к росту средней точности поиска от 4% до 34%.

11.3.5. Эксперимент по использованию WordNet в рамках языковой модели информационного поиска

Результаты по улучшению информационного поиска с использованием WordNet и информации о совместной встречаемости слов в рамках языковой модели информационного поиска получены в работе (Cao и др., 2005).

Авторы работы подчеркивают, что классическая языковая модель информационного поиска основана на независимости слов в текстах друг от друга, что не соответствует реальному положению дел.

Информацию о взаимосвязи слов можно получить из двух источников:

- во-первых, подсчитывая совместную встречаемость слов в некотором текстовом окне.
- во-вторых, извлекая вручную описанные отношения из WordNet, поскольку некоторые указанные лингвистами отношения между словами может быть невозможно извлечь из рабочей коллекции. При этом отношениям из WordNet предлагается приписывать вес также на основе их совместной встречаемости в текстовом окне заданной величины.

Таким образом, оценивая вероятность порождения запроса из документа, предлагается использовать три источника информации по следующей формуле:

$$P(q|d)=\prod_{I=1}[\lambda_L P_L(qi|d) + \lambda_{CO} P_{CO}(qi|d) + \lambda_U P_U(qi|d)], \quad (11.18)$$

где $P_U(qi|d)]$ – вероятность, полученная по классической униграммной языковой модели, - далее модель UM.

$P_L(qi|d)$ – вероятность порождения запроса из документа, полученная на основе отношений лингвистического ресурса WordNet, - далее модель LM

$P_{CO}(qi|d)$ -– вероятность порождения запроса из документа, полученная на основе совместной встречаемости двух слов в текстовом окне, - далее модель CM.

$\lambda_L, \lambda_{CO}, \lambda_U$ –подбираемые коэффициенты.

Исследовался и другой вариант формулы, который приписывал отдельные веса разным типам связей WordNet: синонимам, гипонимам и гиперонимам:

$$P(q|d)=\prod_{I=1}[\lambda_1 P_{SYN}(qi|d) + \lambda_2 P_{HYPE}(qi|d) + \lambda_3 P_{HYPO}(qi|d) + \lambda_4 P_{CO}(qi|d) + \lambda_5 P_U(qi|d)], \quad (11.19)$$

где $\lambda_{1\ldots5}$ – весовые коэффициенты каждого типа отношений.

В базовой униграммной языковой модели в качестве формулы сглаживания использовалась формула абсолютного дисконтирования (см. раздел 11.1.4).

Совместная встречаемость слов, связанных между собой по WordNet, оценивалась в пределах абзаца. Совместная встречаемость слов, не поддержанных отношениями в WordNet, оценивалась в окне из 7 слов.

Для оценки совместной встречаемости в обоих случаях была также применена формула в духе языковых моделей с типом сглаживания по абсолютному дисконтированию. Так, формула для слов, между которыми описаны отношения в WordNet, такова:

$$P_L(w_i \mid w) = \frac{\max(c(w_i, w \mid W, L) - \delta, 0)}{\sum_{w_j} c(w_j, w \mid W, L)} + \frac{c(*, w \mid W, L)\delta}{\sum_{w_j} c(w_j, w \mid W, L)} P_{add-one}(w_i \mid W, L)$$

$$P_{add-one}(w_i \mid W, L) = \frac{\sum_{j=1}^{|v|} c(w_i, w_j \mid W, L) + 1}{\sum_{i=1}^{|v|}\sum_{j=1}^{|v|}(c(w_i, w_j \mid W, L) + 1)} \quad (11.20)$$

где C (wi,w|W,L) – число совместных встречаемостей слов wi и w, связанных отношениями в WordNet, в пределах окна, C(*, w|W,L) - число уникальных терминов встречающихся в окне W. Данная формула соответствует так называемой битермной языковой модели (Srikanth, Srikari, 2002).

Предложенная модель тестировалась на текстовых коллекциях конференции TREC, общим размером более 1200 мегабайт и состоящих из трех различных подколлекций следующих изданий Wall Street Journal (WSJ), Associated Press (AP), San Jose Mercury News (SJM). В качестве базового уровня использовалась униграммная языковая модель информационного поиска, реализованная на инструментальном средстве Lemur (Ogilvie, Callan, 2001). В качестве параметра оценки качества поиска использовалась классическая мера конференции TREC – средняя точность.

Оба варианта модели показали лучшие характеристики средней точности, по сравнению с базовой моделью. Причем улучшения на подколлекции Associated Press достигли 10%, а на других коллекциях – 5%. Большее увеличение показал второй вариант модели, который использовал разные весовые коэффициенты для разных типов отношений WordNet.

Анализ различных комбинаций подэлементов модели показал, что комбинация всех трех элементов модели (UM+LM+CM) всегда превышает показатели частичных комбинаций моделей. Это подтверждает мысль авторов, что посредством привлечения знаний из WordNet удалось использовать в поиске дополнительные сведения, которые не удалось получить на базе только использования информации о совместной встречаемости слов в текстовом окне.

Сочетание моделей UM+LM, то есть базовой модели и модели, основанной на отношениях WordNet, лучше, чем базовой модели UM. В работе делается вывод, что пригодность WordNet для той или иной коллекции может быть автоматически определена посредством автоматической процедуры настройки параметров, которая приписывает такие веса отношениям, установленным в WordNet, которые наиболее хорошо подходят для данной коллекции.

В таблице 11.1 показаны подобранные веса для каждого элемента модели. Как видно, подобранные веса в значительной мере различаются:

Модель	WSJ	AP	SJM
UM	0.3564	0.3006	0.4858
CM	0.1480	0.5282	0.1588
Синонимы	0.1657	0.0883	0.1392
Гиперонимы	0.1745	0.0491	0.0963
Гипонимы	0.1649	0.0338	0.11968
Всего	1.0	1.0	1.0

Таблица 11.1. Веса компонентов, используемых в модели

11.3.6. Расширение по WordNet на основе параметра «ясности» слова запроса

В работе (Shah, Croft, 2004) исследуется вопрос, насколько величина точность работы систем информационного поиска в смысле обеспечения высокой точности выдачи в первых документах выдачи. Исследуя результаты поиска системы Lemur (Ogilvie, Callan, 2001) по заголовкам запросов TREC они показали, что только в 40 процентах из 150 исследуемых запросов на первом месте поисковой выдачи находился релевантный документ.

Проанализировав причины такой ситуации, авторы работы установили, что это происходит из-за следующих проблем:

- наличия многозначных слов в запросе;
- наличие слов различной значимости в запросе;

- несоответствие слов запроса и коллекции. Так, причиной нерелевантности первого документа в выдаче по запросу «Fiber Optics Equipment Manufacturers" было то, что в релевантных документах коллекции чаще употреблялось слово "producers".

Рассматривая возможности автоматического расширения запроса, авторы отметили, что для обеспечения качественного расширения запроса необходимо определить, какие именно слова можно дополнить близкими по смыслу словами в контексте данного запроса, и какими именно из близких по смыслу слов. Так, включение в запрос многозначного слова может привести к резкому снижению качества поиска.

Для определения критериев расширения запроса близкими по смыслу словами авторы предлагают использовать показатель ясности ("clarity") слов. Вычисление этого параметра основывается на следующих наблюдениях.

Если в ответ на запрос получены релевантные документы, то первые документы выдачи характеризуются относительно высокой частотностью небольшого числа тематических терминов. С другой стороны, если в ответ на запрос выдаются нерелевантные документы разнообразной тематики, то по распределению частот документы выдачи должны быть сходны с коллекцией в целом.

Основные этапы расширения запроса заключаются в следующем:

1) вычислить ясность отдельных слов запроса,
2) все слова запроса делятся по параметру ясности на три группы:
   - слова с высокой ясностью не расширяются и оставляются в запросе;
   - слова с низким показателем ясности исключаются из запроса;
   - синонимы слов со средним показателем ясности используются для расширения запроса.

В результате экспериментов было получено, что при поиске по заголовкам запросов параметр Precision (1) повысился на 16,40% с 40.67% до 46.67%, средняя точность выросла на 0.89%. При поиске по полю описание (description) запроса Precision (1) повысилась на 18,18% с 44.00% до 52.00%, средняя точность выросла на 11.45%.

Таким образом, выборочное расширения запроса синонимами из WordNet привело к значимому улучшению результата поиска как по критерию Precision(1), так и по показателю средней точности.

Заключение к главе 11.

В качестве базовых моделей информационного поиска используется несколько различных моделей: булевская модель, векторная модель, вероятностная модель, языковая модель. Наиболее применяемые в настоящее время модели рассматривают текст как набор независимых слов.

При появлении в открытом доступе в сети Интернет тезауруса WordNet многие исследователи предположили, что использование этого ресурса непременно должно улучшать качество информационного поиска, поскольку WordNet предоставляет большое количество дополнительной информации о словах, их синонимах, значениях, отношениях.

Однако многочисленные первые эксперименты по интеграции WordNet в информационный поиск закончились неудачей. Понадобилось практически 10 лет, чтобы предложить модели, в которых применение WordNet дало значимое улучшение качества информационного поиска. Основной смысл предложенных удачных моделей заключается в том, что информация, полученная из WordNet, должна дополнительно взвешиваться, дополнительно оцениваются на основе особенностей конкретной коллекции, на которой производится поиск. Таким образом, производится как бы настройка WordNet на конкретную коллекцию и типовые запросы к этой коллекции.

Глава 12. Тезаурусы в вопросно-ответных системах

Одним из активно развивающихся направлений в сфере информационного поиска является разработка вопросно-ответных систем.

Исследования в области создания вопросно-ответных систем были начаты в 60-е годы. В то время предполагалось, что ответ на вопрос должен искаться в специально подготовленных базах знаний. Второе рождение вопросно-ответные системы стали переживать с 90-х годов 20 века. Теперь вопросно-ответные системы, в подавляющем большинстве случаев, должны искать ответы в больших текстовых коллекциях. От традиционных информационно-поисковых систем вопросно-ответные системы отличаются тем, что должны предоставить пользователю не набор документов, которые наиболее релевантны поставленному вопросу, но выдать фрагмент текста, содержащий точный ответ на заданный вопрос.

В 1999 году стало проводиться тестирование вопросно-ответных систем («вопросно-ответная дорожка») в рамках конференции TREC (Voorhees, 2004), с 2003 года соревнования вопросно-ответных систем в многоязычном контексте начаты на конференции CLEF (Magnini и др., 2005).

Приведем примеры вопросов из конференции TREC:

What is the brightest star visible from the Earth?
Какая звезда, видимая с Земли, является самой яркой?

Which is the Mozart birth date?
Какова дата рождения Моцарта?

When did Hitler attack Soviet Union?
Когда Гитлер напал на Советский Союз?

С 2001 года в рамках вопросно-ответной дорожки конференции TREC стало уделяться особое внимание не только ответам на вопросы о фактах (фактоидные вопросы), но и вопросам на определения и вопросам, предполагающим в качестве ответов списки. В 2003 году отдельные вопросы сменились тематическими группами вопросов, что может моделировать диалог пользователя с вопросно-ответной системой (Voorhees, 2004). Например, предлагалась такая группа запросов о писателе Франце Каффке.

1. *Where was Franz Kaffka born?*
   (Где родился Франц Каффка? – фактоидный вопрос)
2. *When was he born?*
   (Где он родился? – фактоидный вопрос)
3. *What is his ethnic background?*
   (Кто он по национальности? – фактоидный вопрос)
4. *What books did he wrote?*
   (Какие книги он написал? – вопрос на получение списка ответов)

С 2007 на конференции TREC было предложено новое направление исследований в построении вопросно-ответных систем, а именно, поиск ответов на вопросы по блогам, причем коллекция блогов включает как тексты, написанные на хорошем английском языке, так и тексты с плохим английским, а также спамерские тексты.

12.1. Основные этапы обработки вопросов в вопросно-ответных системах

Основными этапами поиска ответа на вопрос в современных вопросно-ответных системах являются следующие (см. рис. 12.1).

Прежде всего, производится подробный анализ вопроса, в результате которого определяется тип вопроса (вопрос времени, места, количества и другие) и соответствующий тип ответа, а также формируется запрос к информационно-поисковой системе.

На втором этапе производится поиск релевантных документов или абзацев информационно-поисковой системой, формируется упорядоченный список наиболее релевантных документов (абзацев), из которого выбирается первых n (например, n=100-1000) документов (абзацев) для дальнейшей обработки.

На третьем этапе производится подробный анализ полученных абзацев: содержит ли абзац требуемый тип ответа, близость слов ответа и вопроса, сходство синтаксических структур и т.п. В ходе такого анализа полученные абзацы оцениваются по мере возможности вхождения в них ответа на заданный вопрос, и переупорядочиваются на основе полученных оценок.

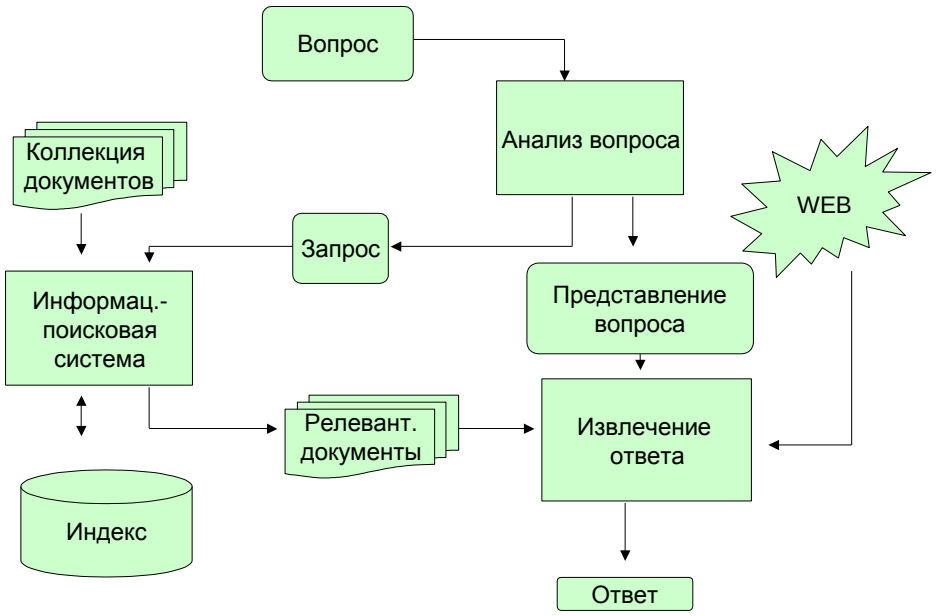

Рис. 12.1 Основные этапы обработки вопроса и формирования ответа вопросно-ответной системой

Обработка поискового запроса в рамках вопросно-ответной системы имеет свою специфику по сравнению с обработкой типичного запроса при поиске в Интернет. Как известно, запросы в глобальных информационно-поисковых системах обычно очень короткие - 2-3 слова, и по ним находятся сотни и тысячи документов. Запросы в форме вопросов обычно значительно длиннее, поэтому если требовать присутствия в документе сразу всех слов запроса, то чаще всего не будет найдено ни одного документа, что означает, что поисковая система должна автоматически определить, какие слова такого запроса должны быть отброшены или заменены.

Классическая векторная модель на основе сравнения векторов запроса и документа позволяет найти наиболее релевантные документы и по частично совпадающему запросу. слов запроса (Сегалович, Маслов, 2004). Однако при формальном выполнении пословных векторых моделей важные для ответа слова вопроса могут быть автоматически

отброшены, поэтому в некоторых современных исследованиях по вопросно-ответным системам стали использоваться не векторные модели поиска, а выполняется булевский поиск.

Использование булевской модели поиска, которая при выполнении стандартного информационного поиска, считается менее качественной, чем векторная модель, связано с тем, что при выполнении задачи сокращения формулировки запроса необходимо осуществлять дополнительный контроль, какие слова формулировки вопроса обязательно должны присутствовать в тексте ответа, а какие могут быть пропасть в тексте ответа с минимальным ущербом для релевантности ответа (Harabagiu и др., 2000; Kupiec, 1993; Hovy и др., 2001). Так, в своем докладе на семинаре ELECTRA 2005 (Vechtomova и др., 2005) известный американский исследователь в области информационного поиска Брюс Крофт отметил, что тогда как для коротких запросов хорошо работают пословные модели, то для сложных вопросов, значение которых формируется на базе отношений между концептами, важно использовать отношения между словами.

Булевское выражение обычно формируется как конъюнкция всех значимых слов формулировки вопроса. Если проводится морфологический анализ запроса или добавляются синонимы, то они объединяются в дизъюнкцию.

Например, если задан вопрос *When did Shapour Bakhtiar die?*, то может быть образовано следующее булевское выражение:

Shapour ***AND*** *Bakhtiar*

AND *(die* ***OR*** *dies* ***OR*** *died* ***OR*** *dying* ***OR*** *died* ***OR*** *death)*

Поскольку стандартной является ситуация, когда не находится документов, которые содержат все значимые слова вопроса, поэтому при обработке вопроса часто необходимо определить, какие именно слова формулировки вопроса можно отбросить, не включить в поисковый запрос без потери сути вопроса. Например, следующему вопросу *«Кто из великих целителей прошлого написал трактат "О медицине"?»* может частично соответствовать два предложения (выделены слова из исходной формулировки запроса):

1) ЦЕЛЬС (Celsus) Авл Корнелий (I в. до н. э.), древнеримский автор энциклопедических трудов «Artes» (сохранился ***трактат "О медицине"****, книги 1 - 8, с ценными сведениями по гигиене, хирургии, дерматологии)*

2) А.Е. Ферсман приводит отрывок из ***трактата*** *"Сокровищница лекарств",* ***написанного*** *арабским* ***целителем*** *около тысячи лет назад: "Ношение бирюзы…*

Первое из предложений содержит правильный ответ *ЦЕЛЬС,* во втором предложении кандидатом на ответ является *А.Е.Ферсман*, что неверно.

Для более точного определения, какие именно слова могут формулировки вопроса могут быть отброшены, обычно предлагается система модификаций, упрощающих исходное булевское выражение, после каждой из которых опять происходит обращение к поисковой системе для проверки, не появились ли релевантные документы.

Обычно используются два основных способа упрощения булевского выражения.

Во-первых, можно часть конъюнкций переводить в дизъюнкции.

Вторым способом является поочередное исключение членов конъюнкции, на основе некоторого множества эвристик, определяющих значимость членов конъюнкции.

Значимость членов конъюнкции может определяться на основе их грамматических характеристик в формулировке вопроса. Так, наиболее значимыми обычно считаются имена, фразы в кавычках, а наименее значимыми считаются глаголы.

Процесс исключения элементов из конъюнкции прекращается, когда количество документов (абзацев) в выдаче достигает заданного числа (например, 50) или до тех пор, пока не остается заданный процент слов исходной формулировки вопроса.

12.2. Роль лексических ресурсов в работе вопросно-ответных систем

В связи с длинной формулировкой естественно-языкового вопроса и частым отсутствием в самых больших текстовых коллекциях ответов, содержащих все или большинство слов формулировки вопроса, значимой становится роль лексических ресурсов, позволяющих найти ответы в тех предложениях, в которых часть слов заменена на близкие по смыслу слова.

Так, например, ответ на вопрос: *Почему электрические **батареи** быстрее **разряжаются** на **холоде**?* может быть следующим: ***Батарейки** быстрее **садятся** на **морозе**, потому что..»,* при этом ответе три слова исходного запроса были заменены на близкие по смыслу слова. Практически каждое слово вопроса имеет соответствующее слово в ответе, при этом сделано 3 лексические замены.

Таким образом, роль лексических ресурсов, онтологий, тезаурусов при обработке вопросов в вопросно-ответных системах представляется достаточно важной.

Многие современные вопросно-ответные системы используют в качестве лексического источника WordNet. В таких системах WordNet может использоваться для решения следующих задач:

- распознавания типа вопроса;
- классификации типов ответов;
- для реализации лексических и семантических замен.

В следующем разделе рассмотрим принципы работы одной из известных вопросно-ответных систем и применяемые методы использования информации из WordNet при обработке вопросов.

12.2.1. WordNet в вопросно-ответной системе Южного Методистского университета США

Одной из самых эффективных систем в вопросно-ответной дорожке конференции TREC 1999 стала вопросно-ответная система Южного Методистского университета, которая на нескольких этапах обработки вопроса и поиска ответа обращается к информации, хранимой в тезаурусе WordNet.

Лексические и семантические замены в системе осуществляются в момент сопоставления формальной структуры вопроса и ответа. Поиск документов организован на основе обработки булевских запросов, в качестве единиц поиска выступают не целые документы, а абзацы (Harabagiu и др., 2000, Moldovan и др., 1999).

На этапе обработки вопроса WordNet используется для определения типа вопроса и типа ответа. Например, если вопрос начинается со слов «what company» - этот вопрос классифицируется как вопрос об организации. При этом на некоторые типы вопросов, кандидаты-ответы могут получены непосредственно из WordNet. Например, если задан такой вопрос как «What flowers did Van Gogh paint?» (*Какие цветы рисовал Ван Гог),* то может быть извлечен список всех 470 видов цветов, упомянутых в WordNet, и использован для проверки в качестве подходящего ответа.

Для организации поиска ответов была разработана классификация ответов на вопросы конференции TREC, которая включала такие типы, как: время, дата, продукция, организация, деньги, место, язык, человек.

После этого WordNet был преобразован в таксономию ответов, релевантные синсеты были сгруппированы под своим типом ответа, а нерелевантные синсеты были удалены. В результате полученная таксономия ответов включала 8707 синсетов, 20 верхних типов. Было добавлено 129 отношений, отсутствующих в WordNet, но полезных для ответов на вопрос.

Таким образом, в значительной мере для нужд классификации вопросов и ответов на основе информации WordNet был построен новый ресурс, настроенный на вопросы, предлагаемые в рамках конференции TREC.

На основе проделанной работы была достигнута правильная идентификация типа ответа для 79% вопросов на конференции TREC-9.

Как уже указывалось, при формулировании запроса к информационной системе часто возникает необходимость удаления некоторых слов формулировки вопроса. Помимо написания (с большой буквы или нет, использование кавычек) и учета частей речи в данной системе используется иерархия WordNet.

Для этого вводится понятие специфичности, которое подсчитывается как число гипонимов за исключением конкретных имен и гипонимов с тем же главным словом. Если полученное число меньше порога (10), то оно считается специфичным, важным для вопроса и не отбрасывается. По этому правилу из запроса можно исключить слово *город (city),* и нельзя исключить слово *биохимик (biochemist).*

В данной вопросно-ответной системе WordNet совместно с серией булевских запросов используется для подбора необходимых лексических и семантических замен Например, такая замена нужна для ответа на следующий вопрос:

Вопрос: *What is the highest mountain in the world?*
(Какая самая высокая гора в мире)

Ответ: *…first African country to send an expedition to Mount Everest, the world's highest peak.*
(*… первая африканская страна послала экспедицию к горе Эверест, самому высокому пику в мире*).

При обработке формулировки запроса строится синтаксическая структура предложения, которая называется семантической формой запроса, а также создается булевское выражение, состоящее из слов запроса. Выполняется поиск и отбираются абзацы текста, удовлетворяющие запросу и содержащие, по крайней мере, одно языковое выражение, подходящее по типу к требуемому типу ответа.

После этого могут быть инициализированы три цикла расширения запроса.

Первый цикл возникает, если получено слишком мало абзацев. В таких случаях запрос расширяется на основе морфологических форм слов и номинализации глаголов (существительных, являющихся однокоренными к данному глаголу). Такой расширенный запрос опять отправляется в поисковую систему для поиска релевантных абзацев.

Второй цикл расширения возникает, если не удается провести унификацию семантических форм вопроса и ответа. В таких случаях добавляются синонимы, прямые гипонимы и гиперонимы. Расширенный булевский запрос опять отправляется в поисковую систему. Например, при обработке вопроса Who killed Martin Luther King? (Кто убил Мартина Лютера Кинга) делается цепочка вывода *Kill – killer* –гипоним-*assassin*, которая позволяет найти правильный ответ.

Третий цикл расширения возникает, если не удается доказать правильность ответа. На этом шаге делается расширение запроса на основе толкований синсетов WordNet. Например, для ответа на вопрос: Where do lobsters like to live? (Где предпочитают жить лобстеры?) удалось использовать главное слово в толковании глагола prefer – like. Был задан булевский запрос «(lobster OR lobsters) AND (like OR prefer)» и получен правильный ответ.

По материалам оценки поиска ответов на вопросы TREC было подсчитано, что при возвращении короткого 50-байтного ответа были получены следующие улучшения:

- Цикл 1 – 40%
- Цикл 2 – 52%
- Цикл 3 – 8%
- Всего – 76%

12.3. Предметные области вопросно-ответных систем

Современные вопросно-ответные системы можно подразделить на два больших класса.

Первый класс – это вопросно-ответные системы общего назначения, которые должны отвечать на широкий круг вопросов на базе сверхбольших текстовых коллекций, например, информации, хранящейся на интернет-сайтах. Величина используемых текстовых коллекций часто позволяет такой системе воспользоваться избыточностью информации, и находить такой текст, в котором ответ может быть получен системой наилучшим образом. На конференциях TREC и CLEF тестируются общие вопросно-ответные системы.

Второй класс вопросно-ответных систем – это вопросно-ответные системы, созданные для ответов на вопросы в рамках конкретных предметных областей, например, поиска информации в технической документации, в коллекции ответов на частые вопросы пользователей и другие. Такие системы располагают значительно меньшей коллекцией документов. В значительной мере для качественного поиска ответов на вопросы эти системы должны пользоваться знаниями о предметной области, хранимых, в частности, в форме онтологий и тезаурусов (Molla, Vicedo, 2006).

Примерами сфер приложений специальных вопросно-ответных систем являются правовая сфера, а также многочисленные форумы по техническим проблемам, программному обеспечению, куда обращаются пользователи со своими проблемами.

Кажется, что сужение сферы деятельности позволяет точнее настроить вопросно-ответную систему, и это действительно так.

Однако, в предметных областях возникает другая проблема: реальные вопросы пользователей не представляют собой аккуратно построенный в виде одного предложения вопрос. Чаще, вопрос реального пользователя включает предварительное описание проблемной ситуации, своих действий в этой ситуации, может содержать несколько подвопросов с отдельными вопросительными словами, а также может содержать значительно количество вводных слов, и другого рода, бессодержательных слов (*помогите, пожалуйста, поясните, help* и т.п.).

Приведем пример такого вопроса из компьютерного форума:

> *Ноутбук Compaq nx9010, месяц от роду, лицензионная русская XP Home SP1, каждые 3-4 дня загадочно исчезают точки восстановления: просто стираются соответствующие папки. Похоже, что при перезагрузке. Но не уверен. В календаре мастера восстановления - тоже исчезают. На диске свободно 27 Гб, движок стоит на все 12%. На десктопе со времён установки XP ничего подобного никогда не наблюдалось (там без сервиспака). Принятые меры: выключение и снова включение восстановления - ноль внимания. Снесение системы, установка заново - аналогично. Где копать? Машина хорошая, претензий нет. К виндам во всём остальном - тоже. Железо? Винды? Хитрые дрова? Что?*

Пример реального вопроса в правовой области: •

> *Расскажите, пожалуйста, о туристических и транзитных визах в США. Что собой представляют визы, выдаваемые супругам, и визы, связанные с обучением? Сколько стоит оформление визы?*

В работе (Jeon и др., 2005) указывается, что если современные интернет-поисковики демонстрируют достаточно высокое качество обработки 2-3 словных запросов, их способность отвечать на сложные вопросы... является явно недостаточными.

(Liddy и др., 2004) также пишут о том, что исследования вопросно-ответных систем в рамках TREC в наибольшей степени было сконцентрировано на коротких, направленных на поиск фактов, общезначимых вопросов, поиск ответов на многие из которых базируется на избытке информации в интернет. Предложенные подходы достаточно

хорошо работают для вопросов типа TREC, однако хорошие результаты не обязательно обеспечивают успех при обработке вопросов вне конференции TREC**.**

В (Liddy и др., 2004) описывается система обработки реальных вопросов в рамках более широкой области аэрокосмической индустрии. Основные компоненты вопросно-ответной системы включают: 1)обработка документов 2) модуль язык – логика (L2L) 3) поисковая машина и 4) нахождение абзацев с ответом. Когда пользователь спрашивает систему, его вопрос сначала посылается в L2L модуль, который порождает внутреннее представление вопроса и идентифицирует фокус вопроса. Поисковая машина возвращает 50 лучших документов. В качестве ответов возвращается 20 лучших абзацев.

Вопросы NASA отличаются от вопросов TREC в нескольких аспектах. Во-первых, вопросы NASA задаются в реальное время студентом, и вопрос может быть многозначным или предполагает неявное знание, которое не эксплицировано в вопросе. Реальные вопросы обычно пишутся в спешке и могут быть сформулированы с нарушением грамматической структуры или содержать орфографические ошибки.

Кроме того, вопросы NASA часто подразумевают комплексные ответы.

Например, простой вопрос «How does the shuttle fly?» («Как летает космический челнок?») является слишком широким, возможны несколько его интерпретаций.

Вопрос может не специфицировать объект, о котором спрашивается: Do welding sites yield any structural weaknesses that could be threat for failure?”

Еще один тип вопроса, который кажется простым: At what temperatures do liquid metals typically exist? Проблема в том, что для разных металлов в разных условиях эта температура – разная.

Еще один сложный тип вопросов требует сравнения двух различных элементов из двух различных документов, ответ из которых должен быть синтезирован вопросно-ответной системой.

(Liddy и др., 2004) указывают, что проблемы плохо сформулированных вопросов уже описаны библиотекарями. Плохо сформулированные вопросы делятся на следующие категории:

- слишком широкий вопрос;
- вопрос, правильный ответ на который, на самом деле, не удовлетворит пользователя;
- вопрос, который связан с недопониманием системы или предмета поиска;
- многозначный вопрос;
- вопрос, основанный на ошибочной информации.

12.4. Поиск ответов на вопрос в вопросно-ответных сервисах

Отдельным направлением в развитии вопросно-ответных систем может рассматриваться поиск уже существующих ответов в вопросо-ответных сервисах глобальных интернет-поисковиков.

Во многих странах стали популярными вопросно-ответные сервисы, когда пользователь может обратиться к сообществу пользователей или к экспертам за ответом на свой вопрос. Такие службы обычно накапливают большие объемы уже отвеченных вопросов, то есть документов типа «вопрос-ответ». При задании вопроса сервис может, прежде всего, выполнить поиск на предмет того, нет ли уже в его базе вопросно-ответных документов ответа на подобный вопрос.

Вместе с тем такие вопросы, будучи сходными по значению, могут быть сформулированы с помощью совершенно разных лексических средств. (Jeon и др., 2005) приводят такие примеры близких по содержанию вопросов, не содержащих ни одного общего слова:

1. Is downloading movies illegal?

2. Can I share a copy of a DVD online?

Поиск ответов на такие вопросы отличается от основной парадигмы современных вопросно-ответных систем тем, что нужно найти не короткий ответ на относительно ограниченный список типов вопросов, а документ, отвечающий на неограниченный список типов вопросов.

Заключение к главе 12

Разработка вопросно-ответных систем представляет собой очень интересную задачу на стыке информационного поиска и автоматической обработки текстов.

Если коллекция текстов, на которой работает конкретная вопросно-ответная система, достаточно велика и имеется большое количество вариантов представления одной и той же информации, то могут использоваться относительно «легкие» подходы, основанные на статистике и ключевых словах.

Однако для поиска ответов на сложные вопросы требуется использовать достаточно глубокую обработку вопроса и текстов, включая распознавание именованных сущностей, разрешение многозначности, синтаксический и семантический анализ, разные виды логического вывода и др. Также при обработке сложных вопросов велик потенциал использования таких ресурсов как тезаурусы и онтологии.

Глава 13. Тезаурусы в системах автоматической рубрикации текстов

Классификация/рубрикация информации (отнесение порции информации к одной или нескольким категориям из ограниченного множества) является традиционной задачей организации знаний и обмена информацией, рассматривается как одна из классических задач информационного поиска. Распространенность больших информационных коллекций делает необходимым развитие автоматических методов рубрикации.

В данной главе мы рассмотрим основные методы автоматической рубрикации, метрики оценки качества автоматической рубрикации, эксперименты по использованию тезауруса WordNet в данной задаче.

13.1. Методы автоматической рубрикации и оценка их качества

Известны две основных технологии автоматической рубрикации:

- методы, основанные на знаниях (также именуемые "инженерный подход"), при применении которых правила отнесения текстов к рубрикам строятся инженерами по знаниям в форме булевских выражений, правил продукций и т.п.
- методы на основе машинного обучения, при применении которых используется коллекция документов, предварительно отрубрицированная человеком. Алгоритм машинного обучения строит процедуру классификации документов на основе автоматического анализа заданного множества отрубрицированных текстов.

Оценка качества автоматической классификации производится путем сравнения с эталонной («правильной») классификацией набора документов, то есть на основе коллекции документов, отрубрицированных вручную.

Для оценки эффективности работы систем рубрицирования используются такие характеристики, как полнота и точность (Агеев, Кураленок, 2004).

Полнота (r – recall) - это отношение R/Q, где R - количество текстов, правильно отнесенных к некоторой рубрике, а Q - общее количество текстов, которые должны быть отнесены к этой рубрике.

Точность (p – precision) – это отношение R/L, где R - количество текстов, правильно отнесенных системой к некоторой рубрике, а L - общее количество текстов, отнесенных системой к этой рубрике.

Метрика F-мера часто используется как единая метрика, объединяющая метрики полноты и точности в одну метрику. F-мера для данного запроса (рубрики) вычисляется по формуле:

$$F = \frac{2}{\frac{1}{p} + \frac{1}{r}}$$

Также иногда используется метрика аккуратности (accuracy), которая вычисляется как отношение правильно принятых системой решений к общему числу решений. Формально

$$Аккуратность = (R+R-)/D,$$

где R - количество текстов, правильно отнесенных системой к рубрике, R- - число текстов, правильно не отнесенных системой к рубрике, D – общее число документов в коллекции. Таким образом, знаменатель не зависит от рассматриваемой рубрики.

Для оценки эффективности методов машинного обучения для задачи автоматической рубрикации текстов используются стандартные корпусы текстов, классифицированных по заданным рубрикаторам.

Считается, что наиболее эффективными, но и наиболее трудозатратными, является методы автоматического рубрицирования, основанные на знаниях. При рубрицировании текстов на основе знаний используются заранее сформированные базы знаний, в которых описываются языковые выражения, соответствующие той или иной рубрике, правила выбора между рубриками и др., (Goodman, 1991; Hayes, 1992).

Так, например, в классической работе по инженерному подходу к автоматической рубрикации текстов (Hayes, 1992) рубрики определяются на основе сопоставления каждой рубрике совокупности специальных шаблонов. Шаблон определяется как конструкция, состоящая из произвольного количества дизъюнкций, конъюнкций, отрицаний. пропусков слов и операторов необязательности. В такой конструкции могут быть также заданы части речи, способ написания (с большой или маленькой буквы), знаки препинания. Каждому такому шаблону приписан вес, определяющий, насколько сильно этот шаблон соответствует той или иной рубрике. Суммирование весов шаблонов, сопоставленных одной и той же рубрике по тексту, дает величину соответствия этой рубрики тексту. Решение о выборе рубрик для текста принимаются на основе правил, в которых учитывается, какие рубрики были обнаружены в тексте, в какой части текста встречались соответствующие шаблоны, и какой суммарный вес имеет каждая рубрика.

Результаты работы таких систем на тех текстовых потоках, для которых они проектировались, дают очень высокие оценки эффективности автоматического рубрицирования. Например, в работе (Hayes, 1992) приводятся следующие характеристики эффективности работы системы автоматического рубрицирования экономических и финансовых сообщений информационного агентства Рейтер: точность - 84%, полнота - 94%. Объем рубрикатора - 674 рубрики. В работе (Riloff, Lehnert, 1994) сообщается о реализации технологии автоматической рубрикации, достигающей 100% точности при 60% полноты.

Однако разработка систем автоматического рубрицирования, основанных на знаниях, требует больших затрат труда и часто занимает несколько человеко-лет. В таких системах базы знаний и алгоритмы жестко настроены не только на предметную область, но и на рубрикатор, размер и формат текстов. Поэтому изменение рубрикатора или необходимость рубрицирования текстов той же предметной области, но из другого источника информации влечет за собой значительные дополнительные усилия.

В настоящее время можно наблюдать всплеск научных работ, посвященных применению методов машинного обучения для автоматической рубрикации текстов. Приводятся высокие оценки результатов работы таких методов (Dumais и др., 1998; Joachims, 1998; Lewis, 2001; Yang, Liu, 1999).

Однако, как отмечалось в ряде работ (Ageev и др., 2002; Dumais и др.; 2002, Lewis, 2001; Sebastiani, 2001; Rose и др., 2002), для больших рубрикаторов - 500 и более рубрик - из-за трудности формирования качественной непротиворечивой обучающей коллекции единственно работающим подходом в настоящее время является так называемый "инженерный" подход (Wasson, 2001; Hayes, 1992; Добров, Лукашевич, 2002a), подразумевающий ручное описание смысла каждой рубрики. Например, в компании Рейтер, предоставляющей текстовые коллекции, на которых продемонстрированы многие высокие результаты технологий машинного обучения, в собственном бизнес-процессе используется технология, сочетающая работу системы автоматической рубрикации, основанной на знаниях, с последующим просмотром редакторами (Rose и др., 2002).

В следующих разделах мы подробнее опишем достигнутые результаты и проблемы разных технологий автоматической рубрикации текстов, а также на основе материалов семинара Operational Text Categorization («Реально работающая» рубрикация текстов)(Dumais и др., 2002; Lewis, Sebastiani, 2001) рассмотрим, каково состояние дел по применению технологий автоматической рубрикации в реальных организациях на реальных текстовых массивах (в противовес к исследовательским публикациям на научных конференциях и в научных журналах). В заключение будут рассмотрены

подходы к использованию тезауруса WordNet как дополнительного источника информации в методах машинного обучения.

13.2. Результаты автоматического рубрицирования на исследовательских коллекциях

Рассмотрим результаты рубрикации для наиболее популярных англоязычных и русскоязычных корпусов текстов.

13.2.1. Исследование методов рубрикации на коллекции Reuters-21578

Большое число исследований эффективности методов автоматической рубрикации проводится на популярной коллекции финансовых сообщений информационного агентства Рейтер — Reuters-21578, которая была специально создана для тестирования методов автоматической рубрикации текстов (Lewis). Для этой коллекции характерны следующие особенности:

- тексты сообщений небольшие по величине и принадлежат узкой предметной области финансовых и биржевых новостей;
- рубрикатор, включающий 135 рубрик, относительно прост, без иерархии, причем первоначально (Dumais и др., 1998; Debole, Sebastiani, 2004) для тестирования использовались лишь 10 наиболее частотных рубрик;
- присвоение рубрик проводилось с контролем качества работы экспертов. В частности, 40% из имеющихся 21578 документов не рекомендуются к использованию из-за того, что присвоение рубрик к ним признано некачественным. Оставшиеся 12902 документа помечены как «качественно отрубрицированные».

Для 10 наиболее частотных рубрик коллекции Reuters-21578 результаты применения машинного обучения весьма высоки — в среднем около 84% F-меры. Сравнительные исследования эффективности методов машинного обучения на коллекции Reuters-21578 (Dumais и др., 1998; Joachims 1998; Ageev и др., 2002) показали, что наиболее эффективным методом является метод опорных векторов SVM по сравнению с методами Байеса, ближайших соседей, Rocchio, деревьев решений C4.5, нейронных сетей, Байесовских сетей.

Дальнейшие исследования, однако, показали, что для менее частотных рубрик качество рубрикации методовм SVM значительно ниже. В среднем по 50 наиболее частотным рубрикам значение F-меры составляет 56% (Ageev, Dobrov, 2003).

В 2004 году в работе (Debole, Sebasiani, 2004) было представлено детальное исследование качества классификации коллекции Reuters-21578 в зависимости от используемого алгоритма машинного обучения, подмножества рубрик и способа усреднения оценок. Оказалось, что:

- выбор способа оценки и множества рубрик влияет на результат сильнее, чем выбор метода машинного обучения;
- качество классификации частотных рубрик значительно выше, чем низкочастотных;
- усреднение по парам документ-рубрика (микроусреднение) (Агеев, Кураленок 2004) дает более высокий результат, чем усреднение по рубрикам (макроусреднение) — этот вывод формально следует из предыдущего, так как высокочастотные рубрики дают больший вклад в микроусредненную метрику, чем макроусредненную;
- лучший результат для 90 рубрик – всего около 50% F-меры в среднем по рубрикам.

Таким образом, при детальном рассмотрении системы рубрикации, основанные на машинном обучении, имеют серьезные проблемы даже на относительно простом

рубрикаторе: 50% F-меры означает, что только половина документов получило правильные рубрики (Агеев и др., 2008)..

13.2.2. Исследование методов рубрикации на коллекции РОМИП

Среди российских исследователей способом оценки эффективности систем автоматической рубрикации текстов является участие в Российском семинаре по методам информационного поиска РОМИП (http://romip.ru). В дорожках классификации РОМИП использовались 5 коллекций документов, и три рубрикатора объемом 160-240 рубрик:

- «Сайты интернет»: NAROD.RU (~700 000 документов), DMOZ (~300 000 документов) и BY.WEB (~1 500 000 документов).
- «Нормативно-правовые документы РФ»: 2004-2006 годы — ~64 000 документов, 2007 год — ~300 000 документов.

Задачи автоматической рубрикации текстов РОМИП имеют следующие особенности:

- коллекции документов и рубрикаторы имеют широкий спектр тематики;
- значительное число рубрик;
- для оценки рубрики присваиваются документам большим количеством экспертов, зачастую — с низким контролем качества.

Участники дорожек классификации РОМИП 2003-2009 годов применяли разные методы машинного обучения: SVM (во множестве вариаций, с оптимизацией различных параметров), нейронные сети, некоторые модификации метода Rocchio и др., Полученные лучшие результаты по разным типам документам и рубрикаторам составляют 45-55% F-меры, что характерно также и для коллекции Reuters-21578.

13.3. Проблемы методов классификации текстов

Традиционно считается, что несоответствие результатов автоматической классификации ожидаемым, разумным критериям соответствия документов рубрикам вызвано несовершенством самих методов автоматической классификации. Данное предположение является основной мотивацией для разработки более совершенных моделей представления текста и методов автоматической классификации.

Однако определение основной тематики текста и выбор адекватных рубрик является сложной проблемой и для человека. Трудность ручного рубрицирования и неоднозначность выбора адекватных рубрик является проблемой, порождающей многие проблемы автоматического рубрицирования (Агеев и др., 2008).

Поэтому сначала мы рассмотрим проблемы ручного рубрицирования, а затем перейдем к описанию проблем автоматических методов рубрицирования.

13.3.1. Проблемы ручного рубрицирования

Характерными особенностями ручного рубрицирования являются:

- высокая точность рубрицирования. Как показывает практика, процент документов, в которых проставлена явно неправильная рубрика, мал;
- низкая скорость обработки документов;
- низкая полнота рубрицирования. Обычно специалисты по рубрикации проставляют рубрики, характеризующие основное содержание документа, хотя документ может быть отнесен и к ряду других рубрик. В результате получается, что при сравнении результатов рубрикации разными экспертами одних и тех же документов процент совпадения проставленных рубрик может оказаться весьма низким – 60%, то есть похожие документы могут получить достаточно разные наборы рубрик. Такая ситуация усугубляется при увеличении величины и иерархической сложности рубрикатора.

Непоследовательность ручного рубрицирования становится серьезной проблемой для настройки разного типа систем автоматического рубрицирования, поскольку затрудняется построение формальных правил отнесения документов к той или иной рубрике.

Представляется, что основными причинами непоследовательной работы экспертов-индексаторов при рубрицировании по большим классификаторам является:

1) сложность ориентации в большом классификаторе (эксперт может не знать или забыть о существовании более близкой по смыслу рубрики);
2) неуверенность эксперта, который обычно является специалистом по ограниченному кругу вопросов, при необходимости принимать точное решение по вопросам, в которых он менее компетентен (например, специалист по строительству будет менее компетентен в вопросах финансов и наоборот). В этом случае эксперт может поставить более широкую рубрику (что не очень плохо), ошибочную рубрику, или не ставить, на всякий случай, никакой рубрики;
3) сложность в принятии решения о важности/неважности побочных тем для содержания документа;
4) наличие неформализованных ограничивающих правил рубрицирования. Суть проблемы заключается в том, что ограничивающие правила рубрицирования, не связанные непосредственно с формулировкой конкретной рубрики, являются серьезной базой для субъективизма:
   - об этих правилах забывает часть экспертов,
   - для разных рубрик эти правила соблюдаются с разной степенью последовательности,
   - эти правила неизвестны пользователю, в большой степени он опирается на буквальную формулировку рубрики.

Таким образом, на наш взгляд, создание достаточно большой, последовательно отрубрицированной текстовой коллекции является серьезной организационной проблемой.

13.3.2. Проблемы методов машинного обучения

При разработке системы автоматической рубрикации, основанной на машинном обучении, необходима коллекция документов, размеченная экспертами по рубрикам. Для эффективного обучения рубрицированию по большому рубрикатору требуется большее число размеченных документов. Важной особенностью такой размеченной коллекции является то, что разметка должна быть выполнена последовательно, то есть, необходимо, чтобы эксперты применяли одни и те же принципы отнесения текстов к рубрике, чтобы похожие документы получали похожие рубрики.

Однако для многих возникающих на практике задач, где требуется автоматическая классификация текстов, коллекция классифицированных документов либо отсутствует, либо имеет недостаточный объем. В этом случае методы машинного обучения неприменимы, и затраты на создание обучающей коллекции адекватного объема весьма высоки. Кроме того, при низкой степени согласованности проставления рубрик, методы машинного обучения дают весьма низкие результаты.

Проблема создания обучающей коллекции достаточного объема и качества обостряется с увеличением количества рубрик. Распределение количества документов по рубрикам существенно неравномерно, поэтому бóльшая часть рубрик содержит весьма мало документов.

Таким образом, факторами, усложняющими или делающими невозможным применение методов машинного обучения для автоматической рубрикации текстов, являются следующие:

- множество примеров рубрикации отсутствует и не может быть создано в короткое время;
- множество примеров рубрикации существует, но при их создании отсутствовали требования к качеству, например, документы отрубрицированы их авторами, то есть людьми, которые не имеют согласованного взгляда на содержание каждой конкретной рубрики;
- множество примеров противоречиво и (или) недостаточно для большинства рубрик (очень большие классификаторы) – такая ситуация может возникнуть и при едином руководстве ручной рубрикацией;
- множество примеров для обучения взято из близкой, но другой коллекции, для которой значимое количество примеров имеется.

Кроме того, попытки использования методов рубрикации, основанных на машинном обучении, в автоматизированных режимах с участием экспертов- индексаторов сталкиваются с проблемой плохой объяснимости результатов машинного обучения, невозможностью продемонстрировать эксперту конкретные слова или словосочетания, которые привели к выбору данной рубрики.

13.3.3. Проблемы автоматического рубрицирования с использованием экспертного описания рубрик

К достоинствам методов, основанных на знаниях, относится высокая эффективность и "прозрачность" алгоритма — результаты обработки легко интерпретировать, то есть понять, почему документ был отнесен к данной рубрике. Для реализации этих методов фактор непоследовательного рубрицирования коллекции не является существенным. Основным недостатком этого класса методов является высокая трудоёмкость описания рубрик.

Проблемы автоматического рубрицирования с использованием «инженерного подхода» связаны со следующими обстоятельствами:

- для автоматической рубрикации нужно вручную создать образ рубрики, как некоторое выражение на основе слов и (или) терминов реальных текстов, неполный учет вариантов употребления слов в тексте может привести к проблемам автоматической рубрикации
- при автоматической обработке конкретных текстов могут возникнуть достаточно серьезные проблемы анализа языкового материала, контекста употребления того или иного слова, требующие привлечения обширных знаний о языке и предметной области, которые очень трудно описать в действующих программных системах автоматической рубрикации.

Так, серьезной проблемой, приводящей к появлению ложных рубрик или нехватке правильных рубрик, является *многозначность слов*, то есть употребление слова в тексте не в том значении, на которое рассчитывал эксперт, составляя образ рубрики.

Еще одной неприятной проблемой является так называемая проблема *ложной корреляции*. Ложная корреляция может возникнуть в случаях, когда для отнесения текста к рубрике необходимо присутствие в тексте двух логических элементов. Например, для рубрицирования по рубрике «Экономические реформы» необходимо присутствие в тексте двух тематических элементов – темы экономики и темы реформы. Ложная корреляция и, соответственно, неправильное отнесение текста к данной рубрике возникает в тех случаях, когда такие тематические элементы присутствуют в тексте, но не имеют отношения друг к другу. Например, такая ситуация может произойти, если в тексте речь шла о судебной реформе и были упомянуты некоторые экономические вопросы.

Сложной является и ситуация, которую можно обозначить как рубрикация по несущественному элементу. Текст отнесен к рубрике по слову или словосочетанию, которое, по сути, соответствует содержанию рубрики, но в данном тексте это опорное слово или словосочетание употреблено случайно или в каком-то специфическом

контексте, из-за чего текст становится нерелевантным рубрике. Например, текст может быть ошибочно отнесен к рубрике «Средства массовой информации» на основе следующего фрагмента: «Около 40 человек умерли во Франции в результате установившейся в стране жары… Правительство и средства массовой информации следят за ситуацией…».

Таким образом, при инженерном подходе к рубрикации после создания образов рубрик необходимо проводить несколько этапов тестирования сделанных описаний рубрик.

13.4. Системы автоматического рубрицирования при работе с реальными коллекциями

В этом разделе мы рассмотрим, как решаются проблемы автоматической рубрикации текстов в различных коммерческих компаниях, службах, функционирование которых требует автоматической рубрикации больших потоков текстовой информации.

13.4.1. Выводы семинара по Операционным системы классификации

В 2001 и 2002 годах проводились специальные семинары «Operational text categorization", целью которых был анализ ситуации в области автоматической рубрикации текстов, в том смысле, насколько различные методы автоматического рубрицирования используются в реальных условиях обработки больших текстовых массивах.

Рассмотрим подробнее основные мнения докладчиков этих семинаров.

М. Вассон из компании LexisNexis сообщил, что система автоматической рубрикации текстов работает в LexisNexis в течение многих лет. Система включает более 70000 категорий, включая рубрики и именованные сущности. Требования по точности и последовательности рубрикации очень высокие, поскольку среди пользователей много профессионалов.

Системы рубрикации в LexisNexis создавались вручную и итеративно. Чистые походы машинного обучения оказались неэффективными из-за огромного разнообразия используемых источников. Однако технологии обучения на примерах, например, в форме линейной регрессии используются в качестве вспомогательного механизма для ручного описания рубрик и взвешивания слов и групп слов. Также, при использовании технологий, основанных на знаниях, все результаты просматриваются экспертом и могут быть изменены.

Докладчик подчеркнул, что данные по эффективности того или иного метода или продукта по рубрикации текстов не всегда являются хорошими предсказателями эффективности их использования в LexisNexis.

Представители компании Kanisa описали свой опыт использования систем автоматической рубрикации текстов для поддержки интерактивных помогающих систем. Документы состоят из документов типа «часто задаваемые вопросы», руководств, информации о продукции, и их нужно классифицировать по нескольким измерениям, что означает, что должны сосуществовать несколько таксономий (до 150 таксономий, до 2000 категорий на таксономию), которые отражают различные точки зрения.

Большое количество близких по смыслу категорий и нехватка данных по многим категориям (так же как и стоимость разметки) не дают возможности использовать чистые технологии обучения по примерам. Текущий подход состоит в использовании ручного определения и описания рубрик, далее используются обучающие данные для настройки весов.

Также была представлена технология автоматического рубрицирования в рамках поисковой машины Northern Light Technology. Используется таксономия, состоящая из 16 тысяч категорий (9 уровней) для тематического рубрицирования, таксономия 150 типов

документов и др., Таксономии созданы библиотекарями и базируются на существующих таксономиях.

Для автоматической рубрикации используется совокупность подходов, включая:

- линейные классификаторы, обученные на примерах;
- классификаторы, построенные на описываемых вручную правилах,
- метаправила, которые заменяют множество более специфичных рубрик на более общую,
- ограниченную ручную рубрикацию.

Точность рубрикации считается более важной, чем полнота. 90 процентов точности необходимо для удобства пользователей. После значительной настройки система автоматической рубрикации в данной поисковой машине получает 90-95% точности по оценкам пользователей, и 60-65 % точности в соответствии с внутренними строгими оценками. Полнота оценивается как 25%, но многие пропущенные документы представляют собой очень маленькие документы, или документы, созданные исключительно для навигационных целей. Точность и полнота выше на документах, не относящихся к интернету.

Д. Льюис описал проект для Национального центра по благотворительной статистике (charitable), в котором необходимо автоматически классифицировать деятельность неправительственных организаций США. Используемая таксономия - большая и иерархическая. Представлено более 20 тысяч примеров рубрикации. Однако были существенные проблемы с данными рубрикации: качество ручной рубрикации было различным (использовался труд стажеров и профессионалов), некоторая разметка происходила от разных версий рубрикатора и т.п. Несмотря на большой объем примеров, более 70% рубрик имело менее 20 примеров.

Выводы организаторов семинара были следующими: в реальных системах широко используется обучение на примерах, однако редко работает схема: на входе данные – на выходе классифицирующая система. Ручное описание рубрик до стадии обучения или модификация классификаторов после обучения является достаточно распространенным явлением в реально работающих системах. Причины включают как необходимость учета человеческого знания о предметной области, которые могли быть и не обнаружены обучающей системой, так и проблемы отсутствия размеченных данных, стоимость разметки, непоследовательность разметки. Важная роль предметных знаний часто приводит к использованию менее эффективных систем классификации, но позволяющих вмешательство человека.

Меры эффективности, включая полноту и точность, иногда используются. При этом заказчики первоначально имеют завышенные ожидания (100% полнота и точность). Приходится проводить «обучение» по поводу пределов технологии и субъективности классификации, а также рассмотрение действительных потребностей в контексте приложения.

Кроме того, такие меры качества рубрикации как точность и полнота не отражают полной картины. В частности, некоторые ошибки системы рубрикации значительно хуже, чем другие в терминах восприятия пользователя. Приписывание категории, которая ошибочна, но близка по смыслу к правильной категории, рассматривается пользователями как менее плохая ошибка, чем присваивание полностью не соответствующей по смыслу категории.

Многие участники семинара выразили ощущение, что лучше всего использовать автоматизированные системы или автоматизацию совместно с человеческим контролем, что может уменьшить издержки и увеличить последовательность в присвоении рубрик.

13.4.2. Организация рубрицирования в Reuters

Как известно, компания Reuters уже в течение многих лет предоставляет свои отрубрицированные коллекции документов для исследований в области автоматической

рубрикации. Интересно рассмотреть, как организован процесс рубрикации документов в самой компании Reuters (Rose и др., 2004).

Компания Reuters начала применять схему автоматизации проставления категорий документов с конца 90-х годов. Применяется следующая схема классификации:

Все сообщения должны быть классифицированы по теме, региону и сектору производства. Тематические классы представляет тематическую направленность каждого документа. Они организованы в 4 иерархические группы с четырьмя верхними категориями: Corporate/Industrial, Economics, Government/Social, Markets. Всего насчитывается 126 рубрик, однако 103 рубрики применяются для рубрикации сообщений.

Для рубрикации по сектору производства используется рубрикатор из 870 рубрик, из которых 376 реально применяются к классификации документов. Имеется также 366 кодов регионов. Основным принципом рубрикации считается, что документ должен содержать хотя бы одну тематическую рубрику и хотя бы одну рубрику региона.

Первоначально использовалась система рубрикации, основанная на правилах. Однако такой подход имел следующие недостатки:

- создание правил требовало специального знания, что затрудняло добавление новых категорий и адаптацию системы к изменяющемуся выводу,
- правила не обеспечивали меры уверенности в своем выводе, что не позволяло фокусировать труд редакторов на наиболее сложных случаях, а также не позволяло обнаруживать изменения во входных документах, требующих изменений или добавлений в наборе категорий.

Текущая схема обработки документов такова. Сначала тексты проходят через систему рубрикации TIS, основанную на правилах, которая содержит правила для проставления большинства рубрик. Однако было выяснено, что проставление некоторых рубрик трудно полностью автоматизировать. Поэтому эти рубрики проставляются только вручную.

Далее автоматически проверяется соответствие проставленных рубрик правилу наличия хотя бы одной тематической рубрики и хотя бы одного кода региона. Если документ не соответствует данному правилу, то он сразу отправляется к редакторам. Если соответствует, то перемещается в специальную очередь.

В очереди каждый документ подвергается проверке хотя бы одним редактором. Кроме того, каждый месяц старший редактор берет выборку отрубрицированных документов на проверку, результаты этой проверки доводятся до сведения редакторов.

Последовательность проводимого рубрицирования можно в некоторой степени оценить, если вычислить процентное соотношение, сколько раз рубрики, проставленные данным редактором, были исправлены по отношению к числу сделанных решений:

Результаты программы автоматической рубрикации – исправлялись в 77 процентах случаев. Средний процент коррекции по людям-редакторам – 5.16%.

Для оценки последовательности рубрицирования конкретными людьми могут быть сравнены средние величины простановки рубрик людьми. В среднем, коэффициент корреляции составил – 0.968 со стандартным отклонением – 0.018. Наибольшее отклонение показывают начинающие редакторы и автоматическая система.

Таким образом, в компании Reuter для автоматической рубрикации текста и обеспечения качества и последовательности рубрикации применяется достаточно сложная организационная схема.

13.5. Использование тезаурусов в автоматической рубрикации текстов

Подходы машинного обучения для автоматической рубрикации документов используют для своего обучения набор свойств, характеристик исходного документа. Существенной составной частью этих свойств является множество слов (отличных от стоп-слов), упоминаемых в документах.

Одним из направлений в подходах, стремящихся увеличить предсказуемостную мощность обучающего метода, является использование знаний о синонимах и лексических отношениях, описанных в WordNet.

Наиболее популярным направлением исследований привлечения информации из WordNet для автоматической рубрикации текстов является дополнение пословного представления документа в виде векторной модели синсетами из WordNet, после чего применяется тот или иной метод машинного обучения.

Одной из первых работ, в которой авторы пытались интегрировать лексическую информацию из WordNet в набор характеристик для машинного обучения, была работа (de Buenaga Rodriguez и др., 1997). В этой работе было выдвинуто предположение, что обучаемая модель может быть усилена за счет применения синонимов к заголовкам категорий, используемых для рубрикации. Для этого авторы вручную выбрали подходящие синсеты из WordNet. Применялось два метода машинного обучения: метод Rocchio и метод Widrow-Hoff. Сравнение этих методов, обученных только на векторах слов, и с учетом названий рубрик и их синонимов, проводилось на коллекции Reuters-21578.

Для обоих методов интегрированное представление дало значимое улучшение, особенно значительным улучшение было на рубриках с малым числом обучающих примеров (<10).

В работе (Scott, Matwin, 1998) WordNet используется для расширения представления документа на базе всех слов документа. Разрешение лексической многозначности не производится, а берутся все синсеты слов, встретившихся в документе. Кроме того, вектор синсетов дополняется гиперонимами. Это дополнение регулируется параметром h – числом шагов обобщения. Использовался алгоритм обучения Ripper. Тестирование на нескольких коллекциях показало, что ни вектор из синсетов (h=0), ни вектор с одним уровнем обобщения не дали стабильного улучшения на разных коллекциях.

В работе (Jensen, Martinez, 2000) также используются синсеты и гиперонимы, но из всех синсетов многозначного слова выбирается наиболее частотный по коллекции синсет и соответствующий ему гипероним. Три алгоритма машинного обучения использовались для классификации текстов на базе различных комбинаций характеристик: слов, синсетов, синсетов с гиперонимами, биграмм. Эксперименты проводились на трех разных коллекциях.

Авторы делают вывод, что использование гиперонимов привело к улучшению показателей автоматической рубрикации на всех коллекциях, и, кроме того, использование гиперонимов всегда улучшает показатели по сравнению с применением только исходных синсетов.

В работе (Kehagias и др., 2001) сравнивается качество автоматической рубрикации трех алгоритмов машинного обучения, включая Naïve Bayes и k-NN классификаторы, на Брауновском корпусе, который размечен значениями WordNet. Тексты корпуса разделены на 15 категорий, и, собственно, этой классификацию и должны осуществлять классификаторы. Было отмечено, что результаты всех методов улучшились на множестве синсетов по сравнению с пословной базой обучения, однако это улучшение было слишком незначительным.

Влияние трех разных онтологических ресурсов на качество автоматической рубрикации изучалось в работе (Hotho, Bloehdorn, 2004). Исследовались такие ресурсы как WordNet, онтология тезауруса в медицинской области MESH (22 тысячи понятий с синонимами и квазисинонимами) и тезаурус по сельскохозяйственной тематике AGROVOC (17 тысяч понятий). Исследование проводилось на базе метода машинного обучения AdaBoost.

Эксперименты на коллекции Reuters для 50 рубрик с наибольшим числом положительных примеров проводились с использованием синсетов и гиперонимов

WordNet. На комбинированном представлении слова+синсеты+гиперонимы (5 уровней) было получено улучшение меры F1 на 3.29% (макроусреднение) и 2% (микроусреднение), что означает, что увеличение качества рубрикации было больше для рубрик с небольшим числом положительных примеров.

Медицинская онтология применялась для классификации текстов из коллекции OHSUMED. Здесь также использовались 50 рубрик с наибольшим числом примеров. Для обработки этой коллекции использовался также и WordNet. Разные варианты применения WordNet дали увеличение F1 меры от 2 до 7%. Относительное увеличение F1 меры на основе медицинской онтологии дало 3-5% на разных прогонах.

Также увеличение F1 меры было достигнуто на некоторых прогонах для текстов сельскохозяйственной тематики на базе тезауруса AGROVOC (до 10% F1 меры).

В работе (Mansuy, Hilderman, 2006) исследуется влияние различных типов расширения по отношениям WordNet в задаче отнесения множества документов к одной из двух рубрик. 15 пар рубрик взято из нескольких коллекций, используемых для оценки качества автоматической рубрикации: Reuters-21578, USENET, DigiTrad, Newsgroups. Для экспериментов использовались два классификатора: Naïve Bayes и SVM. Были сделаны отдельные прогоны для базовой пословной модели, расширения синонимами, расширения синонимами и гиперонимами, синонимами и гипонимами, синонимами и меронимами, синонимами и холонимами. Все расширения проводились только для существительных. В случае многозначных слов бралось наиболее частотное значение.

Авторы работы получили, что расширение на гипонимы и меронимы (части) дает устойчивое снижение показателя «аккуратности» (accuracy), все остальные расширения не показывают значимого повышения показателя по сравнению с базовым классификатором.

Таким образом, на текущий момент разные исследования расходятся в мнениях по поводу того, насколько WordNet и другие онтологические ресурсы могут улучшить качество автоматической рубрикации при использовании их в качестве источник дополнительных знаний для машинного обучения. Некоторые работы показывают небольшое улучшение качества рубрикации, другие – не выявили никакого улучшения качества или неустойчивое улучшение.

Заключение к главе 13.

При обилии информационных потоков в настоящее время автоматическая рубрикация поступающих документов является необходимым этапом обработки таких потоков.

Инженерное описание содержания рубрик сложного и большого рубрикатора является очень непростой задачей. В то же время и создание обширной и последовательно отрубрицированной коллекции как основы для машинного обучения также является делом достаточно сложным и дорогим, не всегда возможным в конкретном приложении.

При обоих подходах к автоматической рубрикации знания, собранные в тезаурусах и онтологиях, могут в какой-то степени облегчать задачу создания образа рубрики. Поэтому поиск возможностей интеграции тезаурусов и онтологий в различные методы автоматической рубрикации является важным направлением исследований.

Глава 14. Моделирование связности текста

Многие модели обработки текстов в сфере информационного поиска базируются на предположении о независимом употреблении слов (bag of words models) в связном тексте.

Между тем известно, что текст содержит множество связанных по смыслу слов, а также имеет внутреннюю иерархическую структуру.

Существует достаточно много разных приложений автоматической обработки текстов, которые могли выдавать более качественные результаты, если бы можно было бы автоматически выявлять содержательную структуру связного текста. Среди них такие приложения как автоматическое сегментирование текстов, разрешение многозначности, собственно информационный поиск, более качественное определение весов термов в документе, рубрикация текстов, автоматическое аннотирование текстов и др.

Понятие связности текста может быть рассмотрено в нескольких аспектах.

Выделяют когезию или структурную связность и когерентность текста[1]. Фактически речь идет о внутренней (структурной) и внешней (прагматической) связности. Когезией называется связь элементов текста, при которой интерпретация одних элементов текста зависит от других (Кронгауз, 2001). Когерентностью называется связность, привносимая чем-то внешним по отношению к тексту, прежде всего знаниями его адресата. На основании этих знаний адресат может конструировать определенный ожидания и достраивать связи, отсутствующие в тексте в явном виде (Гальперин, 1981, Morris, Hirst, 1991, Кронгауз, 2001, Шевченко, 2003).

С другой точки зрения выделяют глобальную и локальную связность текста. Глобальная связность текста обеспечивается тем, что у текста имеется единая тема. Локальная связность дискурса проявляется во взаимосвязи между соседними минимальными единицами текста (Ван Дейк, Кинч, 1988; van Dijk, 1985).

В следующем разделе мы рассмотрим некоторые положения теории связного текста. Не претендуя на исчерпывающий обзор подходов и моделей к анализу связного текста, мы, прежде всего, будем обращать внимание на те свойства связного текста, которые поддаются компьютерному моделированию в настоящее время.

14.1. Типы связности в связном тексте и их моделирование

14.1.1. Тематическая структура и тематическая связность текста

Определение основной темы текста является важным этапом для многих приложений информационного поиска. Понятие основной (или глобальной) темы текста связано с такими свойствами текста как тематическая связность и тематическая структура. Текст может быть формально связным посредством различных типов связности, но если у него нет единой темы, то он не может рассматриваться как текст (Севбо, 1969).

(Tomlin, 1997) указывает на различие трактовки термин *глобальная тема* у разных авторов. Этот термин может относиться к наиболее центральному участнику ситуации, описываемой в тексте. Также термин *глобальная тема* относится к тому, чему посвящен весь текст – и тогда глобальная тема скорее пропозиция, а не именная группа.

(Brown, Yule, 2001) предлагают называть главный персонаж, объект, идею термином «тематический элемент» (topic entity) и отделять понятие тематического элемента от термина *глобальная тема текста*. Именно так мы и будем употреблять термины *главная (или основная) тема документа* и *тематический элемент* или *элемент главной темы* документа.

[1] В лингвистической литературе при обсуждении проблемы связности , структуры текста употребляют термин «дискурс».

Гипотеза, лежащая в основе многих работ, заключается в том, что содержание текста может быть представлено в виде иерархической структуры пропозиций (Новиков, 1983; Шевченко, 2003; Гальперин, 1984; Van Dijk, 1985; Tomlin, 1997; Жинкин, 1958), самая верхняя пропозиция собственно и представляет собой основную тему документа, а пропозиции нижних уровней представляют собой локальные или побочные темы документа.

Ван Дейк (Van Dijk, 1985) описывает тематическую структуру текста, макроструктуру как иерархическую структуру в том смысле, что тема целого текста может быть описана как единственная макропропозиция. Тема целого текста может быть охарактеризована в терминах подтем, а подтемы в терминах еще более локальных подтем. Каждое предложение текста соответствует той или иной подтеме иерархической структуры текста.

Макроструктура текста определяет его глобальную связность. «Без такой глобальной связности, невозможно было бы осуществлять контроль за локальными связями (local connections and continuations). Предложения могут быть хорошо связанными между собой в соответствии с критериями локальной связности, но они могли бы отклониться в сторону, если бы не было глобальных ограничений на их содержание» (Van Dijk, 1985, стр.115-116).

Мы уже упоминали, что учет иерархической структуры текста имеет сложности даже при ручной обработке текстов экспертами. Так, при ручном индексировании или рубрицировании документов (см. разделы 1.5, 13.3.1.) разная трактовка побочных тем документа разными экспертами является одним из существенных факторов субъективности этих процессов.

При автоматической обработке документов важность слова или термина для содержания текста, их близость к основной теме документа оценивается с помощью специальных весов. Предполагается, что чем выше в иерархии тематической структуры упомянуто слово или термин, чем ближе они к основной теме документа, тем больше должна быть величина присвоенного веса.

Самой простой характеристикой моделирующей такой вес естественно является величина частоты употребления слова (термина) в документе, а также различные ее модификации. Более сложные модели автоматического выявления тематической структуры текста связаны с такими видами связности как риторическая связность текста и когезия, которые мы рассмотрим в следующих разделах.

14.1.2. Риторическая структура и риторическая связность текста

Каждый текст создается автором с некоторой целью. Цель написания каждого высказывания текста некоторым образом соотносится с предыдущими высказываниями и целью написания всего текста в целом. Таким образом, моделирование риторической связности состоит в том, чтобы определить, как конкретное предложение соотносится с предыдущими предложениями, что формализуется установлением некоторого набора отношений между парами предложений.

Одним из наиболее известных подходов к риторической связности текста является теория риторических структур (РСТ) (Mann, Thompson, 1987). Теория риторических структур основана на предположении о том, что любая единица текста связана хотя бы с одной другой единицей данного текста посредством некоторой осмысленной связи. Такие связи называются риторическими отношениями.

Риторические отношения могут быть симметричными и несимметричными. Примерами симметричных отношений являются такие отношения как *сравнение, отличие.* Примерами несимметричных отношений являются такие отношения *как уступка, условие, последовательность* и др., Например, в предложении «1) Иван опоздал на работу, 2) потому что он попал в пробку» между двумя клаузами имеет место риторическое

отношение причины. При несимметричном риторическом отношении главная клауза называется ядром, а зависимая клауза называется сателлитом.

Было предложено множество наборов риторических отношений, включающих в себя от нескольких отношений до нескольких сотен отношений (Hovy, Maier, 1995).

Многие подходы предполагают, что совокупность риторических отношений текста образуют структуру в виде дерева (Кибрик, 2003; Carson и др., 2003; Marcu 2000; Mann, Thompson, 1987; Cristea и др., 1998; Литвиненко, 2001). В узлах дерева размещаются типы отношений между предложениями, например, такие, как elaboration – уточнение или contrast - противопоставление (см. рис. 14.2.). Другие авторы указывают, что в риторической структуре имеет место пересечение ветвей и множественное подчинение, что требует для представления менее жестких структур (Wolf, Gibson, 2005).

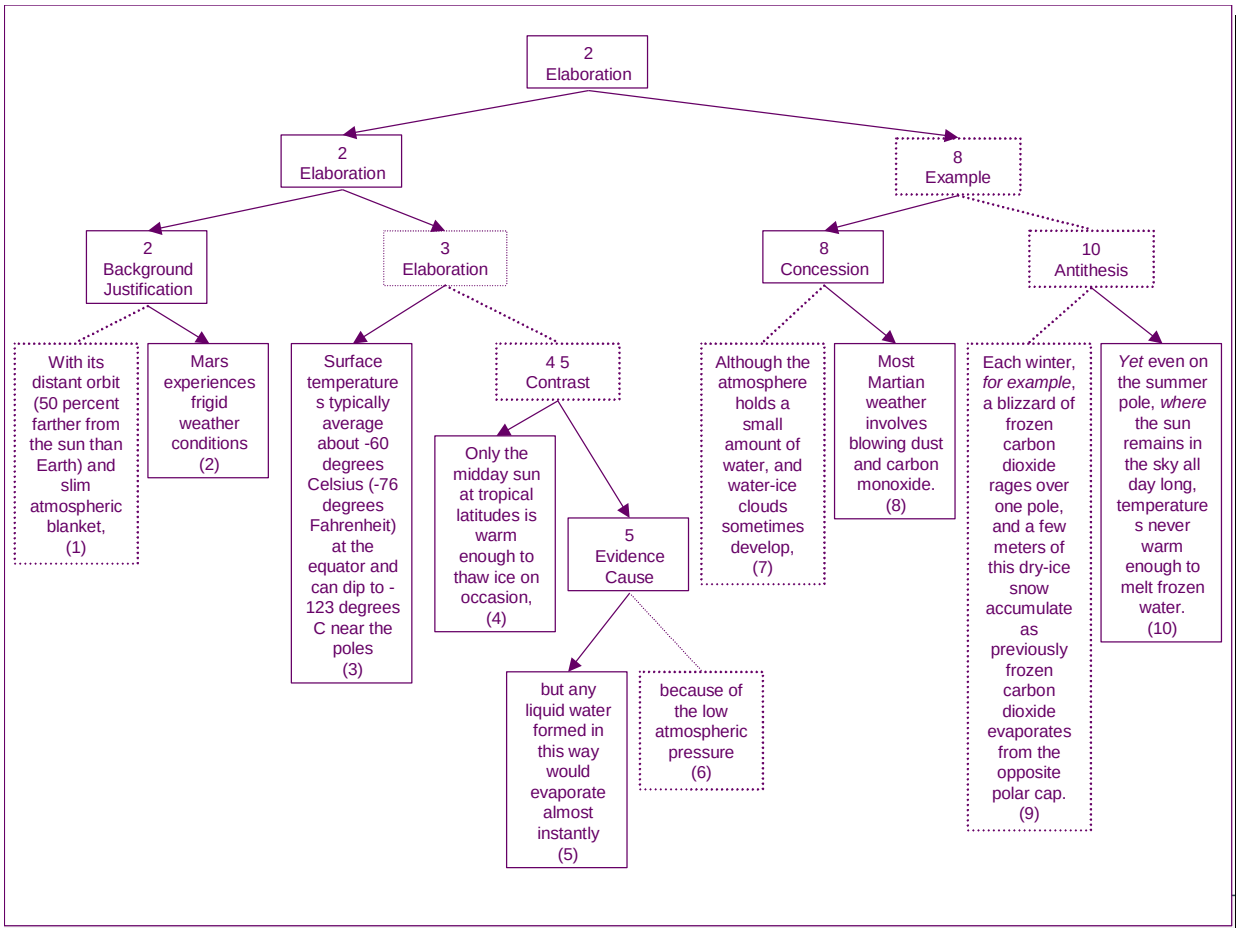

Рис.14.1. Пример построенной риторической структуры документа (Carlson и др., 2003).

Для исследования возможностей автоматического построения риторической структуры текста создаются различные корпуса текстов с разметкой риторической структуры. Первым таким корпусом с разметкой риторической структуры является англоязычный корпус, созданный на базе корпуса Penn Tree Bank (Carlson и др., 2003). Корпус включает 385 документов, для разметки используется 110 риторических отношений, которые объединены в 18 классов.

На основе этого корпуса создан статистический парсер, который позволяет построить структуру дискурса в терминах теории RST (Soricut, Marcu, 2003). Парсер выполняет две задачи. Во-первых, это разделение предложений на элементарные дискурсивные единицы (простые предложения, причастные и деепричастные обороты). Во-вторых, парсер должен построить иерархию выделенных дискурсивных единиц и установить дискурсивное отношение между ними.

Приводятся данные, что парсер разделяет предложения на дискурсивные единицы с полной и точностью порядка 82%, разметка отношений между единицами по 18 классам отношений производится с 49% F-меры, по 110 типам отношений – 45% F-меры. Согласие между экспертами при ручной разметке составляет 77% для разметки по 18 типам отношений, и 71.9% для разметки по 110 отношениям.

Разные типы текстов могут иметь разное риторическое устройство. Значительное число исследований посвящено риторическому анализу научных публикаций.

В работе (Swales, 1981) выделяет 4 основных риторических подструктуры введений в научную публикацию: указание сферы исследования, описание имеющихся результатов в данной сфере, описание собственных усилий в данной области, введение данного исследования

В работе (Teufel, Moens, 2002) рассматриваются 7 риторических отношений для научных публикаций:

Aim – формулировка цели статьи,

Textual – описание структуры статьи,

Own - описание собственных методов, результатов,

Background - общепринятое научное знание,

Contrast – Указание на недостатки других работ,

Basis – Указание на согласие с другими работами или на продолжение других работ

Other – Нейтральное описание других работ.

Предлагаемый в работе риторический анализ не является иерархическим. Авторы работы подчеркивают, что хотя они согласны с авторами теории RST, что в большинство случаев риторическая структура текста является иерархической, но вместе с тем они указывают, что имеется некоторый набор текстовых фрагментов, чей риторический статус может быть определен без анализа полной иерархической структуры текста. Другое отличие предлагаемого риторического анализа состоит в том, что определение риторического статуса текстового фрагмента производится не по отношению к соседним текстовым фрагментам, а по отношению к тексту статьи в целом.

На основе выделенных риторических отношений была разработана аннотационная схема, которая была использована для разметки 80 статей из конференций по компьютерной лингвистике.

Созданная разметка послужила основой для создания автоматической системы разметки научных публикаций на базе машинного обучения с использованием наивного Байесовского классификатора.

Для обучения были выделены следующие характеристики:

- расположение предложения в тексте, измерялось разбиением текста на 10 частей;
- расположение предложения внутри секции;
- относительная позиция предложения в абзаце;
- длина предложения;
- содержание слов заголовка;
- содержание важных слов документа, измеренных вычислением меры tf*idf;
- грамматическое время глагола;
- модальность глагола;
- наличие цитат.

Результаты работы программы были сопоставлены с ручной разметкой. Были вычислены точность, полнота и F-мера. Наиболее сложным для системы оказались категории Contrast (F-мера = 26%) и Basis (F-мера = 38%).

14.1.3. Когезия как структурная связность текста

Еще одним видом связности в тексте является когезия, представляющая собой совокупность лексических и грамматических средств для выражения связей между единицами текста. Когезия может выражаться в тексте несколькими разными способами (Halliday, Hasan, 1976; Кронгауз, 2001; Гальперин, 1984; Селезнев, 1987).

1) Когезия в тексте может осуществляться с помощью специально предназначенных для этого слов, называемых дискурсивными, которые включают чаще всего союзы и частицы, например, *Шел дождь, поэтому на улице никого не было*.

2) Частым видом когезии является лексический повтор или лексическая связность.

Авторы известной работы (Halliday, Hasan, 1976) классифицируют лексическую связность на пять категорий:

- повторение – употребляется одно и то же слово;
- синонимическое повторение;
- связность через обобщение или специализацию (родовидовые отношения);
- связность через отношения часть-целое, например, *Детский сад откроют не раньше понедельника. Еще предстоит просушить все комнаты.* (комнаты как часть детского сада);
- связность через коллокацию, сюда же относится антонимия. Такие отношения могут быть выявлены путем статистики частого совместного упоминания слов. Последние четыре вида лексической связности могут быть названы семантическим повтором.

3) Также распространенным видом когезии является использование анафорических отсылок, например, с помощью местоимений: *Иван поехал на работу. Он сел в трамвай*.

4) Еще одним поверхностным способом выражения когезии следует считать эллипсис. Эллипсисом называется пропуск в тексте подразумеваемой языковой единицы, например, *Врач прописал ему лекарство и отпустил (…) домой*.

Компьютерное моделирование всех этих явлений достаточно сложно.

Наиболее сложно автоматическое восстановление пропуска в виде эллипсиса, и нам неизвестны компьютерные приложения, которые бы в значительной мере учитывали этот вид связности.

Имеется множество работ, посвященных установлению референтов местоимений, однако явления анафоры и кореферентности значительно более разнообразны, чем данная проблема. Многие существительные и именные группы, формально сильно отличающиеся по смыслу, могут иметь одного и того же референта, например,

> *По ошибке медсестры **пациенту** был сделан укол гидроморфона - похожего на морфин по названию и действию... Свою ошибку медики осознали после пересчета наркотических средств и сразу позвонили родственникам **мужчины**.*

Дискурсивные элементы традиционно используются при автоматическом построении аннотаций, особенно аннотаций научных статей см. например, (Саломатина, Гусев, 2006; Toefel, Moens, 2002; Advances in Automatic text summarization 1999; Блюменау и др., 2002)

В настоящее время дискурсивные слова являются одним из наиболее существенных факторов при построении риторической структуры текста (см. раздел 14.1.2.). Однако проблемами использования дискурсивных единиц при построении иерархической структуры текста являются:

- их неоднозначность,
- их отсутствие во многих предложениях для некоторых типах текстов,

- сложность автоматического установления отношения к предшествующему фрагменту текста.

Из всех этих отношений лексическая связность является наименее имплицитной и может быть выявлена с помощью имеющихся лингвистических ресурсов таких, как тезаурусы.

Многие авторы указывают, что лексическая связность – это не просто связи между парами слов текста, а достаточно длинные цепочки слов, близких по смыслу.

Так, Кронгауз (Кронгауз, 2001) пишет, что средством когезии является вообще подбор тематической лексики, то есть лексики, относящейся к одному семантическому полю, и соответственно повтор в тексте интегральных признаков этого поля.

В работе (Morris, Hirst, 1991) указывается, что лексическая связность возникает не только между парами слов, но связывает между собой группы слов текстового фрагмента, посвященного одной и той же теме.

Т.В. Матвеева пишет, что тему текста представляют: первичная тематическая цепочка (прямое название предмета речи, которое обозначается чаще всего нейтральным, общеупотребительным словом) и вторичные (дополнительные), к которым относятся субституты, трансформы, синонимы, местоимения, родовые обозначения вместо видовых и т.д.

В работе (Зубов, Зубова, 2006) рассматриваются цепочки семантически связанных слов в стихотворных текстах такие как «вечер», «утро», «час», «секунда» (имеют семантический признак «время»); «мир», «даль», «расстояние»; «поезд», «путь», «движение»; «тело», «рука», «глаза»; «открытка», «поздравление», «привет» (семантический признак «расстояние»).

В работе (Hasan, 1984) рассматривается понятие «гармонии связности», посредством которого делается попытка формализовать отношения внутри предложения и между предложениями. Гармония связности базируется на цепочках когезии, в том числе лексических цепочках, и семантических отношениях, таких как *агенс, объект, инструмент*, между элементами разных цепочек, устанавливаемыми внутри предложений. R. Hasan указывает, что два языковых выражения должны рассматриваться как единицы одной цепочки, если они более чем один раз выступали в одном и том же отношении в рамках какой-либо ситуации или по отношению к какой-либо третьей сущности. Подчеркивается, что единство текста основывается на том, что «похожие вещи говорятся о похожих или тех же самых сущностях или событиях. Тексты, в которых больше сущностей участвуют в гармонии связности, рассматриваются людьми как более связные.

Алгоритмы автоматического выделения лексических цепочек будут рассмотрены в следующем разделе.

14.2. Моделирование лексической связности на основе тезаурусов

Первой работой, в которой предлагалось использовать имеющиеся тезаурусы для автоматического выявления лексической связности текста в виде лексических цепочек и были предложены алгоритмы построения лексических цепочек на основе тезауруса Роже, была работа (Morris, Hirst, 1991).

В работе указывалось, что лексическая связность возникает не только между парами слов, но связывает между собой группы слов текстового фрагмента, посвященного одной и той же теме. По определению авторов работы лексическая цепочка – это последовательность слов текста, в которой каждое следующее слова связано некоторым отношением с предшествующими словами цепочки.

Лексические цепочки не останавливаются на границах предложений и могут проходить через целый текст. Авторы работы рассматривают лексические цепочки как важный шаг на пути к построению риторической и тематической структуры дискурса.

Эксперименты с использованием тезауруса Роже проводились вручную, поскольку на тот момент не существовало электронных версий тезауруса. С появлением тезауруса WordNet подавляющее число экспериментов по построению лексических цепочек было проведено с помощью этого тезауруса.

В следующих подразделах будут рассмотрены некоторые из походов к построению лексических цепочек.

14.2.1 Подход Hirst and St-Onge

Первой опубликованной работой, которая использовала WordNet как ресурс для построения лексических цепочек, была работа (Hirst, St-Onge, 1998). Авторы предполагали использовать лексические цепочки для обнаружения малапропизмов, то есть ошибок текста, в которых ошибочно написанное слово оказывается реально существующим словом языка, что и затрудняет обнаружение ошибки (Большакова и др., 2006)

Рассмотрим, как предлагается выявлять лексическую связность текста в этой работе.

Все отношения между словами, которые могут быть индикаторами лексической связности, делятся на три группы: экстра-сильные, сильные и средней силы.

Экстра-сильные отношения устанавливаются только между буквальными повторами слов.

Сильные отношения устанавливаются в трех случаях:

- когда два слова описаны как синонимы (*human* и *person*);
- когда два слова связаны горизонтальным отношением (антонимия, подобие);
- если многословное выражение – единица WordNet – включает в себя однословное (*school – private school*).

Сильное отношение имеет меньший вес, чем экстра-сильное и больший вес, чем отношение .средней силы.

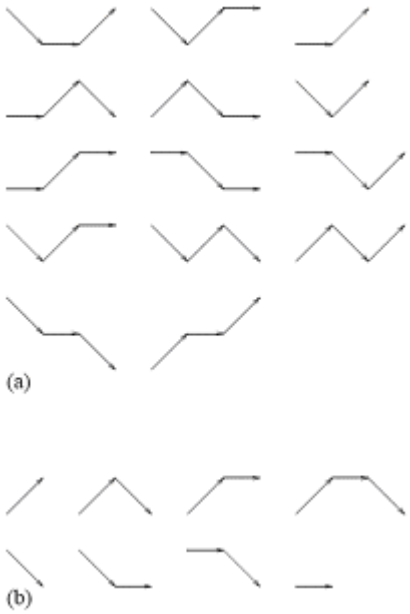

Рис. 14.2. (Hirst, St-Onge, 1998): а) запрещенные пути при построении отношений средней силы, б) разрешенные пути отношений средней силы

Отношения средней силы возникают, когда имеется путь заданной формы между понятиями, к которым относятся два слова. Максимальная длина пути – пять отношений. Не позволяется поворот пути «вниз-вверх». Разрешен только один поворот «вверх – вниз»

и два поворота пути следующего вида: «вверх – горизонтально - вниз». Таким образом, помимо повторов и синонимов рассматриваются как способные участвовать в образовании лексической связности текста:

- слова, являющиеся нижестоящими или частями одного и того же понятия от 1 до 4 уровней;
- слова, лежащие на одной иерархической линии гиперонимов, отношений целое, смешанных отношений гипероним-целое в различных вариантах (см. рис. 14.2).

Предполагается, что лексическая связность текста моделируется совокупностью лексических цепочек слов, чьи значения близки по смыслу. Для выявления этих цепочек предлагается следующий алгоритм:

1) текст просматривается пословно с начала до конца. Просматриваются только существительные.
2) первое слово создает первую лексическую цепочку.
3) для каждого следующего слова проверяется, связано ли оно какими-либо лексически-существенными связями с предшествующими словами (и соответственно, лексическими цепочками):
   - если нет, то слово образует новую цепочку;
   - если очередное слово связано только с одной лексической цепочкой, то туда оно и присоединяется;
   - если очередное слово связано с несколькими лексическими цепочками, то выбирается наиболее сильная связь. Выбирается всегда одна лексическая цепочка.
4) в процессе такого построения цепочек происходит разрешение многозначности слов, поскольку значения, по которым не было подсоединения к существующей цепочке, удаляются.

(Имеются ограничения просмотра – 7 предложений для сильных связей и 3 предложения для связей средней силы).

Авторы данной работы предполагали построить детектор малапропизмов, используя следующую гипотезу: слова, которые не формируют лексические цепочки с другими словами текста, являются потенциальными малапропизмами, поскольку они как бы не соответствуют содержанию текста. Если такое слово обнаруживается, алгоритм подыскивает слова, которые близки по написанию к данному слову и которые удается присоединить к одной из существующих лексических цепочек. Тот вариант, который сильнее всего оказался связанным с существующей лексической цепочкой, считается правильным, то есть именно тем исходным словом, в котором произошла ошибка.

Авторы протестировали свой подход на материале 500 статей Wall Street Journal, в которые были специально внесены малапропизмы, в среднем один малапропизм на 200 слов - всего 1409. Эксперименты показали точность выявления малапропизмов – 12.5. и полноту 28.7. В дальнейшем Буданицким (Budanitsky, 1999) было показано, что обнаружение малапропизмов может быть улучшено на основе более простого алгоритма, который анализировал семантическое расстояние между всеми терминами текста, а не на основании отношения с одной лексической цепочкой.

Тем не менее, работа (Hirst, St-Onge, 1997) оказала сильное влияние на попытки моделирования построения лексических цепочек и применения их в разных компьютерных приложениях при автоматической обработке связного текста.

Оценивая построенные лексические цепочки и анализируя выявленные ошибки, авторы работы отмечали, что значительная часть ошибок в установлении лексических цепочек связана со структурой описаний лексических единиц в WordNet. В частности, отмечены следующие проблемы:

1) отсутствие описаний ситуационных отношений, например, связей вида *Nasdaq – акция, больница – пациент*;
2) недостаточное количество связей между различными частями речи;
3) непоследовательность в мере семантической близости отношений WordNet. Иногда явно лексически связанные в тексте слова соединены слишком длинными путями в WordNet, например, как *steak* и *stew* и наоборот, то, что кажется несвязанным в тексте, имеет короткие пути связи в WordNet (*public - professional*);
4) кроме того, часть проблем была связана с неправильным разрешением многозначности слов.

Описанный в этом разделе алгоритм является так называемым «жадным» (greedy) алгоритмом построения лексических цепочек, поскольку построение цепочек базируется только на словах, которые встречались ранее текущего кандидата. Такой алгоритм может образовать ложные цепочки из-за многозначности слов.

Поэтому предложены также и нежадные алгоритмы построения лексических цепочек, которые предполагают построение полной картины возможных лексических отношений между кандидатами, предварительное разрешение лексической многозначности и только после этого построение лексических цепочек.

14.2.2. Алгоритм Stairmand

Подход к построению лексических цепочек, описанный в работе (Stairmand, 1996), является примером нежадного алгоритма.

Алгоритм сначала выбирает существительные-кандидаты для построения лексических цепочек. На втором этапе устанавливаются все возможные отношения между всеми значениями кандидатов. В данном алгоритме рассматриваются такие отношения как повторы, синонимы, гипонимы, гиперонимы, меронимы, холонимы и антонимы, также используются пути гиперонимических отношений, для которых длина пути не ограничивается. После установления всех возможных связей между словами, порождаются лексические кластеры. Лексические кластеры в данном алгоритме не являются взаимно исключающими, то есть одно и то же слово может относиться к разным лексическим кластерам.

На следующем шаге объединяются все лексические кластеры, относящиеся к одним и тем же значениям слов. Это дает возможность установления транзитивных отношений между словами, которые явным образом не указаны в WordNet.

Полученные лексические кластеры разбиваются на лексические цепочки так, чтобы между соседними элементами цепочки было не более 80 слов и каждая цепочка состояла не менее, чем из 3 слов. Эти цепочкам затем присваивается вес в зависимости от доли текста, которую занимает цепочка (фрагмент цепочки), и плотности цепочки (количество элементов цепочки по отношению к длине фрагмента цепочки).

Stairmand применял свой подход к экспериментам по поиску документов по запросам конференции TREC и сравнивал свой подход с результатами работы известной информационно-поисковой системой, построенной на векторной модели, SMART (Salton, 1989). Эксперименты показали, что система Stairmand находит релевантные документы лучше, если слова запроса относятся к основной теме или важной подтеме документа. Однако система SMART лучше различает между документами, которые частично относятся к теме запроса и нерелевантными документами. Кроме того, полнота работы алгоритма была очень низкой. Автор объясняет данную проблему недостаточным покрытием WordNet реальных текстов, и особенно недостаточным описанием собственных имен в WordNet.

14.2.3 Алгоритм Barzilay and Elhadad

Рассматривая методы построения лексических цепочек с использованием лексических отношений, описанных в WordNet, авторы работы (Barzilay, Elhadad, 1999) указывают на проблему неправильного построения лексических цепочек за счет того, что выбор значений многозначных слов только на основе информации о предшествующих лексических цепочек не является достаточно качественным.

Поэтому в данной работе предлагается выделять все значения слов текста и встраивать их в начатые лексические цепочки. Понятно, что число вариантов цепочек даже для небольшого текста становится слишком большим. Чтобы снизить число вариантов, в процессе обработки текста для каждой начатой цепочки оценивается ее сила, и в тот момент, когда количество вариантов превышает некоторый порог, удаляются наиболее слабые варианты цепочек.

Вес лексической цепочки определяется числом элементов цепочки и весом отношений между элементами цепочки. Для повторов и синонимов установлен вес 10, для антонимов 7, для гиперонимов и холонимов – 4. По завершении обработки текста наилучшая цепочка определяется как имеющая наибольшее число ребер графа цепочки (отношений между элементами цепочки).

В работе было проведено исследование, на основе каких параметров выделенных лексических цепочек, можно отделить более сильные лексические цепочки, то есть более хорошо отражающие основное содержание текста.

Исследовались такие параметры как:

- длина цепочки,
- распределение слов цепочки в тексте,
- плотность цепочки,
- топологию графа,
- число повторов слов в цепочках.

Было выявлено, что наилучшими показателями силы цепочки являются такие показатели как длина цепочки Length, равная числу словоупотреблений в цепочке, и индекс гомогенности Homogeneity Index, вычисляемый следующим образом:

$$Homogeneity\ Index=1-(число\ разных\ слов\ в\ цепочке)/Length$$

Авторы работы, поэкспериментировав с разными формулами вычисления силы цепочки, остановились на следующей формуле:

$$Score(Chain)= Homogeneity\ Index*Length$$

Таким образом, вес цепочки фактически равен числу повторных употреблений слов в этой цепочке, и тем самым имеет прямую аналогию с частотой употребления слова в тексте. Снижение веса для цепочек со слишком разнообразным составом, видимо, позволяет снизить ошибки формирования лексических цепочек.

Для получения статуса сильной цепочки, которая будет использоваться в дальнейшем анализе, необходимо, чтобы для веса цепочки выполнялось следующее соотношение:

$$Score(Chain) > Average\ (Scores)+2*\ StandardDeviation(Scores)$$

Попытка тестирования качества таких лексических цепочек была выполнена в работе (Silber, McCoy, 2002). Предлагаемый метод тестирования основан на использовании аннотаций, созданных людьми.

Предполагается, что если лексические цепочки являются хорошим промежуточным представлением для отражения содержания документа, то можно ожидать, что существительные в таких аннотациях используются в том же самом смысле, что и существительные, сгруппированные в сильные лексические цепочки. Более того, сильные цепочки должны быть достаточно хорошо представлены в ручных аннотациях.

Для оценки использовался корпус из 10 научных статей, которые снабжены авторской аннотацией, а также 14 глав из 10 университетских учебников, для которых также имеются аннотации.

Для каждого документа в корпусе, документ и его аннотация анализировались отдельно, и для каждого из них была построены лексические цепочки. Синсеты (значения) существительных в каждой из цепочек в документе и аннотации были сопоставлены между собой.

Были вычислены следующие метрики:

- число и процент сильных цепочек из оригинального текста, представленные в аннотации, то есть процент слов из сильных цепочек, представленных в аннотации в том же смысле, что и в сильной цепочке документа – (аналогично полноте),
- число и процент сильных цепочек из аннотации, представленных в документе (аналогично точности).

Авторы получили следующие результаты:

- 79.12% существительных из сильных цепочек в документе содержатся в аннотации,
- 80.83% существительных из сильных цепочек аннотации содержатся в документе.

14.2.4 Лексические цепочки: использование частотных ассоциаций

Многие исследователи, исследующие лексическую связность на базе WordNet, отмечали, что серьезной проблемой является недостаточность лексических знаний, описанных в WordNet. В работах (Stokes и др., 2000; Stokes и др., 2004) сделаны усилия для того, чтобы преодолеть эту проблему.

В данных работах предлагается дополнительно использовать следующую информацию:

- статистические ассоциативные связи слов,
- лексические цепочки для собственных имен.

Авторы подчеркивают, что одним из важных назначений учета статистических ассоциаций слов является преодоление уже упоминавшейся теннисной проблемы, то есть проблемы, что в WordNet, слова, относящиеся к одной и той же тематической области, могут располагаться достаточно далеко по иерархии путей. Также авторы отмечают проблему нехватки такой информации, как некоторых значений, а также многословных сочетаний.

Для построения ассоциаций слов авторы использовали текстовый корпус конференции TDT (http://projects.ldc.upenn.edu/TDT/), извлекли из него все существительные и словосочетания WordNet и собрали информацию о совместной встречаемости существительных в пределах текстового окна, состоящего из четырех существительных. Окно было также ограничено границами предложения и документа.

Отфильтровав наименее частотные ассоциации, авторы оставили в работе 25032 пар, что соответствует 3566 существительным, имеющим в среднем 7 ассоциирующихся слов.

Так, например, были получены следующие биграммы:

AIDS: virus 0993, HIV 0.951, patient 0.897, research 0.806, disease 0, 801, infection 0.78 и т.д.

Понятно, что существенной проблемой совмещения построения лексических цепочек на основе WordNet и статистических биграмм, является то, что для биграмм неизвестны точные значения слов, для которых существуют такие ассоциации и, следовательно, статистическая связь может быть применена не к тому значению в тексте, что приведет к неправильному включению элементов в цепочку.

Авторы данной работы применяют систему связей в лексической цепочке предложенных в работе (Hirst, St-Onge, 1998): сверхсильные связи, сильные связи, связи средней силы.

Ассоциативные связи между словами, полученные на основе статистических критериев считаются самым слабым видом отношений между словами и применяются, если более сильных связей не найдено.

Например, для текста, посвященного премьере фильма об убийстве журналистки, получились следующие лексические цепочки (в скобках указывается элемент из цепочки, с которым связан очередной элемент и сила связи):

Film – movie (Film, strong) – premiere (film, medium) – subject_matter (film, strong) – actress (movie, strong) – picture (film, strong) – actor (actress, strong) – approval (subject_matter, strong) – story (subject_matter, medium) – director (actor, Statistical) – tribute (approval, strong)

Investigation – murder (investigation, strong) – killing (murder, strong) – victim (killing, statistical) - crime (victim, statistical) – life (murder, medium) – loss (life, statistical) – murderer (victim, medium)

Для именованных объектов, не входящих в состав WordNet, также предложена система отношений разной силы:

- отношение полного совпадения: *Helmut_Kohl - Helmut_Kohl*
- частичное пословное совпадение: *Hubble_Telescope – Space_Telescope_Science_Institute,*
- частично совпадение по фрагменту слова: *National_Caver's_Association – Irish_Cave_Rescue_Organisation.*

14.2.5. Лексические цепочки: использование информационно-поисковых тезаурусов

О. Медельян (Medelyan, 2007) предлагает использовать недостающее в WordNet ситуативное знание на основе информационно-поискового тезауруса (в работе используется тезаурус AgroVoc). Она указывает, что наиболее известные алгоритмы построения лексических цепочек слишком зависят от порядка слов в тексте, что не соответствует реальной ситуации, когда одно и то же содержание может быть выражено с помощью по-разному упорядоченных последовательностей предложений. Поэтому в работе предлагается сначала собрать цепочки-кандидаты со всего текста, а затем, получив целостную картину лексических цепочек-кандидатов текста, применить разбиение получившегося графа на наиболее связанные фрагменты.

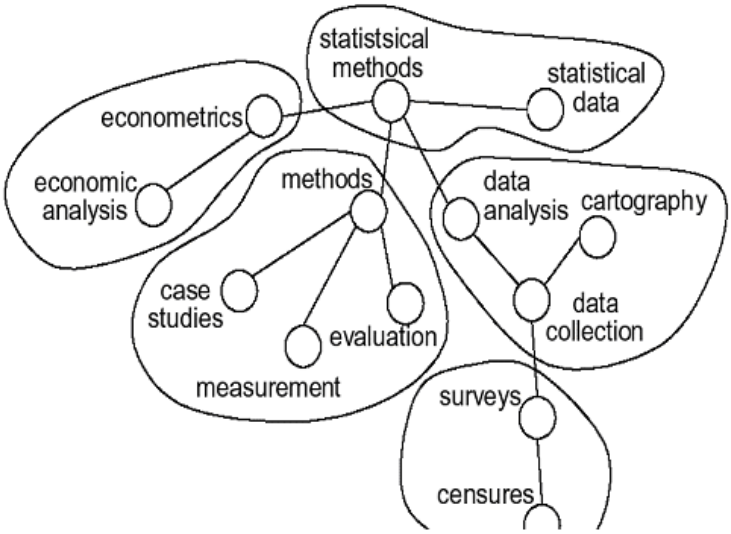

Рис. 14.3. Разбиение графа на лексические цепочке в работе (Medelyan, 2007)

Лексическая цепочка определяется как граф $G = (V, E)$ с узлами $v_i \in V$, представляющими термины тезауруса и дугами графа $(v_i, v_j, w_{ij}) \in E$, описывающими отношения между терминами, где wij – это вес, выражающий силу отношения между терминами.

Такой граф строится следующим образом. Как и в предшествующих алгоритмах, цепочки-кандидаты строятся по порядку движения текста. Различия возникают в том случае, когда очередной термин может быть отнесен к более чем одной лексической цепочке. Тогда эти цепочки склеиваются в единую цепочку, а составные части этой единой цепочки удаляются из списка цепочек.

Получается граф достаточно сложной формы (см. рис. 14.3). Этот граф с помощью алгоритмов кластеризации графа разбивается на фрагменты так, чтобы между каждым элементом подграфа было расстояние не более 3 шагов, тем самым получаются сильно связанные между собой подграфы, которые и предлагается считать лексическими цепочками.

14.2.3. Лексические цепочки в задачах автоматической обработки текстов. Автоматическое аннотирование.

Автоматически выявляемые лексические цепочки используются при решении разнообразных прикладных задач:

- автоматической сегментации текстов (Min-Yen и др., 1998, Mochizuki и др., 2000) ,
- автоматического разрешения многозначности (Galley, McKeown, 2003);
- информационный поиск (Stairmand, 1996);
- автоматическое аннотирование текстов (Barzilay, Elhadad, 1999; Silber McCoy, 2000; Brunn и др., 2001; Stokes, 2004; Reeve и др., 2006);
- распознавание тем текстов (Carthy, Smeaton, 2000),
- построение вопросно-ответных систем (Moldovan, Novischi, 2002) и др.

Одним из самых популярных применений лексических цепочек является автоматическое аннотирование текстов. В следующих разделах мы подробнее рассмотрим особенности этой задачи автоматической обработки текстов, методы ее решения, а также алгоритмы использования лексических цепочек в этой задаче.

14.2.3.1 Виды и методы автоматического аннотирования документов

Современные объемы информации требуют автоматизации процесса краткого изложения отдельных текстов или группы текстов на одну и ту же тематику.

Основной целью составления аннотации является изложение важной информации их исходного текста (текстов) с помощью меньшего количества предложений.

Существуют разные виды аннотаций (Radev и др., 2002). Индикативная аннотация должна передать информацию об общем содержании документа, не сообщая деталей. Информативная аннотация должна сохранить информационную ценность исходного сообщения. Тематически-ориентированные аннотации должны отразить информацию из текста, соответствующую теме, интересующей пользователя, так называемые аннотации по запросу (query-based summaries). Экстрактивная аннотация состоит из фрагментов (предложений) исходного текста, в то время как аннотации в форме абстракта порождаются на основе извлеченного содержания.

Несмотря на существование ряда исследований по созданию аннотаций-абстрактов, основные исследования в настоящее время сосредоточены в сфере экстрактивных аннотаций. Далее мы будем говорить только об экстрактивных аннотациях.

Большинство систем аннотирования используют предложения исходного текста в качестве единиц порождаемой аннотации. Для предложения на основе выделенных характеристик подсчитываются веса, из предложений с наибольшими весами формируются аннотации.

Характеристики, на основе которых может составляться вес предложения, могут быть следующими:

- позиция в тексте,
- частотность слов,
- наличие ключевых фраз вида «Необходимо подчеркнуть»,
- длина предложения,
- именованные сущности,
- повторяемость слов и др.,

Современные подходы используют методы машинного обучения для учета возможных характеристик предложений, включаемых в аннотации (Li и др., 2006).

Одним из относительно новых направлений составления аннотаций является составление аннотации на основе многих документов – обзорного реферата. При составлении такого обзорного реферата необходимо решать такие задачи как:

- борьба с избыточностью информации,
- идентификация важных различий между документами,
- обеспечение тематической связности текста, что усложняется тем, что предложения могут браться из разных источников.

Обзорные рефераты могут делаться для различных наборов документов (Nenkova, Louis, 2008), например, таких как документы, описывающие конкретное событие, документы, обсуждающие одну и ту же тему, документы, обсуждающие биографию одного и того же человека, документы, обсуждающие множество событий одного и того же типа, например, конкретные примеры насилия, документы, представляющие мнения разных сторон на общую тему (например, мнение сената, конгресса, общественности на тему миграции).

Для определения избыточности в порождаемых аннотациях используются различные меры сходства между предложениями. Одним из распространенных подходов является предварительная кластеризация – выделение близких по содержанию кластеров предложений (Radev и др., 2000). Другим подходом к оценке избыточности является сравнение предложений-кандидатов с предложениями, уже попавшими в аннотацию, и

оценка новой (непохожей) информации, например, так называемый подход Maximal Marginal Relevance (MMR) (Carbonell, Goldstein, 1998)

Обеспечение связности изложения является сложной проблемой, поскольку требует реального понимания содержания фрагментов и знаний о структуре связного текста. Многие подходы ограничиваются учетом времени и порядка предложений в тексте (фрагмента из более раннего текста размещаются сначала, в порядке следования в тексте).

14.2.3.2. Оценка качества аннотаций

Оценка качества автоматически порождаемых аннотаций является сложной процедурой, поскольку даже для относительно содержательно простых документов как новостные сообщения, согласие между экспертами может составлять всего 60%.

Оценка качества аннотации может быть внутренней и внешней.

Внутренняя (intrinsic) оценка аннотаций связана с оценкой качества аннотации как собственно текста, сравнения ее с исходным текстом или с аннотациями, порожденными людьми.

При оценке качества аннотации экспертам могут быть заданы такие вопросы с оценкой по 5 бальной шкале:

- является ли предложения аннотации грамматически правильными,
- является ли текст аннотации связным,
- содержит ли аннотация все основные обсуждаемые темы исходного документа (документов) и др.

При оценке аннотаций по многим документам – обзорных рефератов в рамках конференции DUC, эксперты помимо ответа на конкретные вопросы по качеству аннотаций должны проставить и две общие оценки аннотации (Dang, 2006).

Во-первых, эксперты должны были оценить соответствие содержанию кластера, то есть насколько реферат отображает необходимую для пользователя, формировавшего запрос, информацию. При этом не бралась в расчет читабельность реферата, до тех пор, пока она не влияла на объем покрытой в реферате информации.

Во-вторых, эксперты ставили общую оценку аннотации, которая должна отражать как содержательную часть реферата, так и его читабельность. При определении уровня общего соответствия оценщикам не предоставляли доступ к ранее оцененным характеристикам читабельности и соответствия содержанию, вместо этого они должны были «сходу» дать свою оценку. Многие из оценщиков посчитали для себя полезным выставлять уровень общего соответствия исходя из ответа на вопрос: «Сколько я бы заплатил за этот обзорный реферат?». В итоге, плохая читабельность систем занижала их оценку общего соответствия, по сравнению с соответствием содержанию. В то же время, рефераты с высоким показателем читабельности, получали оценки за общее соответствие выше, по сравнению с оценками за соответствие содержания.

Внешняя (extrinsic) оценка аннотации производится в специально поставленной задаче, в которой выясняется, может ли аннотация заменить исходный текст. Такими задачами могут быть классификация документов по его аннотации, или ответы на вопросы по содержанию документа на основе его аннотации.

Один из первых масштабных экспериментов по внешней оценке аннотаций был осуществлен в рамках конференции SUMMAC (Tipster SUMMAC, 1998). В оценку было включено три задачи:

- задача классификации (насколько качество классификации документа по аннотации сравнимо с качеством классификации полного документа),
- ad hoc задача – эксперты должны определить, насколько текст соответствует запросу по аннотации,
- вопросно-ответная задача – эксперты должны ответить на вопросы по основному содержанию документа на основании его аннотации. (см. также п. 22.1)

Важным элементом современной оценки аннотаций является получение автоматических оценок качества аннотаций за счет автоматического сравнения порожденной аннотации с аннотациями, написанными людьми.

В рамках конференции DUC используется метод автоматической оценки качества аннотаций ROUGE (Recall Oriented Understudy for Gisting Evaluation), который подсчитывает число перекрытия (n-граммы слов) автоматической аннотации с «идеальными» аннотациями, составленными людьми (Lin, 2004) (см. также раздел 22.3.4).

14.2.3.3. Использование лексических цепочек для порождения аннотаций

Применение лексических цепочек для автоматического аннотирования позволяет решать несколько задач, возникающих в процессе автоматического аннотирования документов. Они помогают выявлять основную тему документа, и, кроме того, являются дополнительным фактором обеспечения связности создаваемой аннотации. Рассмотрим подробнее некоторые из предлагаемых подходов по использованию лексических цепочек для порождения разного вида аннотаций.

Одной из первых работ, описывающих применение алгоритмов выявления лексических цепочек, к автоматическому аннотированию текстов, была работа (Barzilay, Elhadad, 1999). Как указывалось в разделе 14.2.3, в этой работе был реализован алгоритм построения лексических цепочек на основе WordNet, а также были сделаны усилия, чтобы разобраться, какими свойствами должны обладать так называемые сильные лексические цепочки, то есть цепочки, которые наилучшим образом отражают содержание текста.

Идея применения лексических цепочек для автоматического аннотирования документов состоит в том, что если цепочка отражает важные темы документа, то необходимо для аннотации выбирать предложения, в которых встречались элементы этих важных цепочек. Конкретный алгоритм был следующим: для каждой цепочки выбирается ее представители – элементы цепочки, частотность которых превышает среднюю частотность элементов цепочки. Для составления аннотации берутся первые по порядку текста предложения, которые содержат элемент-представитель для каждой из сильных лексических цепочек. Таким образом, каждая сильная лексическая цепочка представлена, по крайней мере, одним предложением в аннотации.

Для оценки качества предложенного метода автоматического аннотирования было выбрано 40 новостных текстов, каждый в среднем по 30 предложений. Пять ассессоров должны были сделать два вида аннотаций для этих текстов длиной 10% и 20% от длины исходного текста.

На основе этих аннотаций была сформирована «идеальная» аннотация, которая содержала те предложения, которые были выбраны большинством ассессоров.

Автоматически порождаемые аннотации были сравнены с аннотациями, порожденный суммаризатором Microsoft Word (см таб. 14.1), посредством вычисления показателей полноты и точности:

	Microsoft		Lexical chain	
	Prec	Recall	Prec	Recall
10%	33	37	61	67
20%	32	39	47	64

Таблица 14.1. Результаты сравнения аннотаций, построенных на основе лексических цепочек с сумаризатором Microsoft Word.

В таблице (14.1) видно, что аннотации, построенные на базе лексических цепочек, в значительной степени ближе к аннотациям, порождаемым людьми.

В работе (Doran и др., 2004) алгоритм автоматического аннотирования Barzilay&Elhadad тестируется на основе внешней задачи, а именно в рамках задачи автоматического нахождения похожих текстов. Предполагается, что если автоматическая аннотация хорошо отражает основное содержание документа, то аннотации похожих документов будут также похожи, а аннотации разных документов также будут различаться.

Подход Barzilay&Elhadad сравнивался с тремя базовыми подходами: случайным выбором предложения, выбором блока первых предложений, выбором предложений на основе метрики tf.idf. Тестирование проводилось для разных коэффициентов сжатия от 10% аннотации до 60% аннотации. Подход Barzilay&Elhadad уступил базовым подходам только 1 раз: при 10% аннотации лучшими были аннотации, построенные на основе первых предложений исходных текстов.

В работе (Brunn и др. 2001), аннотации строятся на основе другого рода лексических цепочек. Используется «жадный алгоритм» типа (Hirst, St-Onge 1998), который имеет следующие дополнения:

- длина пути между элементами цепочки не более 2 отношений,
- такие отношения должны быть между всеми элементами цепочки.

Наиболее значительное отличие данного подхода от других подходов заключается в том, что делается дополнительный предварительный шаг по отбору существительных – кандидатов для включения в лексические цепочки. В большинстве подходов предварительная стадия построения лексических цепочек включает морфологический анализ и отбрасывание стоп-слов, которые часто дают ошибочные или малоинформативные лексические цепочки. В данной работе проверяется предположение о том, что существительные, находящиеся в подчинительных предложениях, менее информативны, и их можно не включать в процесс построения лексических цепочек.

В работе (Li и др., 2007) исследуется возможность использования лексических цепочек для построения обзорного реферата по запросу. Построение лексических цепочек производится для получения наиболее сильных цепочек, в терминах работы (Barzilay, Elhadad, 1999).

Построение лексических цепочек в этой работе проводится в два этапа. На первом этапе строятся отдельные лексические цепочки, на втором этапе построенные лексические цепочки корректируются.

Построение цепочек происходит, начиная с самых частотных синсетов. В начатую лексическую цепочку вносятся все синсеты, которые могут быть отнесены к синсетам цепочки по принятой мере близости. Этот процесс проводится для наиболее частотной половины из всех синсетов-кандидатов, для которых могут быть построены лексические цепочки. После построения цепочек определяются наиболее сильные цепочки.

На втором этапе сильные цепочки, содержащие хотя бы одно общее слово, сливаются в единую лексическую цепочку.

Для порождения аннотации по запросу из набора документов извлекаются предложения, имеющие наиболее высокий вес по следующей формуле:

$$Score = \alpha P(chain) + \beta P(queries) + \gamma P(nameentity),$$

где P(chain) – это сумма весов лексических цепочек, участники которых были упомянуты в предложениях-кандидатах, P(queries) – это сумма совпадающих слов в предложении-кандидате и формулировке темы запроса, P(nameentity) – это число именованных сущностей, упомянутых как в предложении-кандитате, так и формулировке запроса. В экспериментах были подобраны коэффициенты α=0.2, β=0.3, γ=05.

Заключение к главе 14

Исследователи связного текста выделяют несколько взаимосвязанных между собой видов связности текста. Среди всех видов связности лексическая связность наилучшим

образом поддается моделированию на основе информации, описанной в тезаурусах и онтологиях.

При моделировании лексической связности существенным является не установление пар лексически связанных слов, а цепочек близких по смыслу слов, так называемых лексических цепочек. Получение таких лексических цепочек важно не само по себе, а как шаг к выявлению тематической структуры текста, то есть определению основной темы и побочных тем (подтем) документа.

Алгоритмы, основанные на лексических цепочках, использовались при решении различных задач автоматической обработки текстов. Особенно популярны методы, основанные на лексических цепочках, в задаче автоматического порождения аннотаций для одного и многих документов, поскольку именно в этой задаче особенно важно обеспечить связность порождаемой аннотации. Также лексические цепочки в автоматическом аннотировании помогают снизить излишние повторы в порождаемых аннотациях.

ЧАСТЬ 4. ТЕЗАУРУС РУТЕЗ

Глава 15. Тезаурус РуТез

15.1. Основные принципы разработки лингвистических ресурсов для приложений информационного поиска

Современные приложения информационного поиска работают в широких предметных областях. Если мы хотим создавать лингвистические и терминологические ресурсы для использования в приложениях информационного поиска, то эти ресурсы должны иметь очень широкое покрытие используемой лексики и также иметь возможность применяться в автоматических режимах обработки документов и запросов.

В предыдущих разделах мы рассмотрели различные лингвистические и онтологические ресурсы. Все из них имеют некоторые проблемы при использовании их как ресурсов в рамках решения задач информационного поиска.

Традиционные информационно-поисковые тезаурусы создавались как инструмент для помощи человеку, их структура направлена на предоставление удобств индексатору (удаление слишком конкретных терминов, удаление близких по смыслу терминов, добавление комментариев по употреблению тех или иных дескрипторов). В связи с этим при использовании традиционных информационно-поисковых тезаурусов в автоматической обработке текстовой информации возникают существенные проблемы. В литературе предлагается использовать методы машинного обучения для проставления дескрипторов тезауруса по уже проиндексированному людьми множеству документов, создание которого представляется чрезвычайно дорогой процедурой.

Формальные онтологии, одним из провозглашаемых принципов которых является независимость от конкретного языка, сложно использовать в автоматической обработке текстов для приложений информационного поиска, поскольку для этого единицы формальной онтологии необходимо связать с единицами конкретного естественного языка. Кроме того, стремление к четкой формализации отношений между понятиями к формальной онтологии чрезвычайно трудно соблюсти в ситуации, когда необходимо создавать сверхбольшие ресурсы, и, кроме того, приводит к проблемам при установлении связей «понятие - языковое выражение».

Проблема использования онтологий с большим количеством отношений подобно MikroKosmos или CYC связана с двумя проблемами. Во-первых, для новой предметной области создать такой ресурс чрезвычайно сложно, дорого и требует много времени. Во-вторых, большое количество отношений в таких ресурсах может сослужить и плохую службу при обработке текстов, поскольку в конкретном контексте может быть применима лишь часть описанных отношений, остальные отношения могут приводить к лишним или ложным выводам. При этом автоматически оценить применимость отношений по контексту чрезвычайно сложно.

Ресурсы типа WordNet создаются для описания лексики языка в соответствии с лингвистическими традициями. Но любая информационная система имеет дела не только с общей лексикой, но и с конкретными предметными областями и их терминологиями. Анализируя попытки создать терминологические ресурсы на основе WordNet (см. разделы 3.3.7, 3.4), следует отметить, что структура WordNet не приспособлена для описания терминологий. Раздельное описание частей речи, слишком большой набор несвязанных между собой значений, недостаточная проработанность принципов включения многословных выражений, - все это приводит к проблемам разработки и использования терминологических ресурсов, созданных на базе модели WordNet.

Вместе с тем, в каждом из этих типов ресурсов есть те качества, которые должны присутствовать в большом лингвистическом ресурсе для информационно-поисковых приложений, и, таким образом, мы считаем, что ресурс для автоматической обработки

текстов в информационно-поисковых приложениях в широких предметных областях должен сочетать принципы различных традиций и методологий:

- методологии разработки традиционных информационно-поисковых тезаурусов;
- методологии разработки лингвистических ресурсов типа WordNet (Принстонский университет);
- методологии созданий формальных онтологий.

Поясним необходимость использования этих методологий и их особенности подробнее.

Поскольку важно уметь описывать терминологию широких предметных областей, то необходимо использовать опыт разработки информационно-поисковых тезаурусов, а именно:

- информационно-поисковый контекст;
- единицы ресурса создаются на основе значений терминов;
- описание большого числа многословных выражений, принципы включения (невключения) многословных единиц;
- небольшой набор отношений между понятийными единицами.

Так как предполагается использовать лингвистический ресурс в автоматическом режиме обработки текстов, то необходимо использовать методологию разработки лексических ресурсов типа WordNet, в которой важны следующие положения:

- понятийные единицы создаются на основе значений реально существующих языковых выражений;
- многоступенчатое иерархическое построение лексико-терминологической системы понятий;
- принципы описания значений многозначных слов и выражений.

Из методологии разработки формальных онтологий важны следующие положения:

- разработка лингвистической онтологии как иерархической системы понятий;
- использование для описания отношений формально определяемых отношений с формальными свойствами;
- в качестве аксиом (правил вывода) использование свойств транзитивности и наследования таксономических отношений и транзитивности отношений онтологической зависимости.

Именно эти принципы положены в основу разработки нескольких больших ресурсов для информационного поиска: Общественно-политического тезауруса, Тезауруса русского языка РуТез (Loukachevitch, Dobrov, 2002; Лукашевич, Добров 2002), Онтологии по Естественным наукам и технологиям ОЕНТ (Добров и др., 2005; Добров, Лукашевич, 2006) и ряда других.

Вышеперечисленные ресурсы имеют одинаковую структуру. Они являются онтологиями, поскольку описывают понятия внешнего мира и отношения между ними, которые устанавливаются в соответствии с требованием правомочности расширения запроса по иерархии связей при информационном поиске.

Эти ресурсы принадлежат к особому классу онтологий, так называемым лингвистическим онтологиям (см. раздел 4.4), поскольку введение понятий в значительной мере мотивируется значениями языковых единиц, относящихся к предметной области ресурса.

В то же время они являются тезаурусами, поскольку каждое понятие связано с набором языковых выражений (слов, терминов, словосочетаний), которыми это понятие может быть выражено в тексте, - такой набор текстовых входов понятий необходим для использования онтологий для автоматической обработки текстов.

Основным лингвистическим ресурсом, разработанным на основе упомянутых принципов, является тезаурус русского языка РуТез, и в следующих разделах будут подробно рассмотрены структура и характеристики этого ресурса.

15.2. Тезаурус РуТез: Общая структура

Тезаурус РуТез – это иерархическая сеть понятий. Каждое понятие имеет имя.

Для сопоставления с текстом каждое понятие снабжается набором текстовых выражений («*текстовых входов*», «*терминов*»), значения которых соответствует данному понятию. В качестве таких текстовых входов могут выступать однословные существительные, прилагательные, глаголы, именные и глагольные группы. Количество таких текстовых входов понятий может быть достаточно велико, например, превышать 20 единиц. При вводе нового понятия делаются специальные усилия, чтобы максимально подробно перечислить его возможные текстовые входы.

Каждое понятие связывается отношениями с другими понятиями тезауруса РуТез. Набор отношений тезауруса специально подобран для эффективной работы в информационно-поисковых приложениях.

Особенностью тезауруса РуТез (как и других тезаурусов) является то, что понятия не имеют внутренней структуры в виде атрибутов (фреймовых элементов), то есть свойства понятий описываются только посредством отношений с другими понятиями.

Как уже указывалось, подавляющее число понятий тезауруса РуТез базируются на значениях существующих языковых выражений. В отличие от ресурсов типа WordNet такими выражениями могут не только общеупотребительная лексика и лексикализованные выражения, но и термины в широкой предметной области современной жизни общества, которую мы называем Общественно-политической областью.

Вопросы соотношения лексики и терминологии, причины совмещения их в одном ресурсе будут рассмотрены в следующем разделе.

15.3. Соотношение лексики и терминологии. Общественно-политическая область

15.3.1. Разделение лексики и терминологии.

Подавляющее большинство текстов, хранимых в современных электронных коллекциях и нуждающихся в эффективной обработке и поиске, принадлежат к так называемой деловой прозе и содержат как общеупотребительную лексику, так и терминологию конкретных предметных областей.

Однако общеупотребительные слова и термины изучаются представителями различных научных дисциплин – лексикологами и терминологами. Для описания общей лексики и терминологии создаются различные ресурсы.

Так, предполагается, что ресурсы типа WordNet описывают, прежде всего, общую лексику языка. В Принстонском WordNet можно найти достаточное количество терминов из разных областей, особенно широко представлены термины из биологической систематики. Представляется, что включение терминов в WordNet не носило системный характер, а было связано с тем, что в разных предметных областях существуют иерархии, удобные для внесения в тезаурус.

Это подтверждается тем, что разработчики тезаурусов в рамках проекта EuroWordNet, а также других европейских ворднетов строже ограничивают внесение в свои тезаурусы синсеты, относящиеся именно к общеупотребительному языку. Так, как мы указывали в разделе 3.3.2, разработчики датского ворднета DanNet отказываются вносить в свой ресурс удобный обобщающий синсет, поскольку полагают, что он

соответствует значению термина из сферы страхования. Предполагается, что для терминологии предметных областей должны создаваться отдельные тезаурусы (см. раздел 3.3.7).

Остановимся подробнее на вопросах различия общей лексики и терминологии.

В настоящее время, наиболее общепринятым определением термина является следующее определение: *термин – это слово или словосочетание, номинирующее понятие определенной области знания или действительности* (Суперанская и др., 2003; Лейчик, 1994; Володина, 1996; Шелов, 2003; Гринев-Гриневич, 2008).

Таким образом, первое различие заключается в том, что термин относится к **определенной предметной области**, терминологией владеют профессионалы в данной предметной области, а общая лексика известна многим людям, вне сферы их профессиональных занятий.

Кроме того, определение термина устанавливает связь термина с **понятием предметной области**. Основоположник Венской школы терминологии Э.Вюстер (Wűster,1979) подчеркивал, что одно из существенных различий между методами исследования, используемых лингвистами и терминологами, заключается в том, что терминологи начинают свое рассмотрение с понятия, которое должно быть точно определено и не зависит от своего наименования, а лексикологи начинают с языкового выражения. Поэтому традиционно терминологи говорят о понятиях, а лингвисты о значениях. Х.Фелбер (Felber, 1984) также подчеркивает, что «если в лингвистике содержание слова и его форма рассматриваются как одна единица, то в терминологии понятие и его обозначение ... отделены друг от друга».

Во многих работах подчеркивается, что и понятие, и лексическое значение относятся к категориям мышления, при этом между ними есть существенные различия.

Значение включает в себя помимо понятийного содержания (сигнификативно-денотативного компонента значения), такие компоненты как оценочный, стилистический, сочетаемостный. Значение включает лишь различительные черты объектов, иногда относительно поверхностные, а понятия охватывают их наиболее глубокие существенные свойства.

В связи со значениями общей лексики иногда говорят о **наивных или бытовых понятиях** (Апресян, 1995; Шелов, 2003; Герд, 2005), которые противопоставляются содержательным или научным понятиям. Считается, что наивное понятие включает лишь различительные черты объектов, иногда относительно поверхностные, а научные понятия охватывают их наиболее глубокие существенные свойства.

К важным свойствам термина относят также его **точность и однозначность** (Суперанская и др., 2003; Шелов, 2003; Sager, 1990). Так, в работе (Суперанская и др., 2003) подчеркивается, что термин должен относиться непосредственно к понятию, он должен выражать понятие ясно, значение термина должно быть точным и не должно пересекаться по значению с другими терминами, значение термина не должно зависеть от контекста. Гринев-Гриневич (Гринев-Гриневич, 2008) перечисляет более 10 признаков терминов и требований, предъявляемые к терминам. Таким образом, приводится значительное количество свойств, отграничивающих термин от лексической единицы общего языка.

Как отдельный способ формирования терминов рассматривается превращение в термин общеупотребительного слова - терминологизация, когда общеупотребительное слово получает новое терминологическое значение в конкретной предметной области. В то же время широко распространен и обратный процесс - детерминологизация, когда появившийся в некоторой специальной области термин становится словом общей лексики.

В работе (Суперанская и др., 2003) отмечается, что при этом специальное значение в общей лексике редуцируется, термин приобретает прагматические свойства, которых он прежде был лишен, то есть возникает новое слово с терминологическим значением,

требующее уже не дефиниции, а толкования. Породивший новое слово термин остается в своем терминологическом поле без изменений. Таким образом, считается, что фактически при процессе детерминологизации появляется омоним термина.

В то же время имеет значительное число работ, показывающих относительность вышеперечисленных различий между общей лексикой и терминами.

Так, помимо лексики, которая может использоваться в тексте любой тематики, имеется тематическая общеупотребительная лексика, тесно связанная с терминологией соответствующей предметной области.

С.Д. Шелов (Шелов, 2003) указывает, что теоретическом плане соотношение «специальное понятие» - «неспециальное понятие» и основанное на нем разграничение «термин-нетермин» вряд ли могут считаться совершенно ясными и подлежат дальнейшему исследованию. В.Н. Хохлачева отмечает: «разграничение «специальных объектов и понятий с «неспециальными» - далеко не очевидный факт» (Хохлачева, 1981).

При текущем уровне онтологического моделирования и специальные, и «наивные понятия» моделируются в рамках одних и тех же онтологических структур, что, в частности, как раз и делается в рамках построения иерархических систем значений типа WordNet (Climent и др., 1996; Miller и др., 1990).

Создавая ворднеты для своих языков, лингвисты выстраивают значения слов и языковых выражений в виде иерархических систем, пытаются найти схожие понятия для различных языков, выстраивают верхний независимый от языка уровень таких систем, пытаются использовать созданные структуры для общеупотребительного языка как заготовку для автоматизированного выявления понятийных систем в конкретных предметных областях (Vossen, 2001; Buitellar, Sacalenau, 2001). Таким образом, на текущем уровне представления понятий и понятийных систем, нет возможности представлять по-разному наивные и специальные понятия, если даже в них имеется какая-либо значительное различие.

Многие авторы также приводят многочисленные примеры недостаточной точности и однозначности терминов. Так, С.Д. Шелов (Шелов, 2003) подчеркивает, что утверждения об однозначности и точности терминов в значительной мере преувеличены: «имеется почти регулярная многозначность некоторых технических терминов: *смазка* (вещество) – *смазка* (процесс), *верстка* (первый корректурный оттиск) – *верстка* (процесс), для лингвистической терминологии регулярная многозначность вида «раздел науки – совокупность всех единиц, операций и отношений данного уровня», которая наблюдается в терминах типа *фонетика, фонология, морфология, синтаксис.*

Кроме того, Шелов С.Д. исследует дефиниции терминов, которые включают выражения, допускающие различные интерпретации, так называемые полиморфные определения и, соответственно, полиморфные термины. Например, дефиниция термина паронимы (*слова, имеющие сходство в морфологическом составе, и, следовательно, в звучании, но различающиеся по значению*) содержит слово *сходство*, которое допускает различную интерпретацию.

Из своего рассмотрения Шелов С.Д. делает вывод, что «если учесть описанное свойство полиморфизма, квантитативной и квалификативной мягкости некоторых терминов, то окажется излишне категоричным традиционное мнение о том, что термины – это слова со строго определенным, совершенно точным значением» (Шелов, 2003, с. 71).

Б.Ю. Городецкий (Городецкий, 2006) также подчеркивает, что нечеткость термина присутствует во всех областях научного общения и имеет при этом многообразные проявления. С ней связана принципиальная неполнота любой дефиниции: в ней не может быть отражено все хотя бы потому, что знания постоянно развиваются и обогащаются.

В результате может быть сделан вывод, что разные слова и выражения естественного языка имеют разную степень терминологичности (Городецкий, 2006; Шелов, 2003; Комарова, 1991, Квитко и др., 1986, Гринев-Гриневич, 2008).

Так, в работе (Городецкий, 2006) указывается, что «терминологичность – комплексное свойство, ведущим фактором которого является степень дефинированности, то есть степень конвенциональной закрепленности за фрагментом системы знаний, степень сознательной дефинированности в рамках конкретного подъязыка и конкретной человеческой деятельности. В принципе, степень, дефинированности может убывать плавно. Между терминами и нетерминами имеются, следовательно, различные переходные случаи – квазитермины.»

Также и С.Д. Шелов подчеркивает, что «специальность» понятия, по-видимому, относительна и допускает большую или меньшую степень, традиционную дихотомию «термин-нетермин» естественно изменить и трансформировать в более гибкое и относительное понятие терминологичности (Шелов, 2003).

В ряде исследований описываются способы количественного измерения терминологичности языковых выражений. Один из таких методов, предложенный Шеловым С.Д., будет рассмотрен в следующем разделе.

15.3.2 Степень терминологичности понятия

С.Д. Шелов в работе (Шелов, 2003) предлагает оценивать терминологичность понятия, вычисляя ее как функцию от состава соответствующего термина, а именно:

Если термин членится на свои подтермины, то суммарная терминологичность такого термина вычисляется как сумма терминологичности подтерминов:

Т (ячейка оперативной памяти) = Т (ячейка памяти) + Т (оперативная память),

Т (носитель магнитной записи) = Т (носитель записи) + Т (магнитная запись)…

Если же термин мотивирован (то есть его значение полностью выводимо из каких-либо других терминов), но не членится на свои подтермины, то суммарная терминологичность такого термина – сумма терминологичности всех терминов (а не только подтерминов) необходимых для вывода (мотивации!) соответствующего значения, так что в соответствующих примерах можно записать:

Т (узел вычислительной машины) = Т (узел) +
+ Т (элемент вычислительной машины),

Т (символ блока) = Т (элемент алфавита) + Т (символ) + Т (блок),

Т (ячейка оперативной памяти) = Т (ячейка запоминающего устройства) +
+ Т (оперативная память)…

Каждое полнозначное слово общего языка независимо от его синтаксической функции также вносит «вклад» в терминологичность определяемого термина. Соответственно, общая терминологичность определяемой единицы складывается из терминологичности всех использованных в определении этой единицы терминов, а также общего количества полнозначных слов, появление которых в определяющем выражении не продиктовано число грамматическими требованиями: не подсчитываются некоторые союзы (типа «который») и предлоги, которыми сильно управляют соответствующие существительные, прилагательные, глаголы, т.е. единицы типа предлога «от» в словосочетании «зависеть от». Местоимения заменяются на те слова (или словосочетания), которые они замещают и, в зависимости от этого, их «вклад» в терминологичность определяемой единицы равняется количеству полнозначных замещаемых слов (если замещается слово общего языка) или терминологичности замещаемого термина (если замещается термин).

В работе приводятся примеры оценки степени терминологичности специальной лексики, представляющей правила дорожного движения:

T (автомагистраль)=31, T (велосипед)=29, T (водитель)=48,
T (вынужденная остановка) =70, T (дорога)=29,
T (дорожно-транспортное происшествие)=66, T (железнодорожный переезд)=34,
T (маршрутное транспортное средство)=41, T (мопед)=30, T (мотоцикл)=10,
T (населенный пункт)=6, T (недостаточная видимость) =39, T (обгон)=63, …
T (транспортное средство)=8, T (тротуар)=190.

Таким образом, одним из самых терминологичных слов получается слово *тротуар*. Тротуар в правилах дорожного движения определяется как *элемент дороги, предназначенный для движения пешеходов и примыкающий к проезжей части или отделенный от нее газоном.*

В толковом словаре (БТС, 1998) тротуар *определяется как пешеходная дорожка, идущая сбоку от проезжей части улицы,* то есть самому терминологичному из перечисленных терминов соответствует слово общего языка со значением, которое чрезвычайно трудно отличить от указанного терминологического значения.

Можно ли в этом случае считать, что, согласно утверждению из (Суперанская и др., 2003), на самом деле имеется два значения слова *тротуар* – терминологическое значение и общее значение, и что в таком случае слово *тротуар* является многозначным?

Представляется, что в данном случае разделение терминологического и нетерминологического значения невозможно, поскольку невозможно, чтобы водители и пешеходы, не знакомые с правилами дорожного движения, понимали значение слова «тротуар» по-разному.

Такая ситуация возникает в тех случаях, когда определенная сфера деятельности напрямую контактирует с жизнью вне этой сферы, такого рода пограничная лексика обязана иметь совпадающие терминологическое и общее значение, иначе взаимодействие между профессионалами и населением было бы невозможным. Возможно, именно, прежде всего, такие языковые единицы как *тротуар,* представляют собой примеры квази-терминов, о которых упоминалось в работе (Городецкий, 2006).

Кроме того, следует отметить, что многие термины из сферы дорожного движения должны быть отнесены к таким пограничным терминам, поскольку ими оперируют совершенно разные категории людей: водители-профессионалы, автолюбители, пешеходы.

На наш взгляд, вычисляя терминологичность лексики в рамках той или иной предметной области, необходимо постулировать достаточно низкую степень терминологичности таких пограничных терминов, независимо от того, насколько сложным образом пришлось их определять в рамках системы дефиниций данной предметной области. Так, в работе (Гринев-Гриневич, 2008) подчеркивается, что во всякой терминологии непременно есть некоторое количество лексических единиц, встречающихся как в обыденной, так и в профессиональной речи, - так называемые консубстанциальные термины, которые вызывают ряд трудностей при выделении терминологической лексики из словарного состава языка.

15.3.3. Промежуточный слой между лексикой и терминологией

Если рассмотреть различные предметные области и близкие к ним общеупотребительные слова, то оказывается, что таких общеупотребительных слов русского языка, которые имеют очень близкие по смыслу терминологические соответствия в одной или более областях без серьезного сдвига значения, очень много.

Например, понятия, соответствующие общезначимому слову *здание*, являются необходимым элементом терминологии, по крайней мере, в двух областях: в области строительства и в области коммунального обслуживания. Так, С.В. Гринев-Гриневич (Гринев-Гриневич, 2008) указывает, что строительство является одним из древнейших видов человеческой деятельности, и поэтому в специальные типологии сооружений,

архитектурных элементов зданий и т.д. входят такие понятные всем термины, как *одноэтажные, многоэтажные, панельные здания; двери, окна, стены, лестницы* и т.д.

Если предположить, что всякое терминологическое значение должно быть отделено от общелексического значения, то это означает, что эти значения функционируют в различных контекстах, и совмещение их в одном и том же предложении приводит к такому явлению как «игра слов» (Cruse, 1986). Между тем если такое слово как «здание» употребляется в газетном тексте, то часто невозможно отличить, в каком значении общелексическом или терминологическом это слово употреблено, как, например, в следующем фрагменте текста публикации газеты «Известия» от 18.02.2004 под названием «Кто ответит за Трансвааль»:

> *Допросы строителей, проектировщиков, людей которые отвечали за эксплуатацию* ***здания****, пока мало что дали, - заявили "Известиям" в прокуратуре. Все они, естественно, отрицают свою вину и кивают друг на друга. Будут проверены все, в том числе и поставщики строительства, они могли предоставить материалы не того качества, которые заявлялись. Объективно ответить на вопрос, кто виноват, можно будет только после завершения экспертиз.*

Несомненно, в профессиональной области границы понятия могут быть четче, строитель увереннее непрофессионального носителя языка должен отличать «здания» от других видов строений. При этом для непрофессионала такие «граничные» объекты относятся к области неопределенности - «vagueness» (Шелов, 1998), когда он уже не может четко классифицировать строение и охотно принимает профессиональную классификацию. Поэтому по нашему мнению, понимание слова «здание» у строителя и обычного носителя языка отличается настолько незначительно, что сравнимо с индивидуальными различиями смыслов у разных людей, и этими различиями можно пренебречь.

На наш взгляд, подавляющее большинство общеупотребительных слов, обозначающих артефакты, должны иметь чрезвычайно близкие по смыслу терминологические аналоги, по крайней мере, в двух предметных областях: области производства этого артефакта и профессиональной области его обслуживания.

Кроме того, если человек соприкасается в своей повседневной жизни с профессиональными сферами деятельности, ему необходимо понимать и использовать те же самые значения, что и профессионалам. Например, приходя в банк, клиент должен правильно использовать и понимать соответствующие термины этой сферы, такие как вклад, кредит, иностранная валюта и др.,

Такими же свойствами совместного лексического и терминологического употребления обладают и многие другие общеизвестные слова: названия транспортных средств, должности и профессии, технические устройства, произведения искусства, социальные и природные явления и многие другие. Перечислим лишь некоторые из таких слов, начинающихся на букву «а»: *аборт, аванс, авиабаза, автобус, автопилот, агроном, адвокат, аккредитив, алгебра, алгоритм, алебастр* и т.п.

Наконец, отмечается (Моисеев, 1970), что типичные бытовые слова – *отец, мать, сын, дочь* и т.п. – причисляются к терминологии в качестве терминов родства и свойства.

Мы оцениваем, что до 40 процентов слов, содержащихся в общих толковых словарях, обладают похожими свойствами.

Кроме того, существует значительное количество многословных выражений, которые являясь терминами в специальных предметных областях, понятны носителям языка, например, *военная помощь, авиационная промышленность, внешняя миграция*. Это означает, что взаимопроникновение лексики и терминологии имеет значительно больший масштаб, чем это предполагалось ранее терминологами и лексикологами.

На существование промежуточного слоя между общей и специальной лексикой указывалось и ранее. На рис. 15.1 воспроизведен рисунок из (Rondeau, 1980), на котором показан такой слой. На наш взгляд, эта пограничный лексико-терминологический слой представляет собой достаточно широкую полосу (рис.15.1).

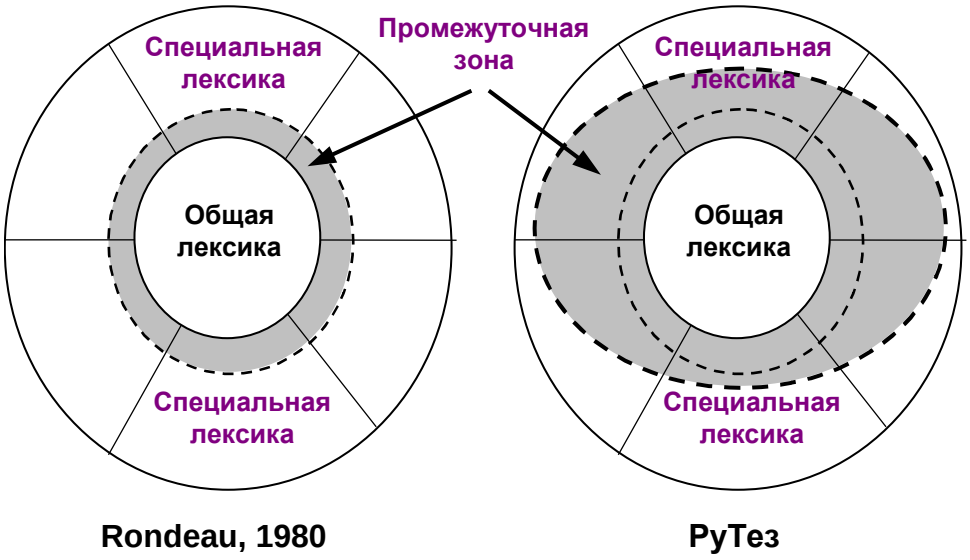

Рис. 15.1. Наличие промежуточного слоя лексики между терминологией (специальной лексикой) предметных областей и общеупотребительной лексикой

15.3.4 Общественно-политическая область

Из предшествующего рассмотрения мы можем сделать следующие выводы (Loukachevitch, Dobrov, 2004d; Лукашевич, Добров, 2004a).

1) В общеупотребительном языке существует лексика, которая может быть употреблена во многих предметных областях, не связана с той или иной предметной областью. Это лексика, связанная с общими процессами, действиями, стадиями, отношениями, оценками. Мы называем такое множество лексики Общий лексикон. При разметке тезауруса WordNet предметными областями такое множество нетематической лексики также было выделено и названо областью Factotum (см. раздел 2.5.3.1.).

2) Мы считаем, что человек разрезает на мир на более узкие или более широкие области, понятийные системы, терминосистемы для удобства.

Понятие в большой степени не зависит от того, какую предметную область мы рассматриваем (см. также концепцию «универсального терминологического пространства в (Мальковский, Соловьев, 2002)).

Так, понятие «ценные бумаги» может входить в разные частично пересекающиеся предметные области такие как

- предметная область «ценные бумаги»,
- предметная область «биржевая торговля»,
- предметная область «инвестиции»,
- предметная область «финансы» и т.п.

Если взять тексты, которые относятся к данным предметным областям, то выяснится, что помимо терминологии этих областей в текстах содержится значительной количество терминологии из более общей предметной области, из «соседних» предметных

областей и другой терминологии. Таким образом, чтобы качественно обрабатывать тексты в той или иной предметной области, нужно описать в тезаурусном ресурсе значительно большее количество языковых единиц.

Поэтому мы не создаем отдельные тезаурусы для большого количества предметных областей, а делаем ресурс на максимально широкую предметную область. Одной из таких областей является так называемая общественно-политическая область.

3) Общественно-политическая область включает в себя лексику и терминологию, которая, с одной стороны, известна достаточно широкому слою населения, с другой стороны, соответствует понятиями профессиональных сфер деятельности.

На такую особенность Общественно-политической области указывают также разработчики Тезауруса Исследовательской службы Конгресса США (LIV, 1994), которые пишут, что для описания широкой области общественных отношений приходится использовать разные типы лексических единиц, в том числе, как специальную терминологию, так и тематическую лексику общего языка (popular terminology).

Рассмотрим состав и особенности общественно-политической области на примере реализации тезауруса РуТез как модели понятийной системы русского языка.

Если представить себе иерархию понятий от более общего к более частному, то наиболее верхние уровни занимает зона общей лексики – Общий лексикон, более нижние уровни занимает Общественно-политическая область (рис.15.2). Одновременно именно в общественно-политической области находятся верхние уровни профессиональных понятийных систем (рис.15.3).

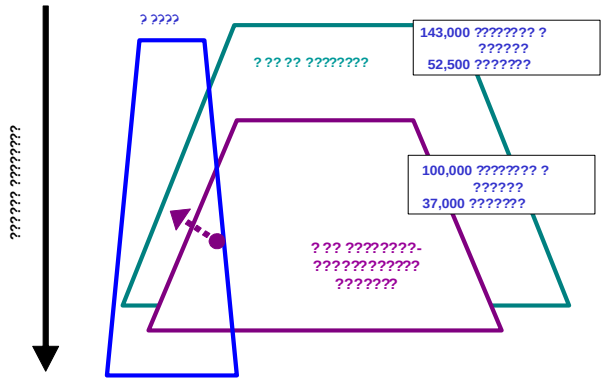

Рис.15.2. Общественно-политическая область vs. Общий лексикон

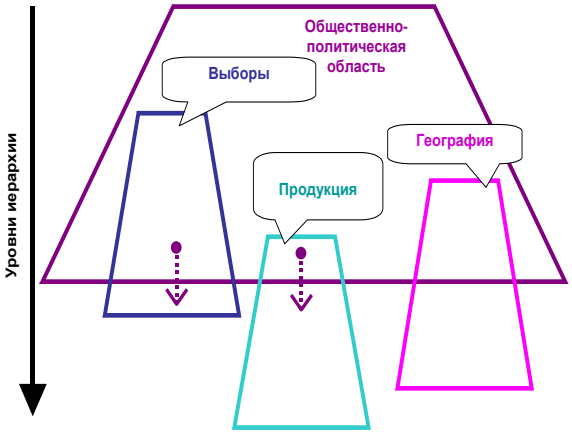

Рис.15.3. Специальная лексика vs.

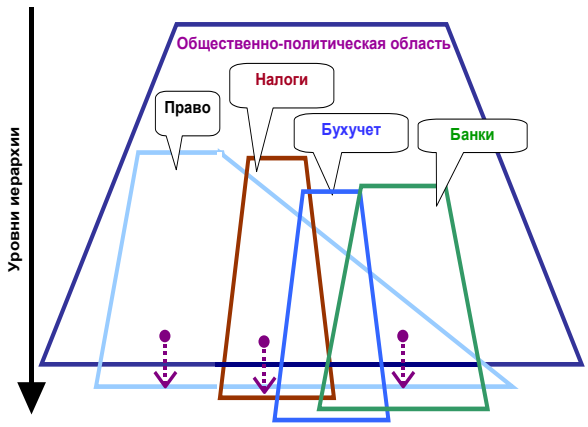

Рис.15.4. Взаимосвязь специальной лексики

Общественно-политическая внутри Общественно-политической области

Разные предметные области имеют различные по величине перенесению с общественно-политической областью. Так, понятийная система предметной области «Выборы» практически полностью находится в Общественно-политической области, в то время как сферы различных промышленных производств пересекаются с общественно-политической областью лишь по небольшому числу понятий (рис. 15.3).

Можно выделить совокупность непроизводственных регулирующих сфер деятельности, которые значимы в повседневной деятельности многих людей и, значит, в значительной степени пересекаются с общественно-политической областью, такие как Налоги, Бухгалтерия, Право, Таможня, Банковская сфера, образуя правовой и финансовый блоки областей (рис.15.4).

Научные понятийные системы пересекаются с общественно-политической областью сложнее. Общественно-политическая область включает основные виды наук, научных учреждений, общенаучные понятия. Однако каждая наука задает свою категоризацию изучаемых явлений, в связи с чем ее верхние уровни классификации могут значительно отличаться от классификации, на базе общей лексики (Рис. 15.5). Наиболее значительно общественно-политическая область пересекается со сферой общественных наук. Общественно-политическая область содержит понятия общественной жизни, которые изучаются общественными науками. При этом каждая общественная наука может иметь свою собственную классификацию рассматриваемых явлений.

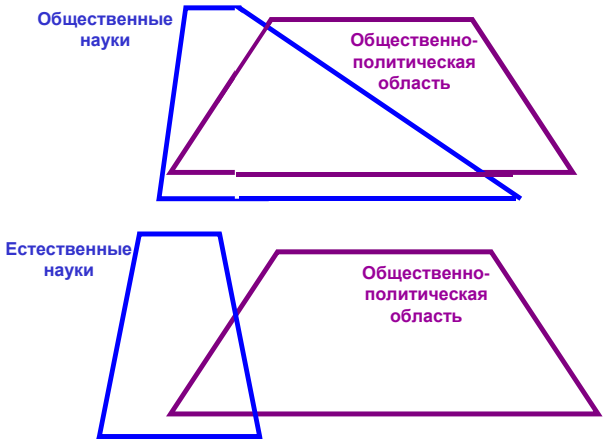

Рис.15.5. Научная лексика vs. Общественно-политическая область

Выделение такой области, а также выделение среди общеупотребительной лексики лексем, принадлежащих этой области, является чрезвычайно полезным для разработки лингвистических ресурсов и технологий автоматической обработки больших электронных коллекций.

Прежде всего, терминология и лексика из этой области активно используется в самых разных по жанру, но значимых для жизни общества текстах, как законы, международные договора, другие официальные документы, газетные сообщения, экономические документы (Loukachevitch, Dobrov, 2002). Таким образом, создание лингвистического ресурса в общественно-политической области может значительно повысить эффективность и содержательность обработки всех этих видов документов.

Поскольку общественно-политическая область содержит наиболее общеизвестные понятия многих профессиональных предметных областей, то лингвистический ресурс, разработанный для общественно-политической области, может стать источником

существенного понятийно-терминологического материала для создания лингвистических ресурсов в конкретных предметных областях.

Одновременно общественно-политическая область – это область общезначимая и содержит значительное количество общелексического материала, который относится к нижним и средним наиболее конкретным уровням языковой системы языка, поэтому понятийная структура общественно-политической области является и существенным базисом, на который можно опираться, например, выстраивая понятийную иерархическую систему языка типа WordNet (Miller и др., 1990).

Кроме того, если рассмотреть количество многозначных общеупотребительных слов внутри общественно-политической области и в общем лексиконе, то многозначных слов в общественно-политической области значительно меньше, а процедура автоматического разрешения многозначности работает эффективнее, поскольку часто значения относятся к различным подобластям общественной жизни, например, в подавляющем большинстве текстов контексты разных значений словоформы *судов* как средства водного транспорта и судебного органа существенно различаются. Это различие можно также эффективно использовать при автоматической обработке текстов, используя, например, комбинированную обработку текстов и запросов при решении информационно-поисковых задач, а именно пытаться разрешать многозначность для слов и терминов, относящихся к общественно-политической области, и использовать пословную обработку для остальной общеупотребительной лексики (подробнее см. разделы 18.2, 20.4).

Таким образом, мы считаем, что описание общей лексики должно сочетаться с описанием терминологии предметных областей общественно-политической области.

В настоящее время Общественно-политический тезаурус интегрирует в себе значительную долю терминологии следующих предметных областей, которая была введена в него в течение деятельности в ряде проектов по автоматической обработке текстов: экономика, право, социология, демография, банковское дело, государственный финансовый контроль, выборы.

15.4. Общественно-политический тезаурус в сравнении с традиционными информационно-поисковыми тезаурусами

К началу 2010 года объем тезауруса РуТез составляет 52.5 тысяч понятий, 143 тысячи разных русскоязычных слов и словосочетаний, 209 тысяч отношений между понятиями. Общественно-политический тезаурус составляет более двух третей от объема тезауруса РуТез и включает в себя 37 тысяч понятий, около 100 тысяч разных русских слов и словосочетаний (рис. 15.2).

Как мы увидим в дальнейшем, Общественно-политический тезаурус в ряде задач применяется отдельно от остального тезауруса и может рассматриваться как информационно-поисковый тезаурус, созданный для автоматического индексирования текстов в широкой общественно-политической области. По широте предметной области Общественно-политический тезаурус соответствует таким тезаурусам как Тезаурус исследовательской службы Конгресса США LIV (LIV, 1994) или тезаурус Европейского сообщества EUROVOC (EUROVOC, 2001). Однако наш Общественно-политический тезаурус во много раз больше упомянутых тезаурусов.

Такое различие связано с тем, что Общественно-политический тезаурус изначально создавался как ресурс для автоматической обработки текстов, когда человека-посредника между информационно-поисковым тезаурусом и языком документов нет. Поэтому достаточно большой объем информации должен быть представлен непосредственно в тезаурусе (см. п. 1.7.).

Общественно-политический тезаурус включает не только термины, которые представляют важные понятия в текстах данной предметной области, но также охватывает широкий круг более специфических терминов, обнаружение которых в конкретном тексте сделает этот текст релевантным запросу по понятиям более высокого уровня.

Синонимические ряды понятий Общественно-политического тезауруса значительно богаче, чем совокупности вариантов дескриптора в тезаурусах LIV или EUROVOC, поскольку синонимы должны описывать различные способы выражения данного понятия в тексте для автоматического процесса, а не для человека. Ряды синонимов включают в себя не только существительные и именные группы, а также прилагательные, глаголы, глагольные группы.

Расширение терминологической базы Общественно-политического тезауруса ведет к необходимости описания многозначных терминов. Общественно-политический тезаурус содержит около 4.5 тысяч многозначных слов и выражений. В традиционных информационно-поисковых тезаурусах нет необходимости аккуратно описывать многозначность употребляемых в текстах слов и выражений, поскольку понимание текста, его основной темы возложено на человека-индексатора.

Расширение понятийной базы Общественно-политического тезауруса ведет к увеличению и усложнению функций отношений между понятиями тезауруса: возникает необходимость логического вывода на отношениях.

Заключение к главе 15.

В данной главе мы представили особенности структуры Тезауруса русского языка РуТез. При разработке тезауруса как ресурса для автоматической обработки текстов были использованы принципы различных традиций и методологий, а именно, методологии разработки традиционных информационно-поисковых тезаурусов, методологии разработки лингвистических ресурсов типа WordNet, методологии созданий формальных онтологий.

Особенностью тезауруса РуТез является то, что в нем выделяются две составные части Общий лексикон и Общественно-политический тезаурус, который содержит тематическую лексику и терминологию, значимую для общества в целом. Такое сочетание в одном ресурсе обычно разделяемых языковых сущностей связано с тем, что граница между лексикой и терминологией представляет собой широкую промежуточную зону. Она содержит лексемы, значения которых совпадают с понятиями конкретных предметных областей, и термины, понятные носителям языка.

Эта зона включает в себя понятия, значимые для общества в целом, поэтому мы называем ее Общественно-политической областью. Лексико-терминологические ресурсы, разработанные для общественно-политической области, полезны для приложений по автоматической обработке разнообразных типов текстов. Знания об общественно-политической области очень важны как для создания лингвистических ресурсов в конкретных предметных областях, так и как основа для описания абстрактной лексики языка.

Общественно-политический тезаурус может рассматриваться как пример информационно-поискового тезауруса в широкой предметной области, созданный специально как ресурс для автоматической обработки текстов в приложениях информационного поиска и поэтому обладающий рядом специфических характеристик по сравнению с традиционными информационно-поисковыми тезаурусами.

Глава 16. Единицы тезауруса: понятия и их текстовые входы

Тезаурус РуТез является лингвистической онтологией, то есть подавляющее большинство понятий в тезаурусе РуТез связаны со значениями реально существующих языковых выражений. В то же время, поскольку тезаурус РуТез является онтологией, то единицы тезауруса должны отвечать правилам представления понятий в онтологиях.

Как мы уже указывали в разделе 5.1, важными принципами представления понятий в онтологии являются следующие:

- необходимо отличать понятие и его имя, разные названия одной и той же сущности не должны приводить к введению отдельных понятий,
- нижестоящие понятия должны отчетливо отличаться от вышестоящих понятий, то есть, например, иметь специфическое отношение или атрибут,
- каждое понятие должно отчетливо отличаться от понятий того же уровня иерархии (понятий-сестер).

Эти рекомендации введения понятий онтологии не просто реализовать, если онтология основывается на значениях реально существующих языковых выражений. Имеется несколько источников таких трудностей.

Во-первых, в некоторых случаях может быть сложно отличить понятие и его различные имена. Как мы видели, в ресурсах типа WordNet отдельные синсеты вводятся для разных частей речи, которые являются деривативами, то есть называют одну и ту же сущность или явление посредством разных частеречных единиц. Также отдельные единицы в ресурсах типа WordNet часто вводятся, чтобы отразить стилистические, географические или диалектные особенности употребления слов.

Во-вторых, серьезную сложность представляет собой представление в виде совокупности понятий значений многозначных слов, особенно в тех случаях, когда эти значения являются очень близкими друг к другу. Часто в таких случаях возникает вопрос, что правильнее с точки зрения как качества описания, так и с точки зрения приложений автоматической обработки текстов: представить такие близкие значения как отдельные, возможно связанные между собой понятия или соединить близкие значения в одно и то же понятие.

В-третьих, непростой проблемой является описание близких значений разных слов. Такие слова могут отличаться посредством множества разных характеристик, особенностей употребления. Разбиение такой совокупности взаимосвязанных значений на совокупность дискретных понятий, каждое из которых должно быть отличимо от других близких понятий, является достаточно сложной процедурой. Но именно такие понятия (несмотря на то, что они мотивированы значениями языковых единиц конкретного языка), приобретают некоторые свойства независимости от конкретного языка: если понятие отличимо от близких понятий, то особенности данного понятия могут тем или иным образом быть сформулированы на разных языках (Nirenburg, Raskin, 2004).

Наконец, непростой вопрос возникает, в каких случаях необходимо или полезно вводить в онтологию понятия, основанные на значениях словосочетаний. Поскольку словосочетаний в языке может быть бесконечное количество, то важным является вопрос, посредством каких принципов должно регулироваться введение в тезаурус понятий, отражающих значения словосочетаний.

В следующих разделах мы подробно рассмотрим решения, принимаемые по всем этим вопросам при разработке тезауруса РуТез.

16.1. Понятия vs. синсеты как единицы тезауруса

Создавая тезаурус РуТез, мы не стремимся отделить лексические знания от знаний о мире, как провозглашалось такими исследователями, как И.А. Мельчук (Мельчук, 1974) или Дж. Миллер (Lenat et.al., 1995). Единицей описания в тезаурусе является не множество синонимичных слов или терминов как в тезаурусе WordNet, а понятие, отражающее значимые классы сущностей, различаемых людьми в мире, в современной общественной жизни, в психической жизни людей. Такие сущности бесконечно разнообразны, обладают индивидуальными особенностями. Учет всех таких особенностей ведет к чрезвычайно запутанной картине мира (Kuznetsov и др., 2007; Cimiano и др., 2004), непригодной для прикладного применения.

Значения слов и выражений, существующие в современных естественных языках, позволяют нам выделить главное, существенное для современной жизни людей. Так, например, созвездия выделяются из других возможных совокупностей звезд, поскольку помогают людям ориентироваться в пространстве и указывать местоположение объектов на звездном небе (Gangemi et.al. 2001).

При этом понятия тезауруса РуТез должны быть отличимы друг от друга, иметь уникальные свойства в системе понятий (см.п.5.1.). Использование в качестве единиц тезауруса таких отличимых понятий позволяет единым образом представлять лексические значения литературного языка и значения терминов предметной области, более последовательно описывать систему отношений между понятиями и тем самым облегчает формальный вывод на отношениях, позволяет отображать единым образом систему значений разных языков (Добров, Лукашевич, 2005; Loukachevitch, 2009a).

Понятию может соответствовать несколько синонимичных текстовых выражений - текстовых входов понятия. Слова и словосочетания, значения которых представлены как ссылки на одни и те же понятия тезауруса, будем называть онтологическими синонимами. От онтологических синонимов не требуется, чтобы они могли заменять друг друга в каких-либо предложениях. Значения онтологических синонимов должны иметь одинаковый набор отношений с другими понятиями тезауруса:

Таким образом, онтологическими синонимами могут являться:

- слова, являющиеся разными частями речи (*стабилизация, стабилизироваться, стабилизационный),* то есть представлять собой дериваты, так называемые синонимы в широком смысле (Апресян, 1995),
- языковые выражения, относящиеся к разным языковым стилям (*коммунальная квартира, коммуналка*),
- однословные выражения, устойчивые выражения, свободные словосочетания, выражающие одно и то же понятие (*аэропорт - воздушные ворота, газ – газообразное вещество)*.

В частности, нейтральные и уменьшительные названия сущностей (*стол, столик*) мы относим к одному и тому же понятию (в отличие от русского WordNet – RussNet (Азарова и др., 2003)), поскольку, на наш взгляд, использование таких названий не приводит к реальному изменению соотношений между понятиями – любой стол может быть назван столиком в некотором контексте. Причем невозможно четко указать причины, по которым было употреблено уменьшительное название: стол был рассмотрен как маленький, как любимая вещь или просто это такая манера разговора. Если нет четкого, независимого от контекста различия между значениями, то отдельное понятие не заводится,

Подобно FrameNet (Fillmore и др., 2003) несовершенный и совершенный виды одного и того же глагола (выбрать, выбирать) также рассматриваются как онтологические синонимы. Глаголы-делимитативы (Зализняк, Шмелев, 2000), описывающие некоторую «порцию» действия, оцениваемую как небольшую и ограниченную по времени, рассматриваются как онтологические синонимы к глаголу, от которого они образованы,

например, *погулять – гулять, почитать – читать, побегать – бегать* и т.п. Однократные и многократные действия (*куснуть – кусать, моргнуть – моргать, плюнуть – плевать*) также описываются как онтологические синонимы.

Таким образом, в тезаурусе РуТез мы пытаемся соблюдать правило разработки онтологии, заключающееся в том, чтобы разные имена одних и тех же сущностей не вели к образованию разных понятий, а были объединены как онтологические синонимы одного и того же понятия.

16.2. Имя понятия и толкование

Для работы с понятиями, анализа результатов автоматической обработки текстов, важно, чтобы понятие имело понятное, однозначное и компактное имя, передающее основной объем этого понятия.

С этой точки зрения, оперирование длинными рядами синонимов как в WordNet не очень удобно. Кроме того, если в WordNet синсет состоит из одного многозначного слова, то пояснить его можно с помощью толкования, что также очень длинно, или с помощью гиперонима, который также может быть неоднозначным.

В тезаурусе РуТез каждое понятие должно иметь однозначное имя, которое построено на базе его текстовых входов, и должно быть понятным носителю языка.

Имена понятий могут быть следующих видов:

- однозначное слово: КАБЕЛЬ;
- однозначное словосочетание, являющееся одним из текстовых входов понятия: КАБИНЕТ ВРАЧА, КАБИНЕТ РЕСТОРАНА;
- неоднозначное словосочетание с пометой подобно пометам, используемым в традиционных информационно-поисковых тезаурусах. В качестве пометы используется по возможности текстовый вход одного из вышестоящих понятий: КАБАЧОК (РАСТЕНИЕ), КАБАЧОК (ПЛОД);
- пара синонимов – текстовых входов понятия через запятую: ИРРАЦИОНАЛЬНЫЙ, ЛОГИЧЕСКИ НЕОБЪЯСНИМЫЙ; ПОТНЫЙ, МОКРЫЙ ОТ ПОТА. В отличие от ресурсов типа WordNet в тезаурусе РуТез пара синонимов в названии понятия должна однозначно идентифицировать суть понятия. Использование таких названий понятий особенно полезно в тех случаях, когда принимается решение совместить в одном понятии значения несколько различающихся слов. Это решение удобно зафиксировать в названии понятия, например, ПАМЯТНИК, МОНУМЕНТ.

Если есть такая возможность, то есть если среди текстовых входов понятия, имеется существительное или именная группа, то имя понятия делается на основе существительного (именной группы).

Понятие может иметь комментарий, который пишется в случае необходимости и не является частью имени понятия. Это также практика, принятая при разработке традиционных информационно-поисковых тезаурусов.

16.3. Ввод понятий для группы близких по смыслу слов

Как известно, в естественном языке есть значительно количество близких по смыслу групп слов – квазисинонимов (см. разделы 5.3.1, 5.5). Выделяя понятия на основе значений таких квазисинонимов, мы пытаемся обеспечить, чтобы введенное понятие имело четкое, независимое от контекста отличие от родового понятия и от так называемых понятий-сестер, то есть видовых понятий к тому же родовому понятию.

Поскольку в настоящее время понятия тезауруса РуТез не имеют внутренней структуры в виде фреймовых элементов или атрибутов, то отличительные свойства понятия могут проявляться в наборе отношений с другими понятиями или в особенностях ассоциированных с понятием онтологических синонимов.

Таким образом, основными принципами работы с квазисинонимами являются следующие:

- необходимо искать различия между квазисинонимами, которые не исчезают в зависимости от контекста употребления квазисинонимов и приводят к формированию разных рядов онтологических синонимов или к разным отношениям с другими понятиями,
- найденные различия между квазисиноними фиксируются вводом понятий с однозначными именами.

Работу с квазисинонимами рассмотрим на примере плохо различимых синсетов из WordNet, отражающих значение сходства (см. раздел 5.5).

На первом шаге необходимо для тезаурусного описания признаки квазисинонимов, то есть такие признаки, в зависимости от которых требуется установление разных отношений с другими понятиями тезауруса.

В совокупности английских слов со значением сходства (similarity), таким элементом значения, например, является сходство по внешним характеристикам:

likeness, alikeness, similitude *-- (similarity in appearance or character or nature between persons or things;* ``man created God in his own likeness" *) – сходство по внешности, характеру или природе между людьми или объектами.*

resemblance *-- (similarity in appearance or external or superficial details).*

Это означает, что в языке значимым является сходство по внешним характеристикам и нужно отразить этот факт соответствующим понятием.

На втором шаге необходимо подыскать подходящее название такому понятию. В качестве названия может выступать однозначное словосочетание, однозначное слово с таким значением, или пара синонимов, пересечение значений которых однозначно идентифицирует данное понятие.

В случае квазисинонимов к слову similarity, таким названием понятия может служить словосочетание *Similarity in appearance* (34700 страниц в поисковой системе Google). Понятие вводится в тезаурус с таким названием.

На третьем шаге необходимо найти разные способы выражения этого же понятия в виде словосочетаний и отдельных слов, например, *resemblance in appearance, similarity of appearance, external resemblance* и др., Все эти варианты добавляются в качестве текстовых входов к понятию.

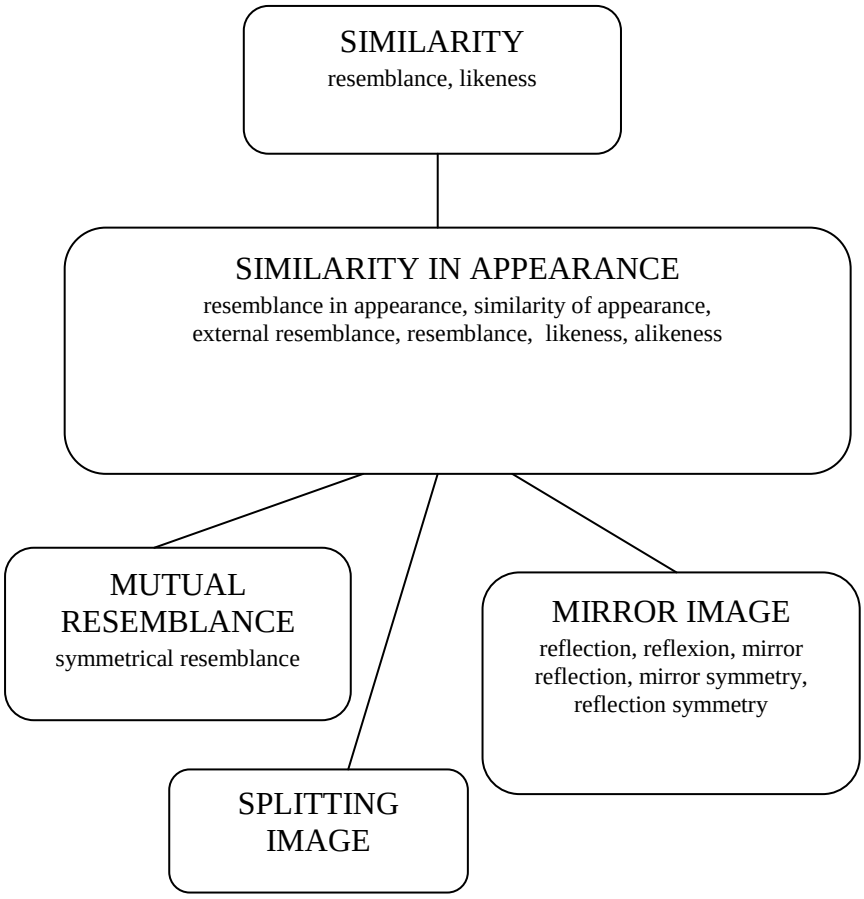

Рис 16.1. Фрагмент совокупности отличимых понятий, отражающих значения квазисинонимов слова *similarity.*

На четвером шаге для отражения значений слов, которые часто выражают именно это понятие, но могут использоваться и для выражения сходства вообще, например, *resemblance*, такое слово указывается как текстовый вход к понятию SIMILARITY IN APPEARANCE и как текстовый вход к более общему понятию SIMILARITY.

В случае если независимых от контекста характеристик для различения значений квазисинонимов, найти не удается, то необходимо представить их в виде одного понятия. Для большей ясности имя такого понятия может быть составлено, как пара соединенных в этом понятии синонимов.

В качестве основы для примера представления значений квазисинонимов на русском языке возьмем синонимические ряды, представленные в синонимическом словаре НОСС (Апресян и др., 2003). Этот словарь интересен тем, что его словарная статья содержит подробный перечень сходных черт и различий синонимов. На основе такой словарной статьи разбора удобно показать, какие различия приводят к представлению синонимического ряда словаря в виде онтологических синонимов одного и того же понятия (то есть такой же синонимический ряд сохраняется и в рамках РуТез онтологии), а для значений каких слов, представленных в данном словаре, как синонимы, введены несколько понятий, и, таким образом, в рамках тезауруса РуТез они синонимами не являются.

В качестве первого примера рассмотрим пару синонимов *памятник, монумент.*

В словаре НОСС (Апресян и др., 2003 стр. 257) указываются следующие различия этой пары слов:

- в память о конкретном человеке обычно ставится памятник, о группе людей – и памятник, и монумент, о событии – монумент; идеи воплощаются в монументах;

- у монументов есть способность увековечивать подвиг живых людей;
- по форме сооружения памятник часто представляет собой изображение увековечиваемого объекта;
- монумент обычно больше по размерам;
- пропагандистская роль больше свойственна монументам.

Анализ примеров употребления этих синонимов показывает, что различия, указанные в п.1, 2, 3, выполняются лишь по умолчанию, имеется достаточное число примеров употребления обоих синонимов в связи со всеми возможными типами увековечиваемых сущностей:

В память о конкретном человеке может быть установлен монумент:

*Монумент выдающемуся исследователю севера Западной Сибири, лесоводу, этнографу Александру Дунину-Горкавичу торжественно откроется в Ханты-Мансийске. (*http://ural.rian.ru/culture/20070614/81566803.html*).*

В память события может быть установлен памятник:

На Пролетарской площади вновь оборудован сквер, в котором установлен памятник Победы (http://www.megatula.ru/site/tulskii_krai/raionnye_centry/67/)

Памятник может быть поставлен идее:

Он сказал, что это не первая акция вандалов в отношении ***памятника*** *русско-армянской* ***дружбы*** *(http://www.patriarchia.ru/db/text/56928.html)*

Памятник может быть поставлен при жизни:

Маргарет Тэтчер, которой в Британии при ***жизни*** *поставили* ***памятник****, узнала, что американцы ее называли "провинциальной матроной". (http://www.rg.ru/2007/10/29/tetcher.html)*

Кроме того, авторы указывают, что различия нейтрализуются при повторной, сокращенной номинации того же сооружения (там же, стр.258):

На площади - установлен первый памятник нашего города - основателю Петербурга. Монумент был открыт в 1782 г.

Таким образом, у слов *монумент* и *памятник* не нашлось ни одного четкого различающего свойства или отношения, которые бы привели к отнесению значений этих слов к разным понятиям, и эти два слова должны рассматриваться как онтологические синонимы.

В качестве второй пары синонимов, которую мы проанализируем с помощью словаря НОСС (Апресян и др., 2003), рассмотрим пару слов *водитель, шофер.*

При рассмотрении этих слов авторы словаря указывают следующее различие: «шофер управляет только автомобилем или автобусом, водитель и другими транспортными средствами (стр.53)». Из этого замечания понятно, что *шофер* и *водитель* не могут быть онтологическими синонимами, поскольку водитель должен иметь отношения с понятиями, соответствующими словам *вагоновожатый, судоводитель*, а *шофер* – нет. Это означает, что для отражения значений этих слов необходим ввод, по крайней мере, двух понятий с названиями *ВОДИТЕЛЬ ТРАНСПОРТНОГО СРЕДСТВА* и *ВОДИТЕЛЬ АВТОМОБИЛЯ.* Видовыми понятиями для понятия *ВОДИТЕЛЬ ТРАНСПОРТНОГО СРЕДСТВА* будут такие понятия как *ВАГОНОВОЖАТЫЙ, СУДОВОДИТЕЛЬ*.

В то же время, носители языка ощущают эти слова как синонимы (см. также Александрова, 1999). Чтобы отразить и это ощущение, и способность расширительного

употребления, необходимо слово *водитель* представить как текстовый вход к двум понятиям *ВОДИТЕЛЬ ТРАНСПОРТНОГО СРЕДСТВА и ВОДИТЕЛЬ АВТОМОБИЛЯ.*

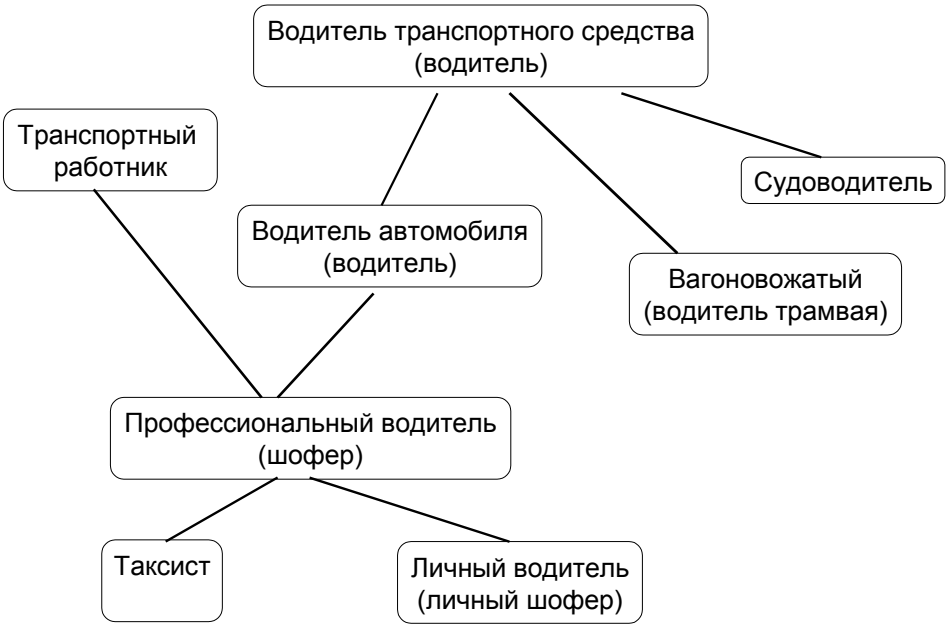

Рис. 16.2. Понятийная структура, соответствующая близким по значению словам *водитель и шофер*

Сначала представляется, что слово *шофер* должно быть отнесено как текстовое выражение к понятию *ВОДИТЕЛЬ АВТОМОБИЛЯ,* но можно заметить, что водители автомобилей могут быть любителями, и профессиональными работниками, а слово *шофер* все-таки относится к профессиональным водителям. Таким образом, онтологический анализ пары синонимов показал, что для адекватного отражения системы понятий, скрывающихся за близкими по смыслу словами *водитель* и *шофер*, нужно использовать три понятийные единицы: *ВОДИТЕЛЬ ТРАНСПОРТНОГО СРЕДСТВА, ВОДИТЕЛЬ АВТОМОБИЛЯ, ШОФЕР (ПРОФЕССИОНАЛЬНЫЙ ВОДИТЕЛЬ)* (см. рис.16.2).

Необходимость принятия решений о представлении значений близких по смыслу языковых выражений посредством совокупности понятий возникает и в конкретных предметных областях.

Так, ситуации кредитования соответствуют такие слова и словосочетания как: *кредитование, кредит, кредитная услуга, кредитное обслуживание, кредитная операция, выделение кредита, выдача кредита, выделение кредитных средств, предоставление кредита* и др. Имеется специфика употребления конкретных выражений из этого списка. Однако неправильным является введение дополнительных понятий онтологии для отражения именно специфики употребления. И в данном случае каждое вводимое понятие должно иметь четкий набор отличительных отношений. До тех пор, пока такие отличия не выделены, все такие выражения должны представляться как онтологические синонимы.

16.4. Ввод понятий для группы близких значений одного слова

С проблемой многозначности слов (лексической многозначностью) сталкиваются как разработчики онтологических ресурсов для автоматической обработки текстов, так и разработчики онтологий для других приложений.

В первом случае разработчики четко понимают, что выделение дополнительных значений в описании ляжет дополнительным грузом на систему обработки, которая должна будет делать автоматический выбор между значениями.

Разработчики понятийных ресурсов, не связанных с обработкой текстов на естественном языке, сталкиваются с проблемой многозначности в процессе анализа предметной области, когда необходимо выделить необходимый набор понятий. Эта процедура как раз и может быть затруднена лексической многозначностью, например, в таких случаях, когда значения слова значительно связаны между собой, поскольку разные значения многозначных слов, представленные как одно и то же понятие, могут некорректно вести себя в приложениях.

Как мы уже указывали в разделе 2.5.2.2., попытки объединить слишком большое количество значений WordNet, чтобы снизить проблему выбора значений при автоматической обработке текстов, не привели к выработке общепринятых критериев такого объединения. В результате исследований способов кластеризации значений WordNet в работе (Gonzalo, 2004) был сделан вывод, что ненужно склеивать, соединять близкие значения многозначных слов, правильнее прописывать отношения между этими значениями, поскольку в разных приложениях автоматической обработки текстов существенны разные типы близости значений.

В любом случае разработчик лингвистической онтологии должен иметь четкие принципы, регулирующие выделение и представление близких значений многозначных слов.

16.4.1. Принципы разделения значений в тезаурусе РуТез

В основу представления значений многозначных слов набором понятий в тезаурусе РуТез используются следующие принципы:

1) Чтобы быть отраженным в отдельном понятии, значение должно иметь независимые от контекста отличия от других значений.
2) Эти отличия выражаются, прежде всего, в наличии специфических синонимов или отношений с другими понятиями тезауруса.
3) В качестве синонимов часто хорошо проявляют отдельное значение многословные синонимы. Наличие разных синонимов является одним из важнейших факторов, делающих необходимым разделение значений и в практике составления традиционных толковых словарей (Апресян, 2006, Atkins, 1993).
4) Если для значения удается найти такие отличающие его синонимы и отношения, мы предпочитаем выделять такое значение в отдельное понятие, даже если имеется относительно близкое значение того же слова. Мы полагаем, что соединение значений с разными синонимами и отношениями в одно понятие единственно ради целей облегчения разрешения многозначности, приведет к проблемам на следующих этапах обработки текста, например, неточное отношение между понятиями может привести к неправильному логическому выводу.
5) Между понятиями, соответствующими близким по смыслу значениям, должно быть установлено онтологическое отношение, которое позволяет смягчить выбор значения в сложных случаях.

Действительно, совмещение разных значений в одном понятии приводит к тому, что у одного понятия описывается несовместимый набор отношений, например, родовидовых отношений. Именно на эту проблему указывал Н.Гуарино (Guarino, 1998), анализируя в онтологии MikroKosmos, понятие ОКНО, которому было приписано два родовых отношения к понятиям АРТЕФАКТ и МЕСТО.

В нашей практике была попытка соединить в одном понятии два значения слова *продавец*. Например, в толковом словаре (БТС, 1998) выделяются два значения слова *продавец:*

Продавец –
1. Работник магазина, отпускающий товар покупателю. Продавец универмага.
2. Тот, кто продает что-то. Продавец цветов, Продавец на рынке.

Близость такого рода значений такова, что возникает желание сопоставить этим двум значениям одну понятийную единицу.

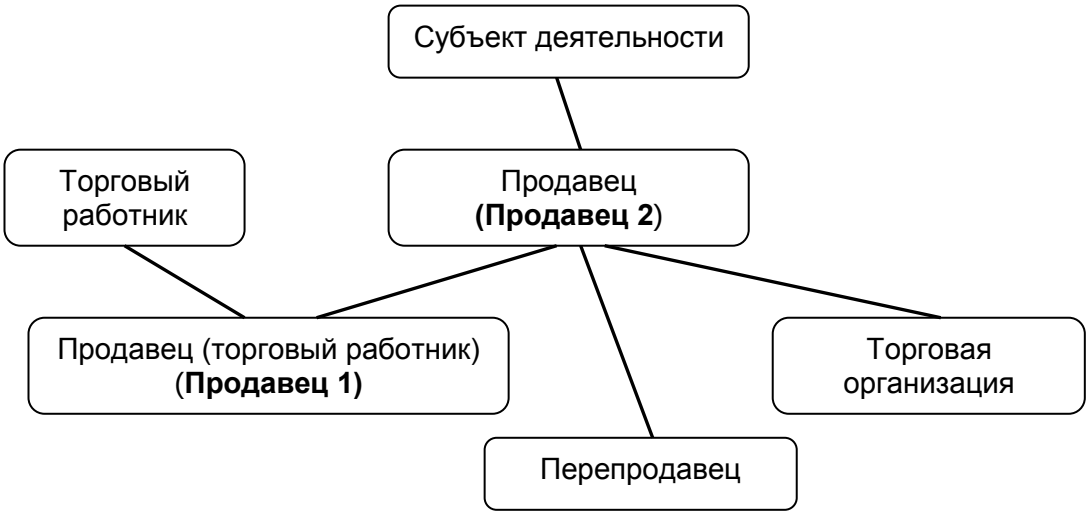

Рис. 16.3. Структура понятий тезауруса РуТез,
соответствующая значениям слова *продавец*

Однако *продавец1* имеет словосочетание-синоним *продавец магазина.* Кроме того, *продавец1* может рассматриваться как вид торговых работников, но *продавец2* не является торговым работником. Зато у *продавец2* могут быть такие виды как, например, *фирма-продавец*, которые невозможны для *Продавец1*. Отображение значений *Продавец1* и *Продавец2* как одного понятия приведет к тому, что понятие ФИРМА-ПРОДАВЕЦ окажется подвидом понятия ТОРГОВЫЙ РАБОТНИК, что приведет к проблемам в различных приложениях, использующих тезаурус. На рис. 16.3. изображено современное описание значений слова *продавец* в тезаурусе РуТез.

Рассмотрим подробнее толкование обсуждаемого в (Guarino, 1998, Nirenburg, Raskin, 2004) значения слово *окно* и применим предлагаемый нами анализ.

В Большом толковом словаре (БТС, 1998) это значение толкуется следующим образом: *отверстие в стене здания или стенке какого-л. транспортного средства для света и воздуха; застекленная рама, закрывающая это отверстие….*

Как видно, в описании одного значения совмещено рассмотрение окна как отверстия и как рамы, то есть артефакта, что имеет свою прямую аналогию в английском языке, и было воспроизведено в описаниях отношений онтологии MikroKosmos. Действительно, многие языки совмещают эти два значения в одном слове. Такое же совмещение происходит и со значениями подобных слов, например, слова *дверь.*

В то же время в языке имеются другие средства – посредством словосочетаний, четко назвать каждое из совмещенных значений, а именно, *окно как отверстие* называется *оконный проем*, *дверь как отверстие* называется *дверной проем*, *окно как артефакт* называется *оконная рама*, *дверь как артефакт* называется *дверная плита*. Совмещение значений в одном понятии делает словосочетание *оконный проем* синонимом словосочетания *оконная рама, а дверной проем* синонимом словосочетания *дверная плита,* затрудняется описание отношений с понятиями проемов и рам.

Таким образом, на наш взгляд, должны быть введены отдельные понятия ОКОННЫЙ ПРОЕМ, ОКОННАЯ РАМА с текстовым входом *окно*, а также понятия ДВЕРНОЙ ПРОЕМ, ДВЕРНАЯ ПЛИТА с текстовым входом *дверь (*см. рис. 16.4*).*

Как мы видели со значениями слова *окно,* для аккуратного описания этого значения посредством понятий и непротиворечивых отношений нам пришлось разбить на два понятия то, что было описано в толковом словаре как подзначения одного и того же значения.

Приведем еще пример значения толкового словаря, требующего при описании в онтологии разбиения на два понятия.

Для описания значения лексемы *покрывало*:

Покрывало –
1. Кусок ткани, предназначенный для покрывания чего-либо, покрывающий что-либо // легкое одеяло, обычно служащее для покрывания постели днем

должны быть введены два понятия *ПОКРЫВАЛО (ПОКРЫВАЮЩАЯ ТКАНЬ) и ПОСТЕЛЬНОЕ ПОКРЫВАЛО,* как вид первого понятия, а сама лексема *покрывало* описывается как текстовый вход к обоим понятиям. Соответствующий фрагмент тезауруса показывает, что это два действительно отдельных понятия:

ПОКРЫВАЛО (ПОКРЫВАЮЩАЯ ТКАНЬ)
с *покрывало*
НИЖЕ *НАКИДКА*
НИЖЕ *ПОПОНА*
НИЖЕ *ПОСТЕЛЬНОЕ ПОКРЫВАЛО*
с покрывало
НИЖЕ *ЧАДРА*

Как мы видим, представление значений многозначного слова посредством совокупности понятий со специфическим набором отношений может приводить к увеличению количества значений, что частично и объясняет тот феномен, что в WordNet среднее количество значений оказалось больше, чем в толковых словарях соответствующей величины.

Мы присоединяемся к мнению авторов работ (Chugur и др., 2000, Gonzalo, 2004), что часть проблем по выбору близких значений многозначных слов может быть снята, если некоторым образом установить отношение между этими значениями. Вопрос заключается в том, какого рода отношения между значениями могут быть описаны в онтологии для автоматической обработки текстов, и как их использовать в случаях неопределенности при выборе значения. Эти вопросы будут рассмотрены в следующем разделе.

16.4.2. Описание отношений между значениями многозначного слова в онтологии для автоматической обработки текстов

Как мы уже упоминали в разделе 2.5.2, многозначность слов разделяется на два основных подвида омонимию и полисемию. В свою очередь, полисемия может быть подразделена на такие подвиды как метафора, метонимия, автогипонимия, а также выделяется регулярная полисемия.

Понятно, что проблемы с разбиением на понятия возникают у полисемичных значений. При этом именно полисемия, в отличие от омонимии, рассматривается как явление коренным образом присущее языку: «Полисемия – это одно из основных средств концептуализации нового опыта. Человек не может понять нового, не имея какого-то «данного», поэтому он вынужден использовать «старые» знаки и приспосабливать их к новым функциям, распространять их на другие ситуации» (Кустова, 2004). Неслучайно,

поэтому, что термины предметной области, которые в идеале должны быть точными и однозначными, демонстрируют массовые примеры полисемических значений.

Для трех и более значений слова могут быть рассмотрены различные конфигурации связей между значениями (Кронгауз, 2001).

Значения могут иметь нетривиальную общую часть. Такой тип связи называется радиальным. Общая часть их значений называется инвариантом (или общим значением). В частном случае общая часть может совпадать с один из значений. Так устроено толкование слова *кромка* СРЯ. Фактически третье значение (кромка – вообще край чего-либо) в той или иной форме присутствует в составе первых двух.

Другим возможным типом связи, объединяющей три значения, является цепочечная связь. Значения А и Б имеют общую часть, значения Б и В имеют общую часть, значения А и В не имеют общей части.

Цепочечная полисемия представлена, например, тремя значениями лексемы *чай:*

1. *Вечнозеленое дерево или кустарник, из высушенных листьев которого приготовляется ароматный напиток*
2. *Ароматный напиток, настоянный на этих листьях*
3. *Чаепитие*

Часто отношения между значениями полисемичного слова могут быть смешанными: **радиально-цепочечными.**

Именно на полезность вышеперечисленных отношений для описания отношений между значениями в лингвистических онтологиях указывалось в работе (Gonzalo, 2004).

Однако здесь нужно не забывать о двух важных обстоятельствах.

Во-первых, онтология описывает отношения не между лексемами, а между объектами и сущностями внешнего мира, и, таким образом, между лексемами как текстовыми входами каких-то понятий могут быть установлены только онтологические отношения между понятиями. Так, например, метафорические отношения между значениями не должны находить отражение в онтологии, поскольку есть ассоциация только между названиями – между объектами никаких отношений нет. Мы не будем, таким образом, устанавливать отношения между понятиями СОТОВАЯ СВЯЗЬ и ПЧЕЛИНЫЕ СОТЫ, только по той причине, что для названия понятия СОТОВАЯ СВЯЗЬ была использована метафора с пчелиными сотами.

Во-вторых, онтология не может описывать все существовавшие ранее миры. Она содержит отношения между сущностями в существующем сейчас мире или существовавшие еще недавно. Тот факт, что когда-то данные сущности были связаны, не может быть отражено в онтологии.

Таким образом, в онтологии естественно могут быть отражены такие отношения между значениями как метонимия, поскольку «метонимическая связь, то есть связь по смежности, имеет место не между смыслами, а между объектами действительности» (Падучева, 2007), именно такое отношение существует между обсуждавшимися значениями слова *окно.* Для описания отношений между понятиями, соответствующими метонимии значений, в тезаурусе РуТез используются отношения *часть-целое* и отношение онтологический зависимости (см. гл.9, п. 17.4. и рис.16.4*).*

Другим типом отношения между значениями, которое может быть отображено посредством онтологических отношений между понятиями, является отношение автогипонии (обобщения), поскольку это отношение соответствует родовидовым отношениям между понятиями онтологии.

Отметим, при этом, что совокупности плохо разделимых значений многозначных слов, чаще всего, связаны именно с этими типами отношений между значениями.

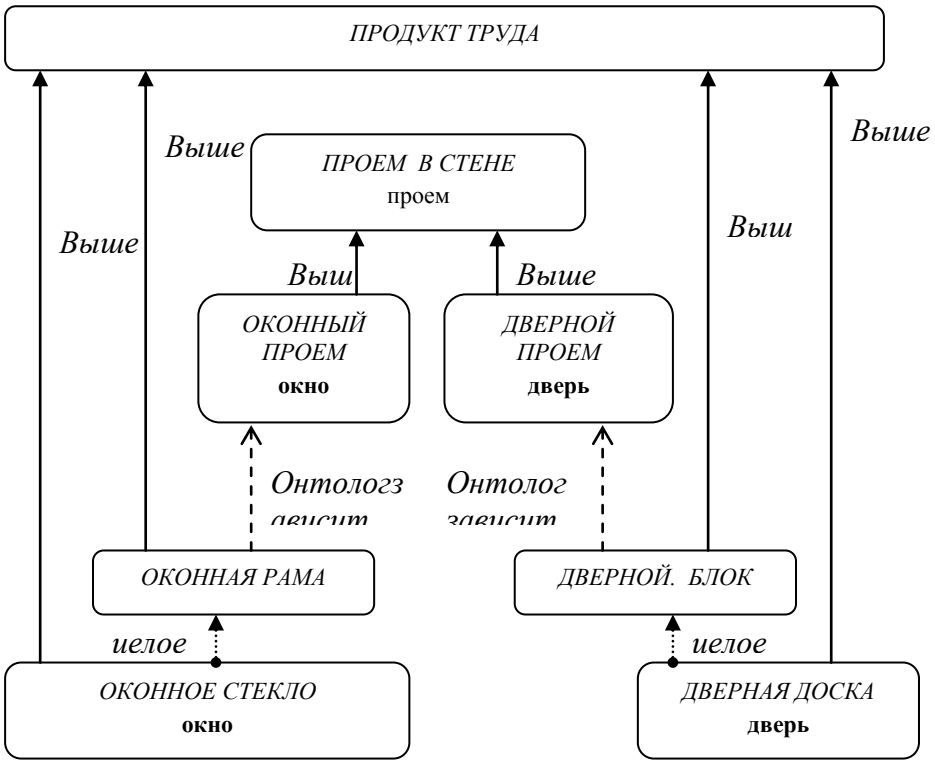

Рис. 16.4. Фрагмент понятийной сети тезауруса, представляющий значения слов *окно* и *дверь*. Над чертой указаны имена понятий, под чертой – текстовые входы.

16.5. Словосочетания как источники понятий в лингвистической онтологии

Одной из серьезных проблем взаимоотношений между понятиями лингвистической онтологии и значениями языка, является вопрос о том, в каких случаях значения словосочетаний должны быть отражены как понятия онтологии.

Проблема описания значений словосочетаний как понятий в лингвистической онтологии относится к более широкой проблеме отбора словосочетаний для описания в компьютерных словарях (Лукашевич, 1995; Bentivogli, Pianta, 2004). С одной стороны, чем больше в компьютерных словарях описано словосочетаний, тем меньше проблем с разрешением многозначности отдельных слов, больше будет зафиксировано специфических случаев сочетаемости. С другой стороны, бесконечное количество существующих словосочетаний все равно требует формулирования каких-либо критериев. Так, в работах Большакова И.А. (Большаков, 2009) предлагается набирать в специальную базу Кросс-лексика все встретившиеся словосочетания вручную. Однако возникают серьезные вопросы о полноте и представительности такой базы.

Традиционным подходом является описание в компьютерных словарях семантически связанных словосочетаний (идиом, фразеологизмов), которые демонстрируют какие-либо отклонения в синтаксическом и/или семантическом поведении (Баранов, Добровольский, 1991; Добровольский, 2005). Спектр таких устойчивых словосочетаний очень широк: от жестко фиксированных словосочетаний, которые могут рассматриваться как "слово с пробелами", до словосочетаний, которые подчиняются практически всем синтаксическим и семантическим правилам языка лишь за некоторым исключением. В последнем случае сразу обнаружить такую особенность может быть весьма сложно.

В работе (Sag и др., 2002) обсуждается еще один важный вид словосочетаний, называемых авторами институциональными выражениями. Для таких выражений характерно то, что по большей части эти выражения выглядят как свободные

словосочетания, однако их компоненты не всегда могут быть заменены синонимами. Кроме того, частотность такого словосочетания очень высока по сравнению с теми словосочетаниями, которые образованы заменой слов-компонентов на синонимы. Примером таких словосочетаний является словосочетание *phone booth (телефонная будка*). Так, и в русском, и в английском языке попытка замены слова *booth (будка*) на другие слова, например, *кабина,* приводит к многократному снижению частотности употребления.

На сложность обнаружения такого рода несвободных словосочетаний указывается в работе (Белоногов, Хорошилов, 2005). Носителям русского языка кажется, что смысл таких словосочетаний, как *электронная вычислительная машина, подводная лодка, теория массового обслуживания, сухопутные войска, военно-воздушные силы*, и смысл сложного слова *пылесос* складываются из смыслов входящих в их состав слов. На самом деле это не так. Например, русский термин *электронная вычислительная машина* обозначается на английском языке словом *computer* (*вычислитель*), в котором ничего «электронного» в явном виде не отмечается; русский термин *подводная лодка* обозначается сложным словом *submarine* (буквально «под морем»), в котором понятие «лодка» отсутствует; термин *теория массового обслуживания* - словосочетанием *querying theory* (*теория очередей*), которое не содержит явных признаков понятия «массовое обслуживание»; термин *сухопутные войска* - сложным словом *land-forces* (*наземные силы*), без признаков “сухопутности”; термин *военно-воздушные силы* - словосочетанием *air forces* (буквально «воздушные силы»), в котором понятие «военный» в явном виде не обозначено; термин *пылесос* - словосочетанием *vacuum cleaner* (буквально «вакуумный очиститель» в составе которого нет понятий «пыль» и «сосать»).

В работе (Белоногов, Хорошилов, 2005) этот феномен объясняется тем, что в словесных формулировках наименований понятий могут быть отображены не все признаки понятий, а только незначительное их число. Часто это бывают не самые важные признаки, характеризующие содержание понятий, а лишь некоторые отличительные признаки, позволяющие выделить эти понятия среди множества других.

В естественном языке, в котором «все связано со всем», понятия, как некоторые социально значимые устойчивые мыслительные образы, могут обладать огромным количеством признаков. Но этим мыслительным образам присваиваются наименования в виде отдельных слов или (значительно чаще) в виде словосочетаний, состоящих из нескольких слов. Наименование понятия, на основе некоторых частичных признаков исходной сущности, приводит к тому, что сущность может быть именована разными способами на основе разных признаков, и тогда возникает синонимия, которая практически никак не следует из значений отдельных слов.

В (Белоногов, Хорошилов, 2005) приводятся следующие примеры таких синонимов-словосочетаний:

Абсолютная жесткость – бесконечно большая жесткость,

Абсолютная температура – температура Кельвина,

Наклонный путь для сортировки вагонов – путь сортировочной горки

Наклоны головы в поперечной плоскости – наклоны головы к правому и левому плечу.

В (Тер-Минасова, 2007) указывается, что имеется большое количество словосочетаний, которые не являются явно заранее данными, но «свобода» образования которых ограничена какими-либо факторами. Поэтому исследование образований этого рода, то есть таких, которые, с одной стороны, не являются фразеологическими единицами, а с другой – не обладают способностью вполне свободно создаваться в речи, связано с большими трудностями. Так, даже простейшие, казалось бы, абсолютно свободные словосочетания *blue sky (голубое небо), white tablecloth (белая скатерть)* явно

социолингвистически обусловлены. В культурном опыте говорящих по-английски предметы мысли tablecloth и white сочетаются вполне естественно, привычно, закономерно, так как скатерти белого цвета общеприняты и широко распространены, кроме того, наличие белой скатерти, как правило, свидетельствует, о торжественном или официальном приеме, праздничном обеде и т.п. Точно также сочетание *blue sky* … употребляется в речи настолько регулярно, что вряд ли можно говорить от свободном сочетании элементов *blue* и *sky* каждым отдельным носителем языка.

В предметных областях вопрос об извлечении словосочетаний обычно обсуждается как вопрос извлечения терминологических словосочетаний. Имеющиеся терминологические списки, относящиеся к текущей предметной области, обычно охватывают лишь малую часть тех терминоподобных словосочетаний, которые встречаются в текстах. Эксперты предметной области могут иметь очень различные мнения по поводу терминологичности того или иного словосочетания (Браславский, Соколов 2007).

Таким образом, спектр словосочетаний с особенностями очень широк, и нужны некоторые формализованные принципы для отражения словосочетания в компьютерных словарях, и, в частности, в структуре понятий лингвистической онтологии.

16.5.1. Принципы, предлагаемые для отбора словосочетаний для включения в словари систем автоматической обработки текстов

В работах (Bentivogli, Pianta, 2004; Calzolari и др., 2002; Pearce, 2001) обсуждается совокупность принципов, которые могут служить (в сочетании) основанием для внесения словосочетания в компьютерный словарь:

- высокая частотность,
- высокая степень ассоциации, то есть более частое употребление друг с другом, чем с другими словами,
- синонимичность лексической единице (например, отдельному слову),
- значительная многозначность компонентов,
- словосочетание обозначает тип объекта, например, *телефонная будка, письменный стол.* Именно типы объектов обладают набором разных свойств, многие из которых могут быть использованы для называния этого типа, в результате чего возникают интересные синонимы, интересные переводы на другой язык (см. предыдущий раздел)

В работе (Pearce, 2001) предлагается использовать для извлечения устойчивых словосочетаний синонимы, описанные в тезаурусе WordNet. Поскольку одним из частых свойств семантически связанных словосочетаний является ограничение на замену одного из слов словосочетания синонимом, то предлагается исследовать сочетания синонимов с одними и теми же словами по корпусу, затем перепроверять в Интернет. Если разница частотностей таких словосочетаний значительна, то можно предлагать частотное словосочетание как устойчивое. Например, сравнивая употребление слов-синонимов *baggage* и *luggage* в сочетаниях с различными словами, можно обнаружить, что только *baggag*e употребляется с таким прилагательным как *emotional.* Таким образом, можно предположить, что словосочетание "*emotional baggage*" является устойчивым.

Как указывалось в разделе 1.1.2, разработчики информационно-поисковых тезаурусов традиционно выделяют особое внимание отбору многословных терминов для включения в тезаурусы.

Так, ГОСТ 7.25 указывает, что допускается включать словосочетания в тезаурус, если в качестве опорного слова они содержат существительное и если выполнено одно из следующих условий:

- значение словосочетания не выводится из значений его компонентов (*черный ящик*);

- хотя бы один из компонентов словосочетания не употребляется в составе других сочетаний или употребляется всегда в другом смысле (*торговля на вынос*),
- для данного словосочетания в словнике ИПТ существуют полные синонимы,
- отдельные слова словосочетания имеют слишком широкое значение,
- имеется общепринятая аббревиатура.

Американский стандарт Z39.19 помимо вышеперечисленных случаев приводит также критерий общепринятости термина профессиональным сообществом, например, *data processing - обработка данных*. Кроме того, этот стандарт указывает, что введение многословного дескриптора позволяет избегать ложных корреляций, например, разбиение термина *Library science* (*наука о библиотеках = библиотековедение*), может привести к нахождению документов о научных библиотеках (*science library*).

Таким образом, мы видим, что различные авторы предлагают различные критерии и соображения для включения многословных конструкций в словари компьютерных систем, что значительно затрудняет принятие решения в конкретных случаях.

16.5.2. Ввод понятий тезауруса РуТез на основе значений многословных выражений

Тезаурус РуТез содержит большое количество понятий, которые соответствуют значениям словосочетаний. Критерием ввода такого понятия является возможность отражения в этом понятии информации, которую невозможно или трудно выразить, используя понятия, соответствующие отдельным словам этого словосочетания. Таким образом, новое понятие в тезаурусе – это точка приложения дополнительной информации, которую система автоматической обработки текстов может использовать в процессе своей работы. При этом частотные и другие статистические характеристики употребления словосочетания не являются, в подавляющем большинстве случаев, решающими основаниями включения соответствующего понятия в тезаурус. Такие характеристики служат обычно лишь дополнительными факторами, заставляющими обратить внимание на словосочетание.

Информацию, которую может фиксировать дополнительно введенное понятие, можно разделить на несколько видов (Добров и др., 2002b).

16.5.2.1. Существует и важно

В любой предметной области существует небольшое число сущностей, которые очень важны в данной ПО. Соответствующие им термины и другие языковые выражения очень частотны в текстах области. Такие сущности должны быть отражены в онтологии. Например, в общественно-политической жизни Российской Федерации важны такие понятия как *ПРЕЗИДЕНТ РОССИЙСКОЙ ФЕДЕРАЦИИ, ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ.*

Если введенные в онтологию понятия имеют фиксированное число видовых понятий, то они должны быть отражены в онтологии.

Еще одним важным видом информации является информация о том, что у двух понятий имеется общий подвид, например,

АДМИНИСТРАТИВНЫЙ ШТРАФ является видом понятий *АДМИНИСТРАТИВНОЕ НАКАЗАНИЕ и ШТРАФ.*

Кроме того, понятия, которые формулируют основания для видового деления некоторого родового понятия, также являются важными для онтологии.

Например, кредиты делятся на рублевые и валютные, краткосрочные и долгосрочные. Поэтому важными понятиями онтологии кредитной сферы являются понятия *ВАЛЮТА КРЕДИТА* и *СРОК КРЕДИТА*.

16.5.2.2. Словосочетание имеет «интересные» синонимы

Обнаружение синонимических текстовых входов, в том числе сокращенных слов, аббревиатур часто приводит к введению нового понятия для фиксации найденной синонимии. При этом разнообразие синонимичных текстовых выражений часто указывает на важность соответствующего понятия.

Например, словосочетание *газовая колонка* имеет синоним *газовый водонагреватель.* Словосочетание *покупательская тележка* имеет синонимы *магазинная тележка и тележка для покупок.*

Словосочетание *работник профсоюза* имеет такие синонимы как *профсоюзник* и *профработник, а лекарство растительного происхождения* - синоним *фитопрепарат.*

Подробнее о видах онтологических синонимов см. раздел 16.6.1.

16.5.2.3. Отношения, которые не следуют из структуры словосочетания

Принципом, используемым для оценки необходимости ввода понятия во многие тезаурусы и онтологии, является то, что многословный термин имеет отношения, которые не следуют из его структуры. Такие отношения могут быть по отношению к нижестоящим по иерархии понятиям:

ФАЗЫ ЛУНЫ – ПОЛНОЛУНИЕ, НОВОЛУНИЕ

ИЗБРАНИЕ ПАПЫ РИМСКОГО - КОНКЛАВ

Или к вышестоящим понятиям:

НАЛОГОВОЕ СТИМУЛИРОВАНИЕ - ЛЬГОТНОЕ НАЛОГООБЛОЖЕНИЕ, НАЛОГОВАЯ ПОЛИТИКА

ПОЛОВАЯ ДОСКА - НАПОЛЬНОЕ ПОКРЫТИЕ

Понятие может иметь как вышестоящие, так и нижестоящие отношения, не следующие из структуры его текстовых входов:

НАПОЛЬНОЕ ПОКРЫТИЕ – ОТДЕЛОЧНЫЕ МАТЕРИАЛЫ, ЛИНОЛЕУМ, ПАРКЕТ

16.5.2.4. Достройка уровней тезауруса

Важным принципом дополнения тезауруса является принцип «замыкания», который имеет два подвида.

Во-первых, если новое, по тем или иным причинам введенное понятие породило новый нижестоящий уровень тезауруса, то он должен быть дополнен другими существенными понятиями того же уровня. Например, если как нижестоящие понятие для понятия *ПРИМЕНЕНИЕ ОРУЖИЯ* вводится понятие *ПУСК РАКЕТЫ,* то необходимо вводить, например, понятие *СТРЕЛЬБА ИЗ ПУШКИ*, как второй важнейший вид применения оружия в данной области.

Этот принцип является одновременно и ограничивающим: если мы собираемся ввести понятие нового уровня, то мы должны оценить последствия этого шага: сколько еще понятий этого уровня мы собираемся ввести; если потенциальных понятий этого уровня слишком много, то нужно сразу оценить, на основе каких принципов мы ограничим ввод.

С другой стороны, может возникнуть и обратная ситуация: найдено и введено несколько понятий, имеющих общие черты, необходимо найти обобщающее понятие. Например, в онтологию вводятся понятия ЗАГОРОДНАЯ *ДАЧА, ВИЛЛА, УСАДЬБА,* характерными особенностями которых является то, что они представляют разные виды

загородного жилья - тогда обобщающим понятием может служить понятие ЗАГОРОДНЫЙ ДОМ.

16.5.2.5. Словосочетание однозначно, а его компоненты многозначны

Фактор неоднозначности слов-компонентов однозначного словосочетания может также приводить к вводу понятий.

Это происходит, например, в тех случаях, когда однозначное словосочетание состоит из слов с большим количеством разных значений, например, о*перация со счетом, ведение счета, снятие со счета, ведение огня, принять в ведение, перевод средств*

В других случаях один из компонентов словосочетания мог бы сочетаться с любым значением другого слова-компонента словосочетания. Тем не менее, в реальности словосочетание употребляется только с одним значением многозначного слова-компонента. Например, в словосочетании *одноразовый станок* слово *станок* может быть только в значении *станок для бритья.*

16.5.2.6. Ввод понятия на основе сочинительной конструкции

В некоторых случаях удобно ввести понятие, которому в языке соответствует сочинительная конструкция, например, *ГОСУДАРСТВЕННАЯ И МУНИЦИПАЛЬНАЯ СОБСТВЕННОСТЬ, ОПЕКА И ПОПЕЧИТЕЛЬСТВО.*

Ввод таких понятий связан с выполнением нескольких критериев:

- у такого понятия имеются характерные свойства, атрибуты, отношения, которые отличают его от вышестоящих и нижестоящих понятий.
- текстовый вход, значение которого послужило основой введения понятия, очень частотен, например, при поиске в поисковой системе Яндекс на запрос "государственная и муниципальная собственность" находится более 17 тысяч документов
- понятие может иметь и другие текстовые входы более традиционной формы, но возможно менее частотные, например, синонимом выражения «государственная и муниципальная собственность» может быть рассмотрен термин «публичная собственность», менее распространенный в современной российской деловой прозе экономики и права,
- существуют частотные термины, в которые входит такое выражение, например, *орган опеки и попечительства,*
- у исходного термина существует перевод на другой язык в виде отдельного слова или именной группы, например, английскими текстовыми входами понятия ОПЕКА И ПОПЕЧИТЕЛЬСТВО являются термины *guardianship* и *legal guardianship.*

Встречаются и обратные ситуации, когда в русском языке, имеется однословный термин, а на другом языке его можно передать только с помощью сочинительной конструкции, которая, действительно, очень часто используется.

Например, в русском языке имеется термин *вексель*, который подразделяется на *простой вексель* и *переводной вексель*. В английском языке *простой вексель* называется *promissory note*, а *переводной вексель* - bill of exchange. Имеется международная конвенция по векселям, в которой обсуждаются оба вида векселей, но не используется никакой обобщающий термин. Эквивалент все-таки имеется, это выражения *bills and notes и bills of exchange and promissoty notes.* На запрос ”*bills and notes*” поисковая система Google находит 140 тысяч документов, а на запрос “*bills of exchange and promissory notes*” 40 тысяч документов.

16.5.2.7 Перестановка слов ведет к разным понятиям

Если перестановка слов или соответствующих понятий в частотном словосочетании ведет к совершенно другому частотному словосочетанию, то необходимо зафиксировать соответствующие понятия. Так, например, *работник профсоюза* при смене порядка слов превращается в *профсоюз работников,* и это является дополнительным основанием для ввода понятия РАБОТНИК ПРОФСОЮЗА.

Так, на дату 8 октября 2009 среди первых 10 документов выдачи поисковой системы Google по запросу *работник профсоюза* только 3 документа из 10 релевантны этому запросу. В поисковой системе Яндекс на первой странице выдачи по этому запросу не было ни одного релевантного документа.

Глава 16.6. Языковые выражения как текстовые входы понятий.

Каждое вводимое понятие должно быть снабжено списком слов и словосочетаний, с помощью которых можно сослаться в тексте на вводимое понятие – текстовых входов. В качестве таких текстовых входов могут быть отдельные слова (существительные, прилагательные, глаголы), а также именные и глагольные группы. Текстовый вход может быть многозначным (иметь другие значения), тогда он должен быть помечен как многозначный. Для лучшего распознавания в тексте текстовые входы тезауруса РуТез снабжаются последовательностью нормализованных форм всех составляющих многословного выражения (мужской род, именительный падеж, единственное число).

Языковые выражения (слова, словосочетания, термины), которые были описаны как текстовые входы одного и того же понятия, становятся неразличимыми с точки зрения РуТез онтологии – онтологическими синонимами.

В тезаурусе РуТез большое значение уделяется работе со словосочетаниями не только как с источниками новых понятий, но и в качестве пополнения синонимических рядов. Поскольку в процессе нашей работы выяснилось, что многие слова имеют многословные синонимы, то такие синонимы специально ищутся и ими пополняются синонимические ряды текстовых выражений, связанных с понятием. Такие многословные синонимы особенно важно найти для многозначных слов, поскольку многословные синонимы уже становятся однозначными (подробнее см. раздел 16.6.3).

Как уже указывалось в разделе 16.1, онтологические синонимы не всегда являются синонимами в том смысле, что не всегда возможны замены в предложении одного онтологического синонима на другой, сохраняющие грамматическую правильность и логическую истинность предложения. Однако онтологические синонимы понятия должны быть эквивалентны относительно отношений этого понятия с другими понятиями тезауруса. Как показала практика, нарушение этого принципа, неаккуратное объединение языковых выражений в рамках ряда онтологических синонимов, рано или поздно проявляет себя в ошибках при автоматической обработке текстов, находится приложение, для которого неучтенное различие языковых выражений оказывается существенным.

16.6.1. Типы онтологических синонимов

Рассмотрим основные типы онтологических синонимов:

1. Лексические синонимы (собственно синонимы):
а) полные синонимы (в том числе синонимы-дублеты):

аванс — предоплата,
космонавт — астронавт,
мятеж — бунт;

б) синонимы, отражающие различные языковые стили:

лошадь — конь,
коммунальная квартира — коммуналка;

в) синтаксические синонимы:

жилищное строительство — строительство жилья,
авария на транспорте — транспортная авария,
контроль за вооружениями — контроль над вооружениями,

г) словосочетания, синонимичные отдельным словам:

болид – космический болид,
болид – гоночный болид;

2. Словообразовательные онтологические синонимы,
а) словообразовательные варианты:

калькуляция — калькулирование,
природоохранный — природоохранительный;

б) дериваты:

приватизация — приватизировать,
охрана природы — природоохранный;

в) видовые пары глаголов:

видеть – увидеть,
снять – снимать;

г) уменьшительные формы существительных:

стол – столик,
двор – дворик;

д) глаголы-делимитативы (Зализняк, Шмелев, 2000):

гулять – погулять,
читать – почитать,
бегать – побегать;

е) однократные и многократные действия:

куснуть – кусать,
моргнуть – моргать,
плюнуть – плевать.

3. Общепринятые в информационно-поисковых тезаурусах условные синонимы:
а) сокращения:

врачебно-трудовая экспертная комиссия — ВТЭК,
автозаправочная станция — АЗС;

б) сложные и сложносокращенные слова:

строительные материалы – стройматериалы
жилищный фонд — жилфонд,
авиационная охрана лесов — авиалесоохрана;

в) некоторые антонимы:

доверие правительству — вотум недоверия правительству,
правовое обеспечение — правовой вакуум;

г) существительные, обозначающие лиц мужского и женского пола:

спортсмен — спортсменка,
владелец — владелица.

4. Другие типы:

а) образные наименования:

авианосец — плавучий аэродром,
взрывные работы — мирный взрыв,
биржевая операция — игра на бирже,
атомная энергетика — мирная ядерная деятельность,
аэропорт — воздушные ворота;

б) фрагменты толкования (используются только в случае реального употребления в текстах):

банковская тайна — тайна банковского счета,
боеголовка — головная часть индивидуального наведения;

в) энциклопедические синонимы, то есть такие языковые выражения, тождественность которых вытекает из "энциклопедических знаний", поскольку известно, что для сокращенного выражения нет других интерпретаций:

альтернативная гражданская служба — альтернативная военная служба — альтернативная служба,
внутренние войска — войска МВД,
космический корабль многоразового использования – корабль многоразового использования – многоразовый корабль;
плавающая процентная ставка – плавающая ставка

г) словосочетания с исключением внутреннего компонента (синонимы типа в) и г) названы в (Гринев-Гриневич, 2008) эллиптическими синонимами:

безналичный порядок расчета — безналичный расчет,
вечерняя форма обучения — вечернее обучение,

д) словосочетания, представляющие собой различные реализации одного из актантов главного слова термина:

встреча на высшем уровне — встреча в верхах,
автомобиль инвалида — автомобиль с ручным управлением,
призыв в армию — призыв на воинскую службу;

е) словосочетания, несущие в себе дополнительную модальность по отношению к основному словосочетанию:

хирургическая операция — хирургическая помощь — хирургическое вмешательство;

ж) словосочетания, совпадающие в одной своей части, а в другой — состоящие из ситуационно связанных слов:

безопасность судоходства — безопасность кораблей — безопасность на море,
защита вкладов — защита вкладчиков.

16.6.2. Формирование синонимического ряда понятия

Понятия в тезаурусе РуТез могут иметь достаточно большие ряды онтологических синонимов. Приведем пример синонимического ряда, включающего несколько типов синонимов для понятия ОХРАНА ПРИРОДЫ (по алфавиту):

Защита окружающей природной среды
Защита природной среды
Защита природы
Защищать природу
Охрана природной среды
Охрана природы
Охранять природу
Природозащита
Природозащитный
Природоохранный
Природоохранительный
Природоохранная деятельность
Природоохранная работа
Природоохранные мероприятия
Природоохранные меры
Сохранение окружающей природной среды
Сохранение природной среды
Сохранение природы
Сохранять природу
Сохранять природную среду

Как видно, синонимический ряд понятия может содержать значительно количество синтаксических вариантов словосочетаний, некоторые словосочетания образуются заменой слова-компонента на синоним. Установление соответствия таких текстовых входов понятию является наиболее простым способом обнаружения понятия в тексте

Хранение таких синтаксических синонимов не предусматривается в традиционных информационно-поисковых тезаурусах, поскольку они были предназначены для ручного индексирования индексаторами, которые легко могут обнаруживать такие варианты в тексте.

Однако, понятно, что автоматически такие варианты обнаруживать может быть сложно, поскольку не все возможные варианты реализуются в тексте, некоторые из них меняют значение. Например, слова *объект* и *предмет* являются синонимами в одном из значений, но словосочетания *учебный предмет* и *учебный объект* имеют разные значения. В английском языке замена слова *forest* на близкое по смыслу слово *wood* в словосочетании *forest fire* (*лесной пожар*), приводит к совершенно другому значению словосочетания: *wood fire* (*дровяное отопление).*

Поэтому при ведении тезауруса РуТез важным правилом является зафиксировать максимальное число реально существующих онтологических синонимов. При вводе нового понятия в онтологию:

- необходимо предложить максимально возможное число разного рода синонимических текстовых входов вводимого понятия,
- проверить реальное употребление предложенных языковых выражений в текстах Интернет. Для ввода выражения необходимо, чтобы данное выражение употреблялось, по крайней мере, в нескольких сотнях разных документов Интернет, относящихся к современной деловой прозе.

В ходе различных экспериментов, при тестировании компьютерных приложений на основе тезауруса при обнаружении языкового выражения, которое может быть рассмотрено как новый текстовый вход существующего понятия, оно обязательно фиксируется в соответствующем синонимическом ряде.

16.6.3. Словосочетания, синонимичные отдельным словам

Большое количество отдельных лексем могут иметь синонимы-словосочетания.

Найденные многословные синонимы могут служить хорошими кандидатами на название понятия, ясно и однозначно выражая содержание понятия. Однозначные словосочетания, синонимичные отдельному многозначному слову, могут в значительной мере помочь в автоматической процедуре разрешения многозначности. Наконец, при анализе значений слов с плавающим значением или группы близких по смыслу слов, использование многословных конструкций позволяет выделить в этой группе сложно связанных значений отчетливые подразделения и зафиксировать эти подразделения в виде совокупности понятий (см. раздел 16.4.1.).

Рассмотрим подробнее типы словосочетаний, синонимичных отдельным словам.

Большинство словосочетаний, синонимичных отдельному слову, включают в свой состав это слово или его дериват.

Известными примерами таких словосочетаний являются, описанные в (Мельчук, 1974), словосочетания с использованием родовых понятий вида Gener (C0)->Q(C0), где Q(C0) – обозначает некоторый дериват от C0, например, *республика = республиканское государство* [C0=*республика,* Gener (C0) = *государство*, Q(C0) = *республиканский*].

Известным видом словосочетаний, синонимичных глаголам и часто являющихся однозначными, являются фразеологические синонимы *оказать помощь=помочь, оказать сопротивление=сопротивляться, принимать решение – решать*.

На самом деле, словосочетания, синонимичные значениям многозначного слова, весьма разнообразны. Часто они образуются из исходного слова или его деривата и из наиболее значимого слова из толкования.

Например, в (БТС, 1998) первое значение слова *агрессия* толкуется следующим образом: «вооруженное нападение государства или группы государств на какое-то государство…». Как синоним этого значения слова *агрессия* активно употребляется словосочетание *вооруженная агрессия.*

Часто у каждого из значений многозначного слова имеется свой однозначный синоним-словосочетание.

Например, слово *болид* имеет два значения (БТС, 1998):

1. Очень яркий крупный метеор
2. Гоночная машина со сверхмощным двигателем.

Соответственно достаточно употребительны словосочетания *космический болид* как синоним к первому значению слова и *гоночный болид* как синоним ко второму значению.

Если рассматривать основные типы структур словосочетаний-синонимов к многозначным существительным, то подавляющее большинство таких словосочетаний представляют собой следующие конструкции (исходное слово C0):

- A(C0)+Gener(C0):
  авангард3 = авангардное искусство, архив1 = архивное учреждение, авиация2 = авиационная техника, экология2 = экологическая система;
- Gener(C0)+C0 в родительном падеже:
  авангард3 = искусство авангарда, авангард4 = произведения авангарда, экспедиция2 = отдел экспедиции, чай3 = настой чая.

Такие конструкции становятся возможными из-за метонимической связи между значениями: внутри словосочетания многозначное слово обычно употребляется в значении, отличном от значения целого выражения.

- C0+(существительное в родительном падеже) или прилагательное+C0.

Зависимые существительные и прилагательные могут в таких словосочетаниях выражать достаточно широкий спектр характеристик значения слова, например:

- его целое (*бородка2 = бородка ключа*),

- происхождение (*болид1 = космический болид, челюсть2 = искусственная челюсть*),
- назначение (*блок1 = подъемный блок, бревно2 = гимнастическое бревно*),
- типы его актантов (*арест2 = арест имущества, адаптация2 = адаптация текста*),
- а также другие значимые характеристики (*карьер2 - открытый карьер, брак1 – зарегистрированный брак*).

Реже встречаются конструкции с предлогами, которые обычно передают назначение предмета: *экран2=экран для показа, штопор1=штопор для бутылок.*

Предложные конструкции синонима-словосочетания также могут основываться на метонимии значений слова*: шахматы1=игра в шахматы, шерсть4= ткань из шерсти.*

Таким образом, явление активного употребления однозначных словосочетаний-синонимов для многозначных слов достаточно распространено. При этом для каждого конкретного многозначного слова достаточно трудно предсказать, существуют ли для его значений однозначные синонимы-словосочетания. Их существование приходится проверять по текстовым корпусам и в сети Интернет.

Поскольку владение такими синонимами-словосочетаниями кажется значимым фактором при автоматическом разрешении многозначности, то мы, разрабатывая Тезаурус русского языка РуТез, предназначенный для автоматической обработки текстов, специально ищем такие однозначные словосочетания и добавляем их в синонимические ряды соответствующих значений. Критерием добавления служит нахождение более 100 интернет-страниц, в которых упомянуто такое словосочетание.

16.6.4. Описание многозначности языковых единиц в тезаурусе РуТез

В Тезаурусе РуТез существуют два основных способа представления значений многозначных терминов.

Первым способом представления многозначности является задание одного и того же текстового входа разным понятиям тезауруса (М-многозначность). Так, например, текстовый вход *пилот* сопоставлен двум разным понятиям понятию *ЛЕТЧИК* и понятию *АВТОГОНЩИК*.

Такое представление используется для задания разных видов лексической многозначности:

- омонимии: слово *брак* соответствует таким понятиям как *СУПРУЖЕСТВО* и *ПРОИЗВОДСТВЕННЫЙ БРАК,*
- терминов из разных предметных областей: слово *прокат* соответствует таким понятиям как *ПРОКАТНОЕ ПРОИЗВОДСТВО* (металлургия), *КИНОПРОКАТ (кинематография), ПРОКАТ ИМУЩЕСТВА* (аренда).
- метонимии: слово *балет* относится к таким понятиям как *БАЛЕТНОЕ ИСКУССТВО (развитие балета), БАЛЕТНЫЙ СПЕКТАКЛЬ (смотреть балет), БАЛЕТНАЯ ТРУППА (приезд балета*). Отметим, что обычно понятия, которым соответствует один и тот же текстовый вход, образованный на основе явления метонимии, связаны между собой тезаурусными отношениями.
- метафоры: слово *сотовый* соответствует понятиям *СОТОВАЯ СВЯЗЬ* и *ПЧЕЛИНЫЕ СОТЫ.*

Второй способ представления многозначности используется в тех случаях, когда слово представлено в Тезаурусе в одном значении, но если известно, что оно может употребляться и в других значениях в целевых текстах, то ему ставится специальная пометка многозначности (А-многозначность).

Например, для слова *уклонист* – Большой толковый словарь дает толкование: *тот, кто уклоняется от участия в чем-либо*. Однако в текстах современной деловой прозы имеется практически единственное употребление в смысле «уклонист от призыва в

армию». В таких случаях можно помещать слово *уклонист* как текстовый вход к соответствующему понятию с пометкой многозначности: превалирующее значение отражено, а при появлении этого слова в другом контексте, соответствующее понятие выводиться не будет.

Пометка многозначности часто используется для отметки географических названий, которые могут совпадать с фамилиями и именами людей, сокращениями и др., например, *Львов* (город), *Владимир* (город), *Павлово* (город в Нижегородской области).

В настоящее время тезаурус РуТез содержит более 15 тысяч многозначных единиц, из них для более 11 тысяч слов представлено несколько значений (М-многозначность), многозначность остальных отмечена пометкой.

В составе Общественно-политического тезауруса насчитывается около 6.5 тысяч многозначных терминов. Для 2204 терминов представлено два и более значений.

В качестве примера покрытия газетного текста единицами тезауруса РуТез и Общественно-политического тезауруса рассмотрим следующий фрагмент статьи из «Независимой газеты» от 23 ноября 2003 года под названием «Первый бриллиант Александра Волошина»:

> В ***понедельник*** на ***сцене Большого театра сверкали*** "***Бриллианты американского балета***". ***Концерт*** был ***посвящен*** 70-***летию установления дипломатических отношений*** между ***Россией*** и ***США***. В ***зале сидели*** все ***мыслимые*** и ***немыслимые дипломаты*** с обеих ***сторон***. В этот ***вечер спектакль разыгрывался*** по обе ***стороны рампы, точнее, оркестровой ямы.*** И второй, надо ***сказать***, был ничуть не менее ***захватывающим***. Пока на ***сцене звезды американского балета показывали чудеса хореографической техники***, в ***противоположной стороне партера***, в ***царской ложе, светила*** другая, куда более ***загадочная звезда.***

Полужирным шрифтом выделены слова, которые включены в качестве единиц в тезаурус РуТез. Видно, что практически вся содержательная лексика включена в анализ.

Подчеркнутые слова входят в тематический подтезаурус – Общественно-политический тезаурус. Фрагмент содержит группы единиц тезауруса, относящихся к зрительному залу: *сцена, зал, рампа, оркестровая яма, ложа, партер*, а также к искусству*: концерт, балет, Большой театр, хореографический*, что дает возможность использования этой информации для разрешения многозначности.

Относительно Общественно-политического тезауруса фрагмент содержит 25 тезаурусных единиц, из них 15 многозначных. Такие слова, как *звезда (небесное тело), техника (техническое устройство), зал (общественное помещение), партер (зрительного зала)* представляют пример А-многозначности, то есть их другие значения не входят в состав Общественно-политического тезауруса, а многозначность отмечена только специальной пометкой.

Относительно Тезауруса РуТез все многозначные слова имеют М-многозначность, за исключением слова *партер,* другие значения которого на момент обработки еще не были описаны.

Заключение к главе 16.

Развивая тезаурус РуТез как лингвистическую онтологию, мы пытаемся следовать двум, вообще говоря, противоречивым критериям.

С одной стороны, мы формируем понятия тезауруса максимально близко к значениям языковых выражений, поскольку считаем, что чрезмерное обобщение, кластеризация значений ведет к искажению системы отношений, проблемам в приложениях автоматической обработки текстов.

С другой стороны, мы стараемся, чтобы понятие тезауруса было действительно понятием, то есть было отличимо от близких по смыслу понятий.

Во многих случаях использованием реально существующих многословных выражений позволяет нам смягчить эти противоречивые требования. Введение понятия на базе значения многословного выражения не меняет суть лингвистической онтологии, но во многих случаях позволяет ввести более отчетливо отделимые понятия.

Использование в качестве единиц тезауруса таких отличимых понятий позволяет единым образом представлять лексические значения литературного языка и значения терминов предметной области, более последовательно описывать систему отношений между понятиями и тем самым облегчает формальный вывод на отношениях.

Для понятия онтологии, которое четко отделимо от других близких понятий, значительно легче найти эквивалентные названия на языках, отличных от исходного языка лингвистической онтологии. Таким образом, хорошо отличимые понятия делают лингвистическую онтологию более языково-независимой. В то же время учет переводных эквивалентов в других языках позволяет лучше увидеть недостаточную отделимость понятий лингвистической онтологии.

Онтологические синонимы, то есть текстовые выражения, сопоставленные одному и тому же понятию, не всегда являются синонимами в том смысле, что не всегда возможны замены в предложении одного онтологического синонима на другой, сохраняющие грамматическую правильность и логическую истинность предложения. Однако онтологические синонимы понятия должны быть эквивалентны относительно отношений этого понятия с другими понятиями тезауруса. Как показала практика, нарушение этого принципа, неаккуратное объединение языковых выражений в рамках ряда онтологических синонимов, рано или поздно проявляет себя в ошибках при автоматической обработке текстов, рано или поздно находится приложение, для которого неучтенное различие языковых выражений оказывается существенным.

Ряды онтологических синонимов формируются с максимальной степенью подробности. Эквивалентность некоторых типов словосочетаний может показаться человеку очевидной, однако практически нет правил, которые работают со стопроцентной точностью. Некоторые словосочетания, полученные в результате «очевидных» трансформаций, почему-то в реальности не употребляются, другие употребляются совсем в другом смысле, чем исходное словосочетание.

Онтологические синонимы демонстрируют огромное разнообразие лексико-синтаксических схем. Особенно интересными оказались однозначные словосочетания, которые достаточно часто употребляются как синонимы однозначных многозначных слов. Эти словосочетания выглядят иногда тавтологичными, однако польза их в том, что в случае необходимости они позволяют называть сущности совершенно однозначно.

Глава 17. Отношения между понятиями в тезаурусе РуТез

Отношения между понятиями, описываемые в онтологическом ресурсе, предназначенном для автоматической обработки текстов в рамках информационно-поисковых приложений должны выполнять разнообразные функции.

Во-первых, эти отношения должны использоваться в классических функциях информационно-поисковых тезаурусов для расширения поискового запроса или вывода рубрики документа.

Во-вторых, отношения важны для разрешения многозначности языковых единиц, включенных в ресурс, поскольку естественным методом реализации автоматической процедуры разрешения многозначности является сопоставление контекста употребления многозначной единицы в тексте и контекста соответствующего понятия в онтологическом ресурсе.

В-третьих, отношения в онтологическом ресурсе могут использоваться для выявления лексической связности в текстах, и использованию выявленной структуры текста для улучшения качества обработки текстов.

Для реализации любой из этих функций необходимо осуществление своеобразного логического вывода: встретив вхождение некоторого понятия в тексте, нужно делать многошаговые проходы по отношениям.

В первых главах мы рассматривали различные онтологические ресурсы, которые в большей или меньшей степени используются при автоматической обработке текста в рамках различных приложений информационного поиска. Эти ресурсы характеризуются разными наборами отношений между своими единицами.

В исходном наборе отношений Принстонского WordNet многие исследователи отмечали нехватку отношений, что проявлялось, например, в возникновении «теннисной проблемы» (см. п. 2.5.3.1). Сделанная впоследствии жесткая разметка синсетов WordNet областями-доменами до некоторой степени смягчает, но не решает эту проблему.

Такие отношения WordNet как *часть-целое (мероним-холоним)* описаны так, что позволяется одновременная принадлежность синсета-части многим синсетам-целым (см. п. 8.6.2.). Это означает, что прежде, чем использовать такого рода отношения для автоматического логического вывода, необходимо установить, о каком целом идет речь в данном контексте, что не всегда возможно.

В большинстве информационно-поисковых тезаурусов используется очень небольшой набор отношений между дескрипторами: отношение ВЫШЕ-НИЖЕ и отношение АССОЦИАЦИИ.

Как указывалось в разделе 1.2.2, отношение АССОЦИАЦИИ часто рассматривается как проблемное отношение по следующим причинам:

- по принципам установления это отношение является симметричным, а часто обозначаемые им отношения явно не симметричны,
- это отношения часто устанавливается субъективно,
- с этим отношением возникают серьезные проблемы при использовании в автоматических режимах расширения запроса, вывода рубрики и т.п.

Поэтому в литературе имеется много предложений по замене отношения АССОЦИАЦИИ на более подробные наборы отношений, что была реализовано в ряде тезаурусов, например, медицинской тематики.

В последнее время активно обсуждается вопрос о преобразовании существующих информационно-поисковых тезаурусов в более формализованные онтологические ресурсы, с более подробной системой отношений, с возможностью логического вывода на базе аксиом, связанных с каждым отношением (см. п.4.5.3).

Однако, на наш взгляд, существуют серьезные проблемы на пути преобразования информационно-поискового тезауруса в такого рода онтологию и использование в

приложениях информационного поиска, поскольку при автоматическом анализе текста далеко не всегда можно быть уверенным в том, что в тексте упомянуто именно определенное отношение между сущностями, а это значит, что сложные онтологические формализмы, построенные на шатком базисе, не смогут работать эффективно.

Таким образом, мы полагаем, что среди потенциального множества отношений понятия наиболее стабильно можно опираться на те отношения, которые не исчезают, не меняются в течение всего срока существования любого или подавляющего большинства экземпляров понятия (Loukachevitch, Dobrov, 2004a; Лукашевич, Добров, 2004b; Добров, Лукашевич, 2008). Например, любой лес всегда состоит из деревьев.

Наиболее известным типом отношения, которое выполняется для всех экземпляров, является таксономическое отношение. Так, если *C1* упомянуто в тексте, и *C1* является видом *C2*, это означает, что в тексте упомянуто и *C2*. Если данный текст релевантен запросу о *C1*, то он будет релевантен и запросу о *C2*.

В условиях невозможности использования сложных правил вывода для осуществления вывода по тексту важно найти и описывать в тезаурусе другие типы отношений, которые, с одной стороны, минимально зависят от контекста упоминания понятия, с другой стороны, обладающие свойствами транзитивности и наследования, подобно таксономическим отношениям.

17.1. Принципы описания отношений

В результате исследований и экспериментов мы пришли к набору отношений ресурса, предназначенного для эффективной автоматической работы в информационно-поисковых приложениях.

В тезаурусе РуТез имеется четыре основных типа отношений.

Первый тип отношений – родовидовое отношение НИЖЕ-ВЫШЕ, представляет собой отношение таксономии, обладает свойствами транзитивности и наследования.

Второе тип отношений – отношение ЧАСТЬ-ЦЕЛОЕ. Используется не только для описания физических частей, но и для других внутренних сущностей понятия, таких как свойства или роли для ситуаций. Важным условием при установлении этого отношения является то, что понятия-части должны быть жестко связаны со своим целым, то есть каждый пример понятия-части должен в течение всего времени своего существования являться частью для понятия-целого, и не относиться к чему-либо другому.

В этих условиях удается выполнить свойство транзитивности введенного таким образом отношения ЧАСТЬ-ЦЕЛОЕ, что очень важно для автоматического вывода в процессе автоматической обработки текстов.

Еще один тип отношения, называемого несимметричной ассоциацией АСЦ2-АСЦ1, связывает два понятия, которые не могут быть связаны выше рассмотренными отношениями, но когда одно из которых не существовало бы без существования другого. Например, понятие *САММИТ* требует существования понятия *ГЛАВА ГОСУДАРСТВА*. В онтологических исследованиях такое отношение называется отношением онтологической зависимости (см. пп.9.2, 17.4).

Последний тип отношений – симметричная ассоциация связывает, например, понятия очень близкие по смыслу, но которые разработчики не решились соединить в одно понятие (см.п. 17.5).

Отношения ВЫШЕ-НИЖЕ, ЧАСТЬ-ЦЕЛОЕ и несимметричная ассоциация являются иерархическими отношениями. Таким образом, на основе свойств иерархичности, транзитивности и наследования для каждого понятия может быть определена совокупность понятий, которые являются для него нижестоящими понятиями по иерархии – так называемое «дерево-вниз», а также может быть определена совокупность понятий, которые являются для него вышестоящими по иерархии – так называемое «дерево-вверх». Эти иерархические деревья не обязательно являются деревьями в строгом математическом смысле слова.

Рассмотрим далее принципы описания отношений в тезаурусе РуТез более подробно.

17.2. Описание родовидовых отношений в тезаурусе РуТез

17.2.1. Принципы описания родовидовых отношений

Отношения ВЫШЕ-НИЖЕ, устанавливаемые в информационно-поисковых тезаурусах, не обязательно являются таксономическими отношениями в смысле онтологического моделирования. Так, например, в некоторых тезаурусах в качестве отношений ВЫШЕ-НИЖЕ могут записываться отношения ЧАСТЬ-ЦЕЛОЕ (см. например, AGROVOC, EUROVOC).

При разработке ресурсов для автоматической обработки текста, пригодных для логического вывода, важно, чтобы отношения, называемые одинаково, обладали одинаковыми свойствами. В тезаурусе РуТез мы используем отношение ВЫШЕ-НИЖЕ для обозначения онтологических отношений, который обладают свойствами онтологических отношений класс-подкласс, описанных в главе 6, а именно:

- каждый пример видового понятия в любой момент своего существования должен быть примером родового понятия,
- видовое понятие должно относиться к тому же семантическому классу, что и родовое понятие,
- видовое понятие должно наследовать основные свойства родового понятия.

Помимо отношений класс-подкласс такими же свойствами обладают отношения между ролевым понятием и понятием-классом в тех случаях, когда экземпляры только этого понятия-класса могут выступать в данной роли (*РАБОТНИК - ЧЕЛОВЕК*).

Другим типом отношений, обладающим такими свойствами, является отношение между фазой какой-либо физической сущности и собственно этой сущностью (*ЩЕНОК – СОБАКА).*

Таким образом, мы предполагаем, у отношения ВЫШЕ-НИЖЕ свойства несимметричности и транзитивности:

ВЫШЕ(X,Y) $\wedge$ ВЫШЕ (Y,Z) $\rightarrow$ ВЫШЕ(X,Z)

НИЖЕ (X,Y) $\wedge$ НИЖЕ (Y,Z) $\rightarrow$ НИЖЕ(X,Z)

ВЫШЕ (X,Y) $\rightarrow$ НИЖЕ (Y,X)

Одной из серьезных проблем описания таксономических отношений в онтологиях является их смешение с описанием ролевых отношений (см. главу 7.). В следующем разделе мы рассмотрим причины возникновения этой частой проблемы и методы описания ролевых отношений в тезаурусе РуТез.

17.2.2 Принципы описания ролевых отношений в Тезаурусе русского языка РуТез

Проблема смешения таксономических и ролевых отношений связана с тем, что в текстах эти отношения часто выражаются сходными языковыми конструкциями. При разработке ресурса для автоматической обработки текстов приходится много информации вводить в тезаурус на основе знаний, полученных из текстов (Лукашевич, 2007b; Лукашевич, 2007c).

Например, следующий фрагмент (http://www.giord.ru/0705211117391.php):

наиболее используемыми консервантами являются: поваренная соль, этиловый спирт, уксусная, сернистая, сорбиновая, бензойная кислоты и некоторые их соли

может показаться хорошим источником информации для того, чтобы описать виды консервантов.

Определение электролита:

Электролит - проводник второго рода; вещество, обладающие ионной проводимостью. Электролитами являются:

- *расплавы солей, оксидов или гидроксидов;*
- *растворы солей, кислот или оснований в полярных растворителях;*
- *а также твердые электролиты.*

может показаться основанием, например, для установления отношения, что соль (как химическое соединение) является видом электролита.

Однако в таких случаях нужно помнить, что *консервант и электролит* являются ролями веществ - вещество становится консервантом или электролитом только, если попадает в некоторые условия. А поваренная соль и соль как химическое соединение являются типами веществ.

Устанавливая родовидовую связь от типа к роли, мы сообщаем системе некорректное знание, состоящее, например, в том, что любое вещество, относящееся к классу солей, в любой момент времени своего существования в любой ситуации, является электролитом, что далеко не так.

Возникает вопрос, можно ли отразить полученную из вышеприведенных фрагментов информацию, выразив ее набором более «надежных» отношений. В тезаурусе РуТез мы обычно пытаемся применить несколько способов.

Во-первых, если мы предполагаем, что в нашей предметной области большинство примеров того или иного типа будут использованы в некоторой роли, то все-таки устанавливается родовидовое отношение от типа как вида к роли как роду, которое снабжается пометкой В – что означает « возможно по умолчанию» (см. п.17.6).

Так, например, мы можем установить такое отношение между понятием *СОРБИНОВАЯ КИСЛОТА* и *КОНСЕРВАНТ*, если посчитаем, что это основное применение сорбиновой кислоты в нашей предметной области, и вероятность встретить в текстах обсуждение сорбиновой кислоты в других применениях (например, в органическом синтезе) в нашей области не слишком велико:

СОРБИНОВАЯ КИСЛОТА

ВЫШЕ_В *КОНСЕРВАНТ*

Однако не рекомендуется устанавливать такое отношение между понятиями *ПОВАРЕННАЯ СОЛЬ* и *КОНСЕРВАНТ*, поскольку основное применение поваренной соли совсем другое. Даже если бы мы установили такое отношение (ввели бы еще пометку для неосновных ролей), то нужно учитывать, что для автоматической системы обработки текстов невозможно качественно учитывать контекст употребления поваренной соли в тексте, чтобы разобраться, можно использовать это отношение или нет.

Таким образом, в некоторых случаях мы все-таки размещаем понятия-роли выше по иерархии, чем понятия-типы, однако отмечаем такое отношение специальной пометкой. Мы применяем это отношение только для описания знания о предметной области, которое верно по умолчанию, то есть, с одной стороны, оно может пригодиться при обработке текстов, с другой стороны, относительно редко может привести к ошибке вывода. Для каждого типа может быть описано максимум одно такое отношение, а описания многих понятий-типов не включают такие отношения, поскольку могут выступать в самых разных ролях.

Именно с использованием отношения ВЫШЕ_В может быть отражено критикуемое Н. Гуарино отношение *яблоко – пища,* если будет известно, что в рабочей предметной области использование в пищу – это основная роль яблок.

В перечисленных в разделе 7.5 способах представления иерархических отношений между типами и ролей такое решение соответствует способу 2.1, однако корректируется дополнительной пометкой

На примере описания понятия *ЭЛЕКТРОЛИТ* может быть продемонстрирована еще одна возможность описания отношений между ролями и типами в тезаурусе РуТез.

Мы можем попытаться ввести дополнительное понятие для ситуации соли в роли электролита. Если это важно для данной сферы, то это наше желание обычно поддерживается и языком предметной области – для такого понятия существует одно или более употребительных языковых выражений. И в нашем случае существует и активно употребляется такое словосочетание как *солевой электролит.*

Таким образом, мы можем ввести понятие СОЛЕВОЙ ЭЛЕКТРОЛИТ и установить следующие отношения:

СОЛЕВОЙ ЭЛЕКТРОЛИТ
ВЫШЕ *СОЛИ*
ВЫШЕ *ЭЛЕКТРОЛИТЫ*

Тем самым мы корректно отражаем знание, полученное нами из прочитанного определения. В перечисленных в разделе 7.5 способах представления иерархических отношений между типами и ролей такое решение соответствует способу 3, однако иерархии типов и ролей пересекаются не только на примерах понятий, которые относятся к обеим иерархиям, но в специально введенных понятиях.

Если рассмотреть такое решение для отражений отношение между понятиями *РАБОТОДАТЕЛЬ, ЧЕЛОВЕК* и *ОРГАНИЗАЦИЯ,* то нужно ввести два дополнительных понятия, например, *РАБОТОДАТЕЛЬ-ФИЗИЧЕСКОЕ ЛИЦО* и *РАБОТОДАТЕЛЬ-ЮРИДИЧЕСКОЕ ЛИЦО.*

Тогда можно сделать следующие описания (Рис.17.1)*:*

РАБОТОДАТЕЛЬ-ФИЗИЧЕСКОЕ ЛИЦО
ВЫШЕ *РАБОТОДАТЕЛЬ*
ВЫШЕ *ЧЕЛОВЕК*

РАБОТОДАТЕЛЬ-ЮРИДИЧЕСКОЕ ЛИЦО
ВЫШЕ *РАБОТОДАТЕЛЬ*
ВЫШЕ *ОРГАНИЗАЦИЯ*

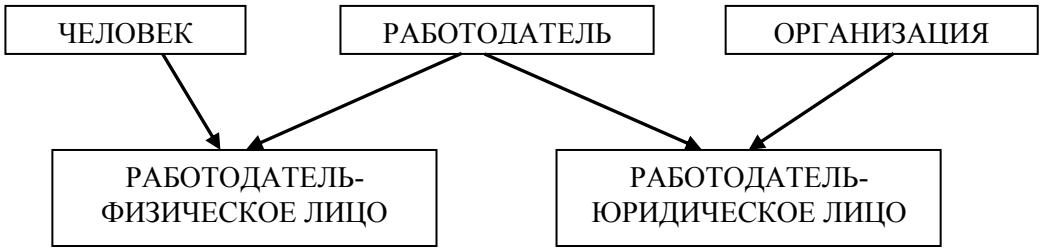

Рис.17.1 Отражение взаимоотношений между понятиями-типами (*ЧЕЛОВЕК, ОРГАНИЗАЦИЯ*) и понятием-ролью (*РАБОТОДАТЕЛЬ)*

Поскольку введение дополнительных понятий может серьезно усложнять описание понятий в ресурсе, такой способ используется лишь в тех случаях, когда такие дополнительные понятия действительно используются в предметной области, как в случае понятия *СОЛЕВОЙ ЭЛЕКТРОЛИТ.* Интересно отметить, что введенные дополнительные понятия *РАБОТОДАТЕЛЬ-ФИЗИЧЕСКОЕ ЛИЦО* и *РАБОТОДАТЕЛЬ-ЮРИДИЧЕСКОЕ*

ЛИЦО также имеют реальное основание в правовой области, поскольку отношения разных типов работодателей с работниками по-разному регулируются законодательством.

Понятия-типы *ЧЕЛОВЕК* и *ОРГАНИЗАЦИЯ* могут равным образом выступать во многих ролях, поэтому в тезаурусе РуТез действительно введено такое понятие как *СУБЪЕКТ ДЕЯТЕЛЬНОСТИ,* как обсуждается в методе 2.1 раздела 7.5, и в качестве нижестоящих к нему понятий размещены многие ролевые понятия, в которых могут выступать и примеры понятия *ЧЕЛОВЕК*, и примеры понятия *ОРГАНИЗАЦИЯ*.

Таким образом, в реальном ресурсе, создаваемом для работы в приложениях в широкой предметной области, приходится применять несколько разных подходов к описанию иерархий типов и ролей, обсуждаемых в литературе как альтернативные.

В литературе обсуждаются и более сложные представления для более адекватного описания взаимоотношений между типами и ролями, однако при создании достаточно больших онтологических ресурсов важно сохранить относительно простую схему описания. Кроме того, сложные схемы описания отношений затруднительно использовать при автоматической обработке текстов.

17.3. Отношение ЧАСТЬ-ЦЕЛОЕ

17.3.1. Принципы описания отношения

Одним из важных свойств, которые часто постулируются у отношения ЧАСТЬ-ЦЕЛОЕ, является транзитивность этого отношения. В то же время многие исследователи указывают на нарушения транзитивности этого отношения. В главе 8 мы подробно обсуждали разные точки зрения, высказываемые по поводу этой проблемы.

Если обсуждать наследование свойств по отношению ЧАСТЬ-ЦЕЛОЕ в ресурсе, предназначенном для автоматической обработки текстов в информационно-поисковых приложениях, то наиболее важной операцией, которую необходимо обеспечить, является релевантность обсуждения частей обсуждению целого. То есть необходимо описывать отношения ЧАСТЬ-ЦЕЛОЕ так, что если текст или его некоторый фрагмент посвящен обсуждению части, то с большой вероятностью этот текст (или его фрагмент) будет релевантен и обсуждению целого (Лукашевич, 2007а).

Здесь может быть приведено следующее возражение: если в тексте говорится о покупке деталей автомобиля, это не означает, что текст обсуждает покупку автомобиля. Мы этого и не утверждаем. Ясно, однако, что текст, обсуждающий покупку деталей автомобиля, релевантен поиску по обобщенному запросу «Автомобили».

Важным условием для обеспечения такого наследования, на наш взгляд, является зависимость (см. п.8.5) существования части от существования целого. Действительно, если все существование некоторой части связано с существованием целого, то и тексты, обсуждающие эту часть, будут иметь непосредственное отношение и к целому, даже если это целое в тексте явно не упомянуто.

Этим требованием мы обеспечиваем выполнение рекомендации тезаурусных стандартов в том, что описание иерархических отношений должно быть независимо от контекста их упоминания. Описание таких независимых от контекста, «надежных» отношений в ресурсах, предназначенных для автоматической обработки текстов, имеет большое значение, поскольку в автоматическом режиме часто невозможно использовать контекст для подтверждения существования того или иного отношения.

Зависимость части от целого не влечет эксклюзивность части по отношению к целому, то есть того, что у части ровно одно непосредственное целое. Так, например, локоть является частью руки человека, и одновременно частью костной системы, при этом локоть является зависимой частью и для руки человека, и для костной системы.

Накладывая условие зависимости частей от целого, мы не ограничиваем подвиды частей (табл.18.1).

ЧАСТЬ – ЦЕЛОЕ		
автопилот	–	летательный аппарат
горбушка	–	хлеб
член партии	–	политическая партия
член предложения	–	предложение
балкон	–	зрительный зал
бородка ключа	–	ключ
ветка	–	растение
железнодорожная ветка	–	железнодорожный путь
персонаж	–	сюжет
гипотенуза	–	прямоугольный треугольник
голень	–	нога
гондола	–	аэростат

Таблица 18.1. Примеры отношений ЧАСТЬ-ЦЕЛОЕ из тезауруса РуТез.

При таком разнообразии отношений ЧАСТЬ-ЦЕЛОЕ важно отделять это отношение от других видов отношений. Для этого мы применяем не лингвистические тесты, которые часто неоднозначны и контекстно зависимы, а используем принципы, упоминавшиеся в п. 8.4. При анализе сложных случаев транзитивности отношения ЧАСТЬ-ЦЕЛОЕ таких, как *ручка двери – дверь – дом,* а именно: если уничтожение (изменение) предполагаемой части оказывает влияние на предполагаемое целое, то между ними устанавливается отношение ЧАСТЬ-ЦЕЛОЕ, если нет, то данное отношение описывается с помощью других типов тезаурусных отношений.

Рассмотрим, например, понятие, соответствующее слову «айсберг». Большой толковый словарь русского языка дает следующее толкование значения слова «айсберг»: *Айсберг – плавучая ледяная гора, отколовшаяся от прибрежного ледника, большая часть которой находится под водой.* Возникает вопрос, стоит ли описывать *АЙСБЕРГ* как часть ледника, как часть океана? По предложенному критерию, если проанализировать последствия разрушения, раскалывания конкретного айсберга, то понятно, что ни в каком леднике изменений не произойдет, окружающее море (океан) также не изменятся, поэтому отношение понятий *АЙСБЕРГ – ЛЕДНИК, АЙСБЕРГ – ОКЕАН* должно описываться с помощью других отношений, а не посредством отношений ЧАСТЬ-ЦЕЛОЕ.

Рассмотрим другой пример, можно ли использовать отношение ЧАСТЬ-ЦЕЛОЕ для описания отношений между понятиями *ГОРОЖАНЕ -ГОРОД, ГОРОДСКОЙ СУД - ГОРОД*. Город – это, прежде всего, населенный пункт, поэтому если что-то происходит с его жителями, горожанами, то это имеет воздействие и на город: так, если исчезнут все жители, то город перестанет быть населенным пунктом. Таким образом, если в тезаурус вводится понятие *ГОРОЖАНЕ*, то от этого понятия к понятию *ГОРОД* должно быть установлено отношение ЦЕЛОЕ.

С понятием *ГОРОДСКОЙ СУД* ситуация иная: есть в городе суд, или нет его, переехал ли этот суд в другой город, само по себе не несет изменений в конкретный город. Поэтому отношение *ГОРОДСКОЙ СУД – ГОРОД* должно описываться не отношением *ЧАСТЬ-ЦЕЛОЕ*, а другим отношением, например, отношением ассоциации.

Интересно рассмотреть пример из книги (Cruse, 1986), приведенный в п.8.4, о том, что не стоит описывать, отношение между Вестминстерским аббатством и Лондоном как

часть целое, поскольку Вестминстерское аббатство - это здание, а Лондон – это географическое место.

Но на самом деле, для Лондона как города существенными частями являются здания, постройки, которые составляют его существенные части. Если уничтожать здания в городе, то и сам город может быть уничтожен. Таким образом, городские здания должны рассматриваться как части города, и, следовательно, Вестминстерское Аббатство должно быть описано как часть Лондона.

Таким образом, описывая отношения ЧАСТЬ_ЦЕЛОЕ в информационно-поисковых ресурсах, предназначенных для автоматической обработки текстов, мы опираемся на три основных принципа:

1) Часть должна быть зависима от целого;
2) Уничтожение или изменение части влечет изменение целого;
3) Свойство релевантности наследуется от части к целому: если в тексте обсуждается часть, то этот текст обсуждает и целое.

Два первых принципа заменяют в тезаурусе для автоматического индексирования правила, установленные стандартами для традиционных информационно-поисковых тезаурусов: независимость отношения от контекста и соответствие семантических типов части и целого.

Однако этим принципам соответствуют также свойства, которые зависят от своих носителей, а также роли, зависимые от своих ситуаций. В качестве примеров таких зависимых свойств можно привести следующие:

грузоподъемность – транспортное средство,
калорийность – пища,
водоизмещение - судно,
октановое число - моторное топливо.

В качестве примеров зависимых ролей можно привести следующие роли:

инвестор – инвестирование,
дирижер - дирижирование,
дубильщик, дубитель – дубление кожи.

Соответственно, такие отношения мы также описываем как ЧАСТЬ-ЦЕЛОЕ.

Таким образом, мы описываем как части разнообразные внутренние сущности и характеристики объекта, проявляющие зависимость своего существования от существования целого объекта.

На основе таким образом определенного отношения ЧАСТЬ-ЦЕЛОЕ естественно решается «теннисная» проблема, возникшая перед разработчиками тезауруса WordNet. Все сущности, относящиеся к той или иной сфере деятельности, описываются как ее части.

ТЕННИС
ЧАСТЬ *ТЕННИСИСТ*
ЧАСТЬ *ТЕННИСНЫЙ КОРТ*
ЧАСТЬ *ТЕННИСНЫЙ МАТЧ*
ЧАСТЬ *ТЕННИСНЫЙ ИНВЕНТАРЬ*

Такое решение «теннисной» проблемы не требует наложения искусственной и жесткой системы доменов-областей, подобной системе, созданной для WordNet.

Как уже было указано, обобщенные части соответствуют разного рода внутренним характеристикам сущности. Такое решение согласуется, например, с позицией Джона Совы (Sowa 2000), который объединяет физические части, участников, стадии, а также свойства в одну категорию внутренних сущностей, то есть сущностей исчезновение или

изменений которых меняет структуру или существование другой сущности. Достаточно широко трактуется отношение ЧАСТЬ-ЦЕЛОЕ и в онтологии CYC (см. п. 8.6.3).

17.3.2. Транзитивность отношения

Поскольку в тезаурусе РуТез отношение ЧАСТЬ-ЦЕЛОЕ обуславливается дополнительными условиями на зависимость существования части от целого, то возникает вопрос, насколько правомерно рассматривать транзитивность такого отношения. Как было показано в п.8.4, наложение дополнительных условий на транзитивное отношение может приводить к ограниченному действию этого свойства.

Вместе с тем, в число аксиом, которые обычно постулируются для отношения онтологической зависимости, входит и аксиома транзитивности (Varzi, 2006). Таким образом, транзитивны и базовое отношение ЧАСТЬ-ЦЕЛОЕ и дополнительно накладываемое на него условие, что дает возможность использования этого отношения для логического вывода в процессе обработки текстов на основе тезауруса РуТез. Такой логический вывод полезен при решении многих задач информационного поиска, таких как автоматическое рубрицирование, автоматическое расширение запроса, поиск ответа на вопрос.

За счет использования транзитивности отношений онтологической зависимости формируются достаточно длинные цепочки вывода (цепочка слева - направо соответствует отношениям от части к целому):

ОБВИНЯЕМЫЙ ПО ДЕЛУ → СУДЕБНОЕ ОБВИНЕНИЕ →
СУДЕБНЫЙ ПРОЦЕСС → СУДОПРОИЗВОДСТВО → СУДЕБНАЯ СИСТЕМА →
ПРАВОВАЯ СИСТЕМА

ДЕНЕЖНАЯ БАЗА → ДЕНЕЖНОЕ ОБЕСПЕЧЕНИЕ →
ДЕНЕЖНОЕ ОБРАЩЕНИЕ → ДЕНЕЖНАЯ СИСТЕМА →
ФИНАНСОВАЯ СИСТЕМА → ЭКОНОМИКА

АПТЕКАРЬ → АПТЕКА → ЛЕКАРСТВЕННОЕ ОБЕСПЕЧЕНИЕ →
МЕДИЦИНСКАЯ ПОМОЩЬ → МЕДИЦИНА → ЗДРАВООХРАНЕНИЕ

Такие цепочки интерпретируются следующим образом: если в тексте обсуждается обвиняемый по делу, то этот текст релевантен и таким темам как *судебное обвинение, судебный процесс, судопроизводство, судебная система, правовая система***.**

Как видно, отношение ЧАСТЬ-ЦЕЛОЕ с дополнительным условием зависимости работает не только для таких наиболее часто ассоциирующихся с этим отношением типов сущностей как физические объекты, но и для весьма сложных для описания абстрактных сущностей.

Используемый в настоящее время набор свойств отношения ЧАСТЬ-ЦЕЛОЕ таков:

ЧАСТЬ (X,Y)<-> ЦЕЛОЕ (Y, X)

ЦЕЛОЕ (X,Y) ∧ ЦЕЛОЕ (Y, Z) -> ЦЕЛОЕ (X, Z) – транзитивность отношения

ВЫШЕ (X,Y) ∧ ЦЕЛОЕ (Y, Z) -> ЦЕЛОЕ (X, Z) – наследование отношения ЦЕЛОЕ по отношению ВЫШЕ-НИЖЕ.

В настоящее время цепочки отношений ЧАСТЬ-ЦЕЛОЕ используются при применении тезауруса РуТез в задачах автоматической рубрикации, при расширении запроса пользователя, поиска ответов на вопросы, разрешении лексической многозначности, построении тематического представления текста.

Таким образом, при моделировании отношения ЧАСТЬ-ЦЕЛОЕ в тезаурусе РуТез основными задачами являлись следующие:

- обеспечение наследования свойства релевантности от части к целому: если в тексте обсуждается часть, то этот текст релевантен и обсуждению целого;

- обеспечение транзитивности отношения ЧАСТЬ-ЦЕЛОЕ как основы для логического вывода в процессе обработки текстов.

В качестве основных принципов моделирования был выбран не лингвистический подход с его опорой на языковые тесты, которые часто неоднозначны и контекстно зависимы, а онтологический анализ отношений, который строится на рассмотрении отношения онтологической зависимости существования понятий и влияния ситуации разрушения предполагаемой части на состояние целого.

17.3.3. Как описать отношение ЧАСТЬ-ЦЕЛОЕ, если часть не является зависимой

В случае если понятие-часть может принадлежать нескольким целым, то можно использовать несколько возможностей для описания такого отношения, которые обычно связаны с введением дополнительных понятий.

Первый способ подходит в тех случаях, если у исходной части есть подвид, который является зависимой частью исходного целого.

Так, например, неправильно описывать в тезаурусе, что ДВИГАТЕЛЬ – это часть АВТОМОБИЛЯ, поскольку не все двигатели являются частями автомобиля. Необходимо ввести дополнительное понятие АВТОМОБИЛЬНЫЙ ДВИГАТЕЛЬ как вид понятия ДВИГАТЕЛЬ и описать понятие АВТОМОБИЛЬНЫЙ ДВИГАТЕЛЬ как часть АВТОМОБИЛЯ.

Тем же способом можно воспользоваться для описания отношения *ДЕРЕВО – ЛЕС*: для этого могут быть дополнительно введены понятия *ЛЕСНОЕ РАСТЕНИЕ, ЛЕСНОЕ ДЕРЕВО (ДЕРЕВО В ЛЕСУ).*

Другим способом является введение обобщающего понятия для всех целых, к которым может принадлежать часть и установить отношение между частью и именно этим целым.

Здесь можно привести пример из химии: альдегидная группа входит в такие соединения как *альдегидокислоты, альдегидоспирты* и т.п., но имеется такое обобщающее выражение как *альдегидное соединение.* Таким образом, можно ввести понятие *АЛЬДЕГИДНОЕ СОЕДИНЕНИЕ* описать как его виды понятия *АЛЬДЕГИДОКИСЛОТЫ* и *АЛЬДЕГИДОСПИРТЫ*, а понятию *АЛЬДЕГИДНАЯ ГРУППА* установить отношение *ЦЕЛОЕ с* понятием *АЛЬДЕГИДНОЕ СОЕДИНЕНИЕ (рис. 17.2).*

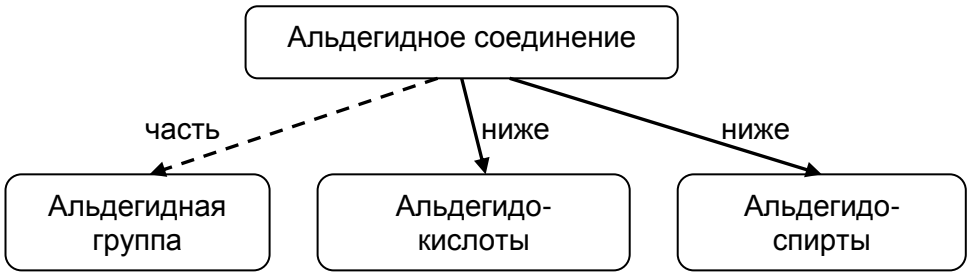

Рис. 17.2 Описание взаимоотношений между понятиями *АЛЬДЕГИДНАЯ ГРУППА, АЛЬДЕГИДОКИСЛОТЫ, АЛЬДЕГИДОСПИРТЫ*

В некоторых случаях можно воспользоваться обоими методами. Так, для описания отношения *РЕАКТИВНЫЙ* САМОЛЕТ – *РЕАКТИВНЫЙ ДВИГАТЕЛЬ*, может быть введено дополнительное понятие *АВИАЦИОННЫЙ РЕАКТИВНЫЙ ДВИГАТЕЛЬ* и/или дополнительное понятие *РЕАКТИВНЫЕ СРЕДСТВА ПЕРЕДВИЖЕНИЯ*.

При таких операциях ввода дополнительных понятий мы стараемся использовать те языковые выражения, которые реально существуют в описываемой предметной

области. Если необходимого языкового выражения не существует, то может быть принято решение не описывать такое отношение часть-целое.

Имеется только одна возможность «смягчения» позиции по поводу описания зависимых частей, которая возникает в тех случаях, когда некоторая часть входит в целое в подавляющем большинстве случаев, может устанавливаться по умолчанию. В таких случаях, такое отношение вводится в тезаурус, но помечается специальной пометкой «В» (Лукашевич, Добров, 2001).

17.3.4. Сложные случаи описания отношений ЧАСТЬ-ЦЕЛОЕ

Существует ряд факторов, усложняющих описание отношений ЧАСТЬ-ЦЕЛОЕ.

Одним из таких факторов является возможная отделимость части от целого в некоторый момент времени. Так, например, в широкой общественно-политической области, в состав которой входит как сельское хозяйство, так и сфера торговли, такая проблема отделимости возникает с отношением между понятиями ЯБЛОКО и ЯБЛОНЯ.

Пока яблоко растет, оно является частью яблони. После созревания яблоко срывается, может стать товаром и достаточно долго существовать без связи со своей яблоней. Так, для описания таких случаев в онтологии DOLCE для отношения ЧАСТЬ-ЦЕЛОЕ вводится аргумент времени.

Мы не считали возможным вводить время в качестве аргумента описания отношений ЧАСТЬ-ЦЕЛОЕ, поскольку считаем, что фактор времени очень трудно учесть при обработке текстов широкой предметной области, отношение начинает зависеть от контекста, что затрудняет его применение на практике.

Другим фактором является проблема, называемая нами *фокусная отделимость*. Она находит свое проявление, например, в отношении между понятиями *ДЕПУТАТ ГОСУДАРСТВЕННОЙ ДУМЫ – ГОСУДАРСТВЕННАЯ ДУМА*. С одной стороны, *ДЕПУТАТ ГОСУДАРСТВЕННОЙ ДУМЫ* является частью *ГОСУДАРСТВЕННОЙ ДУМЫ*. С другой стороны, становясь депутатом, человек получает особый социальный статус, который начинает упоминаться в текстах в ситуациях (например, автомобильная авария), которые не имеют отношения к функционированию Государственной думы. Таким образом, при упоминании в таких ситуациях отношение ЧАСТЬ-ЦЕЛОЕ не исчезает, но и не находится в фокусе сообщения.

Еще одной группой усложняющих факторов является сложность структуры самих объектов. Так, например, дверь дома состоит из двух основных частей: проема в стене и доски, вращающейся на петлях. Ручка крепится на одной части - доске, а непосредственно в состав дома входит другая часть – проем.

Другой пример – приток реки. Решение о том, как правильно описать отношение притока реки с главной рекой (с использованием отношений ЧАСТЬ-ЦЕЛОЕ или присоединения (Noy, Wallace, 2005)), затрудняется тем, что приток реки состоит из потока воды и берегов, при этом берега остаются на месте и не становятся частью берегов главной реки, а воды притока вливаются в основную реку и становятся частью вод основной реки.

Для описания таких усложненных отношений ЧАСТЬ-ЦЕЛОЕ используется пометка «А» («аспект»), обозначающая то, что с некоторой точки зрения установленное отношение может не иметь значения, не быть важным (см. п.17.5).

17.4. Отношение онтологической зависимости в тезаурусе РуТез

В предыдущих разделах (см. пп.1.2.2, 1.7.2, 9.4) мы обсуждали проблемы отношений ассоциации традиционных информационно-поисковых тезаурусов, а также полезность учета в ресурсах, предназначенных для информационного поиска, отношений онтологической зависимости.

При обсуждении отношения онтологической зависимости в разделе 9.2 мы видели, что исследователи рассматривают значительное разнообразие таких отношений онтологической зависимости и указывают на сложность выбора конкретных типов этого отношения для онтологического моделирования. При этом ряд авторов (Н.Гуарино и др.) неоднократно подчеркивали важность изучения и описания в онтологиях отношений внешней онтологической зависимости.

Действительно, при построении наших ресурсов мы можем учесть внутреннюю онтологическую зависимость с помощью отношения ЧАСТЬ-ЦЕЛОЕ. А дополнительное отношение нам нужно именно для того, чтобы описывать отношения между сущностями, которые являются отдельными сущностями по отношению друг к другу, то есть важно ввести еще одно отношение именно для представления внешней онтологической зависимости.

При этом даже отношения внешней онтологической зависимости могут различаться. Мы можем изучать экзистенциальную, то есть зависимость существования отдельной сущности, или концептуальную внешнюю зависимость, (то есть зависимость существования понятия), строгую (т.е. зависимость от конкретного экземпляра другой сущности) или родовую зависимость (т.е. зависимость от класса сущностей).

В следующих подразделах мы рассмотрим небольшой эксперимент, демонстрирующий различное поведение разных типов отношений онтологической зависимости при расширении запроса, и опишем правила представления отношения онтологической зависимости в виде несимметричной ассоциации в тезаурусе РуТез.

17.4.1. Влияние типа отношения онтологической зависимости на качество информационного поиска при расширении запроса

Нетрудно видеть, что различия в типах онтологических отношений понятий должны некоторым образом проявляться и в текстах, которые упоминают эти понятия.

Так, при строгой зависимости зависимое понятие не может быть оторвано от конкретного экземпляра главного понятия, поэтому если возникает, существует, обсуждается конкретный пример такого строго зависимого понятия, то существует и обсуждается пример главного понятия. В случае родовой зависимости конкретный пример зависимого понятия может быть оторван от главного понятия, с ним может происходить что-то не связанное с главным понятием, но обычно недолго и в относительно небольшой доле примеров зависимого понятия. При исторической зависимости пример зависимого понятия может достаточно долго существовать без главного понятия и участвовать в самых разных ситуациях, например, *сельскохозяйственная продукция* создается в процессе *сельскохозяйственного производства*, затем продукция значимое время живет «своей жизнью»: перевозится, продается, хранится.

Различия в «жесткости связей» между понятиями для разных подтипов отношений онтологической зависимости ведут к различным видам поведения этих отношения в информационно-поисковом контексте. Рассмотрим эти различия на основе анализа поисковых результатов так называемых элементарных запросов.

Запросы в информационной системе могут состоять из различного числа терминов и слов. С точки зрения онтологии простейшим запросом является запрос, ссылающийся на одно понятие онтологии. Все другие запросы, ссылающиеся на два или более понятий, должны обрабатываться как функция от элементарного запроса.

Мы предполагаем, что потенциальное качество расширения запроса на базе отношений онтологии может изучаться на простых запросах. Если поисковые характеристики расширения элементарных запросов являются низкими, то качество расширения сложных поисковых запросов не может быть лучше. Если онтологические отношения дают возможность эффективного расширения запроса для простых случаев, то это является важным шагом для изучения способов расширения сложных запросов.

Смысл такого рода элементарных запросов таков: «найти все о *C*», и мы будем обозначать его как SQ(*C*).

Рассмотрим два понятия *C1* и *C2*, между которыми установлено отношение R. Выполняя простой запрос SQ(*C1*), мы хотим узнать, может ли отношение R с понятием *C2* быть использовано для расширения этого простого запроса. При этом в выдачу по запросу SQ(*C1*) с некоторыми весами добавятся документы, содержащие *C2*. Следовательно, чтобы проверить полезность такого расширения для запроса SQ(*C1*), не нужно выполнять реальное вычисление запроса с расширением, а нужно рассмотреть документы, содержащие *C2*, и выяснить, какой процент документов релевантен SQ(*C1*).

Мы будем изучать потенциальную эффективность расширения простого запроса для главного понятия *M* в отношении концептуальной зависимости текстами, в которых упомянуто зависимое понятие *D*. Для этого мы проанализировали 50 первых текстов, полученных по простому запросу SQ(*D*).

В качестве запроса задавались выражающие понятие слово или выражение. Тексты в выдаче упорядочивались на основе стандартной векторной модели *tf*idf* (Callan и др., 1992). Поиск был выполнен на коллекции Университетской Информационной Системы РОССИЯ (www.cir.ru), содержащей более 800 тысяч документов. Результаты поиска представлены в Таблице 18.2.

Зависимое понятие *D*	Тип зависимости	Главное понятие *M*	*nD50*	*nM50*
ЛЕС	Строгая	*ДЕРЕВО*	49	12
САММИТ	Строгая	*ГЛАВА ГОСУДАРСТВА*	49	20
ПИАНИСТ	По классу	*ПИАНИНО*	44	16
ГАРАЖ	По классу	*АВТОМОБИЛЬ*	43	1
АВТОМОБИЛЬ	Историческая	*АВТОМОБИЛЬНЫЙ ЗАВОД*	18	44

Таблица 18.2. Зависимость качества расширения запроса от типа онтологической зависимости между сущностями.

Здесь:

- *nD50* – число текстов, содержащих *D*, релевантных *D* и релевантных SQ(*M*),
- *nM50* – число текстов, содержащих *M*, релевантных *M* и релевантных SQ(*D*).

Таблица демонстрирует корреляцию между типом зависимости и поисковыми характеристиками для простых запросов:

- в случае строгой зависимости для практически всех текстов выполняется, что если текст релевантен зависимому понятию, то он релевантен и простому запросу для главного понятия;
- в случае зависимости по классу число текстов, содержащих зависимое понятие и релевантных простому запросу для главного понятия в отношении концептуальной зависимости, меньше;
- в случае исторической зависимости число текстов релевантных обоим понятиям значительно убывает.

Поисковые характеристики для обратной ситуации в первых четырех случаях (т.е., когда выполняем поиск по главному понятию и смотрим, какие из текстов релевантны зависимому понятию) низки, так как имеется множество текстов, упоминающих главное понятие и не имеющих никакого отношения к зависимому понятию. Одновременно наблюдается отсутствие зависимости понятия *M* от понятия *D*.

В пятой строчке таблицы мы видим, что значительная доля текстов об автомобильных заводах релевантны простому запросу об автомобилях. При этом нужно заметить, что здесь имеется отношение концептуальной зависимости: автомобильный завод строится, чтобы выпускать автомобили – имеется отношение концептуальной зависимости по классу понятия *АВТОМОБИЛЬНЫЙ ЗАВОД* от понятия *АВТОМОБИЛЬ*.

Таким образом, рассмотрев 10 вариантов расширения запроса на основе 5 пар понятий, мы видим корреляцию между эффективностью использования отношения при расширении простого запроса и типом этого отношения в рамках теории онтологической зависимости.

17.4.2 Критерии установления отношения онтологической зависимости в тезаурусе РуТез

После многих экспериментов мы пришли к выводу, что в онтологии, предназначенной для автоматической обработки текстов, прежде всего, для приложений информационного поиска, необходимо, прежде всего, отражать внешнюю концептуальную зависимость, то есть зависимость существования понятия от существования другого понятия (Добров, Лукашевич, 2008). Напомним пример внешней концептуальной зависимости (п.9.2.):

гараж концептуально зависит от автомобиля

Гараж как постройка не перестанет существовать, если в мире исчезнут все автомобили, но ее свойство «быть_гаражом» зависит от существования класса сущностей «автомобили».

«Физическое» объяснение нашего выбора подтипа отношения внешней онтологической зависимости связано с тем, что в текстах широких предметных областей, мы больше всего имеем дело не с конкретными сущностями, а с понятиями, концептами, классами, и, обнаружив в тексте то или иное понятие, должны уметь предположить наиболее близкие к нему понятия, которые можно использовать для разного рода логического вывода. Такими близкими понятиями как раз и являются понятия, которые являются зависимыми от текущего понятия и понятия, от которых это понятие концептуально зависит.

Отношение внешней концептуальной зависимости является несимметричным, и мы используем для его описания отношение несимметричной ассоциации АСЦ1- АСЦ2. Отношение АСЦ1 ведет от зависимого понятия к главному понятию отношения концептуальной зависимости, а отношение АСЦ2 является к нему обратным отношением.

Возникает правомерный вопрос, насколько сложно, вводя новое понятие в тезаурус, онтологию, понять, каковы должны быть отношения концептуальной зависимости, и правильно, отразить их в ресурсе.

Здесь можно рассмотреть два случая.

Во-первых, если вводимое понятие базируется на относительно свободно построенном многословном выражении, как, например, словосочетание *автомобильный завод,* то одно из слов обычно указывает на родовое понятие (*ЗАВОД*), а второе слово в частности может указывать и на отношение концептуальной зависимости. Действительно, понятие *АВТОМОБИЛЬНЫЙ ЗАВОД*, не могло бы возникнуть, если бы не было понятия *АВТОМОБИЛЬ*.

Во-вторых, вводимое понятие может основываться на термине, который имеет определение. Здесь необходимо опереться на отношение онтологической зависимости по определению, введенное в работе (Fine, 1995) (см. п.9.2).

При анализе определений необходимо различать следующие типы понятий, упоминаемых в определении посредством соответствующих слов и терминов:

- родовые и видовые понятия,
- понятия-части и понятия-целые,
- понятия, от которых концептуально зависит определяемое понятие,
- другие понятия.

Принципы установления родовидовых отношений подробно описаны в главе 6 и п. 17.2.1, принципы установления отношений часть-целое - в главе 8 и п. 18.3.1. Для различения отношения концептуальной зависимости от отношений с понятиями типа 4)

помогает применение диагностического высказывания, подобного лингвистическим тестам (см.п.2.1):

Возникновение понятия С0 зависит от существования понятия С1.

17.4.3. Свойства несимметричной ассоциации

В настоящее время в приложениях используются следующие свойства отношения внешней концептуальной зависимости, обозначаемой как несимметричная ассоциация:

АСЦ1 (X,Y)<-> АСЦ2 (Y, X)

Наследование отношения несимметричной ассоциации на виды и части:

ВЫШЕ (X,Y) ∧ АСЦ1 (Y, Z) -> АСЦ1 (X, Z)

ЦЕЛОЕ (X,Y) ∧ АСЦ1 (Y, Z) -> АСЦ1 (X, Z)

17.5 Симметричные ассоциации в тезаурусе РуТез

Существует несколько ситуаций, когда оправданно представление отношений между понятиями в виде симметричной ассоциации. При этом предполагается, что степень ассоциации между понятиями достаточно высокая, то есть если два понятия С1 и С2 связаны отношением симметричной ассоциации, то тексты, содержащие понятие С1 часто релевантны запросам, выражающим понятие С2, и наоборот.

Симметричные ассоциации используются для отражения отношения между понятиями, которые являются взаимозависимыми, но между которыми невозможно поставить отношение ЧАСТЬ-ЦЕЛОЕ, например,

РОДИТЕЛИ – ДЕТИ (СЫНОВЬЯ И ДОЧЕРИ)

Другим случаем такого рода является отношение между растением и его плодом в широкой предметной области, когда плод отрывается от своего «материнского» растения и долго существует отдельно (продается, перерабатывается) (см. п. 17.3.4):

ЯБЛОКО - ЯБЛОНЯ

Симметричной ассоциацией описывается также отношение между близкими по смыслу понятиями, относящимися к одному и тому же родовому понятию, текстовые входы которых используются как квазисинонимы.

Например, есть близкие понятия *АВИАЦИОННАЯ МЕДИЦИНА* и *КОСМИЧЕСКАЯ МЕДИЦИНА*, также имеется множество контекстов употреблений словосочетаний *авиакосмическая медицина, авиационная и космическая медицина.* В некоторый момент развития тезауруса отношение между такими понятиями может быть отражено в виде симметричной ассоциации. В качестве других примеров можно привести такие пары как *МИЛИЦИЯ – ПОЛИЦИЯ, ПОТРЕБИТЕЛЬСКАЯ ЦЕНА – РОЗНИЧНАЯ ЦЕНА*.

Наконец, некоторые виды антонимов могут быть представлены в тезаурусе в виде симметричной ассоциации между соответствующими понятиями. Отношением симметричной ассоциации представляются обычно отношения между антонимами, содержащими указание на разную степень, меру одного и того же качества, свойства

В настоящее время используются следующие свойства отношения симметричной ассоциации:

АСЦ (X,Y) -> АСЦ (Y, X) - Симметричность отношения

Наследование отношения ассоциации на виды и части:

ВЫШЕ (X,Y) ∧ АСЦ (Y, Z) -> АСЦ (X, Z)

ЦЕЛОЕ (X,Y) ∧ АСЦ (Y, Z) -> АСЦ (X, Z)

17.6 Модификаторы отношений: нарушение условий надежности

Помимо основных отношений в тезаурусе РуТез имеются так называемые отношения с модификаторами, то есть отношения ВЫШЕ-НИЖЕ и ЧАСТЬ-ЦЕЛОЕ могут быть снабжены дополнительными пометками (напомним, что разработчики EuroWordNet также предлагали использование дополнительных атрибутов на отношениях (см. п.3.2.1.)).

В тезаурусе РуТез имеется два вида модификатора отношений:

модификатор В (возможно, выполняется по умолчанию) и

модификатор А (аспект, точка зрения).

Отличие между двумя этими модификаторами следующее: модификатор В используется, когда подавляющее большинство экземпляров понятия имеют такое отношение, то есть помеченное этим модификатором отношение имеется у данного понятия по умолчанию. Именно с помощью это модификатора отмечается отношение роли, которую выполняют обычно экземпляры данного понятия (см. гл. 7 и п. 17.2.2.).

Модификатор А используется в тех случаях, когда все экземпляры данного понятия имеют то или иное отношение, но это отношение характеризует экземпляр понятия частично и поэтому может быть не существенным в том или ином контексте.

Первоначально родовидовое отношение с модификатором А (ВЫШЕ_А) часто использовалось для указания двух родовых понятий при совмещении в одном понятии двух несколько отличающихся значений, обычно являющихся проявлением так называемой регулярной многозначности (см.п.2.5.2.1), например, =горная порода= – =строительный камень= *(доломит),* =шоу= - =театр= *(варьете*), =растение= – =плод= *(вишня) и др.,* В настоящее время в случае такой регулярной многозначности (см. п.16.4.1.) мы вводим два связанных между собой понятия, и необходимость в отношении ВЫШЕ_А в таких случаях отпадает.

Именно посредством отношения ВЫШЕ_А описывается в настоящее время отношения экземпляр-класс: *МОСКВА – СТОЛИЦА, УЧИТЕЛЬ – ДОЛЖНОСТЬ, ГОРОДСКАЯ ЗЕМЛЯ – ГОРОДСКАЯ СОБСТВЕННОСТЬ, ПУДЕЛЬ – ПОРОДА СОБАК.* Это не противоречит установленным свойствам отношения ВЫШЕ_А (см. ниже).

Примером применения обоих модификаторов в описании понятия могут служить отношения понятия ПЕНСИОНЕР. По умолчанию пенсионер – это старый человек, поскольку подавляющее большинство пенсионеров являются пожилыми людьми. Кроме того, все пенсионеры являются составными элементами пенсионной системы. Наконец, пенсионер – это также социальный статус, поэтому пенсионеры могут упоминаться вне всякой связи с пенсионной системы – для отражения этого ставится модификатор А.

ПЕНСИОНЕР	
ВЫШЕ_В	*СТАРЫЙ ЧЕЛОВЕК*
ЦЕЛОЕ_А	*ПЕНСИОННАЯ СИСТЕМА*

Отношения с модификаторами являются более слабыми отношениями, не обладающими свойством транзитивности.

Основные свойства отношений с модификаторами таковы:

1. Отношения с модификаторами поглощают отношения без модификаторов:

$ВЫШЕ_{A,B}$ (X,Y) ∧ ВЫШЕ (Y, Z) -> $ВЫШЕ_{A,B}$ (X, Z)

ВЫШЕ (X,Y) ∧ $ВЫШЕ_{A,B}$ (Y, Z) -> $ВЫШЕ_{A,B}$ (X, Z)

$ЦЕЛОЕ_{A,B}$ (X,Y) ∧ ЦЕЛОЕ (Y, Z) -> $ЦЕЛОЕ_{A,B}$ (X, Z)

ЦЕЛОЕ (X,Y) $\wedge$ ЦЕЛОЕ (Y, Z) $_{A,B}$ -> ЦЕЛОЕ$_{A,B}$ (X, Z)

2. Наследование отношений по отношениям с модификаторами:

ВЫШЕ$_{A,B}$ (X,Y) $\wedge$ АСЦ1 (Y, Z) -> АСЦ1 (X, Z)

ВЫШЕ$_{A,B}$ (X,Y) $\wedge$ АСЦ (Y, Z) -> АСЦ (X, Z)

ЦЕЛОЕ$_{A,B}$ (X,Y) $\wedge$ АСЦ1 (Y, Z) -> АСЦ1 (X, Z)

ЦЕЛОЕ$_{A,B}$ (X,Y) $\wedge$ АСЦ (Y, Z) -> АСЦ (X, Z)

3. Отсутствие транзитивности по отношениям с модификаторами

17.7. Примеры описания отношений

17.7.1. Типовые примеры описания отношений

Рассмотрим примеры тезаурусных отношений ассоциации, приводимых в стандартах на разработку информационно-поисковых тезаурусов (см. п.1.2.), и покажем, какими отношениями представляются такого рода отношения в тезаурусе РуТез. Для удобства изложения повторим примеры ассоциаций из первой главы курсивом. Примеры описаний этих же отношений в тезаурусе РуТез даны прописными буквами.

1) научная дисциплина – объект изучения или специалист в этой дисциплине:

математика - математик

неврология - нервная система

Специалист по дисциплине описывается как понятие-часть для своей дисциплины, поскольку специалист является концептуально зависимым от дисциплины и является необходимым внутренним элементом этой дисциплины:

МАТЕМАТИКА
ЧАСТЬ *МАТЕМАТИК*

Часто объект изучения, ни каким образом, не зависит от изучающей его дисциплины, тогда как дисциплина концептуально зависит от своего объекта изучения. Дисциплина не является каким-то существенным элементом функционирования объекта изучения. Поэтому дисциплина связана со своим основным объектом изучения несимметричной ассоциацией:

НЕВРОЛОГИЯ
АСЦ1 *НЕРВНАЯ СИСТЕМА*

2) операции или процессы и их агент или инструмент

контроль температуры – термостат

охотник – охота

Агент и инструмент некоторого процесса, а также другие специальные названия ролей процесса представляют собой существенно важные сущности для функционирования этого процесса. Поэтому, если выполняется условие концептуальное зависимости этого понятия-роли от понятия-процесса (то есть такое же название роли не употребляется для других процессов), то используется отношение ЧАСТЬ:

ОХОТА
ЧАСТЬ *ОХОТНИК*

КОНТРОЛЬ ТЕМПЕРАТУРЫ
ЧАСТЬ *ТЕРМОСТАТ*

3) объекты или процессы и их контрагенты

растения – гербициды

Если контрагент создан специально для противодействия некоторому объекту или процессу, это означает, что он является концептуально зависимым от этого объекта (процесса). При этом этот контрагент не является элементом нормальной жизни функционирования этого объекта (процесса). Поэтому для отображения этого отношения используется несимметричная ассоциация. В случае с отношением понятий *РАСТЕНИЕ – ГЕРБИЦИД*, необходимо уточнить пару понятий в отношении: гербициды создаются, чтобы бороться с сорными растениями, поэтому:

ГЕРБИЦИД
АСЦ1 *СОРНОЕ РАСТЕНИЕ*

4) действия и их продукты:

ткачество – ткань

слезоотделение – слеза

переплетное дело – книга

Действия и процессы, направленные на порождение некоторой сущности, обычно бывают концептуально зависимыми от этих сущностей. Сама сущность в полной мере появляется, когда этот процесс создания завершается. Поэтому для отображения такого отношения используется отношение несимметричной ассоциации:

ТКАЧЕСТВО
АСЦ1 ТКАНЬ

СЛЕЗООТДЕЛЕНИЕ
АСЦ1 *СЛЕЗА*

ПЕРЕПЛЕТНОЕ ДЕЛО
АСЦ1 *КНИГА*

5) объекты и вещества и их свойства (уникальные свойства – unique):

яды – токсичность

жидкость – поверхностное натяжение

Если одно понятие представляет уникальное свойство другого понятия, то есть именно то свойство, по которому эта сущность отличается от других сущностей, то это свойство является концептуально зависимым от этой сущности, представляет собой ее существенную характеристику и для представления этого отношения используется отношение ЧАСТЬ:

ЯД
ЧАСТЬ *ТОКСИЧНОСТЬ*

ЖИДКОСТЬ
ЧАСТЬ *ПОВЕРХНОСТНОЕ НАТЯЖЕНИЕ*

6) понятия, связанные причинно-следственной связью:

смерть – оплакивание

Причинно-следственные связи с точки зрения отношения концептуальной зависимости могут носить разный характер.

Но в данном конкретном примере если считать, что ситуация оплакивания (*Оплакивать – слезами выразить скорбь по поводу чье-то смерти* (БТС, 1998)) возникает в связи с кончиной человека, то имеет место внешняя концептуальная зависимость:

ОПЛАКИВАНИЕ
АСЦ1 *СМЕРТЬ*

7) Понятия и единицы их измерения

электрический ток - ампер

Поскольку единицы измерения создаются, чтобы измерить некоторую существующую сущность, явление, то имеет место внешняя концептуальная зависимость:

ЭЛЕКТРИЧЕСКИЙ ТОК
АСЦ2 *АМПЕР*

17.7.2. Описание отношений между ролевыми понятиями и понятиями контекста

В качестве иллюстрации описания сложных взаимосвязей между понятиями рассмотрим примеры отношений, приписанных понятиям-ролям (см. главу 7).

Как указывалось, в разделах 7.3, 7.4 важнейшей характеристикой ролевых понятий является их зависимость от некоторого контекста: отношений, процессов, ситуаций. В (Loebe, 2005) реляционные и процессуальные роли определяются как части соответствующего отношения или процесса.

В Тезаурусе Русского языка РуТез для описания отношения между ролевым понятием и понятием, соответствующим отношению или процессу, также используется отношение ЧАСТЬ-ЦЕЛОЕ, например,

ОРУДИЕ ПРЕСТУПЛЕНИЯ
ЦЕЛОЕ *ПРЕСТУПЛЕНИЕ*

ИНВЕСТОР
ЦЕЛОЕ *ИНВЕСТИРОВАНИЕ*

Если в ситуации или отношении участвует несколько лиц, например, *ученик, учитель*, и ясно, что для реализации роли учителя необходим ученик, а для реализации роли ученика необходим учитель, то может возникнуть вопрос, как лучше представить взаимозависимость этих ролей.

Учитывая, что взаимозависимость частей одного и того же целого является обычной характеристикой частей (см. главу 8), такие роли представляются как части понятия, обозначающего объемлющий процесс, и отдельно их взаимозависимость не описывается:

ОБУЧЕНИЕ
ЧАСТЬ *УЧИТЕЛЬ*
ЧАСТЬ *УЧЕНИК*

Понятие, соответствующее отношению или процессу для той или иной роли, может не иметь употребительного лексического или терминологического выражения в языке и поэтому отсутствовать как элемент Тезауруса. Если второй участник этого отношения является типом, то устанавливается отношение онтологической зависимости (АСЦ1) от ролевого понятия к понятию-типу. Например, если в тезаурусе нет отдельного понятия *ВОЖДЕНИЕ ТРАКТОРА,* то можно записать:

ТРАКТОРИСТ
АСЦ1 *ТРАКТОР*

Особенностью социальных ролей является то, что контекст их реализации является более сложным, чем для реляционных и процессуальных ролей.

Рассмотрим это различие на группе процессуальных и социальных ролей, связанных с понятием-процессом *ВЕРХОВАЯ ЕЗДА*: *ВСАДНИК, НАЕЗДНИК, ПИКАДОР, ЖОКЕЙ, КАВАЛЕРИСТ, КОННИК (СПОРТСМЕН).*

Ролевое понятие *ВСАДНИК* представляет собой процессуальную роль, поскольку это название субъекта в процессе верховой езды. Роль всадника возникает только в конкретной ситуации верховой езды. Понятие *ПИКАДОР* также является процессуальной ролью, это вид всадника в ситуации боя быков.

Понятие *КАВАЛЕРИСТ, ЖОКЕЙ, КОННИК (СПОРТСМЕН)* являются социальными ролями, поскольку можно назвать человека кавалеристом, жокеем или конником, даже если он не сидит на коне в данным момент времени. Тем не менее присутствует онтологическая зависимость этих понятий от понятий ВЕРХОВАЯ ЕЗДА и ВСАДНИК, поскольку если бы в некотором гипотетическом мире не было бы лошадей или других животных для верховой езды, то не возникло бы и этих понятий.

В соответствии с (Loebe, 2005) социальные роли являются элементами некоторых социальных сущностей. Так, для понятия КАВАЛЕРИСТ такой социальной сущностью является КАВАЛЕРИЯ, а для понятия *КОННИК (СПОРТСМЕН)* в качестве такой социальной сущности выступает сфера деятельности *КОННЫЙ СПОРТ*.

Слово *наездник* может употребляться и как процессуальная роль как синоним слова *всадник,* и как социальная роль в значении близком к слову *жокей.*

Приведем примеры описания отношений для некоторых из вышеупомянутых понятий:

ВЕРХОВАЯ ЕЗДА
ЧАСТЬ *ВСАДНИК*
АСЦ2 *КОННЫЙ СПОРТ*
АСЦ2 *КАВАЛЕРИЯ*

ВСАДНИК
НИЖЕ *ПИКАДОР*
АСЦ2 *КОННИК (СПОРТСМЕН)*
АСЦ2 *КАВАЛЕРИСТ*

КАВАЛЕРИСТ
ВЫШЕ *ВОЕННОСЛУЖАЩИЙ*
ЦЕЛОЕ *КАВАЛЕРИЯ*
АСЦ1 *ВСАДНИК*

КАВАЛЕРИЯ
ВЫШЕ *ВОЙСКО*
АСЦ1 *ВЕРХОВАЯ ЕЗДА*
ЧАСТЬ *КАВАЛЕРИСТ*

КОННИК (СПОРТСМЕН)
ВЫШЕ *СПОРТСМЕН*
ЦЕЛОЕ *КОННЫЙ СПОРТ*
АСЦ1 *ВСАДНИК*

Интересной особенностью социальных и процессуальных ролей является наличие многочисленных примеров, когда одно и то же слово может в разных контекстах выражать как социальную роль, так и процессуальную роль.

Так, например, слово *учитель* может означать субъект в конкретной ситуации обучения, а может означать должность (что является социальной ролью) школьного учителя. Или, например, слово *баскетболист* может обозначать людей, играющих в баскетбол в данный момент (процессуальная роль), а также профессиональных

спортсменов, которые в баскетбол сейчас не играют, а идут, например, по улице (социальная роль).

17.8. Тезаурус РуТез как структура

Таким образом, тезаурус РуТез представляет собой сочетание двух иерархических структур: таксономии (иерархии родовидовых отношений) и партономии (иерархии отношений ЧАСТЬ-ЦЕЛОЕ). И таксономия, и партономия являются структурами с множественным наследованием, то есть понятие в этих структурах может иметь более одного понятия более высокого уровня. Кроме того, между понятиями могут быть установлены понятия симметричной и несимметричной ассоциации.

Отношения тезауруса имеют свойства направленности: направленность вверх, направленность вниз и горизонтальная направленность (см. табл.18.3.):

Отношение	Направленность
ВЫШЕ	вверх
ЦЕЛОЕ	вверх
АСЦ1	вверх
НИЖЕ	вниз
ЧАСТЬ	вниз
АСЦ2	вниз
АСЦ	вверх, вниз

Таблица 17.3. Направленность отношений тезауруса РуТез

Тезаурус РуТез может быть представлен как связный двунаправленный граф

Для приложений автоматической обработки текстов, в которых используется тезаурус РуТез, важными являются следующие понятия.

Путь по иерархии вверх: Между понятиями С0 и С00 существует путь по иерархии вверх, если существует такой набор понятий Ci, что последовательность понятий С0, С1…Ci, С00 образует путь и посредством применения свойств отношений совокупность отношений R(C0, C1), R(C1, C2)…R(Ci-1,Ci), R(Ci, C00) может быть преобразовано в одно отношение, имеющее значение свойства направленность «вверх».

Путь по иерархии вниз: Между понятиями С0 и С00 существует путь по иерархии вниз, если существует такой набор понятий Ci, что последовательность понятий С0, С1…Ci, С00 образует путь и посредством применения свойств отношений совокупность отношений R(C0, C1), R(C1, C2)…R(Ci-1,Ci), R(Ci, C00) может быть преобразовано в одно отношение, имеющее значение свойства направленность «вниз».

Иерархический путь. Иерархический путь – это путь по иерархии вниз или путь по иерархии вверх.

Дерево-вверх. Дерево-вверх понятия С0 представляет собой набор понятий Ci, для которых существуют пути по иерархии вверх от понятия С0 до понятий Ci.

Дерево-вниз. Дерево-вниз понятия С0 представляет собой набор понятий Ci, для которых существуют пути по иерархии вниз от понятия С0 до понятий Ci.

Тезаурусная окрестность понятия. Окрестность понятия С0 представляет собой набор понятий тезауруса, принадлежащих дереву-вверх понятия С0 или дереву-вниз понятия С0 (см. рис. 17.3)

Путь с перегибом-вверх. Между понятиями С0 и С1 существует путь с перегибом-вверх, если понятие С1 не принадлежит к тезаурусной окрестности понятия С0, и понятие С0 не принадлежит тезаурусной окрестности понятия С1, и существует понятие Ci такое, что Ci принадлежит дереву–вверх понятия С0 и дереву-вверх понятия С1.

Путь с перегибом-вниз. Между понятиями C0 и C1 существует путь с перегибом-вниз, если понятие C1 не принадлежит к тезаурусной окрестности понятия C0, и понятие C0 не принадлежит тезаурусной окрестности понятия C1, и существует понятие Ci такое, что Ci принадлежит дереву–вниз понятия C0 и дереву-вниз понятия C1.

Перед началом любой обработки текста на основе тезауруса для каждого понятия тезауруса строится дерево-вниз, которое позволяет быстро устанавливать отношения между понятиями анализируемого текста, связанных между собой иерархическими путями.

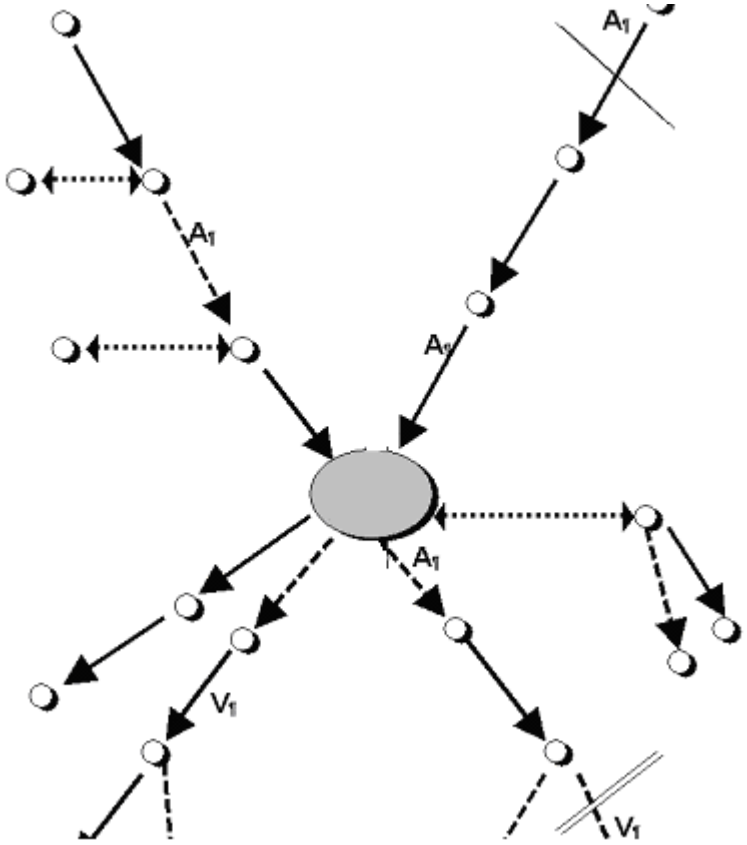

Рис. 17.3. Структура тезаурусной окрестности понятия

Заключение к главе 17

Отношения в ресурсах, предназначенных для использования в приложениях информационного поиска, могут использоваться в разных функциях, таких как расширение поискового запроса, вывод рубрики документа, разрешение многозначности, выявление структуры текста на базе моделирования структуры его лексической связности.

Для качественного выполнения всех этих функций важно обеспечить многошаговый логический вывод, что может быть достигнуто на базе свойств транзитивности и наследования. Кроме того, при описании отношений необходимо добиться того, чтобы отношения были максимально «надежными», не зависели от контекста упоминания понятия.

Для обеспечения этих свойств мы предложили использовать небольшой набор отношений, сопоставимый с набором отношений в традиционных информационно-поисковых тезаурусах. Однако мы ввели более строгие онтологические определения используемых отношений:

1) Для установления родовидового отношения используются онтологические критерии;

2) Важное в некоторых случаях отношение от типа к роли может быть установлено, но должно быть снабжено специальной пометой;
3) При установлении отношения ЧАСТЬ-ЦЕЛОЕ требуется, чтобы существование части зависело от существования целого, тогда удается обеспечить транзитивность отношения ЧАСТЬ-ЦЕЛОЕ.
4) Тезаурусное отношение ассоциации разделено на два отношения: симметричную ассоциацию и несимметричную ассоциацию. В качестве несимметричной ассоциации устанавливается онтологическое отношение внешней концептуальной зависимости.

Такая система отношений отражает наиболее существенные взаимосвязи между сущностями, может применяться (и применяется в наших ресурсах) для описания отношений между понятиями, не относящимися к конкретным предметным областям, а также в самых разных предметных областях.

Предложенная система отношений более формализована, чем система отношений в традиционных информационно-поисковых тезаурусах, и в ресурсах типа WordNet. Каждое отношение связано со своим набором правил вывода, которые используются во всех приложениях, в которых используется тезаурус РуТез.

При начале работ с новой предметной областью большое преимущество получается от того, что сразу понятно, какой минимальный набор отношений нужно использовать для вводимых понятий. Даже если в дальнейшем для конкретного приложения будет необходимо использовать более разнообразный набор отношений, описываемые отношения настолько важны для предметной области, что несомненно сохранятся при любой системе отношений, только могут получить новые имена.

Заключение к части 4

Таким образом, мы предлагаем модель описания знаний о мире, знаний в некоторой предметной области в форме лингвистической онтологии, предназначенной для использования в приложениях информационного поиска, требующих автоматической обработки текстов.

Модель построена на сочетации принципов трех различных традиций и методологий разработки компьютерных ресурсов:

- методологии разработки традиционных информационно-поисковых тезаурусов;
- методологии разработки лингвистических ресурсов типа WordNet (Принстонский университет);
- методологии созданий формальных онтологий.

Сходство с методологией разработки традиционных информационно-поисковых тезаурусов заключается в следующих решениях:

- формирование однозначного имени для понятия тезауруса подобно дескрипторам традиционных информационно-поисковых тезаурусов,
- работа с многословными выражениями, ввод понятий на основе значений многословных выражений подобно принципам ввода дескрипторов традиционных информационно-поисковых тезаурусов,
- небольшой набор отношений между понятиями тезауруса; набор отношений пригоден для широких неструктурированных предметных областей. Также и система отношений традиционных информационно-поисковых тезаурусов (отношения выше-ниже, ассоциация) при всех их недостатках были хороши тем, что могут применяться для многих предметных областей.

Сходство с методами разработки тезаурусов типа WordNet заключается в подробной работе с лексическими единицами, тщательной работой со значениями многозначных слов.

Сходство с методологией разработки онтологий заключается в том, что единицы тезауруса должны быть отличимы от близких единиц в сети тезауруса. Кроме того,

большие усилия прикладываются к тому, чтобы набор отношений устанавливался по формльным правилам, с использованием онтологических принципов. Используются процедуры логического вывода, в частности, активно используется транзитивность отношений часть-целое.

Предложенная модель позволяет в короткие сроки создавать онтологические ресурсы в неструктурированных предметных областях. При этом созданный ресурс, с одной стороны, будет содержать подробное описание терминологии предметной области, а также необходимые общелексические единицы, и, с другой стороны, будет иметь внутреннюю структуру, соответствующую современным онтологическим принципам разработки онтологий в виде отличимых понятий и формальных отношений между понятиями. Эксперименты по применению созданных по данной модели ресурсов в различных задачах информационного поиска будут рассмотрены в следующей части книги.

ЧАСТЬ 5. ТЕЗАУРУС РУТЕЗ В КОМПЬЮТЕРНЫХ ПРИЛОЖЕНИЯХ

Глава 18. Построение тезаурусного индекса, автоматическое разрешение лексической многозначности

Применение тезаурусов и онтологий в информационном поиске требует высокого качества разрешения многозначности слов (см. главу 10). Так, в работе (Sanderson, 1994) обосновывалось, что для того, чтобы в информационном поиске мог проявиться положительный эффект от разрешения лексической многозначности, точность разрешения многозначности должна быть не меньше 90%, в работе (Gonzalo и др., 1998) на основании результатов проведенных экспериментов указывается необходимая величина точности разрешения многозначности – 70%.

В данной главе мы рассмотрим, как проводится сопоставление текста с тезаурусом РуТез, как осуществляется автоматическое разрешение многозначности тезаурусных единиц, и какова точность этой процедуры.

18.1. Построение тезаурусного индекса и тезаурусной проекции

На первом этапе обработки текстов на основе тезауруса производится сравнение единиц текста с единицами Тезауруса.

Сравнение текста и Тезауруса происходит на основе морфологического представления единиц текста и единиц Тезауруса. Последовательности лемм, сопоставленные тезаурусному входу, сопоставляются с последовательностями лемм документа.

При необходимости в процессе сопоставлении текста с РуТез онтологией могут быть применены методы неточного сопоставления (с появлением лишних слов внутри словосочетания, сменой порядка слов, применение словообразовательных вариантов и т.п.) или сопоставление на основе синтаксических структур. Но нужно учитывать, что в первом случае упадет точность сопоставления, во втором дополнительно возрастет сложность сопоставления.

Из множества найденных в конкретном месте текста единиц Тезауруса выбирается единица, имеющая максимальную длину. Если один и тот же фрагмент текста соответствует разным единицам Тезауруса, то фиксируется многозначность термина.

В результате сопоставления с Тезаурусом текст отражается в последовательность понятий Тезауруса. Все синонимы (варианты) одного и того же понятия отображаются в соответствующий номер понятия и далее не различаются. Для каждого понятия Тезауруса фиксируется частота его встречаемости в тексте. Таким образом, после разрешения многозначности языковых выражений (см. п.18.2.) создается так называемый концептуальный индекс документа, в котором синонимы сведены к одному и тому же понятию, а разные значения разведены к разным понятиям.

Для учета отношений между понятиями, найденными в тексте, для всех понятий, связанных иерархическими путями (см. п.17.8) устанавливаются непосредственные отношения, которые выводятся на основе этих иерархических путей. Такая процедура осуществляется за счет заранее построенного дерева-вниз для всех понятий тезауруса.

Совокупность связанных между собой понятий текста, полученных в результате применения процедуры вывода, называется проекцией Тезауруса на текст (тезаурусной проекцией).

Следует отметить, что в подавляющем числе описываемых в дальнейшем приложений обработка текста производится не на полном объеме тезауруса РуТез, а на базе Общественно-политического тезауруса, к понятиям которого в случае необходимости с помощью специальной разметки добавляются те понятия Общего Лексикона, которые важны для данного приложения. В дальнейшем эту расширенную совокупность понятий мы все равно будем называть Общественно-политическим тезаурусом.

Такое решение связано с двумя факторами.

Во-первых, многозначность текстовых входов в рамках Общественно-политического тезауруса значительно ниже, чем текстовых входов Общего лексикона, и, как мы увидим в дальнейшем, точность разрешения многозначности для текстовых входов Общественно-политического тезауруса значительно выше.

Во-вторых, производится в основном тематическая обработка текстов, для которой важно упоминание тех или иных тематически-определенных сущностей в тексте, а не отношений между ними, основные понятия, соответствующие таким сущностям, сосредоточены именно в Общественно-политическом тезаурусе.

Для большинства текстов, тезаурусная проекция представляет собой сложную сеть отношений, которая может распадаться на несколько несвязанных фрагментов, а может содержать достаточно много различных связанных между собой понятий.

Рассмотрим пример текста Постановления Правительства РФ от 26 июня 1995 г. № 604:

О порядке оказания **безвозмездной финансовой помощи** на **строительство (покупку) жилья** и **выплаты денежной компенсации** за **наем (поднаем) жилых помещений военнослужащим** и **гражданам, уволенным** с **военной службы**

Во исполнение **Закона Российской Федерации** "О **статусе военнослужащих**" и в целях обеспечения **прав** на **жилище военнослужащих** и **граждан, уволенных** с **военной службы, Правительство Российской Федерации** п о с т а н о в л я е т :

1. **Утвердит**ь прилагаемое Положение о порядке оказания **безвозмездной финансовой помощи** на **строительство (покупку) жилья** и **выплаты денежной компенсации** за **наем (поднаем) жилых помещений военнослужащим** и **гражданам, уволенным с военной службы.**

2. **Министерству обороны Российской Федерации** и иным **федеральным органам исполнительной власти,** в которых предусмотрена **военная служба**:

в месячный срок разработать и утвердить формы и перечень **документов,** необходимых для принятия решения об оказании **военнослужащим безвозмездной финансовой помощи** на **строительство (покупку) жилья** и о **выплате денежной компенсации** за **наем (поднаем) жилых помещений;**

расходы, связанные с оказанием **военнослужащим безвозмездной финансовой помощи** и **выплатой денежной компенсации** за **наем (поднаем) жилых помещений,** производить за счет и в пределах **средств,** выделяемых из **федерального бюджета** по **сметам** этих **федеральных органов исполнительной власти.**

3. **Органам исполнительной власти субъектов Российской Федерации**:

оказывать **безвозмездную финансовую помощь** в **избранном постоянном месте жительства гражданам, уволенным** с **военной службы,** осуществляющим **строительство (покупку) жилья,** за счет и в **пределах средств федерального бюджета,** выделяемых на **жилищное строительство** для этой категории **граждан;**

Полужирным шрифтом показаны те сущности, которые были найдены в качестве текстовых входов Общественно-политического тезауруса. На рис. 18.1 показан фрагмент тезаурусной окрестности для этого текста, который включает взаимосвязанную совокупность понятий тезауруса: *СООРУЖЕНИЯ – ЖИЛЬЕ – СТРОИТЕЛЬСТВО ЖИЛЬЯ – ЖИЛИЩНО-СТРОИТЕЛЬНЫЙ КООПЕРАТИВ – ПОКУПКА – ПРОДАЖА*.

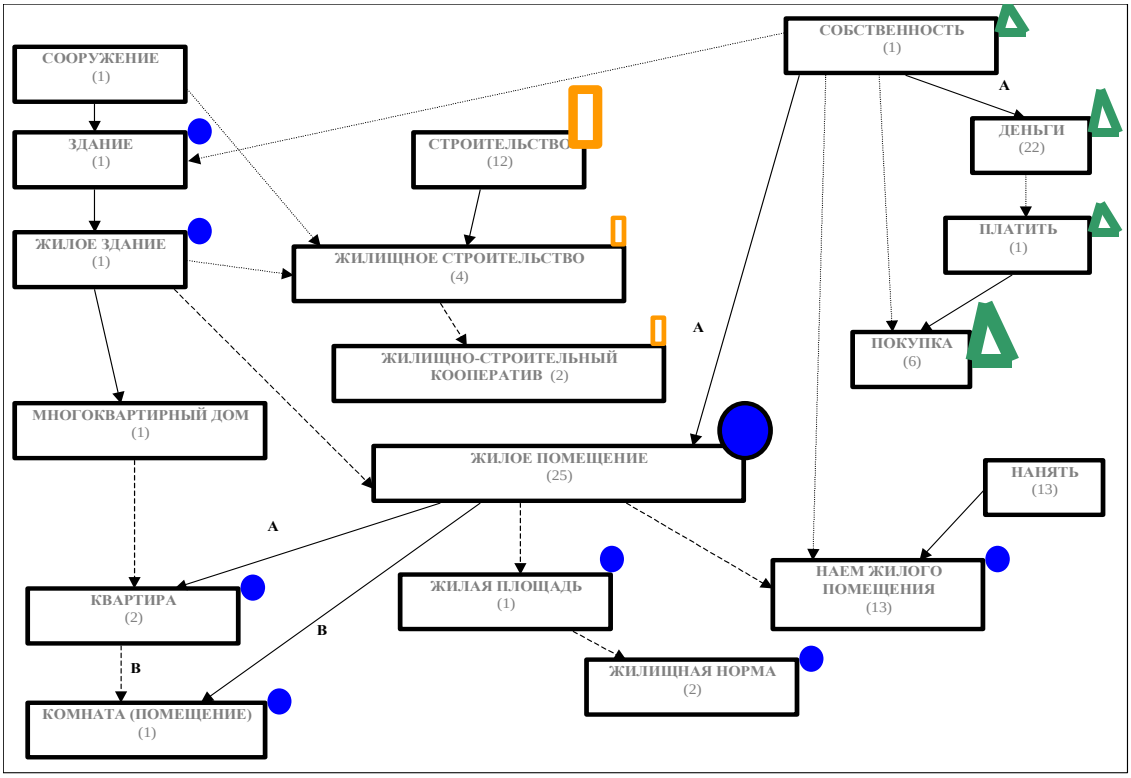

Рис. 18.1. Фрагмент понятийной сети (тезаурусной проекции) для текста Постановления Правительства РФ от 26 июня 1995 г. № 604

18.2. Автоматическое разрешение многозначности

При автоматической обработке текста на основе тезауруса РуТез первым этапом является сопоставление текста с единицами тезауруса и создание концептуального индекса, в котором указываются те понятия, которые встречались в тексте. Многозначность в этом индексе проявляется либо в сопоставлении одной и той же языковой единице разных понятий, либо в специальной пометке понятия, означающей, что текстовая единица, по которой было проведено сопоставление, является многозначной (см. п. 16.4).

Как указывалось в предыдущем разделе, на втором этапе строится так называемая проекция тезауруса для анализируемого текста. Проекция включает в себя понятия индекса и тезаурусные отношения между такими понятиями, которые входят в тезаурусную окрестность друг друга.

В тезаурусную проекцию текста включаются и все варианты понятия, соответствующие многозначным текстовым входам тезауруса. Для них также выявляются все понятия, упомянутые в тексте и входящие в их тезаурусные окрестности.

Для разрешения многозначности текстовых входов тезауруса было предложено и экспериментально проверено два метода: метод глобального подтверждения и метод взвешивания подтверждения от локального и глобального контекстов, которые мы рассмотрим в следующих разделах.

18.2.1. Метод глобального подтверждения

Метод глобального подтверждения заключается в том, что все понятия, вхождения которых обнаружены в тексте, могут оказывать влияние на выбор значения многозначного языкового выражения. Рассмотрение глобального контекста учитывает такое свойство связного текста как лексическую связность текста, то есть повторяемость

одних и тех же лексических единиц и совокупностей семантически близких лексических единиц в связном тексте (Лукашевич, 1996; Лукашевич, Добров, 2007).

Для каждого варианта многозначного выражения собираются те понятия текста, которые "поддерживают" этот вариант. "Поддержка" текста проявляется двумя способами:

- в тексте встречается однозначный вариант помеченного понятия, например, упоминание в тексте словосочетания *расследование преступлений* поддерживает именно это значение у многозначного слова *следствие*.
- в тексте встречается понятие из тезаурусной окрестности неоднозначного термина, например, упоминается понятие *ОБЩЕСТВЕННАЯ ДЕЯТЕЛЬНОСТЬ* из тезаурусной окрестности неоднозначного термина *партия*.

Далее собственно и производится выбор варианта понятия для многозначного термина. Как указывалось в п. 16.4, многозначность в тезаурусе РуТез может быть задана двумя способами: с помощью пометы и с помощью отнесения текстового выражения к разным понятиям тезауруса. Процедура автоматического выбора значения в этих случаях несколько различается:

- неоднозначность задана с помощью пометы. Если текст "поддерживает" описанное в тезаурусе значение неоднозначного термина, то соответствующее понятие включается в понятийный индекс как однозначный. В противном случае, неоднозначный термин исключается из понятийного индекса.
- неоднозначность проявляется в соответствии одного текстового выражения нескольким понятиям. Сначала проверяется, какие из вариантов термина поддерживаются понятиями всего текста, и оставляются только "поддержанные" варианты. Если ни один из вариантов не поддерживается, то все они удаляются из понятийного индекса.

После удаления "неподдержанных" вариантов может остаться только один вариант, и, таким образом, неоднозначность разрешена.

Если же поддержано более одного варианта, то производится выбор значения именно для конкретного вхождения неоднозначного термина: выбирается тот вариант, для которого "поддерживающее" понятие находится ближе всего по тексту. Расстояние измеряется в количестве выявленных понятий между текущим вхождением неоднозначного термина и "поддерживающим" понятием.

Далее этот метод разрешения многозначности мы будем называть Glob.

Данный алгоритм очень прост, однако в нем есть некоторые проблемы.

Во-первых, в этом методе для учета концептуальной близости используются только пути, состоящие из иерархических отношений одной направленности, то есть без перегибов, таким образом, семантически близкими считались только понятия, находящиеся в иерархических отношениях между собой. Это приводило к явным проблемам на относительно коротких текстах, таких как новостные сообщения, когда необходимые для подтверждения иерархически расположенные понятия не входили в состав анализируемого текста.

Во-вторых, нет ограничений на длину пути между понятиями, что приводило, например, к тому, что многозначность очень конкретного понятия могла быть разрешена на основе нахождения в тексте очень абстрактного понятия.

В-третьих, не имеется весовой оценки семантической близости между понятиями на основе путей между ними или каких-либо других: подтверждение производилось на основе принципа «да-нет».

В-четвертых, приоритет отдавался глобальному контексту, то есть сначала проверялось, если ли подтверждение для того или иного значения по всему тексту. Если несколько значений имели подтверждение в глобальном контексте, то проверялся

локальный контекст: выбиралось то значение, подтверждение для которого находилось ближе всего к исследуемому многозначному вхождению.

Поэтому был предложен другой алгоритм разрешения многозначности, который должен более аккуратно учитывать разные характеристики путей между понятиями тезауруса.

18.2.2. Метод взвешивания подтверждения от локального и глобального контекстов

Основой для разработанного алгоритма разрешения многозначности является оценка семантической близости между возможными значениями, с одной стороны, и окружающим текстовым контекстом, с другой стороны. При этом рассматривается как локальный контекст, который задается в виде некоторого окна – линейной окрестности многозначного вхождения слова, так и глобальный контекст, в который входят все слова текста (Лукашевич, Чуйко, 2007).

18.2.2.1. Учет локального и глобального контекста

В качестве локального контекста рассматривается фиксированная линейная окрестность многозначного вхождения слова, измеряемая в количестве найденных элементов тезауруса, - исследовался размер окна окрестности от 1 до 5 элементов в обе стороны.

Также мы исследовали задание локального контекста как «динамического» окна N+N, то есть сначала происходит попытка выбора значения слова в окрестности длиной N, если это удается, то обработка данного вхождения заканчивается. Если не удается, то происходит расширение окрестности еще на N элементов и процедура выбора значения продолжается. Тестировались такие динамические окна как 1+1, 2+2, 3+3.

При использовании глобального контекста возникает вопрос о том, насколько в достаточно длинном тексте правомерно использование полного текста как базы для выбора значения, не нужно ли вводить некоторые ограничения, например, на расстояние (в абзацах, предложениях) между данным многозначным вхождением и упоминанием семантически близкого понятия в тексте. Так, в работе (Galley, McKeown, 2003) разные типы связи имеют разную сферу действия и разный вес в зависимости от такого рода расстояния, измеряемого в абзацах и предложениях.

В процессе экспериментов нами была выбрана следующая специфика учета глобального контекста.

В качестве элементов глобального контекста учитываются только однозначные вхождения тезаурусных единиц. Мы не накладываем никаких ограничений на расстояние между вхождением многозначного слова и семантически близкими словами не вводится. Предполагается, что возможное неправильное подтверждение от далекой части текста должно преодолеваться правильным подтверждением от локального контекста и более близкой части текста.

Поскольку локальный контекст достаточно ограничен, а глобальный контекст может достигать весьма большой величины, то необходимо сбалансировать свидетельства в пользу того или иного значения, получаемые от локального и глобального контекста. Прежде всего, вес подтверждения значения, получаемый от некоторой лексической единицы в локальном контексте всегда выше, чем от той же единицы, расположенной вне локального контекста. Кроме того, мы тестировали возможность применения коэффициента, уменьшающего вес подтверждения от глобального контекста при увеличении длины текста (точнее при увеличении максимальной частотности лексической единицы в тексте).

18.2.2.2. Семантическая близость понятий как функция от особенностей пути отношений между ними

Семантическая близость между двумя понятиями *C1* и *C2* оценивается на основе рассмотрения пути отношений, который существует между этими единицами тезауруса.

Между понятиями в тезаурусе могут существовать пути разной конфигурации, тезаурус связен и всегда существует путь отношений от одного произвольного понятия тезауруса до другого понятия тезауруса. Однако подобно подходу (Hirst, St-Onge 1998) мы ограничиваем конфигурации путей между понятиями *C1* и *C2*, которые рассматриваются при оценке семантической близости понятий, а именно, либо путь должен состоять из совокупности иерархических отношений, направленных в одну сторону, например, последовательность отношений от вида к роду (иерархический путь - см.п.17.8), либо такой путь должен включать ровно один перегиб, то есть изменение направления движения (путь с перегибом). При этом рассматриваются перегибы двух видов: перегиб-сверху, например, сначала несколько отношений от видовых понятий к родовым, затем несколько отношений от родовых понятий к видовым, так и перегиб-снизу.

Как мы описывали в предыдущей главе, в тезаурусе РуТез имеется три вида иерархических отношений ВЫШЕ-НИЖЕ, ЧАСТЬ-ЦЕЛОЕ и несимметричная ассоциация АСЦ1-АСЦ2. Таким образом, три отношения (ВЫШЕ, ЦЕЛОЕ, АСЦ1) направлены по иерархии вверх, а три отношения (НИЖЕ, ЧАСТЬ и АСЦ2) – по иерархии вниз.

Для родовидового отношения ВЫШЕ-НИЖЕ определены свойства транзитивности и наследования, отношение ЧАСТЬ-ЦЕЛОЕ также рассматривается как транзитивное отношение.

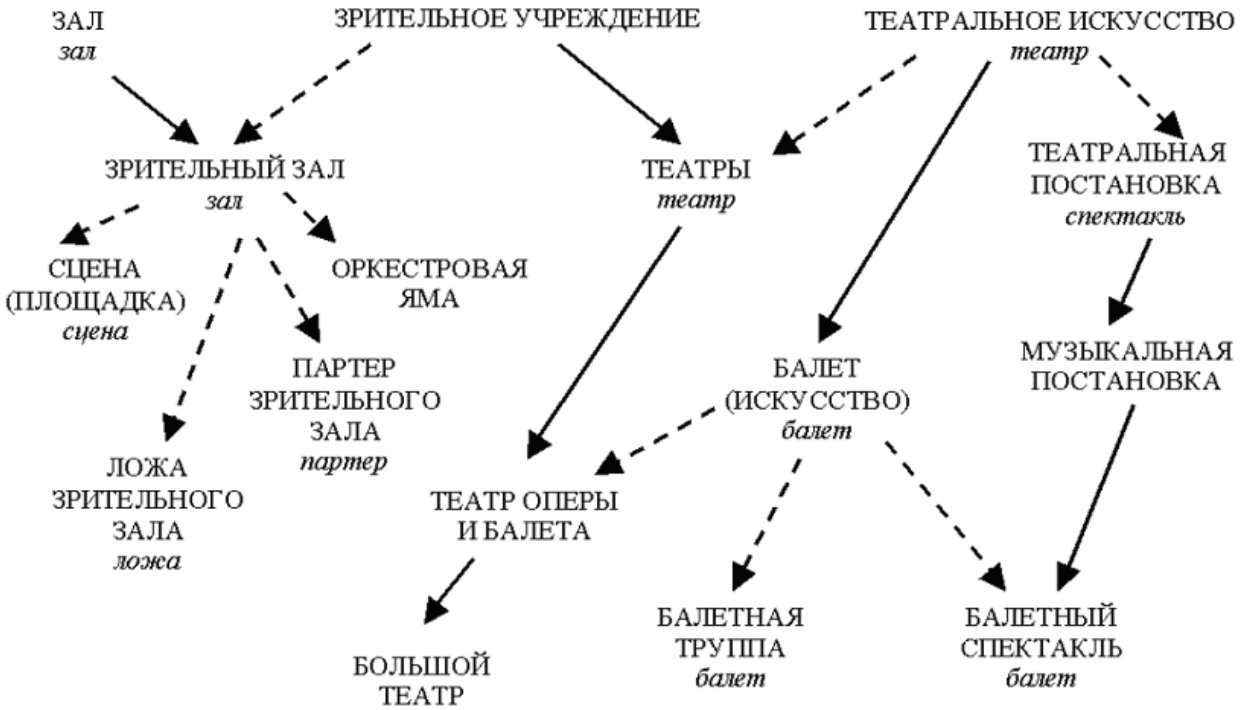

Рис. 18.2. Фрагмент тезаурусной сети понятий текста примера в главе 17 с многозначными текстовыми входами

На рис.18.2 примером иерархического пути является путь

БОЛЬШОЙ ТЕАТР

-- (ВЫШЕ) – ТЕАТР ОПЕРЫ И БАЛЕТА

-- (ЦЕЛОЕ) -- БАЛЕТ (ИСКУССТВО),

Примером пути с перегибом сверху является путь:

ОРКЕСТРОВАЯ ЯМА

-- (ЦЕЛОЕ) -- ЗРИТЕЛЬНЫЙ ЗАЛ

-- (ЧАСТЬ) -- ПАРТЕР ЗРИТЕЛЬНОГО ЗАЛА,

Примером пути с перегибом снизу является путь:

ТЕАТРАЛЬНАЯ ПОСТАНОВКА

-- (НИЖЕ) -- БАЛЕТНЫЙ СПЕКТАКЛЬ

-- (ВЫШЕ) -- МУЗЫКАЛЬНАЯ ПОСТАНОВКА.

Построение разрешенных путей осуществляется следующим образом.

Для каждого понятия тезауруса можно определить совокупность иерархически вышестоящих понятий – так называемое «дерево-вверх». «Дерево-вверх» понятия С0 включает те понятия тезауруса, к которым от С0 может быть проведен путь, состоящий из отношений одной направленности, и который с помощью правил наследования и транзитивности может быть сведен к одному отношению. Схожим образом, на основе иерархических отношений, направленных вниз, определяется совокупность иерархически нижестоящих понятий - «дерево-вниз».

Так, например, на рис.18.2 для понятия *БОЛЬШОЙ ТЕАТР* можно видеть следующие вышестоящие по иерархии понятия (понятия из «дерева-вверх»): *ТЕАТР ОПЕРЫ И БАЛЕТА, ТЕАТР, ТЕАТРАЛЬНОЕ ИСКУССТВО, ЗРИТЕЛЬНОЕ УЧРЕЖДЕНИЕ*.

Таким образом, между двумя понятиями существует путь разрешенной структуры, если либо одно из понятий входит в дерево-вниз или в дерево-вверх другого понятия, либо если между их деревьями имеется непустое пересечение.

18.2.2.3.Числовая оценка семантической близости

Семантическая близость понятий, связанных путем заданной конфигурации, зависит от особенностей пути между понятием-значением и подтверждающим понятием:

- чем длиннее путь между понятиями, тем слабее семантическая близость;
- наличие перегиба на пути ослабляет семантическую близость;
- разные типы перегибов на пути могут по-разному влиять на семантическую близость;
- перегиб пути на высоком уровне иерархии хуже, чем на более низком уровне.

Кроме того, учитывался тот факт, что подтверждение от лексической единицы, которая в свою очередь многозначна, возможно, должно быть слабее. Например, в тексте примера во фрагменте «*светила другая, куда более загадочная звезда»* нахождение рядом слов *светила* и *звезда,* приводит к трактовке обоих слов как небесных тел.

Для учета такого рода рассуждений была применена следующая формула:

$$Sim_{new}(C1, C2) = максимальный\_балл - длина\_пути - цена\_многозначности - цена\_перегиба - цена\ глобальности. \quad (18.1)$$

Максимальный балл представляет собой максимально возможную оценку подтверждения, связанную с тем, что встретился однозначный синоним рассматриваемого многозначного термина. В настоящее время, величина максимального балла равняется 10.

Параметр *цена_глобальности* составляет величину, большую нуля, в случае оценки глобального контекста и величину, равную нулю, при анализе локального контекста.

18.2.2.4. Этапы алгоритма

Поступающий текст проходит через процедуру графематического и морфологического анализа. Далее на основе цепочек лемм, полученных в результате морфологического анализа, происходит сопоставление с тезаурусом. Для каждой сопоставившейся тезаурусной единицы отмечается ее статус: однозначное сопоставление, сопоставление с пометкой многозначности (А-многозначность), сопоставилось несколько единиц тезауруса (М-многозначность). Отметим, что если одна из сопоставленных тезаурусных единиц, полностью включается в другую тезаурусную единицу, то эта ситуация многозначной не считается, сопоставленной считается более длинная тезаурусная единица

Процедура разрешения многозначности начинается с анализа глобального контекста. Для каждого значения неоднозначных единиц текста анализируется, упоминались ли в тексте понятия, семантическая близость которых к текущему понятию, составляет число баллов, большее 0, по формуле (18.1). Все набранные баллы понятий-значений многозначных единиц суммируются и запоминаются.

Далее происходит анализ локального контекста. Для каждого вхождения многозначной тезаурусной единицы просматривается заданная текстовая окрестность, выбираются упоминаемые понятия, связанные с понятиями данной многозначной единицы тезаурусными путями разрешенной конфигурации, и подсчитываются баллы по формуле (6). Баллы, полученные при глобальном анализе и локальном анализе, суммируются.

Для каждого вида многозначности задается свой порог. Если понятия-значения, получили баллы, меньшие, чем заданный порог, то считается, что ни одно значение не подтвердилось, возможно, в тексте использовано какое-то другой значение.

Если понятие единицы с А-многозначностью получает количество баллов, большее чем установленная пороговая величина, то это значение подтверждается и, соответственно, выбирается.

Среди понятий для текстовой единицы с М-многозначностью выбирается значение, получившее максимальное количество баллов.

Если понятия единицы с М-многозначностью получили одинаковое количество баллов, превышающее пороговое, то выбирается вышестоящее по иерархии понятие, так, например, для значений слова *балет* таким понятием является понятие *БАЛЕТНОЕ ИСКУССТВО* (см. рис. 18.2*)*. В случае если такой иерархической связи не имеется, то в настоящее время не выбирается ни одно из понятий – многозначность остается неразрешенной. Если бы на основе разметки корпуса было бы известно наиболее частотное значение, то можно было бы в таких случаях выбирать именно это частотное значение.

Далее на этот алгоритм разрешения многозначности мы будем ссылаться LocGlob.

18.3. Организация тестирования алгоритмов разрешения многозначности

Для определения качества разрешения лексической многозначности необходимо было выполнить эталонную разметку найденных терминов по значениям. Для каждого документа экспертами-лингвистами были созданы эталонные файлы, с правильной разметкой значений.

После получения эталонных файлов они были автоматически сопоставлены с результатами работы программы разрешения многозначности. Были выделены следующие

случаи соответствия (несоответствия) эталонной разметки и результирующего файла работы программы:

1) Значение было выбрано правильно;
2) Значение не было выбрано, и это было правильно;
3) Значение было выбрано неправильно;
4) Значение не было выбрано, и это было неправильно;
5) Система выбрала один из правильных вариантов.

В качестве правильных решений системы рассматривались виды соответствия 1), 2) и 5). В качестве основной характеристики работы алгоритма оценивалась точность разрешения многозначности, которая рассчитывается как отношение между числом правильных решений и числом всех решений.

Число всех решений – это количество обнаруженных в тексте единиц тезауруса, отмеченных как многозначные. Таким образом, при сопоставлении одного и того же текста с Общественно-политическим тезаурусом количество решений, которое необходимо принять, меньше, чем при сопоставлении с объемлющим тезаурусом РуТез.

Тестировались следующие параметры алгоритма:

- максимальная длина дерева, то есть насколько далеко в одном и то же направлении иерархических отношений от исходного понятия можно искать подтверждающие значение понятия - длина дерева может быть различной для локального и глобального контекстов,
- строение (статическое или динамическое см. п. 18.2.2.1) и размер окна локального контекста,
- в локальном контексте: учитывать ли в полном объеме подтверждение от многозначного термина. Если снижать вес подтверждения в таких случаях, то каким образом: вычитать баллы, делить на коэффициент и т.п.,
- цена глобальности – насколько баллы, полученные от одного и того же подтверждения, меньше в глобальном контексте, чем в локальном.
- веса различных перегибов путей для локального и глобального контекста,
- пороги для видов многозначности: А-многозначности и М-многозначности.

Мы тестировали отдельно точность разрешения многозначности по Общественно-политическому тезаурусу, то есть определяли качество разрешения многозначности тематической лексики и терминологии, и по тезаурусу РуТез, то есть тестировалось качество разрешения многозначности для всех знаменательных слов текста. Последняя задача соответствует задаче тестирования «все слова текста», проводимой в рамках конференции Senseval (см. главу 10).

18.3.1. Тестирование алгоритмов разрешения многозначности на основе Общественно-политического тезауруса

Тестирование алгоритмов разрешения многозначности для терминов общественно-политического тезауруса проводилось на материалах газет и наборе новостных сообщений. Предварительно, случайным образом было выбрано несколько дат. Из коллекции Университетской информационной системы РОССИЯ (www.cir.ru) были выгружены газетные публикации, относящиеся к выбранным датам. Набор газетных публикаций включает полные номера газет «Известия», «Ведомости», «Независимая газета», «Комсомольская правда». Каждый номер содержит несколько десятков статей. Средний размер статьи около 5 Кб. За те же даты были взяты новостные сообщения из коллекции новостей Яндекса (данная коллекция распространяется в рамках экспериментов семинара РОМИП).

В процессе эксперимента вручную было размечено 197 документов, что соответствует полным номерам газет «Известия», «Независимая газета», «Ведомости», «Комсомольская правда» от 19 ноября 2003 года, а также было размечено 30 новостных

сообщений за ту же дату. Взятие полных номеров обеспечивает достаточно большое разнообразие тематики документов.

Результаты работы алгоритмов разрешения многозначности по каждому из источников показаны в Таблице 19.1, где N_{doc} - число документов, N_{amb} - число вхождений неоднозначных терминов, $P_{locglob}$ - точность по алгоритму LocGlob, P_{glob}- точность по алгоритму Glob.

Источник	N_{doc}	N_{amb}	$P_{glob+loc}$, %	P_{glob} , %
Известия	44	2525	**75.23**	72.00
Ведомости	62	2697	**77.89**	73.41
Независимая газета	42	2776	**68.14**	66.50
Комсомольская правда	49	2240	**66.74**	63.04
Яндекс-Новости	30	450	**75.05**	68.00
Всего	227	10688	**73.37**	68.77

Таблица 19.1. Точность разрешения лексической многозначности по источникам публикаций

Совокупная точность работы системы по более гибкому алгоритму LocGlob в процессе тестирования составила 73,37% и выросла на 6.7% относительно точности разрешения многозначности, полученной по алгоритму Glob.

Как и предполагалось, наибольший рост точности алгоритма, более гибко учитывающего конфигурации путей отношений тезауруса, а также локальный и глобальный контекст, удалось получить на относительно коротких текстах новостных сообщений. Рост точности разрешения многозначности на этих типах текстов составил более 10%.

Для получения лучших результатов тестировались разные наборы параметров алгоритма LocGlob.

К особенностям наилучшего набора параметров можно отнести следующие закономерности.

Были выбраны разные пороги для разных видов многозначности: 4 балла для А-многозначности, и 2 балла для М-многозначности. Такой результат является предсказуемым, поскольку при М-многозначности между собой «соревнуются» несколько значений, а при А-многозначности значение-контрагент находится вне зоны тезауруса.

Выяснилось, что подтверждение от многозначного термина в локальном контексте значимо так же, как и от однозначного термина. Эта закономерность не была очевидна, при ручном анализе было видно, что между парами многозначных терминов иногда возникают ложные корреляции, приводящие к выбору неправильных значений для обоих терминов.

Наилучшей оказалась динамическая окрестность локального контекста 3+3.

Лучший результат был получен для высоты деревьев 2 как для локального, так и для глобального уровня, то есть при поиске семантически близких терминов в среднем лучше использовать как подтверждение понятия, отстоящие от понятий, соответствующих многозначному выражению, общая длина пути не более 4 отношений.

Из всех типов перегибов «наихудшими», получившими максимальные баллы штрафа, оказались перегибы типа: *видовое_понятие1 – родовое понятие – видовое понятие_2*, что ожидалось, а также перегиб-внизу типа: *родовое понятие_1 – видовое понятие – родовое понятие_2*.

При анализе результатов работы алгоритмов, изложенных в Таблице 19.1, нужно подчеркнуть важное обстоятельство. Тезаурус содержит много однозначных словосочетаний, в состав которых входят многозначные слова, например, *министр обороны, уголовное дело, дополнительный отпуск*. При анализе текста эти многозначные

слова попадают внутрь многословных терминов, и задача разрешения их многозначности не возникает.

Однако если бы словосочетаний не было, то пришлось бы разрешать многозначность этих слов алгоритмически. Было подсчитано, что если учесть те многозначные слова, многозначность которых снимается за счет объемлющих словосочетаний, то точность разрешения многозначности на основе комплекса «многословные термины тезауруса + алгоритм разрешения» возросла бы в среднем на 5 процентов.

Также мы исследовали вопрос, насколько точность разрешения многозначности зависит от частотности многозначной единицы в тексте. Была выявлена интересная корреляция, что разрешение многозначных слов, встретившихся в тексте один раз, во всех подколлекциях на несколько процентов ниже, чем в целом по коллекции. Это означает, что точность разрешения для слов с большей частотностью выше, чем приведенная в таблице.

18.3.2. Тестирование алгоритма разрешения многозначности на запросах из правовой области

Исследуя эффект нового алгоритма по разрешению лексической многозначности для коротких текстов, мы сделали небольшую коллекцию 40 длинных запросов в области права из коллекции семинара по информационному поиску РОМИП (www.romip.ru), например, таких как *компенсация подоходного налога при приобретении недвижимости.* Для этой коллекции разрешение многозначности терминов Общественно-политического тезауруса по алгоритму LocGlob достигло величины 82.02%, в то время как точность прежнего алгоритма Glob на этих запросах составляла величину 48.31%.

Для такой коллекции параметры алгоритма LocGlob настраивались отдельно. Параметры, на которых были получены лучшие результаты для коллекции запросов, оказались совершенно иными, чем для коллекции статей: это максимальные величины деревьев – 7 шагов, минимальные пороги для обоих видов многозначности, минимальные цены перегибов.

Такие результаты привели к мысли, что можно сделать систему автоматической настройки параметров алгоритма в зависимости от длины обрабатываемого текста.

Был проведен следующий эксперимент: та же тестовая коллекция статей (см. раздел 18.3.1) была разделена на пять подколлекций по величине текстов. Мы пытались подобрать лучшие параметры для каждой группы текстов и выявить функцию изменения основных параметров. Однако в этом эксперименте четкой корреляции, позволяющей реализовать самонастройку параметров, не было выявлено. Группа самых коротких текстов статей давала неожиданно низкий результат разрешения многозначности, причем лучший результат - 71.02% был получен на параметрах более близких к параметрам всей коллекции, чем к лучшим параметрам, полученных для запросов.

18.3.3. Тестирование алгоритма разрешения многозначности по Тезаурусу РуТез

Для тестирования алгоритма разрешения многозначности по всему Тезаурусу РуТез, что соответствует задаче «все слова текста» конференции Senseval, было взято по 2 статьи из газет «Известия», «Комсомольская правда», «Независимая газета», «Ведомости». Количество многозначных единиц – 1120. Меньший объем коллекции объясняется значительно большими трудозатратами по подготовке эталонной разметки. Для алгоритма LocGlob была получена точность разрешения многозначности - 57.14%, с учетом разрешения за счет попадания в словосочетания, описанные в тезаурусе – 63.4%.

Для лучшего набора параметров этой коллекции характерна большая величина окна - используется динамическое окно 4+4.

Точность разрешения многозначности, показанная реализованным алгоритмом для задачи «все слова текста», не использующая размеченного корпуса, приблизительно соответствует результатам работы лучших систем на конференции SENSEVAL.

Мы получили этот результат без использования дополнительной информации о наиболее частотных значениях, без использования размеченного корпуса и т.п. Наилучший известный авторам алгоритм, использующий только WordNet, имеет точность - 50.89% на данных SENSEVAL-3 (напомним еще про 10% однозначных слов в тестовой коллекции этой конференции – см. п. 10.1).

Заключение к главе 18

Реализованные алгоритмы автоматического разрешения многозначности показали максимальную среднюю точность разрешения многозначности 73.37% для тематической лексики и терминологии Общественно-политического тезауруса, и 57.14% для всех знаменательных слов текста, то есть по тезаурусу РуТез в целом.

Возникает вопрос, много это или мало, и какое качество разрешения многозначности нужно обеспечить для качественной работы тезауруса в приложениях автоматической обработки текстов.

Качество разрешения многозначности для задачи «все слова текста» значительно превышает показатели, достигнутые для алгоритмов, работающих на основе WordNet в тех же условиях, то есть без учета информации из размеченного корпуса, и, в частности информации о самом частотном значении. Это, на наш взгляд, в значительной мере связано с более богатой отношениями структурой Тезауруса РуТез.

Однако представляется, что полученные результаты точности разрешения многозначности для задачи «все слова текста» даже лучших методов недостаточны для того, чтобы использоваться в реальных приложениях информационного поиска. Так, в начале этой главы мы приводили данные о том, что в экспериментах было показано, что для того, чтобы получить новое качество поиска по сравнению с пословными моделями необходимо обеспечить, по крайней мере, 70% точности разрешения многозначности.

С разрешением многозначности тематической лексики и терминологии Общественно-политического тезауруса ситуация принципиально другая. Достигнуты значительно более высокие результаты разрешения многозначности. Эти результаты потенциально могут быть увеличены за счет использования дополнительной информации (например, о самом частотном значении, которое можно выбирать при величинах оценки значений ниже пороговых или близких к пороговым).

Поэтому во многих приложениях мы более полагаемся на Общественно-политический тезаурус, а также исследуем комбинированные методы, сочетающих пословные методы обработки текстов и обработку по тематическим понятийным ресурсам, таким как тезаурусы и онтологии.

Глава 19. Общественно-политический тезаурус как средство построения тематического представления текста

19.1. Проблемы автоматического построения лексических цепочек

Как мы указывали в п.14.2, описания языковых выражений в тезаурусах, могут использоваться для выявления лексической связности текста, что обычно делается посредством построения так называемых лексических цепочек – совокупностей языковых выражений текста, близких по смыслу.

Основными критериями для построения лексических цепочек в большинстве подходов являются следующие:

- наличие и сила связей между лексемами, описанных в некотором ресурсе,
- расстояние между вхождениями лексем в тексте, измеряемое обычно в предложениях. Если расстояние от текущего слова до предшествующих вхождений лексической цепочки больше некоторого порога, то лексическая цепочка прерывается и начинается новая.

Возникает вопрос, достаточно ли вышеперечисленных критериев для построения лексических цепочек.

Второй вопрос, возможно связанный с первым, заключается в том, что являются ли лексические цепочки такими уж очевидными, поскольку, как мы увидим ниже, эксперименты по сравнению лексических цепочек, выделенных разными людьми, показали достаточно серьезное расхождение в представленных лексических цепочках. Второй вопрос связан с первым, так как важно понять, является ли такая субъективность неизбежной, или не учитывается какой-либо важный критерий построения лексических цепочек.

В следующих разделах мы рассмотрим вопросы критериев и субъективности выделения лексических цепочек подробнее.

19.1.1. Субъективность выделения лексических цепочек

Авторы работы (Hirst, Morris, 2003) указывают на субъективность рассмотрения лексической связности в тексте. Они рассматривают пример небольшого текста:

> *(*) How can we figure out what a text means. One could argue that the meaning is in the mind of the reader, but some people think that the meaning lies within the text itself."*

Отвечая на вопрос, каковы лексические цепочки, которые можно выделить в данном тексте, один автор статьи полагает, что видит две цепочки: «понимание», которые включают такие слова как *figure out, means, meaning, mind, think, meaning* и цепочка «текст», включающая слова *text, reader, text.* Второй автор также выделил две цепочки, но соотнес слова *means, meaning* с цепочкой «текст».

Действительно, при построении лексических цепочек текста (*) слова «значение», «значить» близки по смыслу как лексеме «текст», так и лексемам «думать», «узнать». Можно ли определить, кто из авторов статьи прав, или, может быть, слова «значение» и «значить» входят в две лексические цепочки?

Также в (Hirst, Morris, 2003) описывается следующий эксперимент по изучению согласия между читателями по выявлению лексической связности текста.

Пять участников эксперимента читают полуторастраничный текст из Reader's digest на тему роли киноактеров и киноперсонажей в формировании неправильных моделей ролевого поведения для детей.

Участники сначала должны прочитать статью и отметить каждую связанную по смыслу группу слов разным цветом. Затем каждая выделенная группа должна быть

перенесена на новый лист, и в группе близких слов нужно выделять пары слов и устанавливать между ними тип отношения.

Эти данные стали основой для оценки соответствия между восприятием текста каждым участником. Для каждой пары участников было вычислен коэффициент согласия, который определялся как процент слов, которые встретились в рассмотрении обоих участников, по отношению к общему числу слов, которые они использовали. В среднем для лексически связанных слов этот коэффициент составил 63%.

В работе (Hollingsworth, Teufel, 2005) описывается эксперимент по сравнению лексических цепочек, создаваемых разными людьми, на примере научной статьи Lee Lilian “Measures of distributional similarity”, опубликованной в трудах 37 конференции ACL (pp.25-32). В эксперименте участвовали 3 человека, которым было дано неограниченное время, чтобы создать наборы терминов, которые им кажутся близкими по смыслу в контексте исследуемой статьи.

Участникам были даны следующие инструкции:

- термин может состоять из одного слова или комбинации слов, взятых непосредственно из текста;
- слова, используемые в терминах, могут быть существительными, прилагательными или наречиями;
- возможные отношения между словами в цепочке близких слов могут быть следующими: разные формы одного и того же слова, синонимия, гиперонимия-гипонимия, меронимия или холонимия;
- не накладывались ограничения на размер или количество лексических цепочек.

Каждому аннотатору были даны список всех слов статьи, упорядоченные по мере частотности и максимальные именные группы, извлеченные из текста. Использование этих материалов носило вспомогательный характер.

В статье (Hollingsworth, Teufel, 2005) приводятся лексические цепочки, полученные двумя аннотаторами. В каждой цепочке выделен наиболее частотный элемент, который является как бы представителем цепочки.

Один аннотатор создал 12 лексических цепочек, второй аннотатор создал 22 лексические цепочки, причем имеется совпадение главных элементов лексических цепочек только в четырех случаях (с точностью до единственного/множественного числа): *similarity, probability, cooccurrance, distribution.*

Таким образом, в экспериментах были выявлены значительные расхождения в формировании лексических цепочек людьми, и возникает вопрос, является ли эта ситуация стандартным проявлением субъективности человеческих решений или при рассмотрении лексических цепочек не учитываются какие-то дополнительные факторы.

19.1.2. Построение лексических цепочек с учетом ситуативных отношений

Стандартным базовым ресурсом для построения лексических цепочек является тезаурус WordNet. Однако набор отношений в этом тезаурусе невелик. Многие авторы, занимавшиеся автоматическим построением лексических цепочек, указывали на одну из проблем построения лексических цепочек по WordNet – нехватку ситуативных отношений (см.п.14.2.4). Но появление такого рода отношений в ресурсе (в тезаурусе РуТез такие отношения есть), опять ставит вопрос о критериях выделения цепочек.

Рассмотрим следующий текст на медицинскую тему:

(**)

*Канадские **врачи** убили **пациента** передозировкой **наркотика***

*В Канаде начато расследование **несчастного случая** в **больнице** города Ред Дир, где **медики** по ошибке ввели **пациенту смертельную** дозу **опиоидного наркотика**, сообщает газета The Globe and Mail. 69-летний **пациент** поступил в **приемное отделение больницы** после **травмы грудной клетки**, которую он*

*получил во время конной прогулки. **Врач** назначил ему 10 миллиграммов **морфина** в качестве **обезболивающего** и отпустил домой.*

*По ошибке **медсестры пациенту** был сделан укол **гидроморфона** - похожего на **морфин** по названию и действию. Однако этот **препарат** гораздо сильнее - доза в 10 миллиграммов **смертельна**. Свою ошибку **медики** осознали после пересчета **наркотических средств** и сразу позвонили родственникам мужчины. Однако состояние **пациента** быстро ухудшилось, и он **умер** после возвращения в **больницу.***

*Расследование этого случая завершится в течение 10 дней. Как сообщают в **больнице, укол** сделала опытная **медсестра**, которая полностью признает свою ошибку. Однако есть вероятность, что после расследования ее все же признают невиновной. По заявлению министра **здравоохранения** провинции Альберта, главное - сделать, чтобы такая ошибка не повторилась. (*Источник: Mednovosti.ru)

В тексте содержится множество слов и словосочетаний, имеющих отношение к медицине: наркотики, больница, пациент, травма, морфин, обезболивающее, гидроморфон, медик, врач и др., По тезаурусу РуТез многие из этих терминов достаточно тесно связаны между собой, и возникает вопрос, должны ли все эти слова собраться в одну лексическую цепочку или несколько. Если разбивать на несколько лексических цепочек, то нужно понять, какие формальные критерии должны быть применены.

Следствием более богатой системы отношений в тезаурусном ресурсе является и то, что одно и то же слово может быть отнесено к разным лексическим цепочкам, хотя как указывалось в разделе 14.2. основополагающим принципом подавляющего большинства подходов, в которых изучается автоматическое построение лексических цепочек, является отнесение очередного слова только к одной лексической цепочке. Рассмотрим следующий фрагмент текста:

Президент Украины Виктор Ющенко готовит указ о переносе парламентских выборов. Теперь, предположительно, они пройдут в июне.

*Первоначально Ющенко назначил их на 27 мая. Депутаты отреагировали на это решение обращением в Конституционный суд. Тот обещал спешно рассмотреть вопрос, но до сих пор так и не начал слушания. После подписания нового указа, суд не сможет начать дело, пока 45 депутатов не пришлют ему новое обращение. Так Ющенко затягивает решение главного украинского вопроса: имел ли право глава государства распустить **Верховную раду.***

*По мнению парламента, который Ющенко рассчитывает переизбрать, президент не уверен в своей правоте. Потому и начал сложную игру. Политические страсти, утихшие на Украине во время Пасхи, разгорелись с новой силой. (*Источник: Российская газета)

Очевидно, что словосочетание *Верховная рада* должно быть в равной степени отнесена к двум лексическим цепочкам – цепочке парламента *(парламентских выборов, депутаты, депутатов, Верховную Раду, парламента) и* цепочке Украины *(президент Украины, украинского, Верховную Раду, Украине*).

То, что в реальной ситуации одно и то же слово может быть отнесено к разным цепочкам одновременно, значительно усложняет алгоритмы автоматического построения лексических цепочек.

Мы нашли только одну работу (Hollingsworth, Teufel, 2005), в которой авторы указывают на то, что их алгоритм построения лексических цепочек позволяет относить одно и тоже слово или словосочетание к разным лексическим цепочкам, и при этом они указывают на проблему порождения слишком большого количества лишних лексических цепочек (overgeneration).

При этом авторы подчеркивают, что в проведенном ими эксперименте все эксперты-аннотаторы, по крайней мере, одно слово (словосочетание) отнесли более, чем к одной лексической цепочке.

19.2. Автоматическое построение тематического представления текста

19.2.1. Лексические цепочки и тематическая структура текста

Во всех подходах автоматического моделирования лексических цепочек построение этих цепочек не является самоцелью – лексические цепочки выделяются для того, чтобы «приблизиться» к автоматическому построению тематической структуры текста, то есть уметь выделять, что в тексте главное, что второстепенное, как текстовые сущности связаны друг с другом.

С целью выделения наиболее значимых для содержания текста лексических цепочек, рассматриваются различные параметры лексических цепочек, такие как частотность ее элементов, текстовое покрытие и другие. В лексических цепочках выделяются наиболее частотные элементы цепочки в качестве наиболее важных тематических элементов текста.

Поскольку целью автоматического выделения лексических цепочек является автоматическое построение тематической структуры текста, рассмотрим на методы построения лексических цепочек и вышеописанные проблемы их построения с точки зрения роли лексических цепочек в тематической структуре текста.

Многие исследователи указывают на то, глобальная связность текста проявляется в том, что текст имеет единую тему. Тематическая структура текста представляет собой иерархическую структуру тем и подтем. Каждому предложению текста имеется некоторое соответствие в этой тематической структуре (см. п.14.1.1). Каждая тема (подтема) представляет собой пропозицию – предикат P (C1…Cn). Пропозиции тем (подтем) устанавливают отношения между тематическими элементами C1…Cn. В иерархической тематической структуре главная тема P0 ($C_{01}…C_{0n}$) поясняется, характеризуется, дополняется деталями посредством подтем P_1 (C_{11}, ….C_{im}) … Pi (C_{i1}, …C_{ij}...C_{im}).

Что представляют собой тематические элементы подтем Cij по отношению к тематическим элементам основной темы текста?

В силу глобальной связности текста в каждой подтеме по крайней мере один тематический элемент (а часто и больше) должны соответствовать тематическим элементам основной темы текста. Тематические элементы подтем могут представлять собой прямую отсылку на тематические элементы основной темы в виде точного повтора, синонимического повтора, референциальную отсылку, или обозначать некоторую тесно связанную с элементом основной темы сущность, например, ее часть, свойство и др.

Таким образом, на наш взгляд основная роль лексических цепочек относительно тематической структуры текста состоит в обеспечении представительства тематических элементов более высоких уровней иерархии в подтемах более низкого уровня (см. рис.19.1).

Отсюда следует, что в «правильной» совокупности лексических цепочек текста, то есть в лексических цепочках, отражающих тематическую структуру анализируемого текста, каждому тематическому элементу основной темы текста должны соответствовать свои лексические цепочки (которые могут иметь пересечение в некоторых словах).

Кроме того, лексические цепочки действительно имеют наиболее важных представителей - это элемент темы более высокого уровня. Рядовые элементы цепочки – это тематические элементы нижестоящих тем, раскрывающих эту тему.

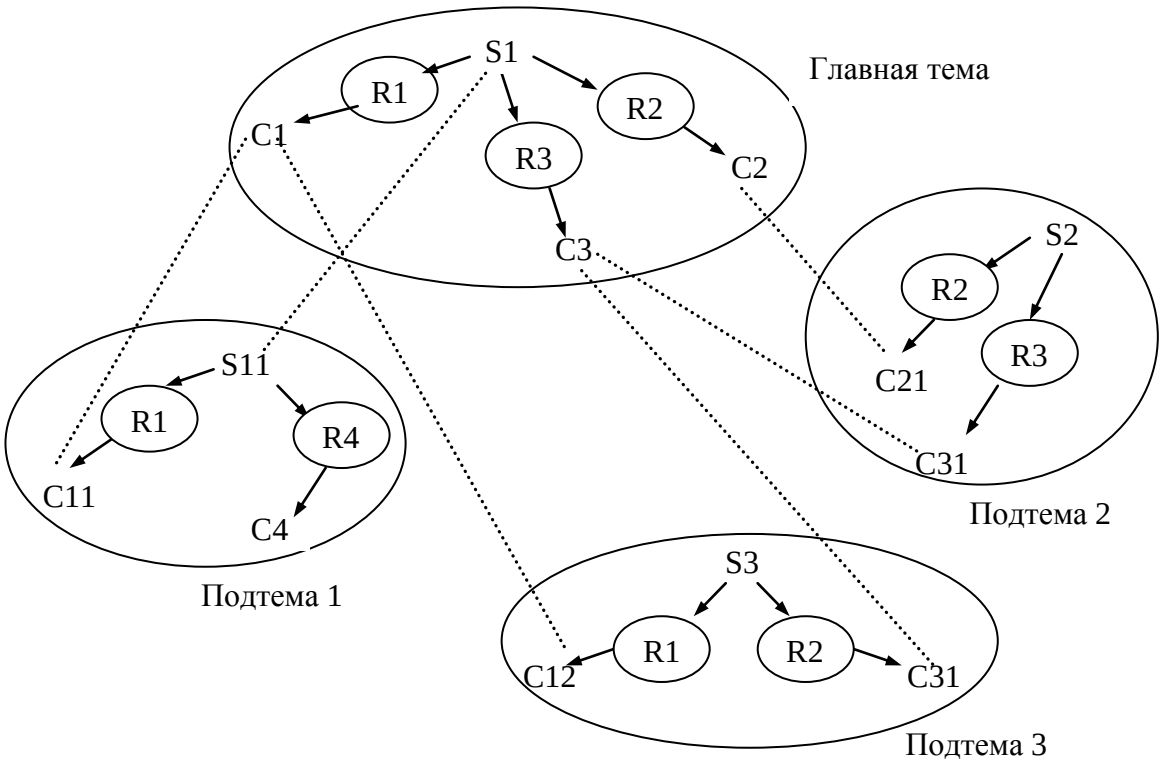

Рис.19.1 Тематическая структура текста как иерархия пропозиций тем

Таким образом, на наш взгляд, по внутренней структуре лексическая цепочка имеет структуру узла с выделенным центральный элементом и некоторой совокупностью лексем, связанных с этим центральным элементом. Назовем лексическую цепочку с такой предполагаемой структурой *тематическими узлом*.

Среди тематических узлов можно выделить основные тематические узлы и локальные тематические узлы. Основные тематические узлы имеют в качестве центра тематические элементы основной темы документа.

С другой стороны, пропозиция основной темы документа, то есть взаимоотношения участников основной темы, также должна находить свое отражение в конкретных предложениях текста, которые должны раскрывать, уточнять взаимоотношения между тематическими элементами. Если текст посвящен обсуждению взаимоотношений между тематическими элементами C1…Cn, то в предложениях текста должны обсуждаться детали этих отношений, что проявляется в том, что сами тематические элементы C1…Cn или их лексические представители должны встречаться как разные актанты одних и тех же предикатов в конкретных предложениях текста.

Отсюда следует практический вывод: если даже очень близкие по смыслу лексические сущности C1 и C2 часто встречаются в анализируемом тексте в одних и тех же простых предложениях, то это означает, что данный текст посвящен рассмотрению отношений между этими сущностями, то есть C1 и C2 соответствуют разным тематическим элементам основной темы или подтемы текста и должны быть отнесены к разным лексическим цепочкам (тематическим узлам).

Таким образом, «правильные» лексические цепочки, отражающие тематическое содержание документа должны отвечать следующим условиям:

1) лексическая цепочка имеет внутреннюю структуру узла – к одному выделенному элементу относятся все другие элементы лексической цепочки;
2) лексическая цепочка не должна содержать слова и словосочетания, которые часто встречались в одних и тех же предложениях текста с главным элементом этой цепочки, поскольку частая встречаемость некоторой лексической единицы

Li с начальным элементом цепочки L0 может означать, что Li и L0 представляют собой равноправные элементы основной или локальной темы анализируемого текста;

3) значимость цепочки для отражения содержания текста определяется не столько длиной, покрытием и другими характеристиками цепочки, а тем, насколько часто элементы этой цепочки встречались с элементами других цепочек в одних и тех же предложениях текста, то есть насколько много пропозиций конкретных предложений текста было посвящено обсуждению отношений между элементами некоторой совокупности лексических цепочек.

19.2.2. Примеры разбора лексических цепочек с учетом тематической структуры текста

Рассмотрим, каким образом выводы предыдущего раздела могут уточнить процедуру выделения лексических цепочек в текстах (*) и (**) из раздела 20.1.

При анализе текста (*) возник вопрос, куда отнести слова *means, meaning* к цепочке *figure out, think…*, или к цепочке *text, reader*.

Учитывая сделанные выводы, можно заметить, что в таком маленьком тексте слова *means, meaning* трижды встретились в одних и тех же простых предложениях со словами *text, reader:*

*what **a text means***
*the **meaning** is in the mind of the **reader***
*the **meaning** lies within the **text** itself*

Это означает, что данный текст посвящен рассмотрению отношения *текст – значение. Текст и значение* представляют собой разные тематические элементы в основной теме текста, и, соответственно, правильная структура лексических цепочек должна отнести слова *текст* и *значение* к разным лексическим цепочкам.

В то же время слова *means, meaning* не стоит относить и к другой лексической цепочке *figure out, think*, поскольку у этих глаголов один из актантов представляет собой клаузы, в которых и упоминаются слова *means, meaning,* то есть опять же это является центральной темой фрагмента, что люди думают по поводу значения текста.

***figure out** what a text **means**…*
***think** that the **meaning** lies within the text itself."*

Таким образом, лексические цепочки данного текста таковы:

1) *text, reader, text.*
2) *figure out, think*
3) *means, meaning, meaning*

В тексте (**) заголовок достаточно подробно называет основные тематические элементы текста*: врач (точнее медицинский работник), убить, пациент, наркотик*. И, действительно, мы видим повторяющуюся встречаемость этих тематических элементов в одних и тех же предложениях текста:

медики** по ошибке ввели **пациенту смертельную** дозу **опиоидного наркотика
***Врач** назначил ему **(пациенту)** 10 миллиграммов **морфина**.*
*По ошибке **медсестры пациенту** был сделан укол **гидроморфона***
*Свою ошибку **медики** осознали после пересчета **наркотических средств***

Таким образом, в тексте (**) должны быть выделены, по крайней мере, три «медицинские» лексические цепочки:

- цепочка «медработники» (врачи, медики, врач, медсестры, медики, медсестра),
- цепочка «пациент» (пациент, пациенту, пациент, пациенту, пациента),

- цепочка «наркотик» (наркотика, наркотика, морфина, гидроморфона, морфин, препарат, наркотических средств).

Кроме того, отдельно может быть выделена лексическая цепочка «больница» *(больнице, приемное отделение, больницы, больницу, больнице*), элементы которой также встречаются в одних и тех же предложениях текста с представителями других медицинских цепочек:

*в **больнице** ..., где **медики** по ошибке ввели смертельную дозу опиоидного **наркотика,***
пациент** поступил в **приемное отделение больницы,
он (пациент)** умер после возвращения **в больницу,
*Как сообщают в **больнице,** укол сделала опытная **медсестра***

Таким образом, анализ предложений текста позволяет выявить, что лучшим представлением для отражения содержания этого текста является не одна медицинская цепочка, а четыре цепочки, каждая из которых соответствует отдельному тематическому элементу данного текста, взаимодействующего с другими тематическими элементами.

Рассмотрим другие примеры текстов и их лексические цепочки.

На примере нижеследующей пары текстов покажем, что одни и те же слова могут попасть в одну или разные цепочки в зависимости от основной темы текста. Тексты представляют собой новостные сообщения середины 90-х годов, касающиеся статуса Чеченской республики:

*(***) Стороны договорились о визите в ближайшее время в **Россию** министра иностранных дел Ирана. Была там тогда достигнута и договоренность о передаче гуманитарной помощи вынужденным переселенцам из **Чечни.***

*Кстати, самолет с 44 тоннами гуманитарного груза на борту как раз и приземлился в тот момент, когда проходила беседа президента и посла в **Бесланском** аэропорту. Они тут же направились к самолету, на "хвосте" которого изображен голубь зеленого цвета. Журналисты уже на ходу задавали свои вопросы, пытаясь выяснить позицию иранского дипломата в отношении военной операции в **Чечне**. На что он ответил: "Это внутреннее дело **России.** Мы лишь хотим, чтобы эта операция имела меньше жертв и поскорее завершилась".*

*(****) Проведен опрос 185 воронежцев. ... Были заданы три вопроса: 1. Считаете ли вы **Чечню** территорией **России?** .. 46,48 процента опрошенных считают **Чечню** территорией **России.** И ровно столько же **ее** не считают территорией **РФ**... О том, что выход **Чечни** из состава **России** может послужить началом развала **Федерации,** никто не задумывается.*

В обоих текстах упоминаются *Россия* и *Чечня.*

В тексте (***) основное содержание текста связано с обсуждением отношений между Россией и Ираном, и Россия представлена в тексте единой лексической цепочкой *Россия, Чечня, Бесланский, Чечня, Россия.*

Во тексте (****) обсуждаются отношения между *Россией* и *Чечней*, эти слова неоднократно встречаются в одних и тех же предложениях текста. Таким образом, объединение их в единую лексическую цепочку склеивает два разных тематических элемента основной темы, что противоречит содержанию документа. Таким образом, в тексте (****) Россия и Чечня должны образовать две разные лексические цепочки:

- лексическая цепочка «Чечня»: *Чечню, Чечню, Чечни*
- лексическая цепочка «Россия»: *России, России, РФ, России, Федерации*

Рассмотрение лексических цепочек через призму их употребления в одних и тех же предложениях текста имеет прямое соответствие с идеей Р. Хазан о «гармонии связности»

(см. п. 14.1.3), которая проявляется в том, что элементы разных лексических цепочек должны выступать по отношению друг к другу в одних и тех же семантических отношениях, и, это значит, в большинстве случаев представители этих цепочек должны упоминаться в одних и тех же предложениях текста (Hasan, 1984). В одном из рассмотренных текстов – тексте (**) элементы четырех медицинских лексических цепочек четко находились по отношению друг к другу в одних и тех же семантических отношениях ‘агент’(медики)-‘пациенс’(пациент)-‘средство’(наркотик)- ‘место‘(больница).

Различие нашего подхода от идеи Р. Хазан заключается в следующих положениях.

Во-первых, мы не требуем, чтобы непременно между элементами лексических цепочек были одни и те же семантические отношения, полагая, что уже частое упоминание элементов разных лексических цепочек в связном тексте не может быть случайным.

Во-вторых, рассмотрение синтагматических отношений между элементами потенциальных лексических цепочек является важным уже на этапе построения лексических цепочек. Это рассмотрение позволяет в сложных случаях употребления в тексте большого количества близких по смыслу слов принимать более обоснованное решение по разделению этого множества слов на лексические цепочки. Кроме того, используя этот принцип формирования лексических цепочек, возможно, формировать цепочки, учитывая достаточно разнообразные отношения между лексемами (заметим, что в своем анализе М. Хэллидей и Р. Хазан обычно ограничиваются небольшим набором рассматриваемых отношений между лексемами: синонимы, родовидовые отношения, отношение часть-целое), а также возможное вхождение одной и той же лексемы в несколько лексических цепочек.

19.2.3 Автоматическое построение тематических узлов

Мы предположили, что лексические цепочки должны связывать не все близкие по смыслу слова текста, но соответствовать тематической структуре текста. Кроме того, лексические цепочки должны иметь форму узла – с главным выделяемым элементов, к которому относятся все другие элементы этой цепочки. Далее таким образом устроенные лексические цепочки будем называть тематическими узлами.

Важно еще подчеркнуть, что поскольку тематические узлы призваны моделировать основное содержание текста, то тематические узлы - это не последовательности близких по смыслу лексем, а совокупности близких по смыслу понятий, то есть, сущностей в которых до какой-то степени устранен фактор лексической синонимии и многозначности.

В предыдущем разделе мы показали, что создать «правильный» (то есть соответствующий тематической структуре анализируемого текста) тематический узел невозможно, используя только локальную информацию о расположении слов в соседних предложениях документа. Нужна совокупная информация о частотности и распределении слов в тексте, которую необходимо сопоставить с имеющимися в тезаурусе знаниями о существующих соотношениях значений слов.

Поэтому лексические цепочки в форме тематических узлов не строятся при движении от предложения к предложению, а производятся из общей картины упоминания понятий в предложениях, полученной по тексту.

Как уже описывалось в предыдущих разделах, на предварительных этапах обработки текст был сопоставлен с тезаурусом:

- текстовые выражения текста были сопоставлены с понятиями тезауруса,
- понятия тезауруса, найденные в тексте, соединены отношениями, описанными в тезаурусе.

На основе созданной таким образом тезаурусной проекции текста произведен выбор значений для многозначных текстовых входов тезауруса.

Для построения тематических узлов существенны два фактора:

- существование пути определенного вида между понятиями тезауруса и

- встречаемость понятий тезауруса в одних и тех же простых предложениях текста.

При изложении методов построения лексических цепочек на базе тезауруса WordNet используются некоторые типы путей между синсетами, в том числе пути, состоящие из отношений различной направленности, то есть пути с перегибами (см. п.14.2.1).

При построении тематических узлов на основе тезауруса РуТез мы отказались от использования путей с перегибами по следующим причинам.

Во-первых, в тезаурусе РуТез имеется больший набор прямых связей между понятиями тезауруса за счет транзитивных отношений часть-целое и отношений направленной ассоциации, описывающих концептуальную зависимость понятий тезауруса друг от друга.

Во-вторых, мы считали важным дать возможность понятию тезауруса входить в несколько тематических узлов,

В-третьих, понятия, соединенные путями с перегибами – виды одного рода, части одного целого и др. – достаточно часто могут выступать как разные, противопоставленные друг другу элементы основной темы.

Таким образом, в основном блоке текущей реализации алгоритма тематические узлы образуются на основе иерархически подчиненных понятий тезауруса, имеющих между собой пути, состоящие из отношений одной направленности (см. п.17.8.).

Для учета совместной встречаемости понятий тезауруса в одних и тех же предложениях текста, для каждого понятия подсчитываются понятия-соседи в линейном контексте внутри предложения. Величина линейного контекста обычно устанавливается величиной 3, то есть для каждого понятия запоминается по три понятия-соседа влево и вправо. Понятия-соседи суммируются по всему тексту, и, таким образом, для каждого понятия получается частотный список понятий-соседей – так называемые текстовые связи понятия.

19.2.3.1. Алгоритм построения тематических узлов

Для построения тематических узлов мы сначала выделяем потенциальные центры тематических узлов. Мы предполагаем, что то понятие тезауруса, которое наиболее точно характеризует развиваемую в тексте тему и которое, соответственно, может стать тематическим центром одного из тематических узлов текста, обычно некоторым образом выделяется в пространстве всех тематически близких понятий, а именно: такое понятие может быть упомянуто в заголовке и/или в начале текста, или имеет максимальную частотность среди других близких по смыслу понятий.

Тематическим центром может стать любое понятие тезауруса, независимо от уровня его общности/специфичности. Единственное условие, которое может быть указано, это общая тематическая принадлежность концепта. При обработке современной прессы, актов законодательства на базе тезауруса РуТез обычно требуется принадлежность начального понятия тематического узла Общественно-политическому тезаурусу, то есть фактически принадлежность понятия к одной из тематических областей общественной жизни.

Таким образом, создание тематического узла начинается с выбора главного понятия тематического узла. Сначала тематические узлы собираются вокруг понятий заголовка и первого предложения текста. Затем тематические узлы собираются для остальных понятий, начиная с самых частотных. Те понятия, которые уже попали в тематический узел некоторого понятия, свой тематический узел не образуют.

Центральное понятие тематического узла С0 присоединяет в создаваемый тематический узел понятия Ci из своей тезаурусной окрестности при выполнении нескольких условий. При присоединении учитываются такие факторы как:

- количество текстовых связей между Ci и С0 (то есть совместной встречаемости Ci и С0 в одних и тех же предложениях) в целом документе – Rtext,
- количество связей между Ci и С0 по предложениям, то есть сколько раз в документе Ci и С0 встречались в текущем предложении и в k (по умолчанию k=7) соседних предложениях, но вне пределов окна установления текстовых связей – Rsent.

В новый тематический узел понятия С0 включаются понятия Ci из дерева С0 при выполнении одного из следующих условий:

- Rsentence (C0, Ci) > 0 и (Rtext (C0,Ci) < 2 или Rtext (C0,Ci) ≤ Rsent(C0,C)), то есть понятия С0 и Ci должны встречаться в тексте в соседних предложениях и при этом либо практически не встречаться рядом друг с другом в одних и тех же предложениях текста, либо частотность встречаемости понятия С0 и С в одних и тех же предложениях текста должна быть меньше, чем частотность встречаемости в С0 и С в соседних предложениях,

Или

- Rsentence (C0,Ci) = 0 и Rtext (C0,Ci) =0 и Rsent (Ct,Ci)>0, где Ct – понятие, уже включенное в тематический узел С0. То есть понятие Ci из дерева С0 включается в тематический узел, если оно нашлось относительно недалеко от понятия Ct, уже включенного в тематический узел С0.

Или

- частотность (Ci)=1

После построения очередного тематического узла выбирается следующее по частотности (заголовку) понятие тезауруса, еще не включенное в тематические узлы, и образует свой следующий тематический узел.

Приведем примеры тематических узлов, созданных в процессе обработки текста (**) из раздела 19.1.2. (главное понятие тематического узла выделено сдвигом влево; указана также частота упоминания понятия в тексте):

```
1) НАРКОТИК                          3
      МОРФИН                         2
      МЕДИКАМЕНТ                     1
2) БОЛЬНИЦА                          4
      ПРИЕМНОЕ ОТДЕЛЕНИЕ БОЛЬНИЦЫ    1
3) ПАЦИЕНТ                           5
4) ВРАЧ                              2
      МЕДИЦИНСКИЙ РАБОТНИК           2
5) КАНАДА                            2
     АЛЬБЕРТА                        1
6) УБИТЬ, ЛИШИТЬ ЖИЗНИ               1
      СМЕРТЬ                         2
      УМЕРЕТЬ                        1
7) ТРАВМА                            1
      НЕСЧАСТНЫЙ СЛУЧАЙ              1
8) МЕДСЕСТРА                         2
```

В этом автоматически полученном наборе тематических узлов можно заметить следующие неточности отражения основного содержания текста.

Во-первых, тематический узел «медицинские работники» разбился на два тематических узла 4) и 7).

Возможно, правильнее иметь единый узел медицинских работников, поскольку текст делает акцент именно на вине медиков в целом:

```
МЕДИЦИНСКИЙ РАБОТНИК                2
   ВРАЧ                             2
   МЕДСЕСТРА                        2
```

Кроме того, словосочетание «несчастный случай» в тексте явно относилось не к травме, а к смерти пациента, то есть более правильным был бы такой узел:

```
УБИТЬ, ЛИШИТЬ ЖИЗНИ                 1
   СМЕРТЬ                           2
   УМЕРЕТЬ                          1
   НЕСЧАСТНЫЙ СЛУЧАЙ                1
```

Но в целом, как мы видим, тематические узлы соответствуют элементам основной темы текста.

При обработке текстов (***) и (****) из раздела 19.2.2. изложенный алгоритм связывает между собой понятия РОССИЙСКАЯ ФЕДЕРАЦИЯ и ЧЕЧЕНСКАЯ РЕСПУБЛИКА по-разному.

Для документа (***) строится тематический узел, объединяющий данные понятия в следующий тематический узел:

```
РОССИЙСКАЯ ФЕДЕРАЦИЯ                2
   ЧЕЧЕНСКАЯ РЕСПУБЛИКА             2
   БЕСЛАН                           1
```

Для документа (****) понятия РОССИЙСКАЯ ФЕДЕРАЦИЯ и ЧЕЧЕНСКАЯ РЕСПУБЛИКА образуют два тематических узла:

```
РОССИЙСКАЯ ФЕДЕРАЦИЯ                4
   ФЕДЕРАТИВНОЕ ГОСУДАРСТВО         1
   ЧЕЧЕНСКАЯ РЕСПУБЛИКА             3
```

Таким образом, изложенный алгоритм формирует тематические узлы так, чтобы каждый тематический узел соответствовал отдельному элементу основной темы документа.

19.2.3.2. Мультиграфы как база для порождения тематических узлов

Как мы уже указывали, построение лексических цепочек в большинстве подходов сводится, в конечном счете, к разбиению графа отношений между понятиями, упоминаемыми в тексте, на подграфы. По сути, та же процедура реализована и в процессе построения тематического представления – граф тезаурусной проекции разбивается на подграфы – совокупности тематических узлов.

Для учета факторов построения тематического представления подходит представление распределения понятий текста в виде мультиграфа, то есть графа с двумя типами дуг между вершинами. Один тип дуг, Rsent, отражает отношения между понятиями в тезаурусе. Другой тип дуг, Rtext, отражает совместную встречаемость понятий в предложениях текста. В вершинах мультиграфа указана частотность упоминания соответствующего понятия в тексте. На дугах Rtext отмечена частота встречаемости данной пары понятий в одних и тех же предложениях текста. Дуги Rsent указывают частотность упоминания данной пары понятия в пределах нескольких предложений (например, 7 предложений), но не в одном предложении текста (Loukachevitch, 2009b).

Таким образом, мультиграф MG тематического представления может быть определен как MG = (*V, fv, Rtext, frtext, Rsent, frsent*) (рис. 19.2).

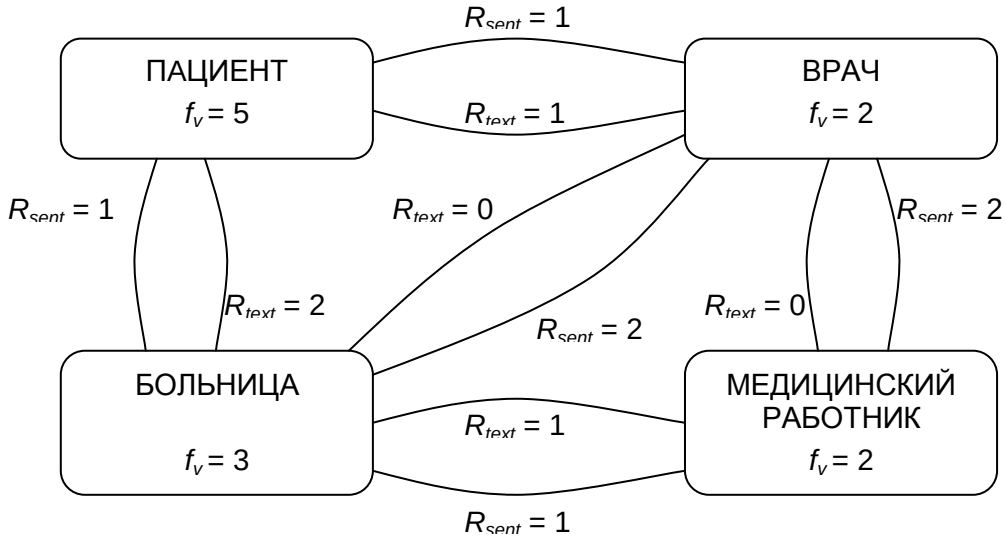

Рис.19.2. Фрагмент мультиграфа для текста (**)

19.2.4. Определение статуса тематического узла

На предшествующем этапе были собраны тематические узлы, каждый из которых включает понятия текста, связанные по Тезаурусу с главным понятием тематического узла. С помощью тематического узла выделяются элементы основных тем и подтем текста, обсуждавшиеся в тексте.

В нашей модели мы предполагаем, что понятия основных тематических узлов постоянно встречаются рядом друг с другом (связаны по тексту) в одних и тех же предложениях текста. Понятно, что реализация проверки такого условия осложняется проблемами правильного выделения простых предложений внутри сложных предложений, построением правильной синтаксической структуры, вхождениями местоимений и использованием эллипсиса (то есть пропусков) в тексте. Поэтому для оценки совместной встречаемости тематических узлов мы используем опять же линейный контекст понятий, называемый нами текстовые связи.

В результате для каждого понятие, упомянутого в тексте, получается совокупность текстовых связей, как, например, для понятия ПАЦИЕНТ из текста (**) (справа указана частота текстовых связей понятия *ПАЦИЕНТ* с другими понятиями текста):

```
ПАЦИЕНТ
  НАРКОТИК                    - 4
  ВРАЧ                        - 1
  УБИТЬ, ЛИШИТЬ ЖИЗНИ         - 1
  НАРКОТИК                    - 2
  НЕСЧАСТНЫЙ СЛУЧАЙ           - 1
  БОЛЬНИЦА                    - 1
  МЕДИЦИНСКИЙ РАБОТНИК        - 1
```

После того как созданы тематические узлы, текстовые связи понятий каждого тематического узла суммируются и определяются текстовые связи между тематическими узлами.

Приведем примеры текстовых связей между тематическими узлами, выделенными в тематическом представлении текста (**). Тематические узлы представлены своими

главными понятиями, число справа - суммарная величина текстовых связей между понятиями тематических узлов, текстовые связи даны для тематического узла, главное понятие которого смещено в примере влево:

ПАЦИЕНТ
  НАРКОТИК – 4
  БОЛЬНИЦА – 3
  ВРАЧ – 3
  УБИТЬ, ЛИШИТЬ ЖИЗНИ – 3
 …

В соответствии с моделью предполагается, что основными тематическими узлами в первую очередь являются такие тематические узлы, которые:

- все связаны между собой текстовыми связями;
- сумма частот текстовых связей между ними максимальна для анализируемого текста (рис. 19.3).

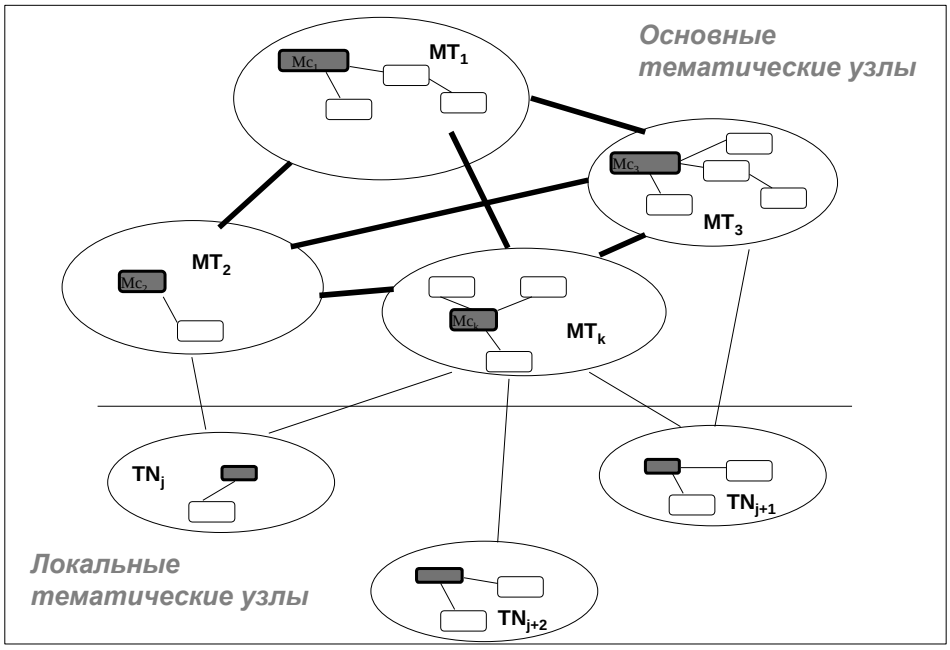

Рис. 19.3. Структура тематического представления. Элементы основных тематических узлов MT_i постоянно встречаются в одних и тех же предложениях текста. Поэтому текстовые связи между ними обозначены жирными линиями

В рассматриваемом примере тематического представления текста (**) основными тематическими узлами стали узлы с главными понятиями *ПАЦИЕНТ, НАРКОТИК, ВРАЧ, БОЛЬНИЦА, МЕДСЕСТРА, УБИТЬ, ЛИШИТЬ ЖИЗНИ, КАНАДА.*

Упомянутый ранее тематический узел *ТРАВМА (несчастный случай)* не прошел в список основных тематических узлов, поскольку не был связан по тексту с тематическим узлом *МЕДСЕСТРА.*

Вычисленные таким образом основные тематические узлы автоматически задают порог, выделяющий среди всех тем, обсуждавшихся в тексте, основные темы текста. Таким порогом считается средняя суммарная частотность основных тематических узлов.

Исходная совокупность основных тематических узлов дополняется теми тематическими узлами, частотность которых превышает вычисленный порог. Это дополнение отражает такую структуру текста, когда некоторая важная тема обсуждается в тексте локализовано, не по всему тексту, но достаточно подробно.

Локальные тематические узлы представляют собой некоторые важные характеристики основных тематических узлов. Тематический узел считается локальным, если этот узел имеет текстовую связь с частотностью большей единицы с одним из основных тематических узлов. Понятия, не вошедшие в состав основных и локальных тематических узлов, объявляются "упоминавшимися" в тексте.

Таким разбиением тематических узлов на основные и локальные задается разбиение понятий, упомянутых в тексте, на следующие пять классов по их важности для анализируемого текста:

- главные понятия основных тематических узлов (основные темы);
- другие понятия основных тематических узлов;
- главные понятия локальных тематических узлов (локальные темы);
- другие понятия локальных тематических узлов;
- упоминавшиеся понятия.

Таким образом, построено тематическое представление текста, в котором понятия тезауруса, упоминавшиеся в тексте, разбиты на тематические узлы. Тематические узлы подразделяются на основные, локальные и упоминавшиеся узлы. Между тематическими узлами фиксируются текстовые связи (Лукашевич, Добров, 1996; Лукашевич, Добров, 2000)

19.2.6. Тестирование качества построения тематических узлов

В работе (Loukachevitch, Dobrov, 2000b) был описан эксперимент по оценке качества автоматического построения основных тематических узлов, соответствующих элементам основной темы текста.

Для каждого текста человеком выбирались его основные понятия, то есть понятия, которые наилучшим образом характеризовали основную тему анализируемого документа. Такие основные понятия выбирались, в основном, из заголовка и первого абзаца документа. Для каждого выбранного понятия автоматически строился тематический узел, состоящий из понятий данного текста. Затем, просматривая текст, мы проверяли, действительно ли включенные в тематический узел понятия относились в данном тексте к исходному понятию.

При этом было принято следующие правило: если отношение между понятиями определено в данном тексте и далее используется для организации связного текста, то это отношение не обязано быть в тезаурусе, и его невключение в тематический узел не считалось ошибкой, поскольку авторы текста не предполагали, что читатель должен знать отношения между понятиями заранее.

В нашем эксперименте все исходные основные понятия были различны и построенные тематические узлы содержали не менее 3 понятий.

На основе анализа 73 тематических узлов для 25 текстов общественно-политической тематики мы получили следующие характеристики качества отражения тематическими узлами лексической связности документов: точность - 89%, полнота - 71%.

Заключение к главе 19

На первый взгляд может показаться, что и человеку, и компьютеру выявить лексическую связность в связном тексте достаточно просто. Однако в экспериментах с людьми-аннотаторами была выявлена высокая субъективность выделения в тексте лексических цепочек близких по смыслу слов текста – такие цепочки являются основным инструментом моделирования лексической связности.

В этой главе мы показали, что для определения лексической связности в тексте недостаточно извлекать совокупности близких по смыслу слов и словосочетаний, для правильного формирования лексических цепочек необходимо учитывать взаимодействие

упоминаемых сущностей в предложениях текста. Данное положение является следствием глобальной связности текста.

Также из глобальной связности текста следует то, что лексическая цепочка имеет внутреннюю структуру узла – все элементы цепочки должны иметь отношение к одному и тому же элементу цепочки – главному элементу цепочки, ее центру.

Об эти фактора позволяют строить лексические цепочки в соответствии с тематической структурой конкретного текста.

На наш взгляд, учет этих факторов в экспериментах с людьми-аннотаторами даст в результате более высокий показатель согласия между аннотаторами при разметке лексических цепочек.

Глава 20. Информационный поиск с учетом тезаурусных знаний

20.1 Концептуальный индекс, веса понятий и отношений

Тематическое представление текста дает возможность построить концептуальный индекс документа, в котором учитывается не только частотность отдельного понятия в документе, но и статус понятия в тематической структуре документа (Добров, Лукашевич, 2001; Лукашевич, Добров, 2001).

Как указывалось в предыдущей главе, в результате построения тематического представления текста все понятия тезауруса, упомянутые в тексте, разделяются на пять базовых классов значимости для текста, каждый из которых имеет свой вес. Задание весов этих классов может осуществляться параметрически. В большинстве случаев, веса классов значимости понятий задаются следующим образом:

- центры основных тематических узлов – 0.95;
- другие понятия основных тематических узлов – 0.85;
- центры локальных тематических узлов – 0.70;
- другие понятия локальных тематических узлов – 0.65;
- упоминавшиеся понятия, не вошедшие в предыдущие классы – 0.20.

Базовый вес понятия получен в качестве интегрального анализа распределения в тексте совокупностей близких по смыслу терминов. Чтобы снизить фактор ошибки вычисления базовых весов, а также сделать веса понятий более дробными, для формирования окончательного веса понятий учитывается также относительная частотность понятий в тексте. Окончательный вес понятия в тексте $\mu(c, \mathrm{D})$ рассчитывается по следующей формуле:

$$\mu(c, \mathrm{D}) = \lambda \cdot \nu^*(c, \mathrm{D}) + (1-\lambda) \cdot freq(c, \mathrm{D}) \cdot [freq^*(\mathrm{D})]^{-1} \qquad (20.1)$$

где $\nu^*(c, \mathrm{D}) = \max_{Th(c,\ \mathrm{D})} \nu(\mathrm{c}, \mathrm{D})$ – максимум базовых весов понятия c в тематических узлах; оптимальная величина $\lambda = 0.7$; $freq(c, \mathrm{D})$ – частота понятия c в документе D, $freq^*(\mathrm{D}) = \max_{d \in \mathrm{D}} freq(d, \mathrm{D})$ – максимальная частота среди понятий документа D.

Таким образом, при загрузке текстов в поисковую систему создается концептуальный индекс текста по тезаурусу, строится тематическое представление текста, каждому понятию присваивается вес по формуле (20.1).

При расширении запроса по тезаурусу необходимо организовать выдачу и таких текстов, в которых нет исходных понятий запроса, но имеются понятия нижестоящие по иерархии – так называемое дерево расширения вниз (Добров, Лукашевич, 2001).

Каждое понятие в дереве расширения имеет свой вес, который зависит от суммарного отношения данного понятия к исходному понятию и не зависит от длины пути до понятия-вершины дерева. В настоящее время используются следующие величины весов Q(t,c), где t - исходное понятие, c – понятие в его дереве расширения:

Q(НИЖЕ(t,c)) = 0.9

Q(ЧАСТЬ(t,c)) = 0.8

Q(АСЦ(t,c)) = 0.6 (20.2)

Q(АСЦ2(t,c)) = 0.5

Эти величины используются как коэффициенты, на которые домножается вес, присвоенный данному понятию при анализе конкретного документа.

Документ может содержать несколько различных понятий их дерева расширения. Для вычисления веса такого документа веса всех понятий из дерева расширения суммируются так, чтобы придать больший вес документам, которые содержат несколько понятий из дерева расширения:

$$W(t) = 0.7 \cdot \max_{c \in Tr(t)} \{V(c) \cdot Q(t,c)\} + 0.3 \cdot \max\left\{V(t), \frac{R(t,D)}{1+R(t,D)}\right\}, \quad (20.3)$$

где $R(t,D) = \sum_{d \in Tr(t) \bigcap D} V(d) \cdot Q(t,d)$

Если документ содержит понятие, которое связано с исходным понятием запроса посредством отношения с модификатором, то используются дополнительные понижающие вес коэффициенты. Это связано с тем, что модификатор сообщает информацию о том, что это отношение недостаточно стабильно и может быть в некоторых контекстах нерелевантно.

Мы считаем, что такое отношение подтверждается, если в документе есть другое понятие из того же дерева расширения, которое связано с понятием-вершиной дерева без дополнительных модификаторов. В этом случае коэффициент отношения с модификатором совпадает с коэффициентом расширения без модификаторов.

Если такое отношение не подтверждается, то используется дополнительное снижение веса отношения в 2 раза.

$$Q(NT_{A(V)}) = 0.45 = (0.9/2) \quad (20.4)$$
$$Q(PART_{A(V)}) = 0.40 = (0.8/2)$$

20.2. Общественно-политический тезаурус как поисковое средство в Университетской информационной системе Россия

Общественно-политический тезаурус используется как поисковое средство в Университетской информационной системе Россия (www.cir.ru), которая создана и развивается как тематическая электронная библиотека и база для исследований и учебных курсов в области экономики, управления, социологии, лингвистики, философии, филологии, международных отношений и других гуманитарных наук (Богомолова и др., 2008).

Пользователь может задать Булевский запрос, включающий как слова, так и понятия Тезауруса. Понятие тезауруса может быть задано без расширения по дереву. Тогда в ответ на запрос будут выданы документы, содержащие хотя бы одно из текстовых выражений, сопоставленных данному понятию. Если понятие тезауруса задано с расширением по дереву, то релевантными считаются документы, содержащие хотя бы один синоним выбранного понятия или (с несколько меньшим весом) хотя бы один синоним понятий из дерева-вниз выбранного понятия. Таким образом, выбор в запрос одного понятия может оказаться равносилен выбору сотен и тысяч слов и словосочетаний.

Поэтому поиск с использованием Тезауруса состоит из следующей последовательности шагов:

- поиск нужного понятия;
- выбор подходящего условия включения понятия в запрос;
- выбор следующего понятия или исполнение запроса.

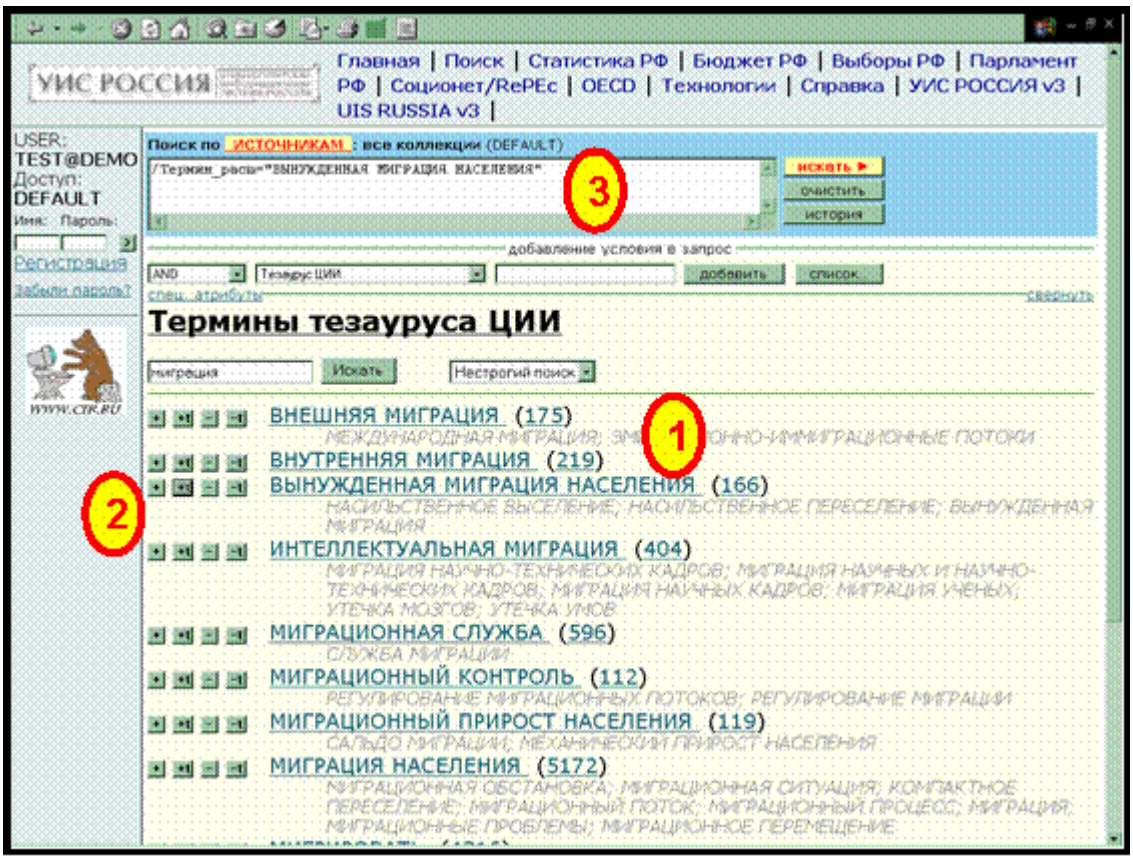

Рис.20.1. Поиск подходящего понятия тезауруса для запроса *“вынужденная миграция населения”*;

Для поиска по Тезаурусу пользователь выбирает из списка общих атрибутов опцию “Тезаурус ЦИИ”, вводит в крайнем правом окне термин (в данном случае было введено слово “миграция”) и нажимает на кнопку “список”. Появляется список понятий Тезауруса, где хотя бы один термин содержит введенное слово (Рис.20.1.).

Пользователь выбирает наиболее подходящее понятие, например, “*Вынужденная миграция населения*” и задает правило учета этого понятия в условиях запроса с помощью клавиш, расположенных слева:

“ + ” - строго данное понятие,

“ +t ” - понятие с расширением по дереву,

“ - ” - исключить термин,

“ -t ” - исключить термин и нижестоящие понятия.

В ситуации, изображенной на Рисунке 20.1-1 пользователь выбрал кнопку “+t”**.** (Рис.20.1-2). В окне запроса появляется новое условие (Рис. 20.1-3):

/Термин_расш=ВЫНУЖДЕННАЯ МИГРАЦИЯ НАСЕЛЕНИЯ

Этот прием избавляет от процедуры ввода длинных строк в условия запроса. Теперь, если пользователь нажимает на клавишу “**искать”**, то система выдает документы, содержащие один из терминов:

- “насильственное выселение”;
- “насильственное переселение”;
- “вынужденная миграция”;

или термины, приписанные подчиненным понятиям “БЕЖЕНЕЦ” (“беженка”, “беженский”), “ВЫНУЖДЕННЫЕ ПЕРЕСЕЛЕНЦЫ” (“вынужденный мигрант”) и т.д.

Использование опции “расширение по дереву Тезауруса” при поиске с использованием географических названий позволяет найти все географические названия и административные единицы. При поиске по термину *ЮГО-ВОСТОЧНАЯ СИБИРЬ* будут выданы также документы, содержащие: *БАЙКАЛ, ЗАБАЙКАЛЬЕ, БУРЯТИЯ, ЧИТИНСКАЯ ОБЛАСТЬ, ПРИБАЙКАЛЬЕ* и т.д.

Особенно впечатляющих результатов удается добиваться, формируя запрос из нескольких понятий с расширением по дереву. В частности, можно эффективно найти документы следующей тематики:

```
/Термин_расш=“ПРЕСТУПНОСТЬ”
and /Термин_расш= “СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ ОКРУГ”
```

или, например,

```
/Термин_расш=“МИГРАЦИЯ”
and /Термин_расш= “АМУРСКАЯ ОБЛАСТЬ”
```

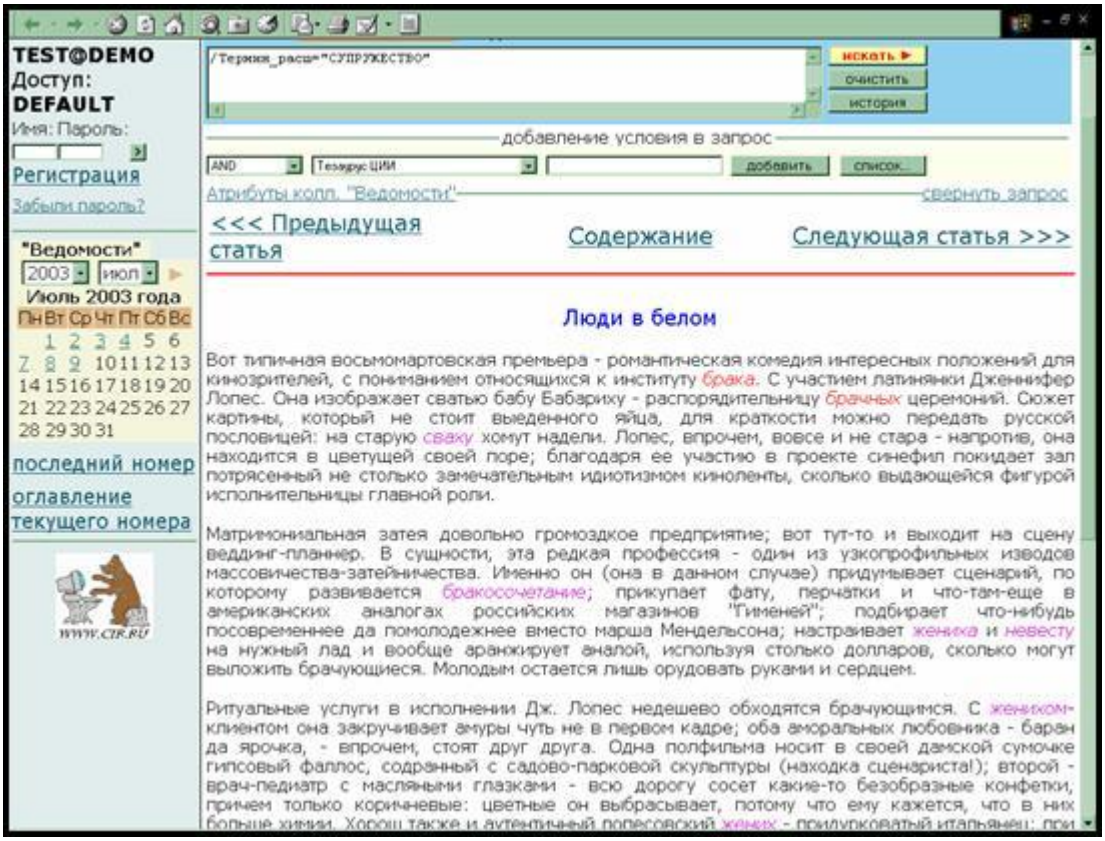

Рис.20.2. Пример статьи,
найденной по понятию *СУПРУЖЕСТВО* с расширением по дереву.

На Рис.20.2 приведен пример статьи, найденной по понятию *СУПРУЖЕСТВО* с расширением по дереву. В статье встречаются термины *брак, бракосочетание, жених* и *невеста, сваха*. При этом само слово *супружество* в документе не встречается. Найденные в документе термины подсвечиваются – красным цветом – синонимы понятия, использованного в запросе, фиолетовым цветом – синонимы подчиненных понятий.

Запрос может быть также уточнен путем просмотра тезаурусной статьи понятия (Рис.20.3), которая получается при переходе по ссылке, связанной с понятием. При этом пользователь, “двигаясь” по связям между понятиями, может выбрать более подходящую ему тематику, тем самым уточнить смысл своего запроса.

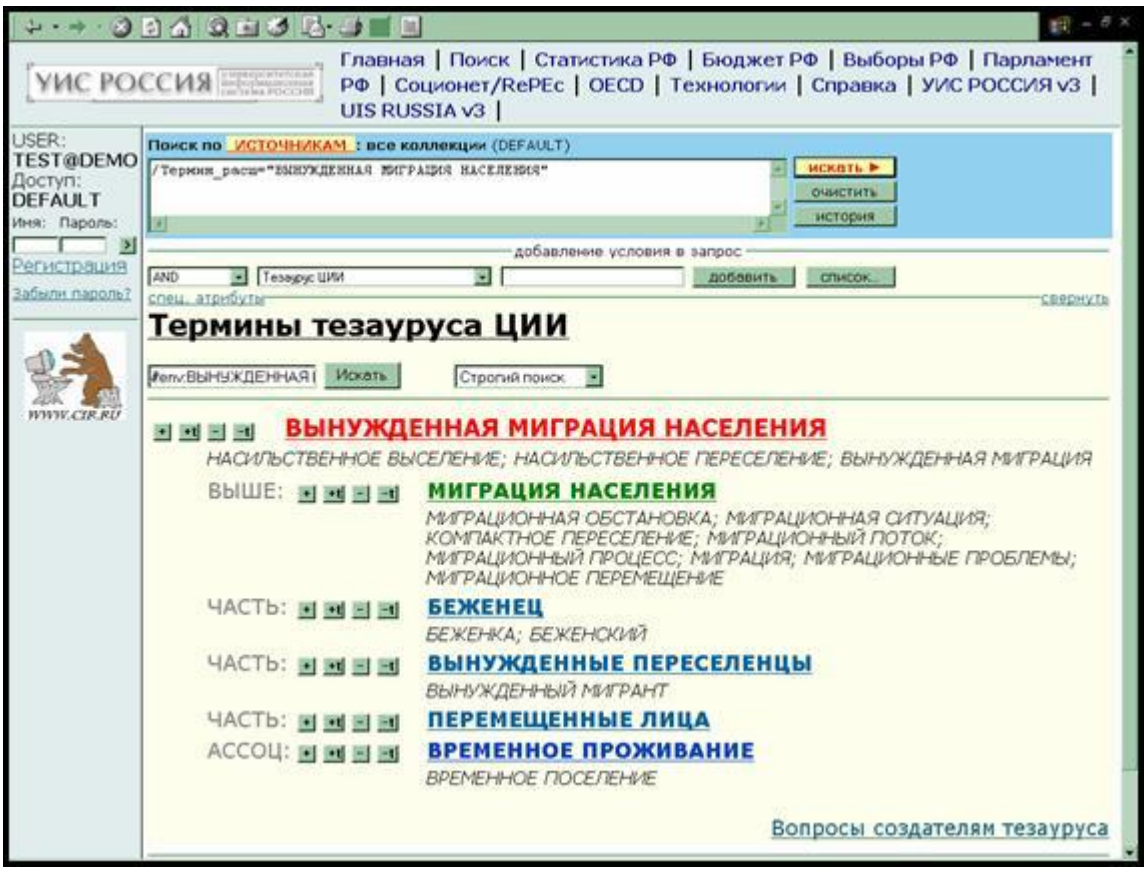

Рис.20.3. Тезаурусная статья для понятия
ВЫНУЖДЕННАЯ МИГРАЦИЯ НАСЕЛЕНИЯ

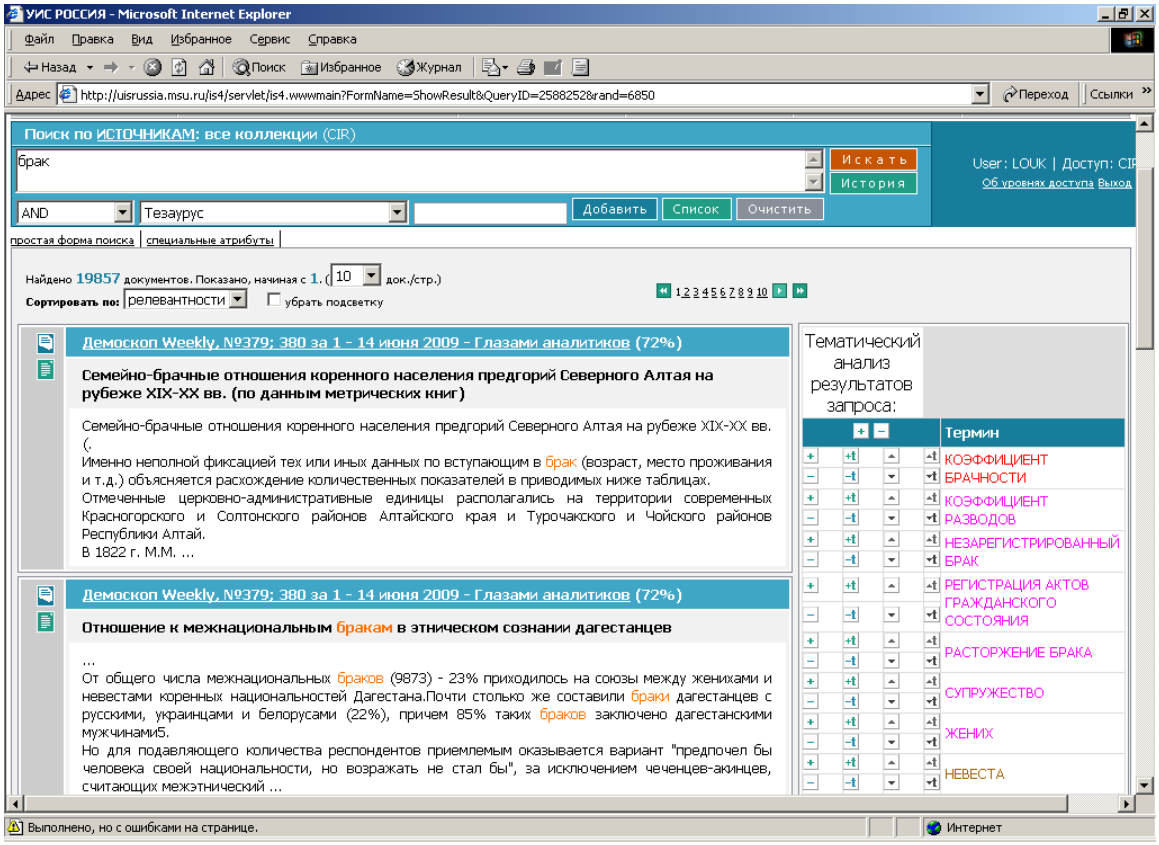

Рис. 20.4. Выдача информера по запросу *брак*

При формировании выдачи документов на запрос, происходит выявление наиболее характерных для данной выдачи понятий тезауруса, которые выдаются в колонку справа (рис. 20.4.). В разных системах выдача такого рода называется ассоциативный контекст, информационный портрет (см. например, (Антонов, Курзинер, 2003)), информер. Понятия тезауруса в информере упорядочиваются на основе веса, полученного по формуле типа tf*idf, когда частотность упоминания понятия в выдаче сопоставляется с частотностью упоминания понятия в коллекции.

Понятия тезауруса в информере также снабжены кнопками добавления в запрос, что позволяет одним нажатием мыши модифицировать запрос так, чтобы учесть в запросе или исключить из запроса данное понятие.

20.3 Тестирование эффективности информационного поиска на основе Тезауруса

В данном разделе мы опишем эксперимент по тестированию качества информационного поиска с использованием тезаурусных знаний в условиях, когда задаваемые запросы хорошо покрываются текстовыми входами Общественно-политического тезауруса. В качестве запросов были выбраны рубрики из Классификатора правовых актов (Указ, 2000). Поиск осуществлялся на коллекции нормативных актов УИС РОССИЯ.

Для тестирования эффективности информационного поиска мы выполнили набор запросов в УИС РОССИЯ. Каждый запрос был сформулирован дважды: один раз как запрос на поиск по словам, второй раз - как запрос на поиск по понятиям тезауруса с полным расширением по дереву. Поиск по словам осуществляется с использованием векторной модели в формулировке системы Inquery (Callan и др., 1992).

При выполнении подавляющего количества запросов количество документов, найденных с использованием деревьев Тезауруса значительно превышало количество документов, найденных по словам. Таким образом, полнота поиска с использованием деревьев тезауруса значительно возросла. Однако, как известно, увеличение полноты поиска часто сопровождается снижением точности поиска, то есть релевантными считается большее количество нерелевантных документов.

Чтобы сопоставить точность поиска по Тезаурусу и по словам, мы использовали методику оценки средней точности по трем заданным значениям полноты, описанную в (Vorhees, 1999). Точность выполнения запроса вычисляется при следующих трех значениях полноты: 0.2, 0.5, 0.8.

Чтобы оценить эффективность поиска, необходимо сначала определить множество релевантных документов, а затем проверить релевантность значительного количества полученных по запросу документов. Для снижения трудозатрат, необходимых на проведение оценок, мы сохранили формулировку запроса, но стали сокращать временной интервал до тех пор, пока не получили как релевантные 30-40 документов. Эффективность поиска на таком количестве документов уже достаточно просто проверить.

Приведем результаты наших оценок для двух запросов.

Мы выполнили запрос «Медикаменты» по нормативным документам во временном интервале 01.09.2000 – 01.01.2001 и получили 40 документов при поиске по Тезаурусу (109 понятий - 243 терминов - в дереве расширения: *антибиотики, аптека, вакцина, витамин* и т.д.) и 8 документов при поиске по словам. Просмотрев все полученные документы, мы выяснили, что имеется 25 релевантных документов.

Точность нужно было вычислить при достижении в списке документов 5-го (5/25=0.2), 12-го (12/25=0.6) и 20-го (20/25=0.8) релевантных документов.

При поиске по Тезаурусу пятый релевантный документ был получен десятым, двенадцатый – двадцатым, двадцатый – тридцатым. Таким образом, средняя точность выполнения запроса: (0.5+0.65+0.66)/3=0.57.

При поиске по словам все восемь документов были релевантны. В первой точке точность равна 1.00, но двух других значений полноты поиск по словам достичь не смог, поэтому точность в этих двух точках равна 0.00 . Средняя точность – 0.33.

По запросу «Пожарная безопасность» по нормативным документам на том же временном интервале было получено 32 документа при поиске по Тезаурусу (26 понятий - 99 терминов - в дереве расширения: *авиапожарная служба, брандспойт, …, пожарная защита* и т.д.), и 20 документов при поиске по словам. Было выявлено 27 релевантных документов. Получены следующие оценки точности:

Тип поиска	Точность при полноте 0.2	Точность при полноте 0.5	Точность при полноте 0.8	Средняя точность
По тезаурусу	1.00	0.78	0.85	0.88
По словам	0.83	0.88	0.00	0.57

Приведем примеры документов, которые были сочтены нерелевантными. Документы о награждении правительственными наградами и документы о подчиненности предприятий тому или иному ведомству были рассмотрены как нерелевантные двум указанным запросам. По запросу «*Пожарная безопасность*» документ об обязательной дактилоскопической экспертизе пожарников был рассмотрен как не имеющий отношения к теме. По запросу «*Медикаменты*» были сочтены нерелевантными 5 документов о психотропных средствах, поскольку в этих документах термин «*психотропное средство*» упоминался наряду с термином «*наркотики*», и документы были посвящены проблеме пресечения незаконного оборота психотропных средств и наркотиков.

Всего было выполнено тестирование 19 запросов – рубрик Президентского рубрикатора. Таким образом, были получены следующие значения точности:

Точность при поиске по терминам:
- Точность по терминам в точке 0.2: -- 0.81
- Точность по терминам в точке 0.5: -- 0.58
- Точность по терминам в точке 0.8: -- 0.46
- Средняя точность: = 0.62
-

Точность при поиске по словам:
- Точность по терминам в точке 0.2: -- 14.76 -- 0.77
- Точность по терминам в точке 0.5: -- 9.77 -- 0.52
- Точность по терминам в точке 0.8: -- 0.36 -- 0.02
- Средняя точность: = 0.44

Отметим, что в условиях эксперимента запросы были небольшой длины и при этом имели достаточно хорошее пересечение с терминами Общественно-политического тезауруса. На практике частой ситуацией является наличие в запросе большого количества слов, не входящих в Общественно-политический тезаурус, имеющих другое значение, чем описано в Общественно-политическом тезаурусе и др.

Данный эксперимент подтверждает, что при совпадении запроса с термином тезауруса расширение поиска по тезаурусу приводит к значительному увеличению эффективности информационного поиска. Кроме того, этот эксперимент подтверждает, что наши усилия описывать наиболее надежные, применимые в разных контекстах, отношения в тезаурусе также дали свои результаты.

20.4. Тезаурус и векторная модель в задаче поиска по коллекции нормативно-правовых актов РОМИП

В реальных условиях задания запросов пользователем запросы по отношению к тезаурусу могут быть весьма разнообразны:

- запрос может быть очень коротким (например, содержать отдельное многозначное слово, значение которого без диалога с пользователем выяснить невозможно),
- запрос может содержать некоторую совокупность слов, в которой не найдены термины тезауруса,
- запрос может быть достаточно длинным, и одна часть запроса может ограничивать контекст расширения для другой части запроса и др.

Для учета разных ситуаций была предложена смешанная модель, основанная на совокупности факторов, включая веса слов по пословной векторной модели, веса понятий тезауруса, нахождение сущностей из запроса в ограниченном числе предложений документа. Модель тестировалась на семинаре РОМИП-2008 в коллекции нормативно-правовых документов (Агеев и др., 2008).

Основной направленностью разработки модели была обработка длинных информационных запросов, то есть запросов, которые имеют длину более 3 слов, и выражают некоторую информационную потребность. Информационные запросы условно противопоставляются навигационным запросам, суть последних в нормативно-правовой коллекции заключается в получении документа путем задания его формальных реквизитов: типа документа, номера документа, даты выхода, заголовка.

Для поиска документов по запросам в нормативно-правовой коллекции использовалась двухшаговая процедура.

На первом этапе исполнялась комбинированная векторная модель, построенная на двух индексах – индексе лемм и индексе понятий Общественно-политического тезауруса.

Понятия тезауруса дают возможность дополнительно учесть три дополнительных фактора:

- синонимию терминов,
- лексическую многозначность – производится предварительный выбор наиболее подходящего по контексту значения слов и выражений,
- близкое расположение в тексте компонентов многословных терминов и выражений.

Поэтому результаты работы двух видов векторных моделей могут достаточно серьезно различаться.

Результаты работы векторных моделей замешиваются с помощью параметра α_1, то есть каждый документ получает вес по следующей формуле:

$$W_d = \alpha_1 W_{word} + (1- \alpha_1) W_{conc}, \quad (21.5)$$

где W_{word} – вес документа по пословной векторной модели, W_{conc} – вес документа по векторной модели, выполненной на основе концептов тезауруса.

Из документов, найденных по смешанной векторной модели, отбирается 100 документов.

На втором этапе обработки запроса найденные 100 документов переупорядочиваются по следующему принципу. Максимальное число элементов запроса (слов и терминов) должно быть найдено не разбросанными по всему тексту, а сосредоточены в двух парах соседних предложений. Коэффициент α_2 оценивает относительную весовую значимость лемм и понятий тезауруса в предложениях.

Получение нового веса документа можно представить как двухпроходный процесс. Сначала подсчитываются веса отдельных предложений, которые получаются суммированием весов лемм и концептов из запроса, найденных в предложении:

$$W_s = \alpha_2 \Sigma w_{wordi} + (1- \alpha_2) \Sigma w_{concj} \quad (21.6)$$

где w_{wordi}, w_{concj} – веса слов и концептов предложения.

На втором проходе вычисляется «усиленный» вес каждого предложения: если не все элементы запроса найдены в текущем предложении, то проверяется, нет ли недостающих элементов в соседнем предложении или в еще одной паре предложений документа. Веса дополнительных элементов найденных в других предложениях домножаются на параметрические коэффициенты α_4 (для присоединения элементов из соседнего предложения) и α_5 (для присоединения элементов из другой пары рядом лежащих предложения).

Таким образом, формула «усиленного» веса предложения имеет следующий вид:

$$W_{s1+} = W_1 + \alpha_4 W_{2-} + \alpha_5 [W_{3-} + \alpha_4 W_{4-}] , \quad (21.7)$$

где W_1 - вес «главного» предложения, W_{2-} – вес следующего предложения, W_{3-} , W_{4-} - веса еще одной пары смежных предложений. Причем для каждого следующего предложения учитываются только те слова и концепты, ассоциируемые с запросом, которые еще не были учтены для предыдущих предложений.

Наконец, **на третьем этапе** исходный вес документа**,** полученный на первом этапе, замешивается с весом документа по предложениям, полученный на втором этапе.

Параметры модели оптимизировались на материалах дорожки нормативно-правового поиска romip-legal-2005. Оптимизировалось максимальное число релевантных документов в первых пяти документах выдачи, то есть показатель Precision(5).

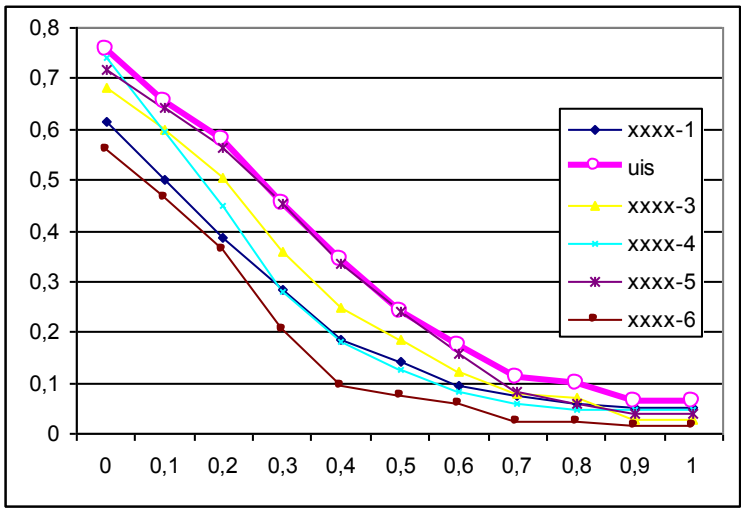

Рис.20.7 Результаты дорожки РОМИП-2008 Legal adhoc (pd35).

В дорожке поиска по нормативно-правовой коллекции представленная модель показала лучший результат из 6 представленных алгоритмов, получив на первых 35 документах, которые были полностью оценены людьми-оценщиками, показатель средней точности MAP (Агеев, Кураленок 2004) – 29.6% (см. рис.21.7), который превышает показатель следующего участника (27.6%) на 7%.

Чтобы проанализировать, насколько хорошо модель отработала на целевом множестве длинных информационных запросов, мы разбили запросы на несколько групп, отдельно выделив длинные информационные запросы, длиной более 3 слов, например, *уплата налога на прибыль организацией при отсутствии затрат* (27 запросов).

Пользуясь этой классификацией, мы разделили все оцененные запросы этой дорожки на соответствующие группы и оценили среднюю точность участников по этим группам. На длинных информационных запросах нами была получена средняя точность MAP – 36%, что значительно превышает наш средний результат (29%), а также результат следующего участника (32%).

Проведенный анализ качества работы системы на разных группах запросов показывает, что важно уметь автоматически классифицировать поступающие запросы, и, в зависимости от класса запроса, применять несколько разные алгоритмы поиска.

20.5. Использование комбинированных моделей для поиска документов по запросам типа «формулировка проблемы» в правовой области

20.5.1. Особенность задачи

Как мы уже указывали в предыдущем разделе, несмотря на то, что подавляющее большинство запросов в поисковых системах относительно небольшой величины (в среднем 2-3 слова), существуют ситуации, когда пользователь задает достаточно длинный запрос. Необходимость в особенно длинных запросах возникает тогда, когда у пользователя есть какая-то проблема, и он обращается в интернет-форумы или вопросно-ответные сервисы, описывает свою проблему и ждет ответа от других пользователей форума или хотел бы найти документ, который помог бы ему справиться с его проблемой. При обращении в форум обязательным условием является то, что перед заданием вопроса людям, необходимо сделать усилия и попробовать найти ответ на свою проблему в предыдущих постах форума.

Задача поиска ответа на вопрос в виде формулировки проблемы значительно отличается от задач, решаемых в стандартных современных вопросно-ответных системах:

- - количество запросов, похожих на вопросы, которые тестировались в рамках конференции TREC (см. главу 12), достаточно мало.
- большинство вопросов представляет собой либо детальное описание ситуации и вопрос, специфичный для данной ситуации, либо совокупность структурно простых подвопросов, которые вместе также задают описание специфической правовой ситуации.
- при этом структурно сложные вопросы состоят из нескольких предложений и/или содержат несколько подвопросов.

При обработке структурно сложных вопросов имеются следующие сложности по сравнению с обработкой простых вопросов:

- автоматически трудно точно определить структуру вопроса – разбить его правильно на подвопросы, определить фокус вопроса;
- если часто можно ожидать, что ответ на структурно простой вопрос может содержаться в одном предложении текста, то ответ на структурно сложный вопрос может «собираться» из нескольких предложений документа.

В связи с этим для структурно сложных вопросов наиболее важным является поиск документов, содержащих описание соответствующей ситуации, при этом часто учет информации о структуре вопроса носит дополнительный характер.

Обработка длинных поисковых запросов в значительной степени отличается от обработки коротких поисковых запросов, которые являются наиболее распространенными запросами к поисковым системам.

Если при поиске по коротким запросам, поисковая система, скорее всего, найдет множество документов, включающих все слова запроса, и ее главной задачей является правильное упорядочение найденных документов, то при обработке длинных запросов к информационной системе в подавляющем большинстве случаев не найдется ни одного документа или найдется всего несколько документов, содержащих все слова запроса. И,

таким образом, основной задачей при обработке такого запроса является поиск и упорядочение документов, содержащих лишь часть слов запроса.

Казалось бы, векторные модели информационного поиска, которые описывают запрос и документы как вектора слов с весами, дают хороший базис для поиска ответов на длинные поисковые запросы, поскольку эта технология дает возможность установления частичного соответствия между запросом и документом.

Однако в реальности оказывается, что при использовании векторной модели часто поиск производится по относительно малозначащим словам запроса, в то время как очень важные слова запроса могут при сопоставлении исчезнуть. Как мы указывали в разделе 12.1 для того, чтобы в некоторой степени управлять формированием поискового запроса предлагается использование многошаговых булевских моделей. В следующем разделе будет описан алгоритм этого типа, который мы назвали «феноменологическая модель».

20.5.2. Алгоритм «Феноменологическая модель»

Феноменологическая модель – методика решения задачи поиска документов по запросу типа «формулировка проблемы» посредством моделирования понятиями тезауруса содержания ситуации вопроса.

Феноменологическая модель преобразует запрос в булевское выражение типа конъюнкция дизъюнкций над понятиями тезауруса:

$$\bigcap_i \bigcup_j c_{i,j},$$

где $c_{i,j}$ — понятия тезауруса.

Элементами дизъюнкции могут быть понятия тезауруса, которые рассматриваются как близкие по смыслу – они связаны между собой тезаурусными путями определенного вида.

Действительно, вопрос не является последовательностью произвольных слов. В длинном вопросе многие упоминаемые понятия связаны между собой, например, принадлежат одной и той же области деятельности или одному и тому же типу.

Запрос типа «формулировка проблемы» описывает некоторую определенную ситуацию. Поэтому, чтобы иметь возможность дополнять булевское выражение понятиями из тезауруса, необходимо иметь дополнительное подтверждение, что то или иное расширение подходит к описанной ситуации. Для этого используются информеры (см. раздел 20.2).

В создаваемое булевское выражение могут быть добавлены понятия тезауруса из дерева-вниз или дерева-вверх одного из понятий запроса, если эти понятия входили в состав информера, то есть принадлежали к множеству наиболее характерных понятий текущей выдачи. Дополнительное понятие вводится в дизъюнкцию к породившему его понятию запроса.

Феноменологическая модель рассматривается нами не как отдельная модель, а как отдельный компонент многошаговой модели. В частности, работа феноменологической модели начинается после предварительной работы векторной модели, которая отбирает 100 наиболее релевантных по запросу документов. Понятия тезауруса из формулировки запроса упорядочиваются по количеству документов, найденных в этой выдаче – так определяются наиболее совместимые друг с другом понятия. Работа феноменологической модели начинается с наиболее частотного понятия в упомянутой выдаче векторной модели, которое становится первым компонентом формируемого булевского выражения.

Рассмотрим работу феноменологической модели подробнее.

20.5.2.1. Обработка исходной формулировки вопроса

Работа модели начинается с того, что формулировка запроса сопоставляется с тезаурусом и составляется список понятий формулировки вопроса. Для многозначных слов проверяется, не разрешается ли многозначность на основе текущего списка понятий. Если есть возможность разрешить многозначность, то производится выбор значения или снятие пометки многозначности.

Для каждого понятия формулировки определяется количество документов предварительной векторной выдачи, в которых оно встречается.

Следующее действие, которое нужно выполнить – построить списки близких по смыслу и поэтому потенциально объединяемых в дизъюнкции понятий запроса, на роль которых подходят понятия, связанные по иерархии тезаурусных связей.

Между понятиями вопроса могут быть выявлены следующие типы взаимосвязей:

Одно понятие находится в дереве другого понятия (это основной тип взаимосвязи)	Тип 1
Деревья-вверх двух понятий пересекаются в основной части тезауруса	Тип 2

Точка пересечения деревьев иерархии может быть расклассифицирована, например, следующим образом:

- пересечение по двум отношениям, одно из которых отношение ЦЕЛОЕ и длина пути не больше 3;
- пересечение по двум отношениям, одно из которых отношение АСЦ1 и длина пути не больше 3;
- пересечение по двум отношениям ВЫШЕ и длина пути не больше 3.
- одно из отношений ЦЕЛОЕ и длина пути не больше 10 до каждого из понятий, и длина пути до одного из понятий не больше 5.

Данная классификация связана с представлениями о близости понятий, не находящимися в непосредственном подчинении в иерархии тезауруса. Типы перегибов упорядочены по предполагаемому снижению семантической близости между исходными понятиями.

Таким образом, для каждого понятия вопроса должна быть вычислена информация:

- нижестоящие по дереву понятия из вопроса;
- вышестоящие по дереву понятия из вопроса;
- понятия из вопроса с взаимосвязью-перегибом (тип перегиба, понятие в точке перегиба).

Данные отношения строятся для всех основных понятий запроса, включая многозначные.

Важной частью обработки формулировки запроса является формирование ядра запроса. Ядро вопроса составляют понятия формулировки вопроса, для которых выполняются два условия:

- они порождаются по однозначным терминам или многозначность терминов была разрешена,
- их частота среди 100 документов, найденных по данному запросу по векторной модели, не менее 5.

Необходимость выделения ядра запроса связана с тем, что в запросе типа «формулировка проблемы» может быть большое количество случайно упомянутых понятий, в том числе, редко встречающихся в коллекции понятий. В таких случаях их относительно малая частотность в целевой коллекции не является критерием их важности для релевантной выдачи.

Остальные понятия формулировки вопроса также запоминаются для последующего уточнения запроса

В ходе поиска документов нужно сформировать такой запрос к поисковой системе, чтобы, он включал все понятия ядра для данной формулировки вопроса. В процессе формирования, найденные документы складываются в копилку документов.

20.5.2.2. Построение формулы описания формулировки запроса

Формула описания запроса наращивается по шагам. Установлены следующие параметры алгоритма:

- doc_num_max – если число документов в выдаче меньше doc_num_max, то найденные на очередном шаге документы складываются в копилку документов (например, doc_num_max=50) в качестве потенциально релевантных;
- doc_num – если число документов в выдаче, меньше этого числа, то запрос начинает расширяться, если больше – то сужаться (например, doc_num=20).

Все действия по расширению и сужению запроса оцениваются относительно первых понятий тезауруса, начавших отдельную дизъюнкцию D0i.

Построение формулы начинается с наиболее частотного в векторной выдаче понятия.

На каждом шаге выполняется сформированный запрос, оценивается количество найденных документов. Рассматриваются две основные ситуации: 1) больше ли количество документов в выдаче, чем doc_num или 2) меньше, чем doc_num.

В первом случае, нужно запрос сужать, то есть увеличивать конъюнкцию новыми элементами. В качестве нового конъюнкта берется понятие из ядра формулировки ядра запроса, не связанное или с наименьшим весом связанное по тезаурусу с начальными понятиями дизъюнкций Di0 текущего булевского выражения. Тем самым более близкие понятия оставляются как ресурс для возможного расширения запроса. Это дает возможность одни и те же понятия в некоторых запросах располагать в разных элементах конъюнкции (то есть использовать для сужения запроса), а в других – как элементы одной и той же дизъюнкции (использовать для расширение запроса).

Если таких (наиболее далеких) понятий несколько, то выбирается первое по списку понятий-кандидатов на добавление.

Во втором случае, необходимо расширять формируемый запрос, дополняя дизъюнкции.

В качестве понятий, которыми могут быть дополнены дизъюнкты, могут использоваться:

- понятия формулировки вопроса, еще не включенные в формируемое булевское выражение и имеющие разрешенные тезаурусные пути к начальным понятиям дизъюнкций Di0,
- понятия, которых нет в формулировке запроса, но которые находятся в дереве-вверх или в дереве-вниз начальных понятий дизъюнкций Di0 и которые были подтверждены информером последнего запроса, как наиболее характерные для последней выдачи документов,
- если таких понятий не имеется и есть еще понятия ядра формулировки, которые не включены в булевское выражение, то последняя дизъюнкция запроса начинает наращиваться этими оставшимися понятиями.

Результат исполнения последнего запроса (который содержит все понятия ядра) заносится в копилку. Отметим, что операции сужения и расширения запроса всегда применимы, пока не все понятия ядра вопроса включены в формулу. Таким образом, алгоритм гарантирует включение всех понятий ядра вопроса в формулу. Документы, полученные работой алгоритма, присоединяются к документам, полученным векторной моделью и направляются на дальнейший анализ, который производится подобно

процедуре, описанной в разделе 20.4, посредством оценки наиболее наполненных элементами запроса и расширением запроса предложений

Приведем пример сформированного феноменологической моделью булевского запроса для следующей формулировки запроса:

Вопрос: *Туристическая фирма (турагент) занимается реализацией путевок сторонних организаций в санаторно-курортные и оздоровительные учреждения. В соответствии с действующим законодательством реализация такого продукта не подлежит обложению НДС. Однако в ходе проверки налоговой инспекцией нам были предъявлены санкции за неуплату налога с суммы агентского вознаграждения. Правы ли в данном случае налоговые органы? ("Консультант бухгалтера", N 7, июль 2001 г.)*

Для данной формулировки выделены следующие понятия ядра, которые необходимо «уложить» в булевское выражение (перечислены по алфавиту):

АГЕНТСКОЕ ВОЗНАГРАЖДЕНИЕ
НАЛОГ НА ДОБАВЛЕННУЮ СТОИМОСТЬ
НАЛОГОВАЯ СЛУЖБА
НАЛОГОВОЕ ОСВОБОЖДЕНИЕ
ОЗДОРОВИТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ПУТЕВКИ НА ОТДЫХ И ЛЕЧЕНИЕ
САНАТОРНО-КУРОРТНОЕ ЛЕЧЕНИЕ
СТОРОННЯЯ ОРГАНИЗАЦИЯ
ТУРАГЕНТ
ТУРИСТИЧЕСКАЯ ФИРМА

Формирование булевского выражения началось с понятия ТУРАГЕНТ. По данному запросу в коллекции найдено 66 документов, что больше установленного параметра doc_num=20, поэтому в конъюнкцию добавляется понятие ОЗДОРОВИТЕЛЬНОЕ УЧРЕЖДЕНИЕ, что приводит к величине выдачи 8 документов. Запрос необходимо расширять. Из формулировки извлекается понятие ТУРИСТИЧЕСКАЯ ФИРМА, являющееся вышестоящим понятием для понятия ТУРАГЕНТ, и вносится в соответствующую дизъюнкцию, получается такой запрос:

```
(ТУРАГЕНТ OR  ТУРИСТИЧЕСКАЯ ФИРМА)
AND
(ОЗДОРОВИТЕЛЬНОЕ УЧРЕЖДЕНИЕ)
```

В результате выполнения такого запроса находится 16 документов. Запрос необходимо расширять дальше. Такую возможность дает информер, сформированный по последнему булевскому запросу. На седьмом месте самых характерных понятий для данной выдачи находится понятие САНАТОРИЙ, который является видом понятия ОЗДОРОВИТЕЛЬНОГО УЧРЕЖДЕНИЕ, и, таким образом, пополняется соответствующая дизъюнкция. Получается следующий булевский запрос:

```
(ТУРАГЕНТ OR ТУРИСТИЧЕСКАЯ ФИРМА)
AND
(ОЗДОРОВИТЕЛЬНОЕ УЧРЕЖДЕНИЕ OR  САНАТОРИЙ)
```

Выдача данного запроса содержит 22 документа, и запрос опять можно уточнять.

В результате последовательности шагов работы алгоритма было сформировано следующее булевское выражение:

```
(ТУРАГЕНТ                          OR
 ТУРИСТИЧЕСКАЯ ФИРМА               OR
 ТУРИСТИЧЕСКИЙ СЕРВИС              OR
```

```
 ПОСРЕДНИЧЕСКАЯ ДЕЯТЕЛЬНОСТЬ          OR
 АГЕНТСКОЕ ВОЗНАГРАЖДЕНИЕ             OR
 ПОСРЕДНИЧЕСКАЯ ОРГАНИЗАЦИЯ           OR
 ПУТЕВКИ НА ОТДЫХ И ЛЕЧЕНИЕ)

AND

(ОЗДОРОВИТЕЛЬНОЕ УЧРЕЖДЕНИЕ          OR
 САНАТОРИЙ                           OR
 ДОМ ОТДЫХА                          OR
 ОТДЫХ                               OR
 ПРОФИЛАКТОРИЙ                       OR
 ДЕТСКОЕ ОЗДОРОВИТЕЛЬНОЕ УЧРЕЖДЕНИЕ  OR
 СТОРОННЯЯ ОРГАНИЗАЦИЯ)

AND

(САНАТОРНО-КУРОРТНОЕ ЛЕЧЕНИЕ         OR
 САНАТОРНО-КУРОРТНАЯ ПУТЕВКА         OR
 ЗДРАВООХРАНЕНИЕ                     OR
 ЛЕЧЕНИЕ)

AND

(НАЛОГ НА ДОБАВЛЕННУЮ СТОИМОСТЬ)

AND

(НАЛОГОВОЕ ОСВОБОЖДЕНИЕ             OR
 НАЛОГОВАЯ СЛУЖБА)
```

По этому запросу был найден 51 документ.

Помимо понятий Тезауруса**,** найденных в исходной формулировке, феноменологическая модель добавила в булевское выражение следующие понятия:

- *ТУРИСТИЧЕСКИЙ СЕРВИС,*
- *ПОСРЕДНИЧЕСКАЯ ДЕЯТЕЛЬНОСТЬ,*
- *ПОСРЕДНИЧЕСКАЯ ОРГАНИЗАЦИЯ,*
- *САНАТОРИЙ,*
- *ДОМ ОТДЫХА,*
- *ОТДЫХ,*
- *ПРОФИЛАКТОРИЙ*
- *ДЕТСКОЕ ОЗДОРОВИТЕЛЬНОЕ УЧРЕЖДЕНИЕ*
- *САНАТОРНО-КУРОРТНАЯ ПУТЕВКА,*
- *ЗДРАВООХРАНЕНИЕ,*
- *ЛЕЧЕНИЕ*

20.5.2.3 Применение феноменологической модели

Знания, описанные в тезаурусе, не могут быть полными, и в очередной формулировке запроса могут потребоваться знания, не отраженные в тезаурусе. Поэтому феноменологическая модель не применяется отдельно, а входит в состав многошаговой модели, описанной в разделе 20.4.

Феноменологическая модель работает после комбинированной векторной модели. Найденные в формулировке понятия тезауруса упорядочиваются по количеству документов, в которых они упоминаются в этих 100 документах для работы феноменологической модели. Таким образом, предполагается, что булевские запросы феноменологической модели будут строиться на понятиях тезауруса, которые наиболее часто упоминаются в связи друг с другом.

В результате работы модели и исполнения построенных булевских запросов «копилка» документов для дальнейшего анализа пополняется дополнительными документами. Кроме того, в процессе своей работы феноменологическая модель расширяет запрос понятиями тезауруса, которые не были упомянуты в запросе, и эти дополнительные понятия будут также придавать дополнительный вес найденным документам.

Суть дальнейшего анализа документов заключается в том, чтобы дополнительно проанализировать все найденные на предыдущих этапах документы (100 документов от смешенной векторной модели и 30-100 документов от феноменологической модели). Наилучшими считаются документы, в которых максимальное число найденных элементов запроса, найдено в 2 парах соседних предложений документа (см. раздел 20.4).

Формула предложения (20.7) дополняется еще и весом понятий тезауруса, которые не были упомянуты в формулировке запроса, но были получены в процессе расширения по феноменологической модели. Таким образом, вес отдельного предложения вычисляется следующим образом по сравнению с формулой в разделе 20.4:

$$W_s = \alpha_2 \Sigma w_{wordi} + (1- \alpha_2) \Sigma w_{concj} + \alpha_3 \Sigma w_{exp} \qquad (20.7)$$

где w_{wordi}, w_{concj} – веса слов и концептов из исходной формулировки, w_{exp} – это вес понятия тезауруса, которого не было в исходной формулировке, но который был добавлен в расширенный запрос на этапе работы феноменологической модели:

$$w_{exp}(C_i) = idf(C_i)$$

«Усиленный» вес за счет дополнительных предложений считается так, как описано в разделе 21.4, в дополнительных предложениях также учитываются дополнительно полученные понятия тезауруса.

Как и указывалось в разделе 20.4, полученный вес предложения замешивается с исходным весом предложения, полученным по векторной модели первого этапа.

Таким образом, выполнение феноменологической модели дает возможность привлечь дополнительное число документов для последующего анализа, и, кроме того, учесть вес понятий, полученных как расширение булевского запроса.

Качество комбинированной модели, включая феноменологическую модель, тестировалось на 165 запросах типа «формулировка проблемы» в юридической области экспертами-юристами на коллекции документов, отвечающих на такие вопросы (40 тысяч документов). Оценка производилась по показателю точности по первым пяти документам - precision (5). В результате было получено, что показатель precision (5) для алгоритма, использующего тезаурусные знания и феноменологическую модель, более чем на 12% превышает работу лучшего алгоритма, работающего только на основе слов (векторная модель + упорядочение по предложениям + замешивание полученных весов).

Заключение к главе 20

Применение тезауруса РуТез в задаче поиска документов, основанное только на тезаурусных описаниях, может оказаться не лучше применения пословных моделей (из-за возможных проблем нехватки информации в тезаурусе, неточности описаний, проблем разрешения многозначности и др.). Однако гибкое сочетание качественной пословной модели и знаний, описанных в РуТез, дает улучшение качества на 10-15 процентов. Поэтому тезаурусные технологии не должны противопоставляться современным технологиям пословной обработки текстов, а органично учитывать последние достижения в этой сфере. При учете таких условий применение тезаурусов может дать улучшение качества решения задачи по сравнению с лучшими пословными методами.

Глава 21. Тезаурус РуТез как ресурс для автоматической рубрикации текстов

21.1. Технология автоматического рубрицирования на основе тезауруса

Как уже указывалось в разделе 13.1, существуют два основных подхода к автоматическому рубрицированию документов – инженерный подход и подход на основе машинного обучения. Традиционным нашим подходом в сфере автоматической рубрикации является инженерный подход, в котором содержание рубрики описывается как булевское выражение над понятиями Общественно-политического тезауруса. Текущий рубрикатор связывается с Тезаурусом посредством небольшого числа опорных понятий, рубрики остальных понятий тезауруса выводятся по связям внутри Тезауруса, тем самым при описании очередного рубрикатора используется большой объем накопленных в тезаурусе знаний.

Процедура рубрикации базируется на автоматически построенном тематическом представлении документов, которое моделирует основную тему и подтемы документа наборами (тематическими узлами) близких по смыслу понятий, упомянутых в документе. Такая основа рубрикации дает возможность обрабатывать тексты разных типов и размеров: нормативные акты, газетные статьи, новостные сообщения, научные публикации в области гуманитарных наук, социологические опросы (Лукашевич 1996; Добров, Лукашевич, 2002а; .Агеев и др., 2008).

Посредством такой технологии рубрикации были разработаны более 15 систем автоматической рубрикации, в частности, такие системы рубрикации как:

- рубрикация законодательных актов по Классификатору правовых актов РФ – 1169 рубрик,
- рубрикация научных статей по экономике по рубрикатору JEL (ссылка – 700 рубрик),
- рубрикация по правовому классификатору Центральной избирательной комиссии (450 рубрик, 4 уровня),
- рубрикация социологических опросов по рубрикатору (300 рубрик) и др.

В следующих разделах рассмотрим подробнее особенности реализации систем автоматической рубрикации на основе тезауруса и тематического представления документов.

21.2. Описание смысла рубрики понятиями тезауруса

При создании лингвистического профиля рубрикатора каждая рубрика R описывается дизъюнкцией альтернатив, каждый дизъюнкт представляет собой конъюнкцию:

$$R = \bigcup_i D_i \;; \qquad D_i = \bigcap_j K_{ij} \;, \tag{21.1}$$

Конъюнкты в свою очередь описываются экспертами с помощью так называемых «опорных» понятий тезауруса. Для каждого опорного понятия задается правило его расширения $f(\cdot)$, определяющее, каким образом вместе с опорным понятием учитывать подчиненные ему по иерархии понятия: без расширения (обозначается символом «N»), полное расширение по дереву иерархии тезауруса (символ «E»), расширение только по родовидовым связям (символ «L»), расширение по всем видам отношений на один уровень иерархии (символ «W»), расширение на один уровень иерархии, не включая отношения НИЖЕ (символ «V»).

Опорное понятие может быть как «положительным», то есть добавлять нижерасположенные понятия в описание конъюнкта, так и «отрицательным», то есть вырезать из описания рубрики свои подчиненные понятия. Последовательность учета

положительных и отрицательных опорных понятий регулируется заданием специального атрибута. Результатом применения расширения опорных понятий является совокупность понятий тезауруса, полностью описывающая конъюнкт:

$$K_{ij} = \bigcup_m f_m(c_{ijm}) \setminus \bigcup_n f_n(e_{ijn}) = \bigcup_k d_{ijk} \ . \qquad (21.2)$$

Отметим, что для рубрикаторов простой структуры, когда рубрики разделяют пространство предметной области на непересекающиеся части, часто возможно обходиться случаем одной альтернативы (одного дизъюнкта) и одного конъюнкта, при этом роль отрицательных опорных понятий может выражать специальная «нулевая» рубрика, задача которой «выедать» ненужные понятия.

Рассмотрим фрагмент представления рубрики 200.020.020 «Встречи на высшем уровне» из Классификатора правовых актов РФ ((Указ, 2000), более 1000 рубрик). Языковые выражения, записанные курсивом, выводятся на основе исходного описания рубрики автоматически (рис.21.1):

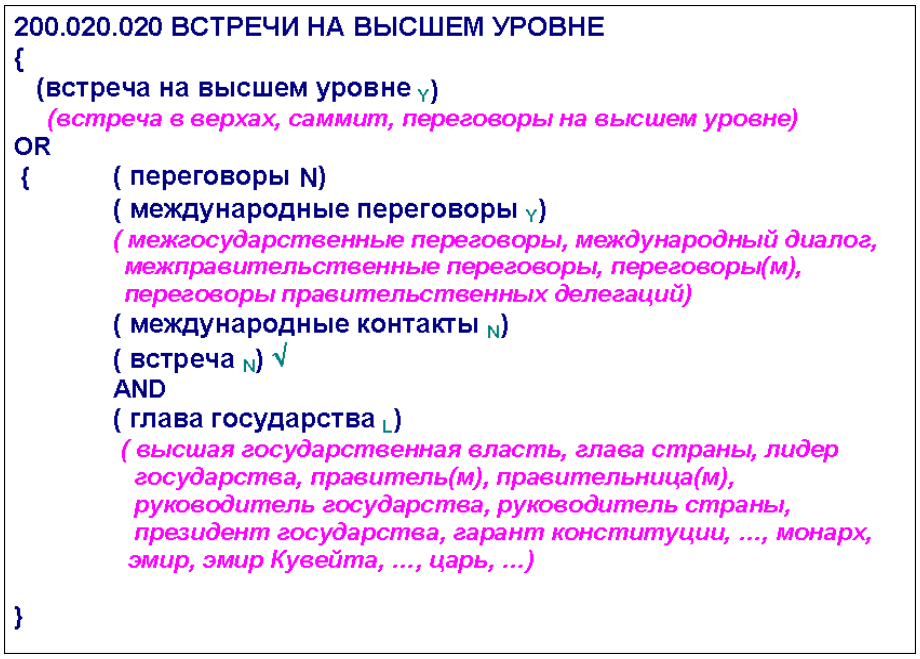

Рис.21.1. Расширенное представление рубрики понятиями тезауруса

Важным атрибутом описания рубрики является пометка о необходимости «подтверждения». Понятия, требующие подтверждения, не могут самостоятельно выводить рубрику, но могут усиливать эту рубрику, если в тексте встречаются понятия, не требующие подтверждения. Например, если в тексте говорится о конфликте двух пенсионеров в очереди, еще не должна выводиться рубрика «Пенсионное обеспечение», так как здесь используется только одно свойство понятия *ПЕНСИОНЕР* - как гражданин преклонного возраста. В нашем описании понятие *ПЕНСИОНЕР* должно иметь пометку о подтверждении для данной рубрики. Однако, если дополнительно в тексте будет сказано, что конфликт произошел из-за маленькой пенсии, низкого жизненного уровня и т.п., то рубрика должна выводиться, причем наличие понятия *ПЕНСИОНЕР* должно усиливать вес данной рубрики.

По умолчанию пометка подтверждения устанавливается для понятия d_{ijk}, если на любом пути от положительного опорного понятия, которому соответствует d_{ijk} , имеется пометка на отношении (см.п.17.6). При этом эксперт, описывающий рубрику, может задать/снять пометку подтверждения вручную, что распространится на все нижерасположенные понятия.

Следует подчеркнуть, что в данной методологии достаточно хранить только опорные понятия, а также понятия, у которых изменен атрибут подтверждения, полное же описание рубрики может быть каждый раз пересчитано заново при изменении тезауруса. Типичные цифры о параметрах описания: на одну рубрику рубрикатора в среднем приходится 1-2 дизъюнкта, 2-3 конъюнкта, 4-8 опорных понятия, 50-100 понятий полного описания, то есть 100-250 текстовых выражений.

21.3. Автоматическое рубрицирование на основе тематического представления

Как отмечалось в предыдущем разделе, рубрика представляется в виде логического условия над понятиями тезауруса:

$$R = \bigcup_i D_i = \bigcup_i \left[\bigcap_j K_{ij} \right] = \bigcup_i \left[\bigcap_j \left(\bigcup_k d_{ijk} \right) \right]. \qquad (21.3)$$

Таким образом, оценка релевантности содержания текста рубрике (вес рубрики) может быть рассчитана на основе информации о весах понятий в тексте, входящих в ее описание.

Вес конъюнкта рассчитывается по формуле:

$$\theta(K_{ij}) = \min\{1.0; \max(\theta(d_{ijk}), \chi \cdot \theta(p_{ijm}))\}, \qquad (21.4)$$

где d_{ijk} понятия, не требующие подтверждения, p_{ijm} – понятия, требующие подтверждения, χ - множитель равный единице, если имеются понятия, не требующие подтверждения, и нулю иначе.

Вес дизъюнкта предназначен учитывать не только сумму весов составляющих его конъюнктов, но и меру близости конъюнктов в тексте:

$$\theta(D_i) = \frac{\sum_{j=1}^{m} \theta(K_{ij}) + \sum_{j<k} S(K_{ij}, K_{ik})}{m + C_m^2}, \qquad (21.5)$$

здесь $S(K_{ij}, K_{ik}) = \min\{1.0; \frac{\sum s(c_{ijq} \in K_{ij}, d_{ikw} \in K_{ij})}{\max s(c \in D, d \in D)}\}$

- сумма всех текстовых связей между понятиями одного конъюнкта и понятиями другого, деленная на значение максимальной текстовой связи между любыми двумя понятиями текста. Этот член равен обычно единице для сильно связанных конъюнктов и принимает малое значение, если понятия различных конъюнктов обсуждались в разных местах текста.

Вес рубрики представляет собой максимум весов входящих в описание рубрики альтернатив. В случае имеющихся иерархических связей между рубриками оценка релевантности нижестоящих рубрик переносится на вышестоящие. Так что при запросе по вышестоящей рубрике будут выходить и документы, к которым были приписаны нижестоящие рубрики.

Алгоритм рубрицирования работает следующим образом. Для всех понятий тезауруса, найденных в тексте, определяется множество рубрик, которые могут быть определены в тексте. Для каждой рубрики происходит расчет ее веса по формулам (21.4) и (21.5). В результирующем множестве остаются рубрики, вес которых превосходит задаваемый заранее для коллекции порог.

Применение описанной технологии для нескольких систем рубрикации для различных текстовых коллекций показали, что описание рубрикатора посредством опорных понятий служит и как основа для соответствующих организационных решений:

- является прообразом свободного от субъективизма комментария к рубрикатору, который может пополняться и уточняться;
- при выводе рубрики всегда можно показать/объяснить, почему была выведена та или иная рубрика, что позволяет быстро уточнять описание рубрик, анализируя замеченные ошибки рубрикации.

21.4. Использование информеров для составления описаний рубрик при инженерном подходе рубрикации

Одним из недостатков инженерного подхода к рубрикации часто указывается сложность использования коллекций, отрубрицированных вручную, в качестве обучающей коллекции. Эта проблема становится особенно важной, если предполагается рубрикация по рубрикатору сложной структуры, и имеется множество различных неявных правил отнесения/неотнесения документа к рубрике.

В таких случаях улучшить и убыстрить построение формул рубрик помогают информеры УИС Россия (см. п. 20.2). Полученная отрубрицированная коллекция документов загружается в поисковую систему, причем предоставляется возможность поиска по проставленным экспертами рубрикам. Выполняя запрос на поиск документов по той или иной рубрике, можно в информере получать и анализировать совокупности наиболее характерных понятий тезауруса для этой рубрики, что помогает составить формулу рубрики.

Опишем алгоритм работы специалиста по рубрикации для решения различных задач поддержки рубрицирования по сложному рубрикатору с использованием информеров.

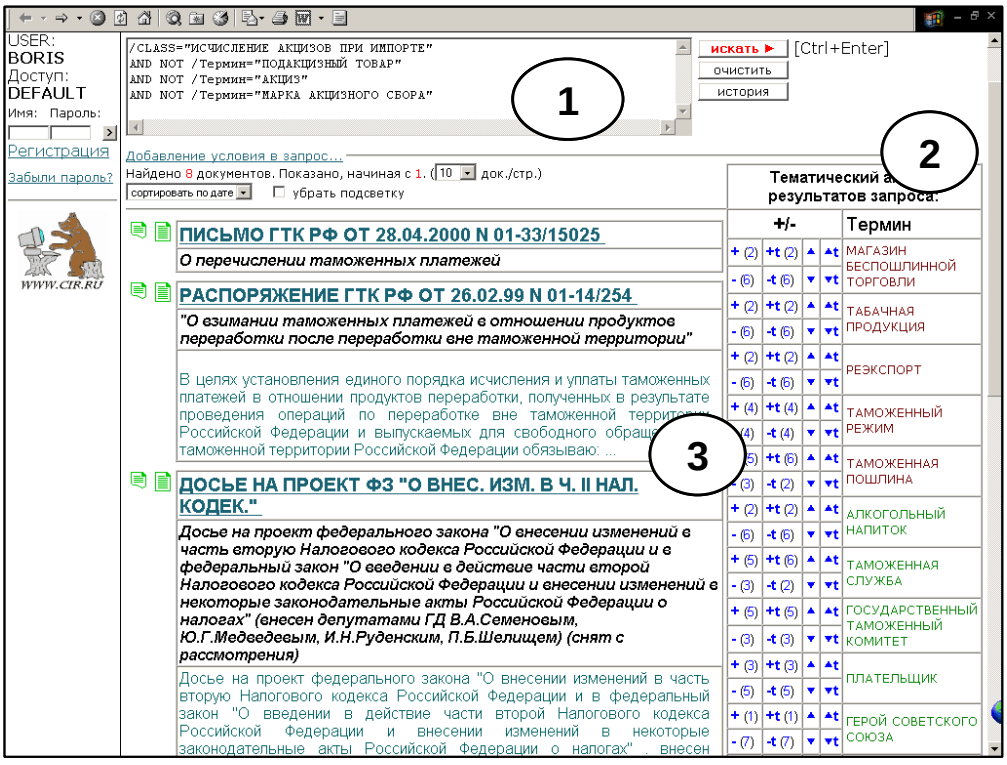

Рис.21.2. Использование информеров УИС РОССИЯ для интерактивного уточнения описания рубрики. (1) окно условий запроса; (2) тематический информер; (3) «ссылки-кнопки» для оперативного добавления условия в запрос

Для того, что составить для рубрики терминологическое описание, необходимо выявить элементарные смыслы рубрики, найти, какими терминами эти смыслы могут выражаться. Далее необходимо записать булевское выражение, в котором термины, выражающие разные составляющие смыслы рубрики, будут соединяться конъюнкцией, а термины, выражающие один и тот же смысл дизъюнкцией.

Для нахождения соответствующих понятий удобно использовать информеры УИС РОССИЯ. Рассмотрим «модельную» рубрику «Особенности исчисления акцизов при импорте». Выполняем поиск по рубрике – получаем набор документов, отнесенных к рубрике экспертами.

Каждый текст, относящийся к этой рубрике, должен содержать термины, относящиеся к сфере импорта, и термины, относящиеся к сфере акцизов.

Выбираем из правой колонки понятия, относящиеся к акцизам: *ПОДАКЦИЗНЫЙ ТОВАР, АКЦИЗ, МАРКА АКЦИЗНОГО СБОРА*. Удаляем из выдачи документы, содержащие эти понятия, чтобы определить, какие еще термины могут относиться к сфере акцизов.

Собираем теперь понятия, относящиеся к *импорту*. Возвращаемся к запросу по рубрике. Изучаем правую колонку – имеется понятия *ИМПОРТ*. Удаляем документы, включающие этот термин, из выдачи.

Информер больше понятий не дает. Начинаем изучать оставшиеся тексты. В текстах содержатся слова *ввоз, ввезти, ввозить, ввозной*. Убираем эти документы – остается 43 документа.

В правой колонке появились понятия *ТАМОЖЕННАЯ ПОШЛИНА, ТАМОЖЕННОЕ ОФОРМЛЕНИЕ ТОВАРОВ, ГОСУДАРСТВЕННЫЙ ТАМОЖЕННЫЙ КОМИТЕТ*. В сочетании с акцизами эти понятия должны указывать на импорт.

Таким образом, мы получили формулу:

```
(ПОДАКЦИЗНЫЙ ТОВАР или АКЦИЗ или МАРКА АКЦИЗНОГО СБОРА)
и
(ИМПОРТ
 или ВВОЗ
 или ТАМОЖЕННАЯ ПОШЛИНА
 или ТАМОЖЕННОЕ ОФОРМЛЕНИЕ ТОВАРОВ
 или ТАМОЖЕННЫЙ КОМИТЕТ)
```

На каждом шаге происходит контроль оставшегося количества документов, процесс уточнения формулы прекращается, если достигнут требуемый уровень ошибки.

Если название рубрики выглядит как состоящее из одного термина, то это часто не означает, что достаточно упоминания этого термина в тексте, чтобы присвоить тексту рубрику. Часто такой текст должен обсуждать какие-то значимые для данного понятия части, свойства и ситуации.

Так, тексты в рубрике «Общества с ограниченной и с дополнительной ответственностью» должны содержать не только термины *общество с ограниченной ответственностью* или *общество с дополнительной ответственностью*, но и обсуждать такие важнейшие аспекты для этих организаций, как создание, регистрация, учредители, уставный капитал, собственность и т.п.

Таким образом, реально рубрика также разлагается на два элементарных смысла, тот что назван в формулировке и что-то вроде «общие вопросы», и описывать рубрику нужно в виде конъюнкции двух частей. Понятия, которые нужно включить во вторую часть конъюнкции, т.е. те которые важны для функционирования первой части, могут быть набраны из информера. Для упомянутой рубрики на правой панели мы увидим: *УСТАВНЫЙ КАПИТАЛ, УЧРЕДИТЕЛЬ, РЕГИСТРАЦИЯ ЮРИДИЧЕСКИХ ЛИЦ, СОВЕТ ДИРЕКТОРОВ*.

Таким образом, в сложных задачах рубрикации существенным становится взаимодействие с экспертами, так как единственным способом решения задачи рубрикации является итерационное уточнение правил рубрицирования.

Для этих целей можно применять методы, основанные на знаниях, которые позволяют легко интерпретировать, почему такой-то документ был отнесен к рубрике. Основным недостатком этих методов является высокая трудоемкость, обусловленная необходимостью привлечения экспертов для составления таких правил. Однако, представляется, что это неизбежно, поскольку в реальных задачах рубрикации отмечена значительная непоследовательность исходных данных ручной рубрикации (см.п.13.3.1.).

21.5. Эксперимент по автоматической рубрикации текстов в рамках семинара РОМИП 2007

Опишем результаты работы системы автоматического рубрицирования, основанной на тезаурусных знаниях в задаче классификации Web-страниц в рамках семинара РОМИП 2007 (Агеев и др., 2008а). Исходный набор данных включал в себя коллекцию страниц с сайтов белорусского интернета BY.web и коллекцию DMOZ, используемую в качестве обучающего множества. Обучающее множество состоит из *сайтов*, но не обязательно все страницы сайта относятся к одной теме. Рубрикация должна была быть выполнена для 247 рубрик рубрикатора DMOZ.

При выполнении данного эксперимента была поставлена задача выяснения, сколько времени нужно потратить на описание заданных рубрик с использованием из информации из тезауруса, и каких показателей качества рубрицирования можно достигнуть.

Работа по описанию 247 рубрик задания была выполнена за 8 часов рабочего времени. В опорных булевских выражениях было использовано около 900 понятий тезауруса, в расширенных булевских выражениях содержится около 40 тысяч понятий (с повторениями). Каждому понятию тезауруса соответствует в среднем 2-3 языковых выражения (слова или словосочетания).

Примером простого описания рубрики может служить описание рубрики 135 «Боевые искусства». Опорное булевское выражение состоит из одного понятия *БОЕВЫЕ ИСКУССТВА* с меткой «Е» полного расширения по тезаурусу. В состав расширенного булевского выражения входят помимо исходного следующие понятия: *АЙКИДО, ДЖИУ-ДЖИТСУ, ДЗЮДО, КАРАТЭ, САМБО, ДЗЮДОИСТ, КАРАТИСТ, САМБИСТ*. Понятия тезауруса, соответствующие людям (*ДЗЮДОИСТ, КАРАТИСТ, САМБИСТ)* входят в рубрику с пометкой подтверждения, поскольку появление соответствующих слов в тексте еще не означает, что текст посвящен боевым искусствам.

Все эти понятия и текстовые входы входили в состав Тезауруса до начала эксперимента. F1-мера по метрике OR для этой рубрики составляет 0.97 (полнота 0.98, точность 0.96).

Для рубрики 60 «Зимние виды спорта» опорное булевское выражение составляет одно понятие *ЗИМНИЕ ВИДЫ СПОРТА*, которое отмечено пометкой полного расширения по иерархии Тезауруса. В расширенном булевском выражении содержится 50 понятий тезауруса. F1-мера по метрике OR для этой рубрики составляет 0.84 (полнота 0.80, точность 0.88).

Рубрика 43 «Домашний ремонт» описана как конъюнкция, состоящая из двух элементов.

Один элемент – дизъюнкция нескольких понятий тезауруса, связанных с темой ремонта без тезаурусного расширения *(РЕМОНТ, КАПИТАЛЬНЫЙ РЕМОНТ, ТЕКУЩИЙ РЕМОНТ, РЕМОНТНО-СТРОИТЕЛЬНЫЕ РАБОТЫ*), второй элемент- дизъюнкция нескольких понятий тезауруса с расширением по видам, относящимся к жилью:

```
(     РЕМОНТ(N)
          OR   КАПИТАЛЬНЫЙ РЕМОНТ(N)
          OR   ТЕКУЩИЙ РЕМОНТ(N)
          OR   РЕМОНТНО-СТРОИТЕЛЬНЫЕ РАБОТЫ(N)  )
AND
         (     ЖИЛОЕ ЗДАНИЕ(L)
          OR   ЖИЛОЕ ПОМЕЩЕНИЕ(L)
          OR   КВАРТИРА(L)  )
```

F1-мера по метрике OR для этой рубрики составляет 0.658 (полнота – 0.71, точность – 0.61).

На выбранных для тестирования 19 рубриках система рубрикации показала наивысшие показатели классификации по F1-мере. По метрике AND, считающей релеватными рубрики документы с учетом мнений обоих асессоров, для оцененных документов величина F1 составила 0.44, что почти на 42% превышает результаты следующей по величине F1 системы рубрикации (0.31). По метрике OR (документ считается относящимся к рубрике, если хотя бы один из асессоров отнес его к данной рубрике) для оцененных документов величина F1 составила 0.72, что более чем на 56% превысило показатели следующей по качеству результатов системы (0.46).

Наиболее низкие по качеству результаты были показаны для рубрики 033 «Сад и огород» - мера F1 по метрике OR составила всего 0.32.

Это связано с тем, что в состав булевского выражения рубрики были введены понятия *ДАЧНЫЙ УЧАСТОК* и *ЗАГОРОДНАЯ ДАЧА*, без дополнительных условий конъюнкции. Это описание недостаточно точно, поскольку появление в тексте соответствующих слов еще не гарантирует то, что текст обсуждает проблемы сада и огорода.

На Рис.21.3 приведены результаты для дорожки классификации Веб-страниц коллекции ROMIP.BY. Достижение показателей качества рубрицирования при нестрогом согласии между экспертами (полнота = 81.7%; точность = 68.2%; F-мера = 72.9%) следует признать весьма успешным для 8 часов трудозатрат экспертов.

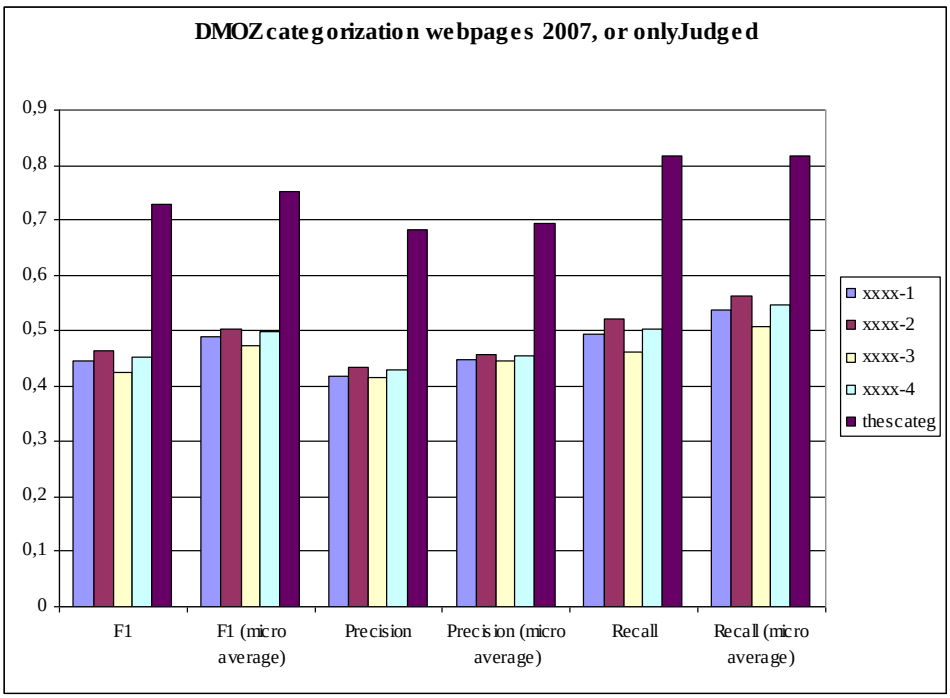

Рис.21.3 Результаты рубрикации веб-страниц на основе тезауруса РуТез в экспериментах РОМИП-2007

Из информации на семинаре РОМИП-2007 известно, что другие методы представляли собой модификации метода машинного обучения SVM – одного из самых успешных методов, применяемых в автоматической рубрикации текстов. Анализ данных коллекции показал, что проблемы методов машинного обучения связаны с серьезной противоречивостью коллекции, а именно, со следующими обстоятельствами. Как база для обучения были представлены данные ручной рубрикации сайтов. Однако внутри сайта могут оказаться достаточно разные по содержанию страницы, что и затрудняет выработку разделяющих правил методами машинного обучения (см. п.13.3.2).

В текущем эксперименте у нас не было возможности сделать предварительный прогон и исправить ошибки и неточности описания. Поэтому имеется очевидная возможность улучшения полученных результатов тематической классификации веб-страниц на основе тезаурусных знаний.

21.6. Тезаурус как база для методов машинного обучения в рубрикации.

Концептуальный индекс, построенный на основе тезауруса и учитывающий синонимы, многозначные слова, словосочетания, отношения между упомянутыми в тексте понятиями, может быть полезен и как база для методов машинного обучения.

В эксперименте на семинаре РОМИП 2004 (Агеев и др., 2004) было показано, что алгоритм машинного обучения SVM, настроенный на основе концептуального индекса, показал результаты, лучше, чем SVM на пословном индексе. Потенциально могут быть интересным и разные комбинации обоих индексов для достижения лучшей эффективности методов.

В том же эксперименте было показано, что лучший результат был получен на основе метода ПФА, который строит формулы описания рубрики в виде булевских формул фиксированной структуры на пословном или концептуальном индексе. В эксперименте РОМИП 2004 был использован концептуальный индекс.

Различные модификации алгоритма ПФА (Агеев и др., 2003) строят формулы вида:

$$\mathrm{U} = \bigcup_{i=1}^{k} \bigcap_{j=1}^{J_i} t_{i,j} \quad \text{(основной алгоритм)}$$

$$\mathrm{U} = \bigcup_{i=1}^{k} \left(\left(\bigcap_{j=1}^{J_i} t_{i,j} \right) \setminus \bigcup_{m=1}^{M_i} t'_{i,m} \right) \qquad (21.6)$$

$$\mathrm{U} = \left(\bigcup_{i=1}^{k} \bigcap_{j=1}^{J_i} t_{i,j} \right) \setminus \left(\bigcup_{i=1}^{k} \bigcap_{j=1}^{J'_i} t'_{i,j} \right)$$

где $t_{i,j}$, $t'_{i,j}$ — множества документов, содержащих некоторое понятие тезауруса (или, в общем случае, некоторый *терм* — элемент векторного представления документов). Конъюнкции, составляющие формулу, имеют длину J_i от 1 до 3.

В 2007 метод ПФА был использован для анализа качества обучающей коллекции задания по рубрикации РОМИП, было показано массовое наличие явно нерелевантных страниц в обучающей коллекции, что позволило спрогнозировать низкие результаты технологий машинного обучения и выполнить задание с помощью инженерного подхода (см. п. 22.5).

Чтобы пояснить характер, возникших проблем, рассмотрим рубрику №135 «Спорт -- Боевые искусства», для которой алгоритм строит следующую формулу с показателями на обучающем множестве (полнота = = 0.82, точность = 0.98, F-мера = 0.96):

```
(   [Тип = L | Имя = КАРАТЭ ])
OR (   {   [Тип = C | Имя = ХОККЕЙНЫЙ КЛУБ ]
        OR [Тип = T | Имя = ОХРАННОЕ ПРЕДПРИЯТИЕ ]}
    AND
        [Тип = T | Имя = БЕДСТВИЕ ])
OR (   {   [Тип = C | Имя = КУЛЬТУРА ]
        OR [Тип = C | Имя = СЕВЕРО-ЗАПАДНАЯ ЧАСТЬ ]}
    AND
        [Тип = C | Имя = ОДЕЖДА ]
    AND
        [Тип = T | Имя = ВЕРОВАТЬ ])
OR (   {   [Тип = C | Имя = МЕДИЦИНСКОЕ УЧРЕЖДЕНИЕ ]
        OR [Тип = C | Имя = КРЫЛАТСКОЕ ]}
   AND
        [Тип = T | Имя = ВОСТОЧНЫЕ ЕДИНОБОРСТВА ])
OR ( [Тип = C | Имя = МАСЛЕНИЦА ])
OR ( [Тип = L | Имя = ДЗЭНИН ])
OR ( [Тип = C | Имя = САМООБОРОНА ]
   AND [Тип = в дереве | Имя = ИСТОРИЧЕСКИЕ НАУКИ ])
```

Причины эффекта переобучения становятся понятны, если более внимательно посмотреть на список сайтов обучающего множества. Среди сайтов типа «www.karate.ru», «aikido.kuban.net», «capoeira.narod.ru» и т.п., встречаются также:

- tornado.spb.ru – не только «спортивный клуб таэквон-до», но и хоккейный клуб, а также охранные услуги и системы сигнализации;
- kryltkd.narod.ru – (Спортивный Клуб "Олимп" в Крылатском и Кунцево) не только «тхэквондо, кикбоксинг, самооборона», но и «развитие гибкости, ОФП».

В результате экспертного изучения качества получаемых формул был сделан вывод о трудности реализации метода машинного обучения без специальных мер по очистке обучающего множества и осуществлено описание рубрик посредством инженерного подхода.

Заключение к главе 21.

Тезаурус РуТез и создаваемый на его основе концептуальный индекс может применяться в обеих технологиях автоматической рубрикации текстов.

При инженерном подходе накопленная в тезаурусе информация дает возможность более быстрого и качественного описания содержания рубрик рубрикатора. Продемонстрированные в экспериментах результаты показывают, что некоторую значимую часть знаний о современной жизни общества и современном языке деловой прозы нам удалось описать и упорядочить в рамках понятийных структур тезауруса.

Кроме того, концептуальный индекс создает дополнительное признаковое пространство для каждого документа, что может позволить повысить качество методов машинного обучения, применяемых для рубрикации. Отношения между понятиями тезауруса могут применяться для снижения проблемы нехватки объема обучающей коллекции.

Глава 22. Общественно-политический тезаурус и автоматическое аннотирование

Как уже указывалось в п. 14.2.3, автоматическое аннотирование, то есть краткое изложение содержания, одного или нескольких документов является одним из важных направлений современных исследований в сфере информационного поиска.

В этой главе мы рассмотрим методы автоматического аннотирования одного или нескольких документов на базе тезаурусных знаний и построенного тематического представления документов.

Тематическое представление одного или группы тематически связанных документов позволяет нам решать типичные проблемы автоматического аннотирования, а именно, обеспечивать полноту представления информации, снижать повторы, обеспечивать связность и понятность аннотации.

22.1. Автоматическое аннотирование одного текста на основе тематического представления

При построении тематического представления текста в виде совокупностей близких по смыслу понятий, упоминаемых в тексте (тематических узлов), мы выявляем основных участников ситуации, описываемой в тексте. Так называемые основные тематические узлы моделируют главных участников описываемой ситуации. Суть текста составляет описание взаимодействия между главными участниками (см. раздел 19.2).

Таким образом, то новое и важное, что несет в себе текст и что должна отразить в себе аннотация, это именно то, каким образом взаимодействуют между собой эти главные участники. Отсюда следует первый принцип составления аннотаций: важными (информативными) и, следовательно, возможно включенными в аннотацию считаются те предложения текста, которые содержат, по крайней мере, два понятия, входящих в состав разных основных тематических узлов текста (Лукашевич, 1997; Loukachevitch, 1998). Напомним, что алгоритмы автоматического аннотирования на основе лексических цепочек и WordNet при извлечении предложений требуют присутствия одного элемента из основных лексических цепочек (см.п.14.2.3.3).

Предложений, содержащих понятия одних и тех же двух основных тематических узлов, в тексте может оказаться достаточно много. Для аннотации необходимо выделить одно предложение, в котором взаимодействие этих двух основных тематических узлов характеризуется “наилучшим образом”.

Не все основные участники начинают обсуждаться в тексте сразу, с первого предложения -- часть из них возникает в последующих предложениях. Чтобы сохранить связность и последовательность изложения текста, автор именно в этом первом предложении новой темы должен наиболее точно указать связь новой темы со всем предшествующим текстом. Таким образом, следуя за автором при вводе нового тематического элемента, можно повысить общую связность аннотации, то есть второй принцип составления аннотации отдельного документа состоит в том, что для каждой пары выявленных основных тематических элементов текста (основных тематических узлов) в аннотацию выбираются предложения, содержащие первое вхождение этой пары, следуя по порядку текста.

Нужно отметить, что при хорошем покрытии предметной области Тезаурусом появление в очередном предложении новой темы выявляется весьма точно, а это означает, что связность получаемой аннотации в среднем весьма высока,

Построение аннотации реализуется следующим образом:

1) Для построения аннотаций сначала формируется множество "аннотационных" фрагментов, которые не являются вопросительными или восклицательными предложениями.

2) Перед построением аннотации создается таблица всех возможных пар основных тематических узлов.
3) Начиная с начала текста, отбираются такие предложения, которые содержат еще не упоминавшуюся пару разных тематических узлов.

Серьезной проблемой автоматического аннотирования является проблема местоимений, которые могут появиться в выбранных предложениях и служить ссылкой на такие предложения текста, которые не вошли в состав аннотации.

В настоящее время в случаях, когда очередное предложение текста подходит для аннотации, но содержит местоимение, принимается одно из следующих решений:

1) если предыдущее предложение входит в состав аннотации, то и данное предложение включается в состав аннотации;
2) если предыдущее предложение не входит в состав аннотации, то проверяется, нельзя ли это предыдущее предложение включить в состав аннотации. Для этого необходимо, чтобы оно не содержало местоимений или следовало за предложением, включенным в аннотацию.
3) в остальных случаях предложение с местоимением не включается в состав аннотации.

Качество технологии автоматической рубрикации тестировалось на конференции SUMMAC (summarization conference) (Tipster SUMMAC, 1998). Программа использовала английский перевод Общественно-политического тезауруса.

Задача, в рамках которой тестировался изложенный метод автоматического аннотирования, состояла в следующем. Каждый участник соревнования получал на две недели 1000 документов и должен был представить две аннотации – аннотацию наилучшей длины (то есть оптимальную длину аннотации нужно было определить автоматически) и 10-процентную аннотацию, т.е. аннотацию, составляющую 10 процентов длины исходного текста.

Тестирование в процессе соревнования относилось к так называемому классу внешних тестирований (extrinsic), то есть проверялось, насколько порожденная аннотация пригодна для решения некоторой внешней задачи.

Внешней задачей в данном случае была задача рубрикации. Все документы, выданные для обработки, относились к двум большим темам «Мировая экономика» и «Налоги». При этом по полному тексту документа, его можно было отнести к более подробным рубрикам. Так, например, для рубрики «Мировая экономика» такими подрубриками были:

- экспорт в промышленности,
- внешняя торговля,
- международная борьба с наркотиками,
- иностранные производители автомобилей.Таким образом, если аннотация сделана правильно и сохраняет основную тему документа, люди-оценщики должны отнести аннотацию документа к той же подрубрике, что и сам документ. При этом каждому человеку-оценщику давался документ, который мог оказаться аннотацией, начальным фрагментов документа или полным текстом.

По ошибкам отнесения можно было оценить качество полученной аннотации.

Качество рубрикации документов по аннотации, и таким образом, собственно аннотаций оценивалось по стандартным метрикам, использующимся при оценивании систем автоматической рубрикации: точность, полнота и F-мера.

Наша система имела лучший показатель F-меры для аннотаций наилучшей длины и показатели 10-процентных аннотаций были лучше, чем средние (SUMMAC Final Report, 1998).

В качестве примера работы описанного метода автоматического аннотирования рассмотрим следующий текст:

Китай и Тайвань установили авиасообщение после 60-летнего перерыва

После почти 60-летнего перерыва открылось регулярное авиасообщение между Тайванем и материковым Китаем. Первый чартерный рейс с 250 пассажирами уже прибыл в столицу Тайваня из китайского города Гуанчжоу, передает «Би-би-си». Ожидается, что аэропорты острова будут принимать рейсы из пяти китайских городов: Пекина, Шанхая, Гуанчжоу, Сямэня и Нанкина. Договоренность о прямых регулярных авиарейсах была достигнута в середине июня 2008 года на переговорах между руководством Тайваня и Китая. Восстановление авиасообщения произошло не в последнюю очередь благодаря победе на выборах главы администрации Тайваня в марте 2008 года сторонников тесного сотрудничества с материковым Китаем. Прямых регулярных авиарейсов между Тайванем и Китаем не осуществлялось с 1949 года, когда Тайвань стал убежищем потерпевших поражение в гражданской войне с коммунистами сторонников партии Гоминьдан. До июля 2008 года прямые рейсы между материковым Китаем и Тайванем осуществлялись только по спецдоговоренности, в основном - в дни праздников, напоминает Лента.ру

Приведем примеры тематических узлов, созданных в процессе обработки этого текста (центр тематического узла выделен сдвигом влево; указана также частота упоминания понятия в тексте):

```
КИТАЙ                          8
   ГАНЧЖОУ                     2
   ШАНХАЙ                      2
   НАНКИН                      1
   ПЕКИН                       1

ТАЙВАНЬ                        7
   ТАЙБЕЙ                      1

АВИАЦИОННЫЕ ПЕРЕВОЗКИ          2
   АВИАРЕЙС                    1
   АЭРОПОРТ                    1

ПОЛИТИЧЕСКАЯ ПАРТИЯ            1
   КОММУНИСТ                   1

ПРАВИТЕЛЬСТВО                  1
   ПУБЛИЧНАЯ ВЛАСТЬ            1
```

В рассматриваемом примере тематического представления основными тематическими узлами стали узлы с главными дескрипторами *КИТАЙ, ТАЙВАНЬ, АВИАЦИОННЫЕ ПЕРЕВОЗКИ, ГОРОД, РЕЙС, БИ-БИ-СИ.*

Для текста примера получаем следующую аннотацию, в которой упомянуты все основные тематические узлы данного документа:

Китай и Тайвань установили авиасообщение после 60-летнего перерыва

Первый чартерный рейс с 250 пассажирами уже прибыл в столицу Тайваня из китайского города Гуанчжоу, передает Би-би-си. Ожидается, что аэропорты острова будут принимать рейсы из пяти китайских городов Пекина, Шанхая, Гуанчжоу, Сямэня и Нанкина.

Отметим, что в аннотации пропущено первое предложение, которое не содержит новой пары тематических узлов по сравнению с заголовком текста.

22.2. Построение структурной тематической аннотации текста

Для некоторых типов текстов хорошая (связная и понятная) аннотация может быть построена не всегда:

- большинство таких документов как законы, президентские и правительственные документы, международные договоры имеют очень сложную структуру, поэтому часто аннотация, основанная на любых принципах выбора информативных фрагментов, может быть слишком длинна, тяжеловесна и неясна;
- автоматические аннотации газетных интервью обычно обрывочны и несвязны;
- в ограниченные по длине аннотации текстов больших размеров могут не уместиться важные темы текста;
- и, наконец, аннотации на исходном языке могут оказаться бесполезными для пользователей многоязычных информационно-поисковых систем. Пользователи таких систем могут быть незнакомы с языком, на котором написаны документы коллекции.

Для краткого представления содержания вышеперечисленных типов текстов требуются другие виды аннотаций (Лукашевич, Добров, 1998; Loukachevitch, Dobrov, 2000a). Такой аннотацией мог бы служить список наиболее частотных терминов текста. Но в таких списках термины, относящиеся к различным подтемам текста, перемешаны, что затрудняет их восприятие. Кроме того, большое количество терминов наиболее представительной темы может занять все предоставленное место, и термины других важных для текста тем будут упущены.

Структурная тематическая аннотация представляет содержание текста посредством описания участников его основной темы, которые моделируются совокупностью понятий, относящихся к этим темам. Структурная тематическая аннотация содержит наиболее информативные фрагменты тематического представления текста, которое включает все понятия текста, разбитые на тематические узлы.

Тематическое представление, содержащее все понятия текста, является слишком подробным, чтобы служить структурной аннотацией текста, предъявляемой пользователю. Поэтому в качестве структурной аннотации может быть использованы выделенные основные тематические узлы тематического представления. Поскольку понятия Общественно-политического тезауруса переведены на английский язык, структурная тематическая аннотация может быть создана для русскоязычного или англоязычного текста и представлена на русском или английском языках.

В значительной мере мы используем структурную тематическую аннотацию для выверки Тезауруса. Сопоставив полученное тематическое представление с текстом, можно оценить, соответствуют ли основные тематические узлы, построенные для данного текста, основной теме текста. Существенные расхождения могут быть связаны с недостаточной точностью описания терминов текста в Тезаурусе, например термин вообще не описан в Тезаурусе; термин описан в Тезаурусе в другом значении; термин включен не в тот синонимический ряд; в тезаурусной проекции текста соответствующий термину дескриптор имеет неверные связи с другими дескрипторами; и тому подобное .

Структурная тематическая аннотация включает в себя следующие части:

- понятия основных тематических узлов, упорядоченных в порядке убывания частотности и расположенных горизонтально;
- отметки об относительно суммированной частотности основных тематических узлов, обозначаемые различным количеством символов “*”;
- отметки об относительной силе взаимоотношений между различными тематическими узлами
  - ”X“-- очень сильное отношение;
  - ”z”-- сильное отношение;

– "." -- отношение.

В качестве примера рассмотрим (Рис.22.1) структурную тематическую аннотацию Федерального закона об информации, информатизации и защите информации Российской Федерации (40 Кб, 164 различных термина).

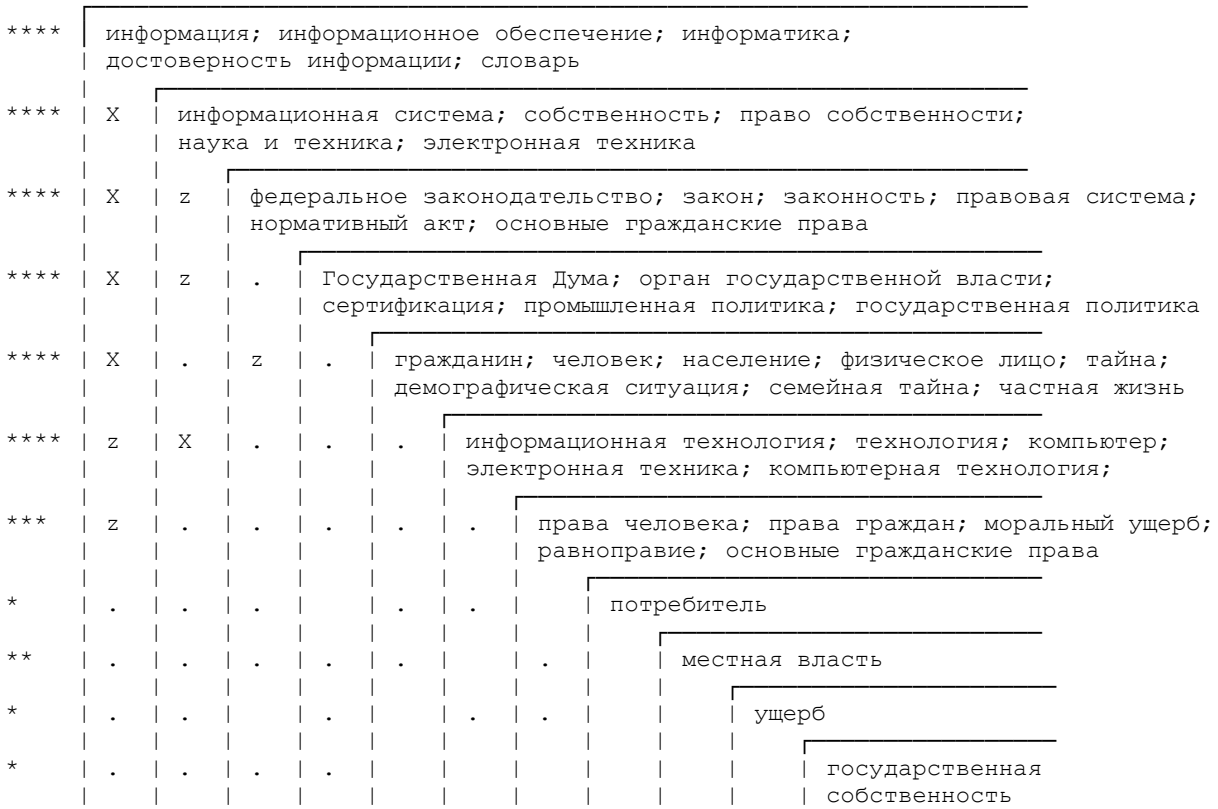

Рис. 22.1

Структурная тематическая аннотация, представленная на Рис.23.2, отражает содержание англоязычного текста (Рис.22.3) - рабочего документа 105-го Конгресса США.

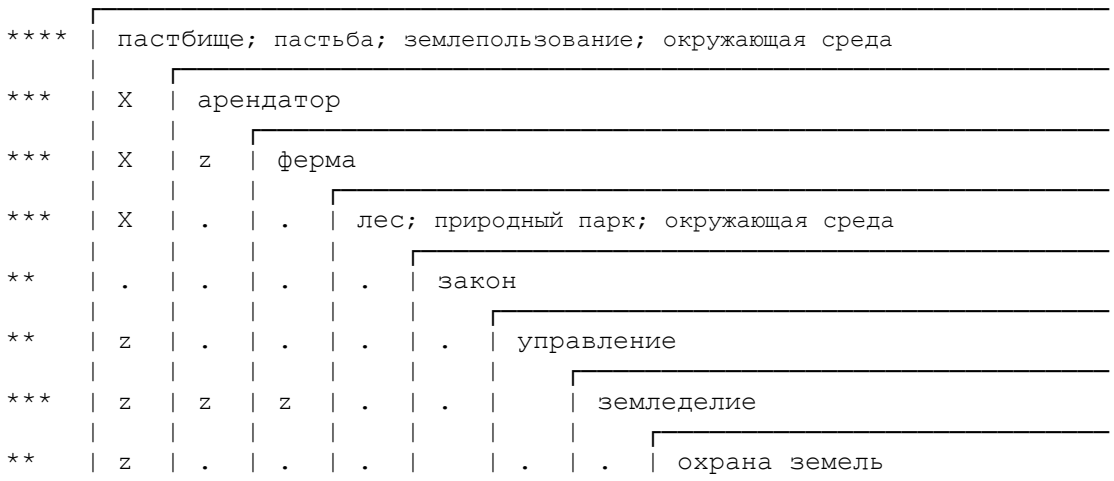

Рис.22.2

"105th CONGRESS
1st Session

S. 749

To provide for more effective management of the National Grasslands, and for other purposes.

...

Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled,

SECTION 1. SHORT TITLE.

This Act may be cited as the ``National Grasslands Management Act of 1997''.

SEC. 2. FINDINGS AND PURPOSE.

(a) Findings.--The Congress finds that--

(1) the inclusion of the National Grasslands within the National Forest System has prevented the Secretary of Agriculture from effectively administering and promoting grassland agriculture on National Grasslands as originally intended under the Bankhead-Jones Farm Tenant Act;

(2) the National Grasslands can be more effectively managed by the Secretary of Agriculture if administered as a separate entity outside of the National Forest System; and

(3) a grazing program on National Grasslands can be responsibly carried out while protecting and preserving sporting, recreational, environmental, and other multiple uses of the National Grasslands.

(b) Purpose.--The purpose of this Act is to provide for improved management and more efficient administration of grazing activities on National Grasslands while preserving and protecting multiple uses of such lands, including but not limited to preserving sportsmen's hunting and fishing and

Структурная аннотация позволяет оценивать содержание текста с одного взгляда, в том числе из-за неявно осуществляемых пользователем-человеком мысленных "связок" между темами.

22.3. Построение аннотации для новостного кластера на основе тематического представления текстов кластера

Современные технологии обработки новостных потоков обычно включают в себя краткое представление содержания новостного кластера в виде аннотации (обзорного реферата). В данном разделе мы рассмотрим автоматический метод создания аннотации новостного кластера на основе тематического представления, построенного для этого кластера.

22.3.1. Построение тематического представления для новостного кластера

Новостной кластер представляет собой совокупность тематически близких документов. Поэтому тематическую структуру новостного кластера так же, как и отдельного элемента можно выявить за счет построения тематического представления этого кластера, и это представление можно будет использовать для управления набором предложений в аннотацию кластера, а именно для решения таких задач как обеспечение полноты, снижения повторов, а также обеспечения связности аннотации кластера.

Построение тематического представления новостного кластера осуществляется простым способом: все тексты кластера склеиваются в единый текст, для которого производится стандартный тематический анализ одного документа и строится тематическое представление.

Результат этой процедуры, а затем и результат построения аннотации в некоторой степени зависит от порядка просмотра документов в кластере. Мы используем следующий метод объединения документов кластера в единый текст, используемый для построения аннотации.

Сначала в новостном кластере определяется «центр кластера» - документ, наиболее близкий к центру тяжести множества документов кластеров в метрическом пространстве нормализованных лемматическом и концептуальном (по тезаурусу) индексов. Определяется «ядро» кластера – документы достаточно близкие к центру (по некоторому порогу). Затем «центр кластера» сдвигается в документ из ядра кластера, который был опубликован последним по времени. Пересчитываются веса связей документов кластера к новому центру. С учетом задаваемого интервала времени по убыванию веса сначала заполняются документы за последнее время, затем все остальные. Так как отбирается всего несколько предложений, то имеется общее ограничение на количество отбираемых в «единый документ» документов.

После порождения «единого документа» кластера для него строится тематическое представление. Так, для кластера, в который входит текст примера из раздела 22.1, основными тематическими узлами становятся следующие совокупности понятий (справа указана частотность понятия в кластере):

КИТАЙ	103
ПЕКИН	21
ГУАНЧЖОУ	13
ГОСУДАРСТВО	9
ЮАНЬ	7
ШАНХАЙ	6
КИТАЙЦЫ	5
НАНКИН	5
ГУАНДУН	1
ТАЙВАНЬ	103
ТАЙБЕЙ	21
АВИАЦИОННЫЕ ПЕРЕВОЗКИ	33
АВИАЦИОННАЯ КОМПАНИЯ	9
САМОЛЕТ	9
АВИАРЕЙС	7
ТРАНСПОРТНАЯ СФЕРА	4
АЭРОПОРТ	3
ТРАНСПОРТНЫЕ ПЕРЕВОЗКИ	2
АЭРОБУС	1
АВИАЛИНИЯ	1
ГОРОД	17
ТЕРРИТОРИЯ, УЧАСТОК	3
НАСЕЛЕННЫЙ ПУНКТ	1
ОСТРОВ	17
ЖИТЕЛЬ ОСТРОВА	1
ЧАРТЕРНЫЕ ПЕРЕВОЗКИ	14
ТУРИСТ	12
ЧЕЛОВЕК	62
ТУРИЗМ	2

```
ПОЕЗДКА                          1
ПАССАЖИР                        10
ПРАВИТЕЛЬСТВО                    6
   РУКОВОДИТЕЛЬ                  6
   ОРГАН ПУБЛИЧНОЙ ВЛАСТИ        3
   РУКОВОДСТВО                   2
   ОРГАН ИСПОЛНИТЕЛЬНОЙ ВЛАСТИ   2
   ПУБЛИЧНАЯ ВЛАСТЬ              1
```

Таким образом, по основным тематическим узлам тематического представления могут быть определены основные элементы, обсуждаемой в кластере темы.

Как видно, тематические узлы включают концепты достаточно разной частотности. Низкочастотные концепты тематического узла могут быть ошибочно включены в тематический узел, кроме того, представительность ими основной темы документа невелика. Поэтому можно задать выделение ядра тематических узлов, которое определяется как коэффициент от 0 до 1. Этот коэффициент определяет, какая доля наиболее частотных понятий от общей частотности понятий в тематическом узле будет включена в ядро.

Так, при значении коэффициента тематического ядра 0.7 получим следующие ядра тематических узлов:

```
КИТАЙ                          103
   ПЕКИН                        21
   ГУАНЧЖОУ                     13
ТАЙВАНЬ                        103
АВИАЦИОННЫЕ ПЕРЕВОЗКИ           33
   АВИАЦИОННАЯ КОМПАНИЯ          9
   САМОЛЕТ                       9
ГОРОД                           17
ОСТРОВ                          17
ЧАРТЕРНЫЕ ПЕРЕВОЗКИ             14
ТУРИСТ                          12
   ЧЕЛОВЕК                      62
ПАССАЖИР                        10
ПРАВИТЕЛЬСТВО                    6
   РУКОВОДИТЕЛЬ                  6
   ОРГАН ПУБЛИЧНОЙ ВЛАСТИ        3
```

22.3.2. Метод построения аннотации новостного кластера по тематическому представлению кластера

Аннотация новостного кластера обычно состоит из заголовка и нескольких предложений из разных документов новостного кластера.

Зная ядра тематических узлов, полноту изложения содержания кластера мы обеспечиваем тем, что должны отбирать для аннотации предложения, содержащие пары этих тематических узлов – именно тогда эти предложения будут описывать взаимоотношения между основными тематическими элементами кластера.

При отборе заголовка для аннотации ищется заголовок, содержащий пару наиболее частотных тематических узлов. Если таких заголовков нет, то ищутся заголовки, содержащие понятия из одного наиболее частотного тематического узла.

Для выбора очередного предложения в списке основных тематических узлов отмечаются все тематические узлы, которые уже были упомянуты. Очередное

предложение должно содержать пару основных тематических узлов: наиболее частотный тематический узел, который еще не упоминался, и какой-нибудь еще основной тематический узел.

Для обеспечения связности требуется, чтобы очередное предложение содержало либо уже упомянутый тематический узел, либо уже упоминавшееся слово с большой буквы.

Кроме того, делается ряд дополнительных проверок:

- предложение не должно являться вопросительным или отрицательным предложением,
- предложение не должно содержать в заданном числе первых слов местоимение,
- начало предложения не должно совпадать с началами заголовка и предложений, уже взятых в аннотацию,
- число слов предложения, совпадающего со словами предшествующих предложений не должно превышать некоторой доли длины предложения.

Понятно, что даже при проверке вышеупомянутых условий может найтись еще достаточно много подходящих предложений-кандидатов. Кроме того, оценка предложений на основе понятий тезауруса не является достаточной без учета упоминаемых именованных сущностей, которые могут быть и не описаны в тезаурусе.

Поэтому вводится еще и общая оценка предложения с помощью вычисления веса предложения, которая складывается из двух компонентов: весов упомянутых понятий Тезауруса, которые были получены в тематическом представлении, а также весов содержащихся в предложении слов с большой буквы, не считая первого слова предложения.

Для вычисления весов слов с большой буквы (далее Слов), сначала вычисляется вес самого частотного Слова W_{max_word} в документе кластера:

$$W_{max\_word} = \min (1{,}0 , W_{max\_conc} * (Fr_{max\_word} / Fr_{max\_conc}))$$

где W_{max_conc} – максимальный вес понятия тезауруса в тематическом представлении, Fr_{max_conc} – частотность в тексте понятия тезауруса с максимальным весом, Fr_{max_word} – частотность самого частотного Слова.

Остальные веса Слов (W_{word}) вычисляются пропорционально их частотности:

$$W_{word} = W_{max\_word} * (Fr_{word} / Fr_{max\_word})$$

Так мы сводим веса понятий и слов к одной шкале.

Просмотр предложений-кандидатов начинается с начала документа кластера, то есть предложения набираются сначала из главного документа кластера и наиболее близких к нему по содержанию. Каждое следующее предложение берется из другого документа.

Для кластера примера была получена следующая аннотация (в скобках указан источник новости и время публикации):

Предложения	Тематические узлы
Китай и Тайвань установили авиасообщение после 60-летнего перерыва *(Новые Известия - лента новостей, 04.07.2008 11:08:45)*	*КИТАЙ, ТАЙВАНЬ, АВИАЦИОННЫЕ ПЕРЕВОЗКИ (авиасообщение)*
Первый чартерный рейс с 250 пассажирами уже прибыл в столицу Тайваня из китайского города Гуанчжоу. (Lenta.ru - главные новости, 04.07.2008 9:47:25)	*КИТАЙ, ЧАРТЕРНЫЕ ПЕРЕВОЗКИ (чартерный рейс), ГОРОД, ПАССАЖИР*
С 4 июля самолеты с материкового Китая на остров	*КИТАЙ, ТАЙВАНЬ,*

Предложения	Тематические узлы
Тайвань и обратно будут летать каждую неделю с пятницы по понедельник. (РегKURSCITY.RU - Курс, 04.07.2008 9:35:34)	*АВИАЦИОННЫЕ ПЕРЕВОЗКИ (самолет), ОСТРОВ*
Перед прибывающими в ближайшие выходные 600 туристами из Китая будет расстилаться красная ковровая дорожка. (BBCRussian.com (Главная), 04.07.2008 1:18:25)	*КИТАЙ, ТУРИСТ*
По завершении в 1949 году гражданской войны в Китае и изгнания правительства Гом-Инь-Дана на Тайвань, отношения между двумя сторонами Тайваньского пролива были заморожены. (РегЛПГАБізнесІнформ - Украины - Новости за рубежом, 04.07.2008 9:14:00)	*КИТАЙ, ТАЙВАНЬ, ПРАВИТЕЛЬСТВО*

В заголовке аннотации мы имеем три основных тематических узла: *КИТАЙ, ТАЙВАНЬ, АВИАЦИОННЫЕ ПЕРЕВОЗКИ:*

- в первом предложении сообщается о конкретных городах, связанных с авиаперевозками, и указывается о том, что перевозки чартерные – таким образом, упомянуты еще два тематических узла – *ГОРОД* и *ЧАРТЕРНЫЕ ПЕРЕВОЗКИ;*
- второе предложение содержит новый тематический узел *ОСТРОВ;*
- третье предложение содержит узел *ТУРИСТ;*
- четвертое предложение содержит тематический узел *ПРАВИТЕЛЬСТВО*

Таким образом, каждое предложение содержит не менее двух разных основных тематических узлов, один из которых новый (выделен подчеркиванием в правом столбце таблицы), а другой был упомянут ранее.

22.3.3. Тестирование предложенной модели аннотации новостного кластера

Предлагая метод аннотирования новостного кластера, мы сделали несколько предположений о внутренней структуре аннотации и о нашей способности выявлять эту структуру на основе создаваемого автоматически тематического представления. Для проверки предложенной модели аннотации новостного кластера был проведён эксперимент по проверке соответствия сделанных предположений ручным аннотациям, составленными экспертами-лингвистами.

Лингвисты создали несколько аннотаций новостных кластеров из предложений этого кластера. Аннотация представляла собой заголовок и четыре предложения. Общее количество разных аннотаций в эксперименте – 13. Для новостных кластеров были получены их тематические представления. Далее ручные аннотации были размечены на предмет наличия основных тематических узлов для данного кластера и именованных сущностей.

Задачей данной разметки являлась проверка описанных выше условий для составления аннотаций, а именно:

1. Действительно ли реальные аннотации должны содержать в себе как минимум два основных тематических узла из тематического представления текста и/или именованные сущности.
2. Используются ли в ручных аннотациях понятия-элементы основных тематических узлов и именованные сущности для организации лексической связности текста, а именно, повторяются ли в последующих предложениях ручных аннотаций понятия уже упомянутых основных тематических узлов или уже упомянутые именованные сущности.

3. Содержат ли очередные предложения элементы новизны в виде нового, еще не упоминавшегося тематического узла или именованной сущности.

Результаты эксперимента представлены в таблице 22.1.

Проверка представленности основных тематических узлов:	
Всего предложений:	**65**
Из них количество предложений с не менее чем двумя тематическими узлами:	**60**
Количество предложений, в которых есть один основной тематический узел и не менее чем одна именованная сущность:	**58**
Оценка связности и новизны:	
Общее количество предложений, не считая первые предложения:	**52**
Количество предложений с новым основным тематическим узлом:	**35**
Количество предложений с новым именем:	**28**
Количество предложений с повтором упоминавшегося тематического узла:	**46**
Количество предложений с повтором упоминавшегося имени:	**38**

Таблица 22.1. Выявление основных тематических узлов и именованных сущностей в ручных аннотациях

Результатом проведённого анализа явился тот факт, что 83% предложений реальных ручных аннотаций (от общего числа предложений), сделанных экспертами-лингвистами, удовлетворяют сделанным предположениям. Особенность оставшихся 17% предложений состоит в том, что все они являлись последними предложениями ручной аннотации. Такая ситуация связана с тем, что основная тема новостного кластера уже изложена, и дальнейшее описание событий «разрывается» на второстепенные темы документы, которых обычно имеется большое количество).

Проведенный эксперимент доказывает, что сделанные предположения в методе автоматического аннотирования новостных кластеров имеют высокую корреляцию со структурой человеческих аннотаций.

22.3.4. Оценка качества аннотаций новостных кластеров

Как мы упоминали в разделе 14.2.3.2, тестирование качества автоматических аннотаций является сложной процедурой. В качестве метрики аннотаций новостных кластеров, позволяющая автоматизировать этот процесс, используется такая метрика как ROUGE, которая подсчитывает число перекрытия (n-граммы слов) автоматической аннотации с «идеальными» аннотациями, составленными людьми (Lin, 2004).

Другой используемой мерой оценки качества аннотаций является Метод Пирамид, который основан на ручном выделении экспертами «информационных единиц» из эталонных аннотаций - Summary Content Units (SCUs) и вычислении процентной доли этих единиц, упомянутых в автоматических аннотациях (Harnly и др., 2005).

Далее рассмотрим подробнее результаты применения этих методов оценки для тестирования наших аннотаций новостных кластеров. Кроме того, будет рассмотрена процедура применения ручных оценок.

22.3.4.1. Тестирование аннотаций новостных кластеров методом ROUGE

Поскольку в разных статьях, описывающих эту метрику, содержатся несколько разные способы ее вычисления, то конкретные используемые нами формулы мы назвали ROUGE-1-cir и ROUGE-2-cir (Лукашевич, Добров 2009) и вычисляли их следующим образом:

$$ROUGE-N-cir(A_i) = \frac{\sum_{M_{ij}} count(Ngram(A_i) \cap Ngram(M_{ij}))}{\sum_{M_{ij}} count(Ngram(M_{ij})} ,$$

где A_i – оцениваемая обзорная аннотация i-того кластера, M_{ij} – ручные аннотации i-того кластера, *Ngram(D)* – множество всех n-грамм из лемм соответствующего документа D. При сравнении отдельных документов в расчет берутся только уникальные n-граммы, присутствующие в обоих документах - не поощряется многократный повтор одного и того же предложения. При рассмотрении нескольких аннотаций, наоборот, повторение одинаковых элементов поощряется. Биграммы в наших оценках учитывались с перестановками.

Для оценки качества построенных аннотаций мы воспользовались данными, любезно предоставленными С.Д. Тарасовым (Военмех, Спб.). В проведенных С.Д. Тарасовым экспериментах группе студентов было предложено построить ручную аннотацию для новостных кластеров, которые брались из системы Google.Новости в период с 01 по 05 декабря 2008 года. Ручная аннотация должна была быть составлена из четырех предложений. Ограничений на выбор предложений из разных текстов не накладывалось.

Мы выбрали достаточно случайным образом из полученных данных 15 новостных кластеров разной тематики, включая новости о погоде, спорте, финансах и политике, для которых имелось от 18 до 40 ручных аннотаций (всего 462).

В качестве «базовой оценки», следуя (Dang, 2006), мы рассматривали следующие варианты искусственных аннотаций:

- первый документ кластера;
- заголовки первых четырех документов;
- первые предложения первых четырех документов;
- последний документ кластера.

В качестве автоматической аннотации рассматривались аннотации из заголовка и трех предложений, взятых из разных текстов.

Мы получили следующие результаты (в таблице приведены результаты для разных параметров ядра кластера – см.п. 23.3.2) :

Вид аннотации	ROUGE-1-cir	ROUGE-2-cir
первый документ кластера	0,219	0,083
заголовки первых четырех документов	0,162	0,056
первые предложения первых 4 документов	0,269	0,107
последний документ кластера	0,278	0,168
автоматическая аннотация с ядром 0,20	0,331	0,150
автоматическая аннотация с ядром 0,40	0,328	0,140

Следует отметить, что некоторые ручные аннотации совпадали с первым или последним документом кластера. Определенным недостатком используемых данных является то, что некоторые кластеры содержали документы за несколько дней, поэтому ручные аннотации чаще содержали предложения из последних документов кластера.

Существует определенная критика использования метрик ROUGE для оценки качества аннотирования. Метрика чувствительна к длинам сравниваемых документов, не

учитывает связность аннотаций. В целом, существует большое разнообразие между ручными аннотациями разных экспертов. В нашем случае нам лишь важно было оценить близость построенных автоматических и ручных аннотаций для оценки перспективности предложенного подхода.

22.3.4.2. Тестирование аннотаций новостных кластеров Методом Пирамид

Метод Пирамид основан на выделении в аннотациях отдельных единиц получаемой информации (SCU) (Harnly и др., 2005). Выделенная информационная единиц» получает вес, равный количеству ручных эталонных аннотаций, где она встречается. Название «Метод пирамид» как раз и связано с тем, что информационные единицы SCU выстраиваются как бы в пирамиду: на вершине небольшое число единиц с большим весом, внизу пирамиды - большое число информационных единиц с маленьким весом. Общая оценка автоматической аннотации складывается из суммы весов SCU, которые она содержит, по отношению к общему количеству SCU для данного текста:

$$\frac{[Суммарный\_вес\_найденных\_SCU]}{[Суммарный\_вес\_всех\_SCU\_для\_данного\_топика]}$$

В качестве примера SCU и её вхождений в тексты аннотаций можно рассмотреть следующие фрагменты предложений новостного кластера:

SCU: Мини-субмарина попала в ловушку под водой.

1. *мини-субмарина... была затоплена... на дне моря...*
2. *маленькая... субмарина... затоплена... на глубине 625 футов.*
3. *мини-субмарина попала в ловушку... ниже уровня моря.*
4. *маленькая... субмарина... затоплена... на дне морском...*

Для сравнения качества предложенного метода аннотирования новостных кластеров в терминах информационных был реализован известный метод аннотирования Maximal Marginal Relevance (MMR) (Carbonell, Goldstein, 1998), показавший высокое качество аннотирования на конференции SUMMAC, а его модификации и на более поздних конференциях. Метод MMR – это итеративный алгоритм выбора предложений в аннотацию. Пусть имеются:

- Q – запрос для аннотирования или в нашем случае общего тематического аннотирования – вектор слов всего кластера,
- S – множество предложений кандидатов,
- s – рассматриваемое предложение кандидат,
- E – множество выбранных предложений.

Тогда на каждой итерации предложение в итоговую аннотацию будет отбираться в соответствии с формулой:

$$MMR = \arg\max_{s \in S} \left[\lambda \cdot Sim_1(s, Q) - (1 - \lambda) \cdot \max_{s_j \in E} Sim_2(s, s_j) \right]$$

Предложения итоговой аннотации сортируются в соответствии с их порядком следования в исходном документе.

Для предложенного нами метода аннотирования новостного кластера и метода MMR была применена пирамидная оценка. Сравнивались аннотации длиной 100 слов. Наш метод аннотирования получил среднюю оценку 0.638, метод MMR - 0.643. Таким образом, по полноте изложения информации предложенный нами метод не показал лучшие результаты. На наш взгляд, это частично связано с тем, что для обеспечения

лучшей связности аннотации требуется некоторая степень повторяемости в предложениях.

22.3.4.3. Оценка связности аннотаций новостных кластеров

Тестирование связности и читабельности автоматических аннотаций может производиться только человеком. Была применена следующая процедура: лингвист должна была читать каждый вид аннотации последовательно от предложения к предложению, и каждому предложению выставить некоторый штрафной балл:

0.0 – если предложение «хорошее» (связано с остальными предложениями, качественно вписывается в данную аннотацию и т.д.),

1.0 – если предложение «плохое» (не связано с другими предложениями, является лишним в данной аннотации и т.д.)

0.5 – в спорных ситуациях.

Таким образом, каждый вид аннотации получил некоторую совокупность штрафных баллов, чем меньше баллов, тем лучше. В среднем аннотации, порожденные методом MMR, получили 0.7 штрафных баллов, нашим методом – 0.3 балла (Алексеев, Лукашевич, 2010).

Заключение к главе 22.

Описанные методы аннотирования отдельного документа и новостного кластера на основе тематического представления позволяют решать такие проблемы методов автоматического аннотирования как обеспечение полноты представления содержания, снижения повторов, обеспечения связности аннотации.

Основная суть предложенных методов автоматического аннотирования заключается в выявлении основных участников обсуждаемой в тексте или кластере ситуации и в предположении, что наиболее информативными являются предложения, в которых сообщается информация о взаимодействии этих сущностей.

Полнота передачи содержания документа (документов) обеспечивается тем, что отбираются предложения, упоминающие основных участников ситуации. Снижение повторов становится возможным, поскольку один и тот же участник ситуации может быть распознан в значительном разнообразии текстовых выражений. Кроме того, снижение повторов обеспечивается обязательным упоминанием нового, еще не упомянутого элемента тематического представления в очередном предложении аннотации. Наконец, связность аннотации обеспечивается повторяемостью тематических узлов и именованных сущностей.

Выявленные закономерности построения аннотаций новостных кластеров не обязательно требуют наличие тезауруса. Нахождение основных участников ситуации может быть смоделировано на основе совершенно других нетезаурусных методов обработки текстов, а фактор необходимости упоминания в предложениях аннотации, по крайней мере, двух основных участников может быть добавлен как фактор в совокупность учитываемых факторов, таких как вес предложения, сходство с заголовком, позиционное расположение и др.

Описанный метод построения обзорных рефератов позволяет в широких пределах варьировать представление кластера при сохранении уровня отображения содержания и связности. Можно задавать как количество документов (исходящих ссылок), отражаемых в аннотации, так и количество предложений из каждого документа. В частности, могут быть смоделированы аннотации, формируемые в ресурсе Яндекс.Новости (до трех-четырех документов по одному-два предложения), или аннотации, формируемые в ресурсе Google.Новости (три-четыре предложения из одного документа и два заголовка из других документов), или Рамблер.Новости (три предложения из одного документа и два-три предложения из других документов).

В случаях когда основной аннотацией кластера служит аннотация отдельного документа, сначала порождаются автоматические аннотации отдельных документов кластера, отбирается лучшая такая аннотация по признаку наибольшего покрытия тематического представления кластера. Дополняющие предложения из других документов новостного кластера выбираются с использованием описанного метода аннотирования новостного кластера.

ЧАСТЬ 6. РАЗВИТИЕ ТЕЗАУРУСА РУТЕЗ И РЕСУРСЫ, ОСНОВАННЫЕ НА ТЕЗАУРУСЕ РУТЕЗ

Глава 23. Развитие и пополнение тезауруса РуТез

23.1. Этапы развития тезауруса РуТез

Развитие тезауруса РуТез началось в 1994 году с разработки Общественно-политического тезауруса (Лукашевич, Салий, 1996; Лукашевич, Салий, 1997). Основой создания Общественно-политического тезауруса стали автоматически извлекаемые из нормативных документов Российской Федерации слова и терминоподобные словосочетания (Лукашевич, 1995). Извлеченные слова и терминоподобные словосочетания просматривались людьми, которые принимали решение о включении или невключении данных выражений в тезаурус, об их статусе (образование нового понятия или включение в существующий синонимический ряд), проставляли отношения между понятиями тезауруса.

Процедура автоматизированного извлечения терминоподобных словосочетаний проработала около 4 лет до тех пор, пока эффективность ее не стала слишком низкой**,** поскольку число выявленных новых терминов на тысячу просмотренных словосочетаний резко сократилось.

В 1995-1996 заработали первые приложения на базе создаваемого Общественно-политического тезауруса: автоматическое разрешение лексической многозначности, моделирование лексической связности посредством автоматической группировки близких по смыслу терминов, упоминаемых в текстах, выявлялись основные понятия текста в процессе автоматического построения тематического представления, была реализована процедура автоматической рубрикации с выводом рубрики по отношениям тезауруса.

Эти процедуры стали серьезной проверкой для наполнения терминологического наполнения тезауруса, качества описания отношений. Выяснилось, что нужно исследовать принципы описания отношений в тезаурусе, на основе которых можно было бы обеспечить качественную реализацию их разнообразных функций в рамках процедур автоматической обработки текстов.

В 1996 встал вопрос о том, можно ли использовать тот же тезаурус, сформированный на базе нормативных документов, для автоматической обработки газетных статей и новостных сообщений. Выяснилось, что и нормативные акты, и сообщения средств массовой информации могут быть отнесены к одной и той же широкой области современных общественных отношений, только различается язык написания этих текстов и существенен разный уровень детализации для разных подобластей. Таким образом, Общественно-политический тезаурус стал использоваться для автоматической обработки сообщений СМИ и соответственно стал пополняться и уточняться на базе анализа результатов работы этих автоматических процедур.

В течение первых лет своего развития Общественно-политический тезаурус пополнялся, в основном, терминологией из разных областей общественной жизни и тематической лексикой, с 1997 года в те же тезаурусные структуры, на основе толкований толковых словарей, стали представляться значения слов и выражений, которые могут употребляться в разных предметных областях – возникло то, что теперь называется Общим лексиконом. Также на основе толковых словарей выявлялась и дополнялась тематическая лексика, расширялось покрытие Общественно-политического тезауруса. Собственно с этого момента Общественно-политический тезаурус стал перерастать в тезаурус русского языка РуТез (Лукашевич 1999, Лукашевич, Добров, 2002).

В 2000 году на основе тезауруса РуТез была реализована автоматическая рубрикация нормативных актов Российской Федерации по так называемому Президентскому классификатору, содержащему более 1000 рубрик. Для этой процедуры понадобилось расширить Общественно-политический тезаурус в некоторые профессиональные подобласти сферы общественных отношений, такие как налогообложение, таможенное дело, бухгалтерский учет, международные отношения и другие.

С 2001 года Общественно-политический тезаурус переводится на английский язык. Перевод позволил увидеть сделанные описания с точки зрения другого языка, что привело к значительному уточнению некоторых описаний. Например, выяснилось, что некоторые значения многозначных слов, которые, казалось, можно представить как одно понятие тезауруса, на самом деле имеют два совершенно отличных друг от друга английских перевода. Кроме того, естественно, нашлись фрагменты тезауруса, плохо переводимые на английский язык, и были выполнены исследования по поводу того, как можно сделать представление менее зависимым от конкретного языка, то есть было сделано движение от тезауруса конкретного языка к лингвистической онтологии.

В настоящее время онтологический комплекс Общественно-политический тезаурус- Тезаурус РуТез продолжает уточняться и пополняться.

Имеется несколько основных источников уточнений и пополнения тезауруса.

Во-первых, уточнение и пополнение тезаурусных описаний происходят в процессе анализа результатов работы тезауруса в реальных приложениях автоматической обработки текстов, информационного поиска, проектов, выполняемых в конкретных предметных областях широкой общественно-политической сферы.

Во-вторых, тезаурус пополняется за счет анализа упорядоченного по частотности списка лемм, полученных по документам Университетской информационной системы РОССИЯ.

В-третьих, тезаурус уточняется и пополняется за счет анализа упорядоченного по частотности списка англоязычных словоформ, полученных для коллекций газетных статей Glasgo Gerald и Los Andgeles Times, полученных в процессе участия в конференции по многоязычному поиску CLEF.

Кроме того, по модели тезауруса РуТез создаются другие лингвистические онтологии. Развитие одного из таких ресурсов Онтологии по естественным наукам и технологиям будет рассмотрено в следующей главе.

В июле 2009 года тезаурус РуТез имеет следующий объем:

- 51.5 тысяч понятий,
- 141.7 тысяч разных русскоязычных слов и словосочетаний,
- 159.5 тысяч отношений понятие – русское языковое выражение, то есть с учетом разных значений языковых единиц,
- 126.7 тысяч разных англоязычных слов и словосочетаний,
- 137 тысяч отношений понятие – английское языковое выражение,
- 204.5 тысяч отношений между понятиями.

В следующих разделах рассмотрим подробнее разные этапы развития тезауруса РуТез.

23.2. Первичное наполнение Общественно-политического тезауруса

Построение любого тезауруса начинается с накопления слов и словосочетаний – кандидатов в тезаурусные единицы. Одним из методов формирования исходной совокупности терминов-кандидатов является автоматическое извлечение слов и словосочетаний из текстов предметной области. Поскольку отбор терминов в тезаурус предполагалось осуществлять вручную, важно, чтобы этот автоматически сформированный список слов и словосочетаний, с одной стороны, представлял предметную область достаточно полно, с другой стороны, число словосочетаний, извлеченных с ошибками, было минимально.

Анализ имеющихся тезаурусов показывает, что основную массу тезаурусных единиц составляют слова-существительные, а также словосочетания из двух-трех слов. Наиболее часто структура словосочетаний основывается на зависимых от главного существительного прилагательных и существительных в родительном падеже.

В то же время, именно такие виды словосочетаний наиболее качественно, с минимальным числом ошибок, извлекаются из текстов. Большой процент неоднозначности при выделении предложных конструкций и относительно небольшое

количество терминов среди таких конструкций обусловили решение отказаться от автоматического извлечения предложных конструкций на первом этапе наполнения Общественно-политического тезауруса.

Таким образом, автоматически из текстов документов извлекались следующие типы словосочетаний (обозначим A - прилагательное, N - существительное):

N	Существительное
A+N	согласованные прилагательное + существительное
N+N	существительное + существительное в род. падеже
A+A+N	согласованные прилагательное + прилагательное + существительное
N+A+N	существительное + согласованное прилагательное + существительное в родительном падеже

Кроме ограничений на типы синтаксических конструкций извлекаемых словосочетаний, были еще введены лексические ограничения. Это было связано с тем, что достаточно большое количество слов, употребляемых в текстах, практически не участвуют в образовании терминов. В число таких слов входят разного рода оценочные слова, эмоциональная лексика и др., Для описания возможности образования терминов с прилагательными и именными группами в родительном падеже был создан специальный словарь сочетаемости (в определенной степени аналогичная система учета сочетаемости слов используется в программе).

На основе категорий, приписанных словам, работают правила, которые приписывают словосочетанию категорию "+" или "-" . Категория "+" для словосочетания означает, что словосочетание будет предъявляться эксперту, "-" - словосочетание эксперту предъявляться не будет. Категория словосочетания со словом категории "0" зависит от категорий других слов, входящих в словосочетание.

Обозначим G - группа "прилагательное + существительное", примеры правил:

A(-)+N(-)=G(-)	*важная проблема*
A(+)+N(-)=G(+)	*внешнеполитическая деятельность*
A(-)+N(+)=N(+) (G=N)	*вчерашняя продажа*

Созданный словарь сочетаемости включал около 30000 слов. При этом считалось, что всякое новое относительно словаря слово, появившееся в тексте, имеет категорию "+". По отношению к зависимой конструкции в родительном падеже новое существительное имеет категорию "-".

В 1994-1997 гг. в системе автоматизированной разработки тезауруса было обработано около 50 тысяч нормативных документов Российской федерации, что составляет порядка 200 Мбайт текстовой информации. Было выявлено более 300 тысяч слов и словосочетаний, которые были просмотрены экспертами. На основе этих словосочетаний была создана первая версия Общественно-политического тезауруса - около 28 тысяч тезаурусных входов.

23.3. Пополнение тезауруса в результате работы в компьютерных приложениях

Важным источником пополнения и исправления тезаурусных описаний в течение всего срока существования тезауруса являлся анализ результатов автоматической обработки текстов, произведенной с использованием тезауруса.

Было создано специальное интерфейсное средство, обеспечивающее возможности для удобного анализа результатов обработки текстов.

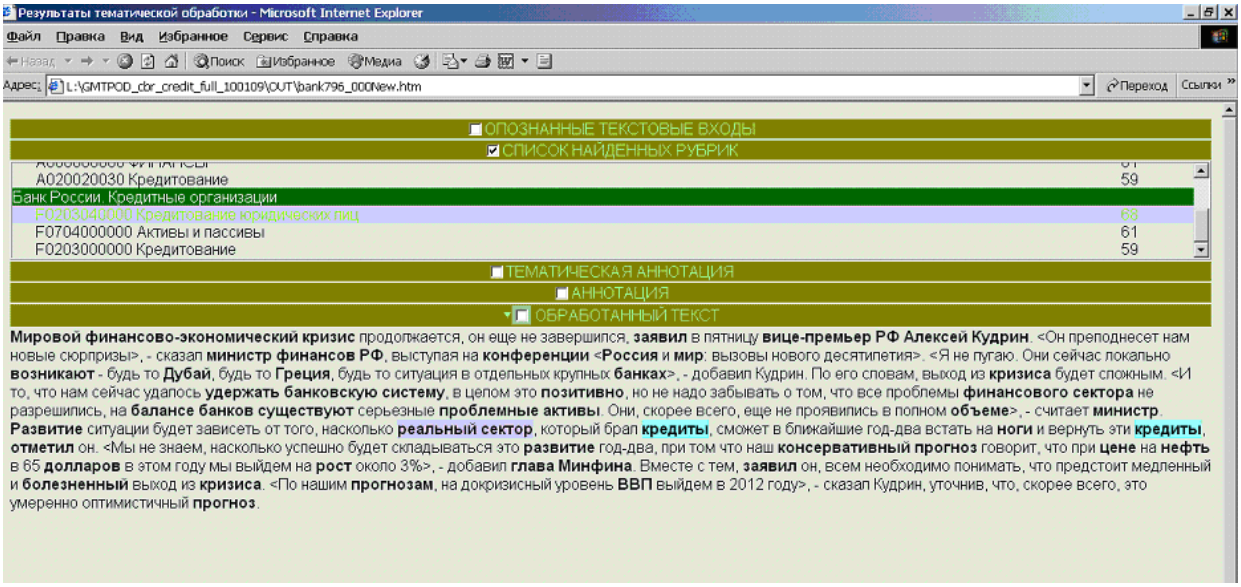

Рис. 23.1. Форма просмотра результатов обработки текстов на основе тезауруса. Показано, что полученная для текста рубрика «Кредитование юридических лиц» была выведена на основе терминов *реальный сектор* и *кредиты.*

Интерфейсное средство позволяет (см. рис.23.1):

- подсвечивать все обнаруженные в тексте текстовые входы тезауруса. Тем самым легко находятся слова и словосочетания, которые еще не внесены в тезаурус;
- показывать значения слов или словосочетаний, описанные в тезаурусе, и значение в конкретном месте текста, выбранное в процессе автоматической обработки текста.
- При процедуре рубрикации может быть показан список понятий, по которым была выведена данная рубрика, а также могут быть подсвечены соответствующие слова и словосочетания в тексте,
- Может быть просмотрена структурная аннотация текста, по которой могут быть выявлены неправильно разрешенные значения слов или неточные отношения между понятиями.

23.4. Пополнение тезауруса на основе анализа списка русскоязычных лемм

В 2001 был выгружен список лемм коллекции Университетской системы РОССИЯ, употреблявшихся в более чем 10 документах. Величина списка составила около 200 тысяч лемм.

Далее производилась вычитка этого списка:

- исключалось из списка то, что уже описано в тезаурусе,
- исключалось то, что не нужно описывать в тезаурусе (личные имена, ошибки и др.),
- вносились описания слов, которые еще не были включены в тезаурус,

- вносились описания значений слов, которые отсутствовали в тезаурусе.

В настоящее время величина данного списка составляет около 22 тысяч лемм. Основная работа идет на первых 8 тысячах лемм, которые в момент выгрузки встречались как минимум в 100 текстах. В исходном списке этим 8 тысячам соответствовали 43 тысячи наиболее частотных лемм.

Для подавляющего большинства этих оставшихся 8 тысяч лемм на текущий момент не представлены некоторые из основных значений, в связи с чем работа продолжается.

При анализе очередного слова для пополнения описания его значений можно встретиться со следующими ситуациями:

- все значения слова, на самом деле, описаны, за счет того, что они были введены при работе с каким-либо другим словом: синонимом, дериватом - тогда это слово просто исключается из списка,
- у слова имеются значения, для которых непонятно, насколько они представлены в тексте современной деловой прозы. В таких случаях производятся проверки употребления данного слова в таком значении в коллекции УИС РОССИЯ и Интернете. Проверки производятся посредством, например, поиска по примерам, приводимым в толкованиях словарных статей толковых словарей,
- у слова имеется несколько значений, между которыми очень трудно почувствовать и описать разницу. В таких случаях помогает поиск синонимичных однозначных словосочетаний, видовых понятий к гипотетическим понятиям, введенным по каждому из значений,
- у слова имеется значение с подзначениями, для адекватного отражения которых возможно нужно будет ввести несколько понятий. Здесь работа производится как в предыдущем пункте. В разделе 16.4.1 мы приводили пример анализа значения слова *покрывало,* для описания которого были введены два отдельных понятия *ПОКРЫВАЛО (ПОКРЫВАЮЩАЯ ТКАНЬ)* и *ПОСТЕЛЬНОЕ ПОКРЫВАЛО*.
- в некоторых случаях слово в данном значении имеет очень узкую сочетаемость. В таких случаях в тезаурус могут вводиться конкретные словосочетания, соответствующие этому значению слова, а отдельно для данного слова значение не описывается. Так, в Большом толковом словаре слово (БТС, 1998) *статья* имеет следующее значение: *5. Разряд, степень звания старшины во флоте*. В тезаурус введены такие понятия как *СТАРШИНА ПЕРВОЙ СТАТЬИ, СТАРШИНА ВТОРОЙ СТАТЬИ и* обобщающее понятие *СТАРШИНСКИЙ СОСТАВ.*

В настоящее время на основе такого просмотра списка в основном пополняется Общий лексикон тезауруса РуТез, некоторое пополнение производится и в части Общественно-политического тезауруса. Так, одно из значений слова *колонка (БТС: 4. Название разнообразных устройств, деталей, конструкций и т.п. обычно удлиненной формы. Стереофоническая колонка. Газовая колонка)* стало источником таких понятий как *АКУСТИЧЕСКАЯ КОЛОНКА, ВОДОГРЕЙНАЯ КОЛОНКА, ГАЗОВАЯ КОЛОНКА.*

23.5. Пополнение Общественно-политического тезауруса за счет проникновения в профессиональные области

В результате использования Общественно-политического тезауруса в различных проектах он пополняется специальной терминологией различных сфер общественной деятельности таких, как банковская деятельность, бухгалтерское дело, судебная практика, гражданское право, таможенное дело и др.,

Характерной особенностью этого направления пополнения Общественно-политического тезауруса является то, что понятия, соответствующие терминам таких

предметных областей, располагаются на нижних уровнях иерархии тезауруса, то есть добавление происходит достаточно органично.

Другой особенностью этих предметных областей является то, что документы этих областей содержат большой объем лексики и терминологии из соседних сфер деятельности, поэтому для качественной обработки документов в этих предметных областях необходимы значительно больший объем понятий тезауруса.

Ряд проектов был связан с обработкой научных текстов в области экономики, социологии, права, что потребовало введения в тезаурус понятий соответствующих наук.

Каждая наука вводит свой понятийный аппарат достаточно высоких уровней абстракции, что приводит к тому, что необходимо встраивать новые понятия внутрь иерархии тезауруса. Примерами таких терминов, которые потребовали введения такого рода понятий, являются экономические термины: *открытая экономика, денежная экономика, факторы производства* и др.; социологические термины: *социальная общность, социальная группа и др.,* При этом научные публикации, создаваемые в рамках этих наук, содержат большое количество терминов из самых разных областей общественной жизни, практической деятельности.

Таким образом, дополнение Общественно-политического тезауруса терминологией научных и практических сфер общественной жизни ведет либо к добавлению понятий нижних уровней иерархии или к введению отдельных понятий внутрь иерархий, то есть изменения структуры тезауруса невелики, при этом вся накопленные тезаурусные знания могут использоваться при анализе соответствующих документов. Если возникает необходимость автоматически обрабатывать документы технических, производственных областей, документов в области естественных наук, такое совмещение терминологии этих сфер и описаний тезауруса РуТез и Общественно-политического тезауруса становится не очень сложным.

В таких случаях по модели тезауруса РуТез создаются отдельные тезаурусы. При этом все полезные для нового тезауруса описания, имеющиеся в тезаурусе РуТез, автоматизировано выгружаются в новый тезаурус. Такая процедура производится на основе сопоставления понятий тезауруса РуТез с текстами предметной области.

Подобные отдельные тезаурусы были созданы в области компьютерной безопасности, авиации (Добров и др., 2002; Добров и др., 2004). В следующей главе будет подробно рассмотрена процедура разработки нового ресурса типа тезауруса РуТез на примере Онтологии по естественным наукам и технологиям – ОЕНТ.

23.6 Тезаурус РуТез: Создание двуязычной онтологии

В главе 16 мы указывали, что стремимся создавать тезаурус РуТез как лингвистическую онтологию, то есть, с одной стороны, подавляющее большинство понятий тезауруса должно быть связано со значениями реально существующих языковых единиц, с другой стороны, понятия тезауруса должны иметь четкие отличия от ближайших понятий в понятийной системе тезауруса. Мы указывали, что в какой-то мере это противоречивые требования, однако это противоречие удается смягчить за счет интенсивного использования существующих в языке словосочетаний.

Одним из следствий формирования отличимых понятий в онтологии является возможность относительно четкой передачи этого понятия на другом языке, посредством отдельного слова, существующих в языке словосочетаний, или, в крайнем случае, словосочетаний, порожденных по правилам этого другого языка. На наш взгляд, формирование онтологического ресурса на базе таких отличимых понятий и делает онтологию по-настоящему независимой от конкретного языка, даже если на первых шагах развития ресурса основой для построения его понятийной системы послужил какой-то конкретный язык (Loukachevitch, Dobrov, 2004; Лукашевич, Добров, 2003).

С 2001 года мы переводим тезаурус РуТез на английский язык, и этот процесс позволяет исследовать возможности создания онтологической понятийной системы на базе языковых значений.

Суть перевода тезауруса состоит не в переводе каждого отдельного текстового входа, а в некотором роде повторении той процедуры образования понятий, которая уже была сделана для русского языка, а именно в выделении понятия, снабжения его однозначным названием и обеспечением максимального ряда текстовых входов для этого понятия.

Поскольку понятия, выделенные на основе русскоязычных значений, уже существуют, то в простейшем случае, если имеются соответствующие слова или выражения на английском языке, то понятие снабжается англоязычным именем и снабжается англоязычными текстовыми входами.

Так, например, в Общественно-политическом Тезаурусе понятие *ЗДРАВООХРАНЕНИЕ* имеет следующий набор вариантов на русском и английском языках (английские и русские текстовые входы даны по алфавиту):

ЗДРАВООХРАНЕНИЕ	PUBLIC HEALTH
защита здоровья *здравоохранительный* *здравоохранительные меры* *обеспечение здоровья* *общественное здравоохранение* *оздоровление граждан* *оздоровление населения* *охрана здоровья* *система здравоохранения*	*community health* *health care* *health care sector* *health care system* *health field* *health of population* *health promotion* *provision of health* *public health* *public health field*

Важной особенностью перевода тезауруса, предназначенного для автоматической обработки текстов, является то, что все возможные текстовые варианты на обоих языках должны быть эквивалентны относительно тезаурусных связей, например, находиться на приблизительно одном уровне иерархии.

При создании традиционных словарей ситуация несколько иная. Основной целью традиционных двуязычных словарей является обеспечение совокупности наиболее частых переводов слова в различных текстах. Переводы даются как бы с запасом. В список переводов включаются и точные переводы, и переводы с более узким значением и с более широким (именно поэтому англо-русские и русско-английские словари не являются обратимыми). Предполагается, что читающий или переводчик разберется по контексту, какой перевод выбрать.

С.В. Гринев-Гриневич (Гринев-Гриневич, 2008) связывает избыточность переводов в терминологических словарях с тем, что в словари включаются окказиональные употребления иноязычных терминов. Окказиональность в употреблении – употребление в несвойственном термину значении может возникнуть по разным причинам: из-за особенностей контекста, по стилистическим причинам, из-за ошибок и неточностей изложения в исходном документе. Во всех случаях неточное значение термина фиксируется автором словаря и приводит к проблемам при переводе.

При переводе онтологического ресурса важным является обеспечить максимальную симметричность, то есть интенсионал и экстенсионал понятия в русском и английском языке должны быть сходными. В то же время тезаурусная сеть позволяет наглядно представить соотношение понятий – более узкое представить нижестоящим понятием, более широкое – вышестоящим понятием. Термин, который не имеет

адекватного перевода, может остаться совсем без перевода, однако по тезаурусной сети можно легко выяснить ближайшие по смыслу более узкие и более широкие понятия. Часто также бывает, что точное значение того или иного понятия может представить достаточно употребительным словосочетанием.

Рассмотрим пример, иллюстрирующий вышеописанные различия двуязычного словаря и двуязычного онтологического ресурса. Электронный словарь Multilex 1.0 представляет значение существительного *sabotag*е следующим образом:

sabotage I n

1. саботаж

2. диверсия; подрывная деятельность; вредительство

В русской части Общественно-политического тезауруса перечисленным словам соответствуют три различных понятия: *САБОТАЖ, ДИВЕРСИЯ, ВРЕДИТЕЛЬСТВО*. Если буквально приписать слово *sabotage* ко всем этим трем понятиям, то мы получим, что у слова три значения. Но это не соответствует англоязычным источникам, которые описывают одно значение этого слова:

*Sabotage - any underhand interference with production, work, etc., in a plant, factory, etc., as by enemy agents during wartime or by employees during a trade dispute (*Random House Webster's Unabridged dictionary, 1998)

Sabotage – any deliberate destruction, disruption, or damage of equipment, a public service, etc., as by enemy agents, dissatified employees, etc. (Collins electronic dictionary, 1992)

Анализ употребления английского слова sabotage и соответствующих слов русского языка приводит к необходимости установления отношений родовидовых отношений между понятиями ВРЕДИТЕЛЬСТВО - САБОТАЖ, ВРЕДИТЕЛЬСТВО – ДИВЕРСИЯ (этих отношений до начала работы со словом sabotage не было – так сопоставительный анализ значений английских слов приводит к более качественному описанию значений русских слов). Понятие SABOTAGE было поставлено в соответствие понятию ВРЕДИТЕЛЬСТВО. Англоязычный ряд для русского понятия САБОТАЖ имеет следующий вид: employee sabotage, labor sabotage, sabotage by employees, silent sabotage, workers sabotage. Англоязычный ряд для русского понятия ДИВЕРСИЯ таков: sabotage attack, enemy sabotage, sabotage by enemy, sabotage explosion.

В качестве другого примера рассмотрим значение англоязычного слова brother-in-law, для которого в русском языке нет ни соответствующего слова, ни употребительного словосочетания. В таких случаях заводится понятие со специальной пометкой #, обозначающей, что русского эквивалента нет. Понятие снабжается русским пояснением. Отношение с другими понятиями тезауруса показывает соотношение русских и английских понятий:

BROTHER-IN-LAW		--	*# ДЕВЕРЬ ИЛИ ШУРИН*
ВЫШЕ	*KINSMAN*	--	*РОДСТВЕННИК-МУЖЧИНА*
НИЖЕ	*BROTHER OF HUSBAND*	--	*ДЕВЕРЬ*
НИЖЕ	*BROTHER OF WIFE*	--	*ШУРИН*

Иногда подобная сочинительная конструкция употребляется и самими носителями языка для заполнения лексической лакуны. Так, в разделе 16.5.2.6 мы уже упоминали отсутствие отдельного слова, значение которого соответствует значению русскоязычного термина *вексель*. Как уже указывалось, векселя делятся на простые векселя (*promissory notes* или просто *notes*) и переводные (*bills of exchange* или просто *bills*). Значению термина *вексель* соответствует конструкция *bills and notes* (80000 употреблений в Google).

В ряде случаев взгляд с точки зрения английского языка помог подобрать более адекватное понятийное представление для значений многозначных русскоязычных слова. В качестве яркого примера можно привести слово *масло*, которое в разных контекстах переводится *butter* или *oil.*

По русским толковым словарям не очень понятно, как лучше представить значения слова *масло,* как они представлены в словосочетаниях *сливочное масло, растительное масло, минеральное масло, топленое масло.*

Русскоязычные толковые словари (Словарь Ефремовой, БТС, словарь Ожегова) подразделяют значения по признаку использования или неиспользования в пищу. Словарь Ефремовой выделяет два подзначения в одном значении по признаку использования или неиспользования в пище:

1.1) Жидкое или твердое жировое вещество, искусственно добываемое из веществ растительного, минерального или животного происхождения.
2) Пищевой продукт животного или растительного происхождения.

Словарь (Ожегов, Шведова, 1995) выделяет по тому же признаку два отдельных значения. Словосочетания *сливочное масло, животное масло* указываются как примеры к первому значению, а словосочетание *бить масло* – ко второму значению.

Словарь (БТС, 1998) также выделяет два подзначения. Второе подзначение несколько отличается и выглядит так:

Пищевой продукт, получаемый путем сбивания сливок; сливочное масло.

Причем словосочетание *топленое масло*, получаемое обычно из сливочного масла, дано как пример к первому подзначению.

Как и сколько понятий тезауруса правильно создать на базе таких толкований, не очень ясно.

При создании двуязычной онтологии такой беспорядок в русскоязычных источниках усложняется тем, что в качестве переводов в этих употреблениях слова *масло* используются два разных слова: *butter* и *oil.* Значение слова *butter* соответствует русскому *сливочное масло,* а значение слова *oil* в словаре Encartha толкуется следующим образом*:*

oil – 1. ***thick greasy liquid:*** *a liquid fat, obtained from plant seeds, animal fats, mineral deposits, and other sources, that does not dissolve in water and will burn.*

Из этого толкования можно понять, что oil1 – это жидкий жир. Мы решили принять представление значений слова *масло* именно на базе их английского перевода, поскольку получившиеся понятия обладают набором характерных свойств, отличающих их от других понятий.

Таким образом, этим толкованиям соответствуют два понятия МАСЛО (ЖИДКИЙ ЖИР) и СЛИВОЧНОЕ МАСЛО, со следующими наборами отношений:

МАСЛО (ЖИДКИЙ ЖИР)

с	*масло, масляный, жирное масло*
ВЫШЕ	*ЖИДКОСТЬ*
ВЫШЕ	*ЖИР*
НИЖЕ	*МИНЕРАЛЬНОЕ МАСЛО*
НИЖЕ	*РАСТИТЕЛЬНОЕ МАСЛО*
АСЦ2	*МАСЛЯНАЯ КРАСКА*

СЛИВОЧНОЕ МАСЛО

С	*масло, масляный, сливочное масло, животное масло*
ВЫШЕ	*ЖИВОТНЫЙ ЖИР*
ВЫШЕ	*МОЛОЧНАЯ ПРОДУКЦИЯ*
АСЦ2	*ТОПЛЕНОЕ МАСЛО*

Еще раз подчеркнем, что такой выбор понятий был сделан не в угоду англоязычной лексикализации, а потому, что английский язык подсказал наиболее адекватное разбиение существующих явлений на различимые понятия.

Для уточнения англоязычной части лингвистической онтологии – тезауруса РуТез (так же как и для русского языка) проводится процедура вычитки значений слов по частотному списку, который был получен на основе коллекций газетных статей Glasgo Gerald и Los Andgeles Times (1994-1995 гг.), предоставленных в процессе участия в конференции по многоязычному поиску CLEF .

Заключение к главе 23

В настоящее время тезаурус РуТез продолжает развиваться. В сфере общей лексики продолжается подбор наилучшего понятийного представления для значений наиболее частотных слов, вводятся словосочетания, позволяющие четче разграничить эти значения.

Общественно-политический тезаурус пополняется за счет вхождения в профессиональные понятийные системы. Также пополнение Общественно-политического тезауруса происходит за счет уточнения значений общей лексики.

Продолжает развиваться и уточняться англоязычная часть тезауруса РуТез.

Глава 24. Онтология по естественным наукам и технологиям

24.1. Проблемы разработки онтологии в сфере естественных наук

Для профессионального, в том числе научно-технического, поиска информации часто требуется обеспечение поиска, основанного на знаниях, – использование синонимов, возможности автоматического расширения запроса, возможностей автоматического анализа результатов запроса и помощь в интерактивном поиске.

Традиционными средствами тематического поиска научной информации в течение многих лет являлись информационно-поисковые тезаурусы. Однако, как мы уже указывали, такие тезаурусы создавались для их использования в процессе ручного индексирования и поиска, и не обеспечивают эффективного информационного поиска в автоматических режимах обработки текстов. Кроме того, отношения между терминами, используемые в традиционных информационно-поисковых тезаурусах считаются недостаточно формализованными, субъективными.

Создание формализованных онтологических ресурсов в сфере естественных наук связано с рядом проблем.

Во-первых, такие ресурсы должны быть достаточно большой величины, включая десятки тысяч понятий, что обычно снижает возможность их формальных описаний.

Во-вторых, формализация ограничивается развивающейся природой науки, что проявляется в существовании различных теорий, частичным пониманием введенных понятий.

В-третьих, (Tsujii, Ananiadou, 2005) указывают на такую проблему, как гипотетическая природа онтологий. В логических онтологиях классификационная схема существует до описания конкретных явлений. В то время как в научных онтологиях классификационная схема должна наилучшим образом объяснить наблюдаемые явления. Нахождение наилучших классификационных схем – это важнейший научный результат, помогающий объяснить и описать явления.

Наконец, в научных сферах понятия неразрывно связаны с терминами, их языковыми представителями.

Все эти факторы дали возможность предположить, что для создания онтологии в сфере естественных наук может быть использована структура лингвистической онтологии тезауруса РуТез, характеризующимся небольшим набором формализованных отношений и серьезной опорой на значения реально существующих языковых единиц – слов и словосочетаний.

В 2004 году были начаты работы по разработки Онтологии по естественным наукам и технологиям ОЕНТ (Добров и др., 2005; Добров, Лукашевич, 2006). Широта выбранной области, сочетание разных наук связано с тем, что для конкретных разделов той или иной естественной науки необходимы знания из разных разделов этой же науки или других наук, а также математики. Действительно, значимой проблемой при структуризации знания в пределах одной области науки является трудность в отграничении данной области от других, либо исследующих те же объекты, либо применяющих аналогичные подходы. С другой стороны, доступ к знанию таких родственных научных подходов был бы крайне интересен каждому исследователю.

Начало работ над Онтологией по естественным наукам и технологиям означало, что было принято решение раздельно разрабатывать две разные онтологии для анализа текстов в общественно-политической сфере (газетные статьи, новостные сообщения, законодательные акты, международные договоры) и научных публикаций.

Решение о разделение онтологий было связано с несколькими серьезными факторами.

Во-первых, обе онтологии достаточно объемны, включают десятки тысяч понятий и отношений, при этом большая часть понятий общей онтологии обычно не используется в текстах естественных наук, и наоборот, научные понятия, по большей мере, не нужны для анализа таких общезначимых документов, как газетные статьи, информационные сообщения, законодательные акты.

Во-вторых, разделение онтологий снижает многозначность описанных слов и выражений.

В-третьих, предполагалось, что существует несоответствие, так называемой, «бытовой» картины мира и научной картины мира. То есть отношения, описанные и правильные в рамках одной онтологии, должны быть изменены в рамках другой онтологии.

И наконец, последнее (по перечислению, но не по важности) эти две онтологии отличаются по способам рассмотрения внешнего мира: онтология РуТез рассматривает мир через призму современного цивилизованного общества: что известно о мире значимому количеству образованных людей современного общества, что важно (воздействует, используется) в существовании современного общества. Онтология в области естественных наук и технологий исключает из рассмотрения аспекты общественного мировосприятия и должна описывать в виде онтологической модели устоявшиеся воззрения современной науки, основываясь на материалах научных публикаций.

Вместе с тем, хотелось бы отметить, что существуют типы текстов, для анализа которых могут понадобиться обе онтологии, работающие одновременно, и поэтому нужно иметь четкое представление об отражении сходных явлений в разных контекстах.

К числу текстов, требующих, как представляется, использования обеих онтологий относятся:

- анализ соответствий между требованиями технического регулирования и описанием производственных процессов;
- документы вида «заявки/отчеты» о научном исследовании,
- инвестиционные заявки, связанные с промышленным внедрением научных исследований.

В следующих разделах будут подробно рассмотрены этапы создания онтологии ОЕНТ**.**

24.2. Этапы создания онтологии ОЕНТ

Основной задачей при создании лингвистической онтологии большого размера силами небольшого коллектива является максимальное использование методов автоматизации, а также фрагментов ранее созданных лингвистических онтологий. Процедура формирования первой версии онтологии ОЕНТ включала интеграцию информации из нескольких разных источников.

24.2.1. Автоматический набор терминологии по текстам

Для каждой науки из рассматриваемого списка (математика, физика, химия, биология, геология) были сформированы коллекции документов (от 3000 до 8000 документов, от 50 до 90 Мб). Источником коллекций являлись документы, доступные в Интернет, следующих основных типов:

- материалы школьных уроков;
- рефераты;
- университетские лекции;
- материалы специализированных сайтов.

Была произведена обработка специальными процедурами автоматического извлечения терминоподобных словосочетаний, что дало возможность проверки

употребимости терминов в материалах, а также нахождения терминов, входящих в состав предметной области.

Для извлечения терминоподобных словосочетаний было использовано два алгоритма.

Первый алгоритм извлечения словосочетаний выделяет существительные, прилагательные, согласованные пары и тройки прилагательных и существительных, а также генитивные конструкции (существительное + существительное в родительном падеже и т.п.) (см. п.23.2).

Второй алгоритм может выделять часто повторяющиеся именные группы, состоящие из нескольких слов, в том числе предложные (Добров и др., 2003).

Кроме того, тексты сопоставлялись с терминами Общественно-политического тезауруса. Полученные терминоподобные слова и словосочетания упорядочивались по убыванию суммарной частотности и убыванию количества содержащих их документов.

24.2.2. Автоматизированное формирование первой версии онтологии

Основной целью при формировании первой версии ресурса являлось быстрое получение приближения предметной области. При этом выбор делался в сторону большей избыточности первого приближения, чтобы в дальнейшем минимизировать по возможности поиск и добавление новых терминов.

Для отбора терминологии по каждой предметной области были образованы верхние части частотных списков терминоподобных слов (по 10 тысяч) и словосочетаний (по 15 тысяч), которые были направлены на быструю разметку экспертам. Отметим, что нижняя часть этих списков соответствовала уровню встречаемости в 5-6 документах.

Эксперты должны были в рамках «своей» науки пометить принадлежность к предметной области того или иного термина. Допускалась пометка термина для нескольких предметных областей, но полнота такого рода разметки не требовалась. После окончания этого этапа списки разных экспертов были объединены – получился список из 32 тысяч помеченных слов и словосочетаний.

Следующим этапом стало использование терминологии, включенной в Тезаурус русского языка РуТез. Этот тезаурус содержит общеупотребительную лексику, лексику и терминологию нормативно-правовых актов и материалов СМИ. Поэтому имеет значимое пересечение с терминологией практически любой значимой предметной области.

Для каждой новой предметной области были заданы несколько понятий верхнего уровня, такие как *=НАУКА=, =РАСТЕНИЕ=* и т.п., касающиеся сущности исследуемых предметных областей и их предметов ведения. Для таких понятий были выбраны способы расширения по иерархии тезаурусных связей (полное расширение или расширение только по таксономическим отношениям). Полученные группы понятий были помечены специальными пометками отнесения к дополнительной предметной области соответствующей науки и к специальной служебной рабочей предметной области «кандидат».

После этого список отобранных экспертами терминов по текстам был сопоставлен с текстовыми входами понятий Тезауруса РуТез. В случае совпадения с текстовым входом из тезауруса, все понятия, ассоциированные с данным текстовым входом, получали дополнительные пометки новых предметных областей – соответствующей науки (наук) и предметной области «кандидат».

Если отобранный экспертами термин был не известен, то заводилось новое понятие, название и единственный текстовый вход которого совпадали с данным термином. Новое понятие получало пометки принадлежности к предметной области соответствующей науки и «кандидат». Кроме того, автоматически вводилось таксономическое отношение ВЫШЕ к специальному временному понятию в каждой науке, например, *=@ГЕОЛОГИЧЕСКАЯ ТЕРМИНОЛОГИЯ=, =@ХИМИЧЕСКАЯ ТЕРМИНОЛОГИЯ=*, и т.п.

Наконец, для отобранных из Тезауруса РуТез понятий (получивших пометку «кандидат») было выполнено «замыкание» - были добавлены понятия, расположенные выше по таксономическим связям. Эти понятия получали аналогичные дополнительные пометки предметных областей.

Таким образом, в результате предыдущих этапов был сформирован «пополненный» терминологический ресурс. Так как все интересующие нас понятия имели пометку отнесения к служебной предметной области «кандидат», то мы использовали стандартную процедуру «экспорта» фрагмента тезауруса для формирования нового ресурса.

24.2.3. Методология работы экспертов

После появления первой версии онтологии с корпусом, а также со словарями предметной области начинают работать эксперты – инженеры по знаниям.

Основными целями их работы являются следующие:

- изучая конкретные языковые выражения, их словарные определения, употребление в конкретных текстах определить, какому понятию соответствует значение данного языкового выражения. Если такое понятие уже существует, данное языковое выражение приписывается этому понятию. Для нового понятия создается отдельная единица в иерархической сети;
- Для каждого понятия по корпусу набирается максимально возможное число различных слов, выражений, значения которых соответствуют этому понятию – текстовых входов понятия;
- Для каждого понятия проводится концептуальный анализ для выяснения его таксономических отношений и отношений онтологической зависимости. Поскольку эти отношения являются наиболее важными для широкого круга понятий, их часто можно выявить на основе анализа определений соответствующих терминов в терминологических словарях, употреблений в текстовых контекстах, сопоставления определений и текстовых контекстов.

Как показывает практика, в связи с многократно описанными проблемами получения знания от экспертов в предметной области (Гаврилова, 2001), наиболее эффективным является максимально полная разработка ресурса на основе анализа текстового корпуса. В многочисленных исследованиях подтверждено, что к описанию предметной области следует привлекать специалистов по знаниям, а также специалистов по языку, чем полностью доверить данную работу специалистам в предметной области. Во-первых, описание структуры научного знания предметной области является другим видом деятельности, чем, собственно, исследование в рамках данной науки. Во-вторых, большинство ученых действует в рамках своей научной школы, недостаточно представляя понятийный аппарат других школ. В-третьих, каждый исследователь видит недостаточную детализацию любого понятия в рамках своей науки (вполне достаточную для информационной обработки результатов исследований).

Далее созданный проект ресурса предъявляется экспертам в предметной области, которые уже достаточно легко находят в нем возможные ошибки и неточности, могут объяснить, почему им не понравилось то или иное отношение.

Следует отметить, что на этапе разработки онтологии в качестве инженеров по знаниям выступают лингвисты, которые имеют опыт работы с текстовыми корпусами, лексическими значениями.

В настоящее время инженерами по знаниям используются следующих три основных источника:

- профильные и общие энциклопедии, терминологические словари – как источник профессиональной информации;
- накопленные списки терминоподобных слов словосочетаний, которые очень эффективны при добавлении синонимов, вариативно отличающихся от указанных в опубликованных энциклопедических источниках;

- каждый текстовый вход должен быть проверен по употреблению в Интернет. Такая проверка производится с использованием глобальных поисковых машин.

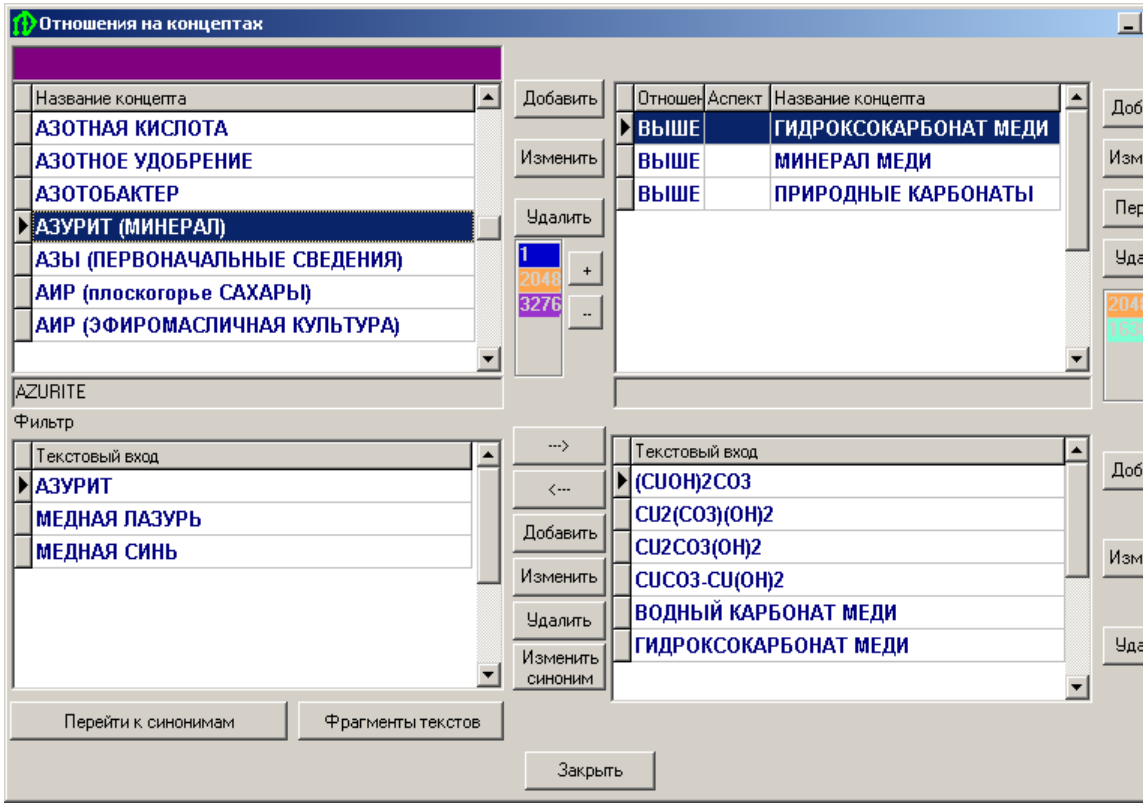

Рис.24.1 Основная экранная форма редактирования отношений и текстовых входов понятий

24.3 Текущее состояние проекта

В настоящее время онтология включает 50 тысяч понятий, 140 тысяч терминов таких областей как математика, физика, химия, геология, экология, биология.

Описанные термины в значительной мере покрывают терминологию этих областей, вводимую в школе и на начальных курсах ВУЗов. Состав терминологии сопоставлен с терминологией рубрикатора УДК.

Рис 24.1 представляет рабочий экран системы ведения онтологии. В левом верхнем углу помещены понятия онтологии, в левом нижнем углу представлены текстовые входы для понятия, на котором установлен курсор *АЗУРИТ (МИНЕРАЛ) – азурит, медная лазурь, медная синь*. В правом верхнем углу показаны отношения этого понятия. Оно описывается как подкласс понятий *КАРБОНАТ МЕДИ, МИНЕРАЛ МЕДИ, ПРИРОДНЫЕ КАРБОНАТЫ*. Правый нижний угол экрана представляет варианты текстовых входов для понятия *КАРБОНАТ МЕДИ*. Видно, что экран отражает отношения между традиционно геологическими и химическими понятиями. Таким образом, отражение понятий, традиционно относящихся к разным наукам, в рамках единого ресурса дает возможность использовать для описания отношений понятий разных наук.

24.4 Изменения в описаниях понятий, полученных из Тезауруса РуТез

Возможность вторичного использования однажды разработанных онтологий в других областях или других приложениях является важной проблемой в онтологических исследованиях (Guarino, 1998b; Kalinichenko, Skvortsov, 2004). Для поддержки процедуры

слияния онтологий и создания на этой основе новой онтологии разработано несколько программных продуктов (McGuinnes и др., 2000; Noy, Musen, 2000).

Отдельное направление исследований составляет использование онтологий верхнего уровня или общезначимых онтологий (онтологий, не ориентированных на конкретную предметную область) для разработки онтологий в конкретных предметных областях. В качестве такой общей онтологии при разработке предметно-ориентированных онтологий для автоматической обработки текстов часто используется лингвистическая онтология WordNet (Magnini, Speranza, 2002; Buitellar, Sacalenau, 2001; Vossen, 2001).

Близкие по смыслу понятия общей и предметно-ориентированной лингвистической онтологии могут состоять между собой в следующих отношениях (Magnini, Speranza, 2002; Buitellar, Sacalenau, 2001; Hovy, 1998):

- синонимы, то есть понятия двух онтологий могут быть склеены между собой;
- понятие конкретной онтологии является видом для понятия общей онтологии;
- понятия конкретной онтологии и общей онтологии являются квазисинонимами, то есть одному понятию общей онтологии соответствуют два понятия частной онтологии, или одному понятию частной онтологии соответствуют два понятия общей онтологии. В случае WordNet наличие в нем двух понятий (синсетов), относящихся к одному понятий предметной онтологии, обычно связано с более детальной трактовкой лингвистических явлений, чем это обычно принято в терминологических ресурсах.

В начале работ над онтологией ОЕНТ мы выгрузили часть Тезауруса РуТез – лингвистической онтологии в предварительную версию новой онтологии. Таким образом, фрагменты общезначимой онтологии были перемещены в другой контекст – область естественных наук. При этом приложение онтологий является одинаковым – информационно-поисковые задачи, такие как индексация и поиск документов, автоматическая рубрикация, поиск ответов на вопросы, поиск похожего документа и т.п.

В 2006 году, через два года после начала проекта было прведено исследование, как изменились описания понятий, выгруженных из тезаруса РуТез в процессе работ над онтологией ОЕНТ.

Для изучения описаний понятий, перенесенных из Тезауруса РуТез (далее онтология-прототип), мы образовали список таких понятий, которые эксперты одобрили для включения в Онтологию по естественным наукам и технологиям, то есть сняли пометку «понятие-кандидат». Таких понятий оказалось 4540.

С описаниями понятий могли произойти следующие типы изменений:

- изменение названия понятия;
- изменение набора текстовых входов понятия:

удаление текстовых входов понятия;
добавление текстовых входов понятия;

- изменение отношений между понятиями онтологии-прототипа:

исчезновение отношений между понятиями онтологии-прототипа;
появление новых отношений между понятиями онтологии-прототипа;

- введение отношений понятий онтологии-прототипа с новыми понятиями:

введение отношений вверх по иерархии;
введение отношений вниз по иерархии.

В следующих подразделах рассмотрим наиболее интересные явления, которые удалось выявить.

24.4.1. Удаление текстовых входов понятия

Изменения набора текстовых входов понятия связано в основном с двумя причинами.

Во-первых, от понятия отсоединяются текстовые входы, носящие метафорический, образный характер, свойственные газетным текстам и неупотребляемые в научной речи, например, *ВЕРБЛЮД – корабль пустыни*.

Во-вторых, (и таких удаленных текстовых входов большинство) часть текстовых входов исходного одного понятия перешло как текстовые входы к новообразованному понятию, то есть практически понятие расщепилось на два (или более) понятий. Например, были разделены в отдельные понятия бывшие синонимы (текстовые входы одного и того же понятия): *ХИМИЧЕСКАЯ РЕАКЦИЯ* и *ХИМИЧЕСКИЙ ПРОЦЕСС*, *СУДОРОГА* и *СПАЗМ, СОЛИ ФОСФОРНЫХ КИСЛОТ* и *ФОСФАТЫ* и т.п.

24.4.2. Замена отношений между понятиями онтологии-прототипа на более длинные цепочки отношений

Авторы (Hovy, 1998; Magnini, Speranza, 2002), работавшие с двумя онтологиями, одна из которых более общая, а вторая относится к конкретной предметной области, предполагали, что набор вышестоящих отношений более общей онтологии не подвергается изменениям.

Однако наше сопоставление показало значимое число удаленных родовидовых отношений между понятиями онтологии-прототипа. Более тщательный анализ показал, что достаточно часто удаленное отношение заменяется на более длинную цепочку отношений, состоящую из двух или трех отношений, то есть между понятиями, перешедшими из более общей онтологии, вклиниваются одно-два понятия из предметной онтологии.

Например, в Тезаурусе РуТез для понятия *АДСОРБЕНТ* было установлено родовидовое отношение к понятию *ВЕЩЕСТВО*, а в новой онтологии создана цепочка понятий *АДСОРБЕНТ - СОРБЕНТ – ВЕЩЕСТВО*.

Отношение между понятиями *БОКСИТ – ГОРНАЯ ПОРОДА* заменилось на цепочку *БОКСИТ – БИОГЕННАЯ ГОРНАЯ ПОРОДА – ОСАДОЧНАЯ ГОРНАЯ ПОРОДА – ГОРНАЯ ПОРОДА*.

Отношение между понятиями *БУЙВОЛ – ЖВАЧНОЕ ЖИВОТНОЕ* заменилось на цепочку *БУЙВОЛ – ПОЛОРОГИЕ – ЖВАЧНОЕ ЖИВОТНОЕ* и т.д.

Количество таких замен одного отношения на цепочку отношений оценивается на текущий момент как более 1000 единиц, что для множества рассматриваемых понятий онтологии-прототипа (4540) представляется значительной величиной.

Важно отметить, что часть из нововведенных отношений может быть перенесена и в исходную онтологию, послужить для уточнения исходных описаний. Вместе с тем значительная часть нововведений не подлежит переносу в онтологию-прототип (см. примеры выше), поскольку введенные понятия соответствуют исключительно научной терминологии и практически не используются в общезначимых текстах.

24.4.3. Несоответствие наивной, бытовой картины мира и научной картины мира

Тезаурус РуТез предназначен для обработки общезначимых документов: информационных сообщений, нормативных документов, газетных статей. Поэтому он должен отражать знания о мире, которыми обладают авторы и читатели такого вида документов. Картина мира, представленная в тезаурусе, может отличаться от картины мира, излагаемой в рамках естественных наук.

Хрестоматийным примером отличия бытовой картины мира и научной картины мира является знание о том, что кит является млекопитающим, а не рыбой (Апресян, 1995). Однако этому вопросу уделяется достаточное внимание в курсе зоологии средней школы. В частности, не удалось найти ни одного такого текста в текстовой коллекции Университетской информационной системы РОССИЯ (www.cir.ru, более миллиона

документов), в котором бы автор считал, что кит – это рыба. Тезаурус РуТез также описывает китов как морских млекопитающих.

Однако удалось выявить ряд несоответствий наивной картины мира, зафиксированной в Тезаурусе РуТез, и научной картиной мира.

Здесь можно выделить два типа различий. Первый тип различий состоит в том, что, то, что в наивной картине мира кажется связанным простым отношением (например, родовидовым), в научной картине мира напрямую не связано. Второй тип различий – то, что представляется несвязанным в наивной картине мира, непосредственно связано между собой в научной картине мира.

Большинство примеров несоответствий находится в сфере биологии. Так, птица эму, которую часто называют *страус эму*, по биологической классификации не является *страусом*.

С другой стороны, по биологической классификации *бледная поганка* относится к *мухоморам*, а *горчица* и *брюква* к *роду капуста*.

Наиболее запутанной ситуацией является ситуация с употреблением слова *орех*. Биологическая наука рассматривает орех как особый вид плода, к которым, например, не относятся грецкие орехи. Одновременно существует «хозяйственный» (по выражению Большой Советской энциклопедии) взгляд на орехи – плоды деревьев и кустарников, «состоящие из сухой деревянистой оболочки и заключённого в ней съедобного и питательного ядра».

Кроме того, существует еще более отличающееся от научного употребление слова *орех*, которое включает в *орехи – арахис, земляной орех*. Это растение по биологической классификации относится к бобовым культурам и не является деревом или кустарником.

Работа с такими несоответствиями связана с двумя видами деятельности: изменение отношений между понятиями на более научно-мотивированные (в том числе и в онтологии-прототипе) и/или ввод разных понятий для разного употребления того или иного слова и описание такого слова как многозначного. Так, видимо, целесообразно иметь два понятия для плода орех – орех как плод ореховых культур (биологическая картина мира) и орех как плод орехоплодных культур («хозяйственная» картина мира).

24.4.4. Смена антропоцентрической картины мира на естественнонаучную картину мира

Наивная картина мира отличается еще и тем, что она ставит в свой центр человека и общество, то есть является антропоцентрической. При переходе к естественнонаучной картине мира эта антропоцентричность пропадает, что находит отражение в отношениях онтологии.

Мы заметили это явление в двух проявлениях.

Есть знание, которое известно и в наивной картине мира, но из-за того, что в повседневной жизни некоторая сущность чаще всего встречается в той или иной форме, то эта форма и считается основной для сущности.

Это явление хорошо видно на примере веществ и их агрегатных состояний и проявляется уже в различиях в толкованиях, которые даются в толковых словарях и энциклопедических словарях.

Так, в толковом словаре (Ефремова, 2006) первое значение слова *вода* таково:

> *1. Бесцветная прозрачная жидкость, представляющая собою химическое соединение водорода и кислорода и содержащаяся в атмосфере, почве, живых организмах и т.п.*

В Большой Советской энциклопедии термин *вода* имеет такое определение:

окись водорода, H_20, простейшее устойчивое в обычных условиях химическое соединение водорода с кислородом (11,19% водорода и 88,81% кислорода по массе),

молекулярная масса 18,0160; бесцветная жидкость без запаха и вкуса (в толстых слоях имеет голубоватый цвет),

Как следствие, в тезаурусе РуТез установлено отношение *ВОДА – ЖИДКОСТЬ*, в Онтологии по Естественным наукам *ВОДА* – это *СОЕДИНЕНИЕ КИСЛОРОДА С ВОДОРОДОМ, ОКСИД НЕМЕТАЛЛА*. Вводится дополнительное понятие *ЖИДКАЯ ВОДА* (вода в жидкой фазе, вода в жидком состоянии), которая и является видом понятия *ЖИДКОСТЬ*.

При этом образованным современникам отлично известно, что *СОЕДИНЕНИЕ ВОДА* бывает в разных агрегатных состояниях, но установить отношение между понятиями *ВОДА* и *ЖИДКОСТЬ* в общезначимом ресурсе удобно, так как жидкое агрегатное состояние воды является наиболее обсуждаемым, другие агрегатные состояния *ПАР* и *ЛЕД* воспринимаются как производные от основного.

Еще один элемент антропоцентрической картины мира в тезаурусе РуТез – это наличие таких оценочных понятий как *СТИХИЙНОЕ БЕДСТВИЕ*, которое оценивает воздействие тех или иных явлений на человеческое существование и включает такие понятия как *ЗЕМЛЕТРЯСЕНИЕ, СМЕРЧ, НАВОДНЕНИЕ* и др. Как представляется естественнонаучная онтология должна избегать таких оценочных понятий как *СТИХИЙНОЕ БЕДСТВИЕ* и должна использовать нейтральные классификации: *СЕЙСМИЧЕСКОЕ ЯВЛЕНИЕ, МЕТЕОРОЛОГИЧЕСКОЕ ЯВЛЕНИЕ* и т.п.

24.4.5. Пример

В качестве примера сравним описание понятия *АЗУРИТ* в составе Тезауруса РуТез и Онтологии по Естественным наукам и технологиям.

Азурит – достаточно известный минерал, используется для получения меди и медного купороса, а также для изготовления синей краски.

Описание понятия АЗУРИТ в тезаурусе РуТез	Описание понятия АЗУРИТ в Онтологии по естественным наукам
АЗУРИТ син АЗУРИТ син МЕДНАЯ ЛАЗУРЬ ВЫШЕ *МИНЕРАЛ* син МИНЕРАЛ син МИНЕРАЛЬНОЕ ВЕЩЕСТВО син МИНЕРАЛЬНЫЙ АСЦ1 *МЕДЬ* син МЕДНЫЙ син МЕДНЫЙ КОНЦЕНТРАТ син МЕДЬ син МЕДЬСОДЕРЖАЩИЙ	*АЗУРИТ (МИНЕРАЛ)* син АЗУРИТ син МЕДНАЯ ЛАЗУРЬ син МЕДНАЯ СИНЬ ВЫШЕ *ГИДРОКСОКАРБОНАТ МЕДИ* син (CUOH)2CO3 син CU2(CO3)(OH)2 син CU2CO3(OH)2 син CUCO3-CU(OH)2 син ВОДНЫЙ КАРБОНАТ МЕДИ син ГИДРОКСОКАРБОНАТ МЕДИ ВЫШЕ *МИНЕРАЛ МЕДИ* син МЕДНЫЙ МИНЕРАЛ син МИНЕРАЛ МЕДИ син ПРИРОДНАЯ МЕДЬ ВЫШЕ *ПРИРОДНЫЕ КАРБОНАТЫ* син КАРБОНАТНЫЙ МИНЕРАЛ син МИНЕРАЛ КЛАССА КАРБОНАТОВ син ПРИРОДНЫЕ КАРБОНАТЫ

Описание понятия АЗУРИТ в тезаурусе РуТез	Описание понятия АЗУРИТ в Онтологии по естественным наукам

Рисунок 24.1 показывает рабочий экран ведения Онтологии ОЕНТ. В левой верхней части экрана помещен список понятий, курсором выделено рассматриваемое понятие - *АЗУРИТ*. В левой нижней части экрана показаны текстовые входы для понятия. Правая верхняя часть экрана представляет список понятий, связанных отношениями с рассматриваемым. Курсор установлен на отношении с понятием *ГИДРОКСОКАРБОНАТ МЕДИ*. Правая нижняя часть экрана показывает текстовые входы понятия, выделенного курсором в правой части экрана.

Рисунок 24.2 показывает верхние уровни иерархии понятия *АЗУРИТ* в Онтологии по естественным наукам и технологиям (за недостатком места не все существующие отношения отражены). Ромбиками помечены понятия, которые были экспортированы из тезауруса РуТез. Мы можем видеть, что прямые отношения понятия АЗУРИТ в тезаурусе РуТез заменились на многоступенчатые структуры, описывающие химический состав минерала.

На рисунке 24.3 для сравнения показаны верхние уровни иерархии понятия *АЗУРИТ* в тезаурусе РуТез.

При рассмотрении различий в описании одинаковых и близких по смыслу понятий в общезначимой онтологии и предметно-ориентированной онтологии на примере Тезауруса РуТез как общезначимой онтологии и Онтологии по естественным наукам как предметно-ориентированной онтологии мы выявили особенности структуры «стыка» между такими онтологиями.

Стык не представляет собой сплошную полосу понятий, принадлежащих обеим онтологиям. Стык онтологий выглядит как совокупность полос, в которых между уровнями, принадлежащими обеим онтологиям, находятся понятия, принадлежащие только одной из онтологий.

Различия в антропоцентрической «наивной» картине мира и естественнонаучной картине мира, проявляются в несоответствиях между описаниями понятий в соответствующих онтологиях.

Полагаем, что сложная картина соответствий между описаниями близких по смыслу понятия в онтологии РуТез и онтологии ОЕНТ объясняются тем, что эти две онтологии отличаются по способам рассмотрения внешнего мира. Онтология РуТез рассматривает мир через призму современного цивилизованного общества: что известно о мире значимому количеству образованных людей современного общества, что важно (воздействует, используется) в жизни современного общества. Онтология в области естественных наук и технологий исключает из рассмотрения аспекты общественного мировосприятия и должна описывать в виде онтологической модели устоявшиеся воззрения современной науки, основываясь на материалах научных публикаций.

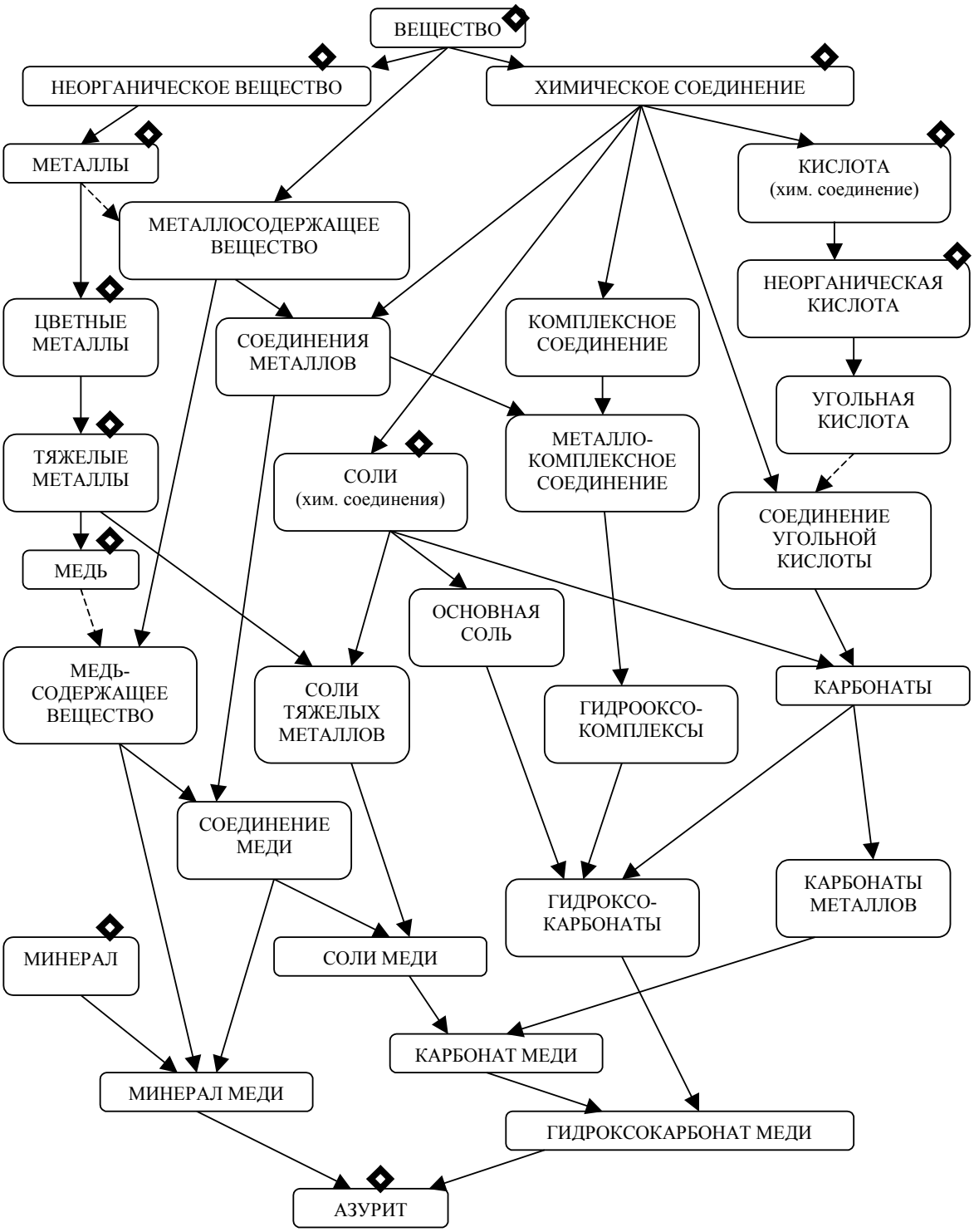

Рис.24.2. Фрагмент Онтологии по естественным наукам и технологиям

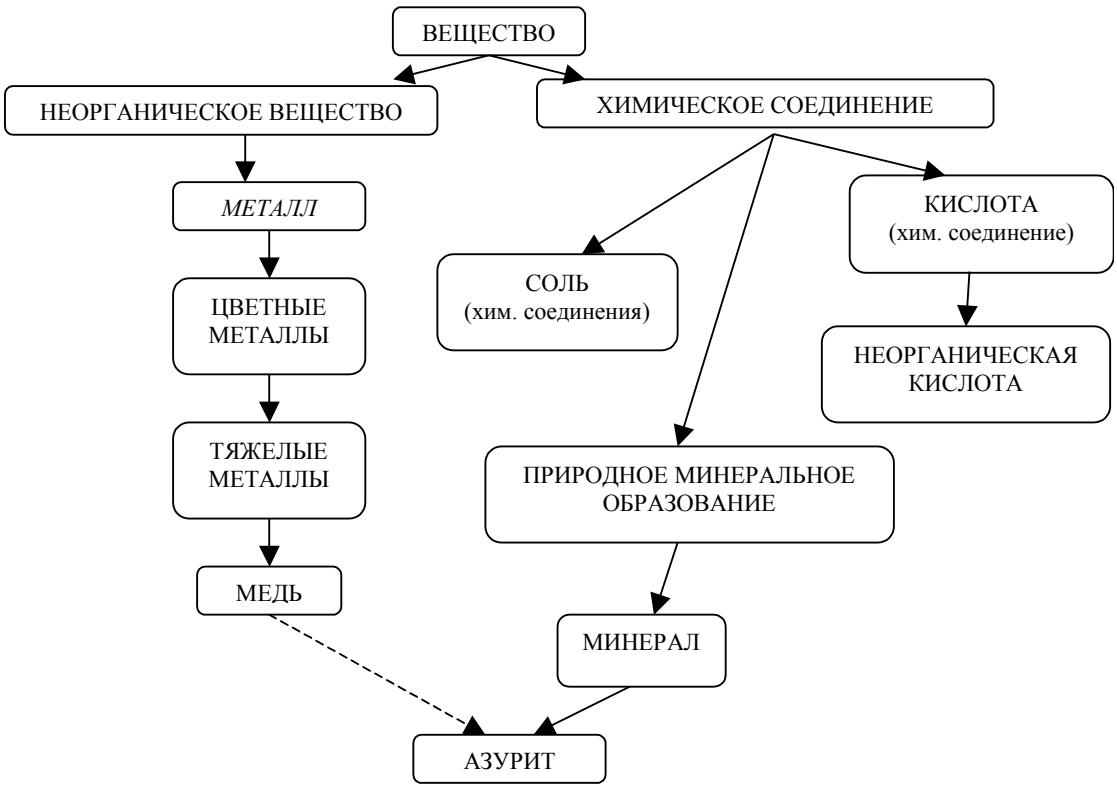

Рис.24.3. Фрагмент Тезауруса РуТез, аналогичный рис. 24.2.

24.4.6. Будущее развитие Онтологии ОЕНТ

Онтология ОЕНТ рассматривается нами как бесплатный ресурс для некоммерческого использования. Понятно, что небольшая группа исслдователей не может учесть все особенности употребления терминологии в такой широкой сфере научных исследований.

Поэтому мы предполагаем, что с некоторого момента онтология ОЕНТ может развиваться при поддержке начного сообщества (технологии Web 2.0), для чего в рамках проекта могут быть созданы специальные средства – сервисы анализа и коррекции понятийно-терминологической сети ОЕНТ, сервисы автоматической классификации и приписывание ключевых слов для пользовательских текстов, сервисы автоматического реферирования и автоматического расширения запросов к поисковым машинам.

Известно, что многие современные терминологические и лексические ресурсы развиваются благодаря усилиям и критике пользователей. Так была создана Википедия. Авторы известного ресурса - тезауруса английского языка Wordnet получали сотни писем с информацией о неточностях и ошибках. Такой ресурс как Онтология генов (Gene ontology) многократно улучшился в процессе обсуждения научным сообществом.

Самым простым методом взаимодействия с научным сообществом является обсуждение содержания и структуры ресурса в форумах и электронной почте. Следующим возможным уровнем взаимодействия является предоставление исследователям средств для описания фрагментов онтологии (пользовательской онтологии), которые можно было вставлять в объемлющую онтологию.

Но как показывает практика, таких простых методов взаимодействия недостаточно для создания терминологического ресурса, предназначенного для автоматической обработки текстов, поскольку:

- исследователь может употреблять термины из разных областей,

- имеется множество словосочетаний, которые непонятно относить к терминам или нет (как известно, имеются очень большие расхождения между экспертами в процессе разметки термин/нетермин),
- важность некоторого словосочетания, его терминологическая природа может быть выявлена не на одном тексте, а на некоторой текстовой коллекции.

Поэтому необходимо обеспечить ряд сервисов, связанных с автоматической обработкой научных текстов и помогающих пользователю определить, насколько его профессиональный язык отражен в онтологии ОЕНТ. В число этих сервисов должны входить такие сервисы как:

- автоматическое сопоставление научной публикации с онтологией ОЕНТ (проецирование ОЕНТ на текст публикации), с подсветкой найденных терминов и известных взаимосвязей, в том числе с учетом иерархии. Просмотрев результаты такой обработки текстов, исследователь может легко увидеть, что не отражены какие-то важные для его области термины;
- автоматическое сопоставление онтологии ОЕНТ с текстовой коллекцией, для составления частотного состава терминов ОЕНТ, употребляемых в этой коллекции, а также выявление частотного состава терминологических словосочетаний - то есть словосочетаний, которые не сопоставились с ОЕНТ, но могут рассматриваться как термины-кандидаты;
- автоматическая рубрикация научных публикаций по одному или нескольким научным рубрикаторам - если публикация относится автоматом к неправильной рубрике, то это означает, что, скорее всего, не учтены какие-то важные термины.

Заключение к главе 24

В этой главе мы рассмотрели использование модели представления знаний, использованной в тезаурусе РуТез, для формирования другого ресурса – Онтологии по естественным наукам и технологиям ОЕНТ.

По сочетанию отличительных особенностей, направленных на максимальную пригодность для задач информационного поиска, онтология ОЕНТ является уникальным ресурсом в научно-технической сфере.

Ожидается, что создание и применение такого ресурса может привести к оживлению научных исследований в области автоматического анализа научно-технической литературы, методов семантического поиска, ускорения обмена научно-технической информации.

Кроме того, ОЕНТ как свободно распространяемый ресурс может служить "образцом" для развития и тестирования методов извлечения знаний из текстов, которые могли бы автоматически извлекать новые понятия и отношения, вводимые в новых научных публикациях.

ЗАКЛЮЧЕНИЕ

В данной книге мы описали ряд широко известных онтологических ресурсов, рассмотрели алгоритмы их применения в различных задачах информационного поиска. Для каждого рассматриваемого алгоритма и системы были приведены данные по оценкам качества решения задач с использованием такого рода ресурсов.

С использованием лингвистических и онтологических ресурсов в решении задач информационного поиска часто связывается обсуждение возможности использования более смысловых, семантических, глубоких методов автоматической обработки текстов, чем при использовании пословных моделей обработки.

При этом в сообществе исследователей информационного поиска существует две противоположных точки зрения.

Большинство исследователей в этой сфере считает, что статистика по текстам и коллекциям и так прекрасно отражает и моделирует семантические явления.

Другие исследователи, обычно пришедшие в эту сферу из других областей компьютерной науки, компьютерной лингвистики, считают, что создание и применение ресурсов могло бы кардинально улучшить информационный поиск. Авторы книги неоднократно слышали на различных конференциях, обсуждениях, высказывания, что глобальным поисковикам специально не занимаются применениям онтологических ресурсов, чуть ли не о заговоре по отношению к этим ресурсам.

На основе рассмотрения особенностей существующих онтологических ресурсов, экспериментов и систем, применяющих их для решения задач информационного поиска, а также на основе создания наших собственных ресурсов и реального применения их в решении практических задач, мы хотели бы сделать следующие выводы.

Во-первых, мы считаем, что невозможно создать такой онтологический ресурс, применение которого в задачах информационного поиска, давало резкое преимущество по сравнению с существующими статистическими пословными методами. Это связано с тем, что любой ресурс всегда неполон, всегда недостаточно настроен на коллекцию.

Однако применение комбинированных методов, сочетающих лучшие современные статистические подходы с использованием знаний, описанных в ресурсах, может давать 10-15 процентов улучшения качества обработки текстов, которые уже невозможно достигнуть на текущем уровне развития статистических пословных методов.

Кроме того, в конкретных предметных областях для обработки поступающих документов, для создания информационно-аналитических систем специалистам необходимы онтологические ресурсы разных типов (рубрикаторы, тезаурусы, формальные онтологии), исследования в этих направлениях безусловно будут активно продолжаться.

Наши собственные решения и эксперименты в сфере разработки и использования онтологических ресурсов в сфере информационного поиска могут быть сформулированы следующим образом:

1) Предложена модель информационно-поискового тезауруса - лингвистической онтологии, предназначенного для автоматического индексирования текстов и автоматического расширения информационно-поискового запроса.

Модель построена на сочетации принципов трех различных традиций и методологий разработки компьютерных ресурсов:

- методологии разработки традиционных информационно-поисковых тезаурусов;
- методологии разработки лингвистических ресурсов типа WordNet (Принстонский университет);
- методологии созданий формальных онтологий.

Предложенная модель позволяет в короткие сроки создавать онтологические ресурсы в неструктурированных предметных областях. При этом созданный ресурс, с

одной стороны, будет содержать подробное описание терминологии предметной области, а также необходимые общелексические единицы, и, с другой стороны, будет иметь внутреннюю структуру, соответствующую современным онтологическим принципам разработки онтологий в виде отличимых понятий и формальных отношений между понятиями. Особенностью предлагаемой модели описания предметной области является то, что она построена с учетом эффективного применения в различных задачах информационного поиска, что показано в целом ряде экспериментов.

На основе этой модели построен Тезаурус русского языка РуТез, Онтология по естественным наукам и технологиям ОЕНТ и некоторые другие. Используемую модель построения ресурсов мы называем РуТез* Онтология.

Особенностью Тезауруса РуТез является также сочетание в одном лингвистическом ресурсе общеязыковых лексических единиц и терминов широкой предметной области современной общественной жизни.

2) Предложена модель разрешения лексической многозначности на основе тезаурусных знаний, сочетающая информацию о локальном и глобальном контексте употребления многозначного слова. Для задачи «все слова текста» результаты алгоритма сопоставимы с результатами лучших систем, достигаемых комбинированными методами с использованием семантически размеченных корпусов и информации о наиболее частотном значении. Для тематической лексики точность разрешения лексической многозначности достигает 75%.

3) Предложена модель лексической цепочки в форме тематического узла как проявление глобальной связности текста. Такая лексическая цепочка имеет следующие свойства:

- лексическая цепочка имеет внутреннюю структуру узла – к одному выделенному элементу относятся все другие элементы лексической цепочки,
- лексическая цепочка не должна содержать слова и словосочетания, которые часто встречались в одних и тех же предложениях текста с главным элементом этой цепочки;
- значимость цепочки для отражения содержания текста определяется не столько длиной, покрытием и другими характеристиками цепочки, а тем, насколько часто элементы этой цепочки встречались с элементами других цепочек в одних и тех же предложениях текста, то есть насколько много пропозиций конкретных предложений текста было посвящено обсуждению отношений между элементами лексических цепочек

4) Предложена и реализована модель тематического представления содержания текстов, учитывающая свойства глобальной тематической связности текста и лексической связности текста. Тематическое представление моделирует основное содержание текста посредством выделения тематических узлов – совокупностей близких по смыслу понятий текста. Выделяются основные тематические узлы, соответствующие основной теме документа и локальные тематические узлы, соответствующие подтемам документа. Построение тематического представления базируется на знаниях о понятиях и отношениях между ними, описанных в ресурсах типа РуТез* Онтология.

5) Предложена модель концептуального индексирования документов для информационно-поисковой системы, базирующаяся на знаниях тезауруса и построенном тематическом представлении документов. Концептуальный индекс по тезаурусу русского языка РуТез используется в Университетской информационной системе РОССИЯ (www.cir.ru).

6) Предложена модель автоматической рубрикации документов, основанная на использовании тематического представления документов и описании рубрик в виде булевских выражений над понятиями тезауруса и способная рубрицировать тексты различных типов (официальные документы, сообщения информационных агентств, газетные статьи). Модель можно легко настроить на новый рубрикатор и новые типы текстов, рубрицирование можно осуществлять сразу по нескольким рубрикаторам.

На основе предложенной модели было реализовано около 20 систем автоматической рубрикации текстов с количеством тематических рубрик от 35 до 3000. Возможности быстрой настройки системы рубрикации на новый рубрикатор и достигаемый при этом уровень качества рубрикации был продемонстрирован на семинаре по информационному поиску РОМИП-2007. Создание системы рубрикации заняло 8 часов, качество рубрикации было оценено как более чем 70% F-меры.

7) Предложена и реализована модель автоматического многошагового построения булевского выражения по длинному запросу на естественном языке, включающая расширение запроса по тезаурусным отношениям, подтвержденным поисковой выдачей.

8) Предложена модель автоматического аннотирования отдельного документа и совокупности тематически близких документов на базе выделения из текстов наиболее содержательных предложений. Модель базируется на тематическом представлении содержания текстов, что позволяет повысить связность создаваемой аннотации. Реализованная система автоматического аннотирования одного документа получила наилучший результат в одной из номинаций на конференции SUMMAC в 1998 году.

9) Предложена модель автоматического аннотирования новостного кластера на основе тематического представления кластера, моделировании лексической связности, что позволяет улучшить связность и полноту аннотации, а также снизить повторы.

Литература

Advances in Automatic Text Summarization. Ed: I. Mani, Inderjeet, Maybury, Mark T., The MIT Press Cambridge, Massachusetts, 1999.

Ageev M., Dobrov B., Loukachevitch N. Text Categorization Tasks for Large Hierarchial Systems of Categories. In Proceedings of SIGIR 2002 Workshop on Operational Text Classification Systems / Eds. F.Sebastiani, S.Dumas, D.D.Lewis, T.Montgomery, I.Moulinier — Univ. of Tampere. 2002 — p.49-52.

Ageev M., Dobrov B. Support Vector Machine Parameter Optimization for Text Categorization Problems. In Proceedings of Information Systems Technology and its Applications (ISTA'2003).Vol 30. 2003. – pp. 165-176.

Agirre E., Rigau G. A Proposal for Word Sense Disambiguation using Conceptual Distance. - In : Proceedings of the First International Conference on Recent Advances in NLP. -- Tzigov Chark, Bulgaria. 1995.

Agirre E., Rigau G. Word Sense Disambiguation Using Conceptual Density. In Proceedings of COLING'96, Copenhagen, Danmark. 1996. – pp.16 – 22.

Agirre E., Lacalle Lopez O. Clustering Wordnet word senses. In Proceedings of RANLP 2003. 2003.

Agirre E., Aldezabal I., Pociello E. Lexicalization and multiword expressions in the Basque WordNet. Proceedings of Third International WordNet Conference. ISBN 80-210-3915-9. Jeju Island (Korea). 2006. – pp. 131-138.

Agirre E., Magnini B., Lacalle O., Otegi A., Rigau G., Vossen P. SemEval –2007 Task 01: Evaluating WSD on Cross-Language Information Retrieval. *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), in conjunction with ACL*. 2007.

AGROVOC Multilingual Agricultural Thesaurus. Fourth Edition. 1999.

Aitchinson Y., Gilchrist A. Thesaurus construction: a practical manual. – 2nd ed. – L.:Aslib, 1987.

Alonge, A., N. Calzolari, P. Vossen, L. Bloksma, I. Castellon, T. Marti, W. Peters. The Linguistic Design of the EuroWordNet Database. In: Nancy Ide, Daniel Greenstein, Piek Vossen (eds), Special Issue on EuroWordNet. Computers and the Humanities, Volume 32, Nos. 2-3. 1998. - pp. 91-115.

Art and Architecture Thesaurus. Second Edition. Toni Petersen, Director. New York: Oxford University Press, 1994. 5 vols.

Artale A., Franconi E., Guarino N., Pazzi L. Part-Whole Relations in Object-Centered Systems: An Overview. // Data and Knowledge Engineering. – V.20. 1996. – pp.347-383.

Asmussen J., Pedersen B., Trap-Jensen L. DanNet: from Dictionary to WordNet. In: Informatik Berichte 336 – 3/2007. GLDV-2007 Workshop: Lexical-Semantic and Ontological Resources. 2007. – pp . 1-10.

Atkins S. Building a lexicon: The contribution of lexicography. In: International Journal of Lexicography 3. 1993. – pp. 167-204.

Atserias J., Climent S., Rigau G. Toward the Meaning Top Ontology: Sources of Ontological Meaning. - Proceedings of International conference Language Resources and Evaluation (LREC-2004). – 2004. - v.1, p. 11-14.

Barzilay R., Elhadad M. Text summarizations with lexical chains / Inderjeet Mani and Mark Maybury, eds,Advances in Automatic Text Summarization. MIT Press, 1999.

Bates M. How to use controlled vocabularies more effectively in online searching. – Online archive V.12, Issue 6, 1988. - pp.45-56.

Bentivogli L., Pianta E. Beyond Lexical Units: Enriching WordNets with Phrasets. In proceedings of EACL-03, Budapest, Hungary. 2003.

Bentivogli L., Pianta E. Extending WordNet with Syntagmatic Information, *In Proceedings of International Wordnet Conference (GWC – 2004)*. – 2004. – pp. 47-53.

Bentivogli L., Bocco A., Pianta E. ArchiWordNet: Integrating WordNet with Domain-Specific Knowledge. In *Proceedings of the Second Global WordNet Conference*, Brno, Czech Republic. 2004. - pp. 39-46.

Bodenreider O, Smith B, Kumar A, Burgun A. Investigating subsumption in DL-based terminologies: A case study in SNOMED CT. In: Hahn U, Schulz S, Cornet R, editors. Proceedings of the First International Workshop on Formal Biomedical Knowledge Representation (KR-MED 2004). 2004. - pp. 12-20.

Bolshakova E. Recognition of Author's Scientific and Technical Terms. In: Computational Linguistics and Intelligent Text Processing. A. Gelbukh (Ed.). Lecture Notes in Computer Science, N 2004, Springer-Verlag. 2001.

Bouaud J., Bachimont B., Charlet J., Zweigenbaum P. Methodological principles for structuring an "ontology". Proceedings of IJCAI-95 Workshop "Basic Ontological Issues in Knowledge Sharing". 1995.

Brewster Ch., Iria J., Ciravegna F., Wilks Y. The Ontology: Chimaera or Pegasus, presented at the Dagstuhl Seminar Machine Learning for the Semantic Web, 2005

Brown G., Yule G. Discourse analysis. – Cambridge University Press, 2001.

Brunn M., Chali Y., Pinchak C. Text Summarization Using Lexical Chains. In the Proceedings of the Document Understanding Conference (DUC-2001). 2001. – pp.135-140.

Buckley C., Allan J., Salton J. Automatic Routing and Ad-hoc Retrieval Using Smart: TREC 2. In Proceedings of the Second Text Retrieval Conference. NIST Special Publication 500-215. 1993. – pp. 45-56.

Budanitsky A. Lexical Semantic Relatedness and its application in Natural Language Processing. PhD Thesis. – Technical Report CSRG-390, Computer Systems Research Group, University of Toronto. 1999.

Buenaga Rodriguez M., Gomez-Hidalgo, J., Diaz-Agudo B. 1997 Using WordNet to complement training information in text categorization // In Proceedings of the 2nd International Conference on Recent Advances in Natural Language Processing (RANLP 1997), Bulgaria. 1997 - pp. 150-157.

Builellar, P. Corelex: Systematic Polysemy and Underspecification. Ph.D. Department of Computer Science, Brandeis University, Boston, USA. 1998.

Buitellar, P., Sacalenau, B. Extending Synsets into Medical Terms. // Proceedings of the NAACL workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations, Pittsburg, USA. 2001.

Burgun A, Bodenreider O, Aubry M, Mosser J. Dependence relations in Gene Ontology: A preliminary study. Workshop on The Formal Architecture of the Gene Ontology - Leipzig, Germany, May 28-29. 2004.

Callan J.P., Croft W.B., Harding S.M. The INQUERY Retrieval System // A.M. Tjoa and I. Ramos (eds.), Database and Expert System Applications. Proceedings of {DEXA}-92, 3rd International Conference on Database and Expert Systems Applications. - Springer Verlag, New York. 1992. - pp.78-93.

Calzolari N., Fillmore Ch., Grishman R, Ide N., Lenci A., MacLeod C., Zampolli A. Towards Best Practice for Multiword Expressions in Computational Lexicons. // Proceedings of LREC. 2002 - pp.1934-1940

Carbonell J., Goldstein J. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In proceedings of the 21st Annual International ACM SIGIR Conference. 1998. - pp.335-336.

Cao G., Nie J., Bai J. Integrating Word Relationships into Language Models. In Proceedings of SIGIR-2005. 2005. - pp.298-305.

Carlson L., Marcu D., Okurowski M. 2003. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In Jan van Kuppevelt and Ronnie Smith, editors, Current directions in Discourse and Dialog, Kluwer Academic Publishers. 2003. - pp.85-112.

Castillo M., Real F., Rigau G. Automatic Assignment of Domain Labels to WordNet. - In Proceedings of International Wordnet Conference (GWC – 2004). – 2004. – pp. 75-82.

Chavez N., Pfeiffer H., Hartley R. Using and Interfacing Background Knowledge in Story Understanding. In Proceedings of SENSE-09 Workshop on Conceptual Structures for Extracting Natural Language Semantics – 2009.

Chen, H., Lynch, K. J., Basu, K., Ng, T. D. Generating, integrating, *and activating thesauri for concept-based document retrieval*. IEEE Expert.1993. – pp. 25--34.

Chen S.F., Goodman J. An empirical study of smoothing techniques for language modeling. Tech.Rep. TR-10-98, Harvard University. 1998.

Chugur I., Gonzalo J., Verdejo F. A study of sense clustering criteria for information retrieval applications. In Proceedings of OntoLex 2000. 2000.

Chugur I., Gonzalo J., Verdejo F. Polysemy and sense proximity in the Senseval-2 Test Suite. In Proceedings of the ACL-2002 Workshop on "Word sense Disambigation:recent successes and future directions". 2002.

Cimiano, P., Stumme, G., Hotho, A., Tane, J.: Conceptual Knowledge Processing with Formal Concept Analysis and Ontologies. In: Eklund P.W. (ed.) Proceedings of the The Second International Conference on Formal Concept Analysis (ICFCA 04). LNCS, vol. 2961, pp.189--207. Springer (2004)

Clark P., Fellbaum Ch., Hobbs J. Using and Extending WordNet to Support Question-Answering. – In Proc. Fourth Global WordNet Conference (GWC'08), Hungary: University of Szeged, 2008. – pp. 111-119

Climent S., Rodriguez H., Gonzalo J., Definitions of the links and subsets for nouns of the EuroWordNet project. - Deliverable D005, WP3.1, EuroWordNet, LE2-4003. 1996.

Corcho O, Gomez-Perez A. A Roadmap to Ontology Specification Languages / Rose Dieng and Oliver Corby (eds). Knowledge Engineering nad Knowledge Management. Methods, Models and Tools. Springer: 2000, 80-96.

Cristea D., Ide N., Romary L. Veins Theory: A Model of Global Discourse Cohesion and Coherence. In Proceedings of Seventeenth Conference of Computational Lingustics (COLING 1998). 1998 - pp.281-285.

Croft B., Metzler D., Strohman T. Search Engines: Information Retrieval in Practice. Addison Wesley, 2009.

Cruse D. Lexical Semantics. Cambridge. University Press. 1986.

Cruse D. Hyponymy and its varieties. In: The Semantics of Relationships: An interdisciplinary Perspective, Information science and Knowledge Management. Springer Verlag. 2002

Cyc Ontology Guide: Introduction. *(http://www.cyc.com/cyc-2-1/intro-public.html).*

Dang H.T. Overview of DUC 2006. National Institute of Standards and Technology (NIST). 2006. http://www-nlpir.nist.gov/projects/duc/pubs/2006papers/duc2006.pdf

Debole F., Sebastiani F. An Analysis of the Relative Hardness of Reuters-21578 Subsets // Journal of the American Society for Information Science and Technology, 2004.

Dijk van T. Semantic Discourse Analysis. In: Teun A. van Dijk, (Ed.) *Handbook of Discourse Analysis*, vol. 2. London: Academic Press. 1985. – pp. 103-136.

Doran W., Stokes N., Carty J., Dunnion J. Assessing the Impact of Lexical Chain Scoring Methods and Sentence Extraction Schemes on Summarization. – Proceedins of CICLING- 2004. 2004. – pp. 627-635.

Dumais S., Platt J, Heckerman D., Sahami M. Inductive learning algorithms and representations for text categorization. In Proc. Int. Conf. on Inform. and Knowledge Management. 1998. – pp. 148 - 155.

Dumais S., Lewis D., Sebastiani F. Report on the Workshop on Operational Text Classification Systems (OTC-02) // SIGIR-2002 — Tampere, Finland, 2002

Edmonds Ph., Hirst G. Reconciling fine grained lexical knowledge and coarse-grained ontologies in representation of near-synonyms. Proceedings of workshop on Semantic Approximation, Granularity and Vagueness, Breckenridge, Colorado. 2002

EUROVOC. Информационно-поисковый тезаурус. Русская версия тезауруса EUROVOC. Том 1. Алфавитно-пермутационное представление. – М.: Издание Государственной Думы, 2001. – 500 стр.

ERIC. Thesaurus of ERIC. Descriptors. 12th Edition. James E. Houston, Ed. Phoenix, AZ, Oryx Press, 1990.

Fan J., Barker K., Porter B, Clark P. Representing Roles and Purpose // Proceedings of 1st Int Conference on Knowledge Capture (K-Cap'01). 2001. – pp. 38-43.

Farreres X., Rigau G., Rodriguez G. Using WordNet for Building WordNets. Use of WordNet in Natural Language Processing Systems: Proceedings of the Conference. 1998. – pp . 65-72.

Felber H. Terminology Manual. – Unesco, Infoterm, 1984. – 426p.

Fellbaum Ch. A Semantic Network of English Verbs. - In: Fellbaum, C (ed) WordNet – An Electronic Lexical Database. – The MIT Press. 1998. – pp. 69-104.

Fellbaum Ch. Parallel Hierarchies in the Verb Lexicon. In Proceedings of 'The Ontologies and Lexical Knowledge bases' workshop (OntoLex 2002). 2002.

Fellbaum Ch., Miller G. Whither WordNet. Presentation of winners of Antonio Zampolli Award. LREC-2006. (http://www.lrec-conf.org/lrec2006/article.php3?id_article=45)

Fillmore C., Atkins B. Describing polysemy: the case of *crawl*. In Ravin, Y., Leacock C., editors, Polysemy: Linguistic and Computational Approaches. Oxford Univercity Press, Oxford. 2006

Fillmore C.J., Miriam R.L. Petruck J.R., Abby W. Framenet in Action: The Case of Attaching, International Journal of Lexicography, 2003. Vol 16.3. – pp. 297-332.

Fine K. Ontological Dependence. Proceedings of the Aristotelian Society 95. 1995. – pp. 269-290.

Fox M.S., Gruningen M. On Ontologies and Enterprise modeling. In Proceedings of International Conference "Enterprise Integration Modeling Technology, 1997.

French J., Powell A., Gey F., Perelman N. Exploiting Manual Indexing to Improve Collection Selection and Retrieval Effectiveness. In Information Retrieval, v.5, n.4, p.323-351.

Galley M., McKeown K. Improving word sense disambiguation in lexical chaining. In Proceedings of IJCAI 2003. 2003

(Gangemi и др., 2001a) Gangemi A, Guarino N., Oltramari A. Conceptual analysis of lexical taxonomies:the case of wordnet top-level. In Proceedings of the international conference on Formal Ontology in Information Systems. ACM Press. 2001.

(Gangemi и др., 2001b) Gangemi A., Guarino N., Masolo C., Oltramari A., Understanding Top-Level Ontological Distinctions. In Proceedings of IJCAI 2001 Workshop on Ontologies and Information Sharing. – Seattle. 2001. – pp. 26-33.

Gangemi A., Navigli R., Velardi P. The OntoWordNet project: extension and axiomatisation of conceptual relations in Wordnet. International Conference on Ontologies, Databases and Applications of Semantics (ODBASE), Catania (Italy). 2003.

Gene Ontology. An Introduction to Gene Ontology. Код доступа: http://www.geneontology.org/GO.doc.shtml.

Gerstl P., Pribennow S. A conceptual theory of part-whole relations and its applications // Data and Knowledge Engineering. 1996. – V.20. – pp. 305-322.

Gomez-Perez A., Fernandez-Lopez M., Corcho O. OntoWeb. Technical Roadmap. D.1.1.2. - IST project IST-2000-29243, 2001. (*http://www.aifb.uni-karlsruhe.de/WBS/ysu/publications/OntoWeb_Del_1-1-2.pdf*)

Gomez-Perez A., Corcho O, Fernandez-Lopez M. Ontological Engineering: with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web. First Edition (Advanced Information and Knowledge Processing). Springer-Verlag. 2004.

Gonzalo J. Sense Proximity versus Sense Relations. In Proceedings of International Wordnet Conference (-GWC – 2004). 2004. – pp. 5-6.

Gonzalo J., Verdejo F., Chugur I., Cigarrán J. Indexing with WordNet synsets can improve text retrieval. Proceedings of the COLING/ACL '98 Workshop on Usage of WordNet for NLP. 1998.

Grenon P. Spatio-temporality in Basic Formal Ontology: SNAP and SPAN, Upper-Level Ontology, and Framework for Formalization: PART I. IFOMIS Report 05/2003, Institute for Formal Ontology and Medical Information Science (IFOMIS), University of Leipzig, Leipzig, Germany. 2003.

Gruber T.R. A translation approach to portable ontologies. Knowledge Acquisition, 5(2). 1993. – pp. 199-220.

Guarino N., Giaretta P. Ontologies and Knowledge Bases: Towards a Terminological Clarification. In N. Mars (ed.) Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing 1995. IOS Press, Amsterdam. 1995. – pp. 25-32.

Guarino N. Formal Ontology and Information Systems. In N. Guarino, editor, Proceedings of the 1st International Conference on Formal Ontologies in Information Systems, FOIS'98, Trento, Italy, IOS Press. 1998. – pp. 3-15.

Guarino N. Some Ontological Principles for Designing Upper Level Lexical Resources. In Proceedings of First International Conference on Language Resources and Evaluation. Granada, Spain, 1998.

Guarino N., Welty C. Ontological Analysis of Taxonomic Relationships. In Proceedings of ER-2000. The international conference of Conceptual Modeling. Springer Verlag. 2000.

Guarino N., Welty C. Evaluating ontological decisions with ONTOCLEAN // Communications of the ACM, 45(2). 2002. – pp. 61-65.

UNBIS Guidelines. Guidelines for Subject Analysis of UN documents and Publications. Код доступа: http://www.un.org/Depts/dhl/unbisref_manual/indexpolicy/guidelines.htm

Halliday M., Hasan R. Cohesion in English. - Longman, London. 1976.

Harabagiu S., Moldovan D., Pasca M., Mihalcea R., Surdeanu M., Bunescu R., Girju R., Rus V., Morarescu P. FALCON: Boosting Knowledge for Answer Engines. In Proceedings of TREC-9. 2000.

Harman D. Towards Query-Specific Customization of IR systems. ACM SIGIR 2005 workshop Predicting Query Difficulty. 2005. Способ доступа http://www.haifa.ibm.com/sigir05-qp/papers/Harman.pdf.

Harnly A., Nenkova A., Passonneau R., Rambow O. Automation of summary evaluation by the pyramid method. Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'2005). - Borovets, Bulgaria, 2005.

Hasan R. Coherence and Cohesive harmony. In J. Flood, editor, Understanding reading comprehension, Newark, DE: IRA. 1984. – pp. 181-219.

Hayes Ph. Intelligent High-Volume Processing Using Shallow, Domain-Specific Techniques. Text-Based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval. New Jersey. 1992. – pp. 227-242.

Hepp M. Possible Ontologies: How Reality Constrains the Development of Relevant Ontologies. IEEE Internet Computing, Vol. 11, No. 1. 2007. – pp. 90-96.

Hirst G., St-Onge D. Lexical Chains as representation of context for the detection and correction malapropisms. In C. Fellbaum, editor, WordNet: An electronic lexical database and some of its applications. Cambrige, MA: The MIT Press, 1998.

Hirst G. Ontology and the Lexicon. - Handbook on Ontologies in Information Systems, Berlin – Springer, 2003.

Hirst G., Morris J. The subjectivity of Lexical Cohesion in Text. In James C. Chanahan, Yan Qu, and Janyce Wiebe, editors, Computing attitude and affect in text. Springer, Dodrecht, The Netherlands.". 2005. pp. 41–48.

Hlava M., Hainebach R. Multilingual Machine Indexing. - Proceedings of The Ninth International Conference on New Information Technology, Pretoria, South Africa, November 11-14, 1996. – pp. 105-120.

Hollingsworth W., Teufel. S. Human Annotation of Lexical Chains: Coverage and Agreement Measures. In: Workshop proceedings ``ELECTRA: Methodologies and Evaluation of Lexical Cohesion Techniques in Real-world Applications", SIGIR 2005, Salvador, Brazil. 2005.

Hotho A., Bloehdorn S. Boosting for Text Classification with semantic features, in Proceedings of the Workshop on Mining for and from the Semantic Web at the 10th International Conference on Knowledge Discovery and Data Mining (KDD-2004). 2004. – pp.70-87.

Hovy E., Nirenburg S. Approximating an interlingua in a principled way. Proceedings of the DARPA Speech and Natural Language Workshop, Hawthorne, NY. 1992.

Hovy E., Maier E. Parsimoniuus or profligate: How many and which discourse relations? Technical report, University of Southern California. 1995.

Hovy E.H. Combining and standardizing large-scale, practical ontologies for machine translation and other uses. In Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC). Granada, Spain. 1998.

Hovy E., Hermjakob U., Lin C.-Y. The use of external knowledge in factoid QA. – In Proceeding 10th Text Retrieval Conference (TREC 2001). 2001.

Hovy E. Methodologies for the Reliable Construction of Ontological Knowledge. – B F.Dau, M.-L. Mugnier, G. Stumme (ред.). Conceptual Structures for Sharing Knowledge. In Proceedings of 13 annual conference «Conceptual Structures (ICCS 2005). Springer Lecture Notes in AI – 3596. 2005. - pp.91-106.

Hovy E., Marcus M., Palmer M. Ramshaw L., Weischedel R. OntoNotes: The 90% Solution. Proceedings of HLT-NAACL-2006. 2006.

ISO 2788-1986 - Guidelines for the establishment and development of monolingual thesauri.

Jacobs P. S., Rau L. F. SCISSOR: extracting information from on-line news, Communications of the ACM 33. 1990. – pp. 88 - 97.

Jacquemin C., Tzoukermann E. NLP for term variant extraction: Synergy of morphology, leaxicon and syntax. - In T. Strzalkowski (Ed.) Natural language Information Retrieval, p.25-74. Boston, MA, Kluwer. 1999.

Jarmasz M., Szpakowicz S. Not As Easy As It Seems: Automating the Construction of Lexical Chains Using Roget's Thesaurus. *Proceedings of the 16th Canadian Conference on Artificial Intelligence (AI 2003)*, Halifax, Canada, June. 2003. – pp. 544-549.

Jensen L., Martinez T. Improving text classification by using coceptual and contextual features // In Proceedings of the Workshop on Text Mining at the 6th ACM SIGKDD Int. Conference on Knowledge Discovery and Data Mining (KDD 00). 2000. pp. 101-102.

Jeon J., Croft B., Lee J.H. Finding Similar Questions in Large Question and Answer Archives. In Proceedings CIKM 2005. 2005. – pp. 84-90.

Jiang J., Conrath D. Semantic similarity based on corpus statistics and lexical taxonomy / J. Jiang, D. Conrath. In Proceedings of COLING 1997. 1997.

Joachims T., Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In Proceedings of ECML-98, 10th European Conference on Machine Learning, 1998.

Johansson J. On the Transitivity of the Parthood Relations // Relations and Predicates. – Frankfurt: Ontos Verlag. 2004. – pp. 161-181.

Jones S. A Thesaurus Data Model for an Intelligent Retrieval System. *Journal of Information Science* 19. 1993. – pp. 167-178.

Kehagias A., Petridis V., Kaburlasos V., Fragkou P. A comparison of word- and sense-based text classification using several classification algorithms. – Journal of Intelligent Information Systems 21(3), 2003. – pp. 227-247.

Kennedy A., Szpakowicz, S. Evaluating Roget's Thesauri. *Proc of ACL-08: HLT.*, Columbus Ohio, USA Association for Computational Linguistics. 2008. – pp. 416-424.

Kilgarriff A., Yallop C. What's in a thesaurus? // *Proc. Second Intnl Conf on Language Resources and Evaluation.* Athens, Greece. 2000. – pp. 1371-1379.

Kilgarriff A., Rosenzweig J. Framework and Results for English Senseval. In A. Kilgarriff, J. Rosenzweig. Computers and the Humanities, V34, 2000. - pp.15–48.

Kluck M. GIRT Data in the Evaluation of CLIR Systems – from 1997 until 2003. In Comparative Evaluation of Multilingual Information Access Systems: 4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003, edited by C. Peters. Lecture Notes in Computer Science 3237, Springer. 2003.

Kumar A., Smith B. The ontology of blood pressure: a case study in creating ontological partitions in biomedicine. 2004.

Kunze C., Wagner A. Integrating GermaNet into EuroWordNet, a multilingual lexical-semantic database. In: Sprache und Datenverarbeitung - International Journal for Language Data Processing. Bonn. 1999.

Kunze C., Naumann K. GermaNet homepage. http://www.sfs.uni-tuebingen.de/lsd

Kupiec J. MURAX: a Robust Linguistic Approach for Question Answering Using On-line Encyclopedia. In proceedings of SIGIR-1993. 1993. – pp. 181-190.

Kuznetsov, S., Obiedkov, S., Roth C.: Reducing the Representation Complexity of Lattice-Based Taxonomies, in Proc. 15th International Conference on Conceptual Structures (ICCS'07), U. Priss, S. Polovina, R. Hill (Eds.), LNAI, vol. 4604, pp. 241--254. Springer (2007)

Landes S., Leacock C., Tengi R.I. Building semantic concordances". In Fellbaum, C. (ed.) WordNet: An Electronic Lexical Database. Cambridge (Mass.): The MIT Press. 1998.

Lassila O., McGuinness D. The Role of Frame-Based Representation on the Semantic Web. Knowledge Systems Laboratory Report KSL-01-02, Stanford University. 2001

Leacock C., Chodorow M. Combining local context and WordNet similarity for word sense identification. In WordNet: An electronic lexical database – The MIT Press. 1998.

Lenat D., Miller G., Yokoi T. CYC, WordNet, and EDR: critiques and responses. - Communications of the ACM. - Volume 38, Issue 11. 1995. – pp. 45 - 48.

Lenci A., Bel. N., Busa F., Calzolari N., Gola E., Monachini M., Ogonowski A., Peters I., Peters W., Ruimy W., Villegas M., Zampolli A. SIMPLE – A General Framework for the Development of Multilingual Lexicons. In: T. Fonetenelle (ed.). International Journal of Lexicography, Vol.13, Oxford University Press. 2000. –pp. 249-263.

Lesk M. Automatic sense disambiguation using machine-readable dictionaries: How to tell a pine cone from an ice cream cone. SIGDOC, Proceedings of the 5th annual international conference on Systems documentation. 1986. – pp . 24 – 26.

Lewis D. Applying Support Vector Machines to the TREC-2001 Batch Filtering and Routing Tasks. Proceedings of TREC-2001 conference. 2001.

Lewis D. Reuters-21578 text categorization test collection. Distribution 1.0 http://www.daviddlewis.com/resources/testcollections/reuters21578/readme.txt

Lewis D., Sebastiani F. Report on the Workshop on Operational Text Classification Systems (OTC-01) // SIGIR-2001 — New Orleans. 2001

Li J., Sun L., Kit C., Webster J. A Query-Focused Multi-Document Summarizer Based on Lexical Chains. In Proceedings of the Document Understanding Conference DUC-2007. 2007.

Li S., Ouyang Y., Sun B., Peking University, Z. Guo, IBM "Peking University at DUC 2006". In Proceedings of DUC-2006. 2006.

Liang A., Lauser B., Sini M., Keizer J., Katz S. From AGROVOC to the agricultural ontology service/concept server: An OWL model for managing ontologies in the agricultural domain // In Proceedings of OWL: Experiences and Directions Workshop. 2006.

Liddy E.D., Diekema A.R., Yilmazel O., Chen J., Harwell S., He L. Finding Answers to Complex Questions. In Maybury, M. (Ed.) New Directions in Question Answering. 2004. – pp . 141-152.

LIV (Legislative Indexing Vocabulary*)*. Congressional Research Service. The Library of Congress. Twenty-first Edition. 1994.

Lin D. An information theoretic definition of similarity. ICML. 1998.

Lin Chin-Yew. ROUGE: a Package for Automatic Evaluation of Summaries. In Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004), Barcelona, Spain. 2004.

Liu Sh., Liu F., Yu C., Meng W. An effective approach to document retrieval via utilizing WordNet and recognizing phrases. In the Proceedings of SIGIR-2004. 2004. – pp. 266-272.

Loebe F. An Analysis of Roles: Towards Ontology-Based Modelling. Onto-Med Report No. 6. Research Group Ontologies in Medicine (Onto-Med), University of Leipzig. 2003.

Loebe F. Abstract vs. Social Roles: A Refined Top-level Ontological Analysis. In Proceedings of the 2005 AAAI Fall Symposium 'Roles, an Interdisciplinary Perspective: Ontologies, Languages, and Multiagent Systems / Guido Boella, James Odell, Leendert van der Torre and Harko Verhagen (ed.). AAAI Press. 2005. pp.93–100.

Loukachevitch N. Text Summarization Based on Thematic Representation of Texts // *AAAI'98 Spring Symposium on Intelligent Text Summarization*. 1998.

Loukachevitch N., Dobrov B. Thesaurus-Based Structural Thematic Summary in Multilingual Information Systems - *Machne Translation Review*, - N 11, December 2000. – pp. 10-20. (http://www.bcs.org.uk/siggroup/nalatran/mtreview/mtr-11/mtr-11-8.htm). 2000a.

Loukachevitch N., Dobrov B., Thesaurus as a Tool for Automatic Detection of Lexical Cohesion in Texts. In Proceedings of 5th JADT. 2000. – pp. 155-162. 2000b.

Loukachevitch N., Dobrov B., Development and Use of Thesaurus of Russian Language RuThes. In Proceedings of workshop on WordNet Structures and Standartisation, and How These Affect WordNet Applications and Evaluation. (LREC2002) / Dimitris N. Christodoulakis - 2002, Gran Canaria, Spain. 2002. – pp. 65-70.

Loukachevitch N., Dobrov B. Development of Ontologies with Minimal Set of Conceptual Relations. In Proceedings of Fourth International Conference on Language Resources and Evaluation / Eds: M.T.Lino и др., – vol. VI. 2004. – pp.1889-1892. 2004a.

Loukachevitch N., Dobrov B. Development of Bilingual Domain-Specific Ontology for Automatic Conceptual Indexing. In Proceedings of Fourth International Conference on Language Resources and Evaluation / Eds: M.T.Lino и др., – vol. VI. 2004. – pp. 1993-1996. 2004b.

Loukachevitch N., Dobrov B. Ontological Types of Associative Relations in Information Retrieval Thesauri and Automatic Query Expansion. In Proceedings of OntoLex 2004: Ontologies and Lexical Resources in Distributed Environments / Eds: A.Oltramari и др., 2004. – pp. 24-29. 2004c.

Loukachevitch N., Dobrov B. Sociopolitical Domain as a Bridge from General Words to Terms of Specific Domains. In Proceedings of Second International WordNet Conference GWC-2004. 2004. – pp.163-168. 2004d.

Loukachevitch N. Concept Formation in Linguistic Ontologies. Conceptual Structures: Leveraging Semantic Technologies. Proceedings of ICCS-2009. Eds Sebasian Rudolph, Frithjof Dau, Sergei O. Kuznetsov, Springer Verlag, LNAI-5662, pp. 2-22. 2009a.

Loukachevitch Natalia. Multigraph representation for lexical chaining. In Proceedings of SENSE workshop, 2009. pp. 67-76. 2009b

Lowe E.J. Ontological dependence / Stanford encyclopedia of Philosophy. 2005. (http://plato.stanford.edu/entries/dependence-ontological/)

Magnini B., Cavaglia G. Integrating Subject Field Codes into WordNet. – In proceeding of the Second International Conference on Language Resources and Evaluation LREC 2000, Athens, Greece. 2000.

Magnini B., Speranza M. Merging Global and Specialized Linguistic Ontologies. – In Proceedings of OntoLex 2002. 2002.

Mahesh K., Nirenburg S. A Situated Ontology for Practical NLP. In Procedings Workshop on Basic Ontological Issues in Knowledge Sharing, International Joint Conference on Artificial Intelligence (IJCAI-95), Montreal, Canada. 1995.

Mann W.C., Thompson S.A. Rhetorical Structure Theory: Description and Construction of Text Structures. Natural Language Generation. 1987.

Manning Ch., Shutze H. Foundations of Statistical Natural Language Processing, MIT Press. Cambridge, MA. 1999.

Manning Ch., Raghavan P., Shutze H. Introducion to Information Retrieval. – Cambridge University Press. 2008 (http://www-csli.stanford.edu/~hinrich/information-retrieval-book.html)

Mansuy T., Hilderman R. A characterization of WordNet Features in Boolean Models for Text Categorization. In proceedings Australasian Data Mining Conference (AusDM-2006), vol. 61. 2006. – pp. 103-109.

Marcu D. Rhetorical Parsing of Unrestricted Text: A surface-based Approach - *Computational Linguistics*, 26 (3). 2000. – pp. 395-448.

Marinelli R., Tiberi M., Bindi R. Encoding Terms from a Scientific Domain in a Terminological Database: Methodology and Criteria . - Proceedings of the Sixth International Language Resources and Evaluation (LREC'08). 2008.

Margolis E., Laurence S. Concepts. Stanford Encyclopedia of Philosophy. – 2006. Код доступа http://plato.stanford.edu/entries/concepts/#ClaThe.

Masolo C., Vieu L., Bottazzi E. Catenacci C., Ferrario R., Gangemi A., Guarino N. Social roles and their descriptions. In Proceedings of the Ninth International Conference on the Principles of Knowledge Representation and Reasoning. AAAI Press. 2004.

Masolo C., Borgo S., Gangemi A., Guarino N., Oltramari A., Shneider L. WonderWeb. Final Report. Deliverable D18. 2003.

Mauldin M. Retrieval performance in Ferret a conceptual information retrieval system. – In Proceedings of 14th SIGIR Conference. 1991. – pp . 347 – 355.

McCarthy D. Relating WordNet Senses for Word Sense Disambiguation. In Proceedings of NACCL Wordkshop on Making Senses of Sense. 2006.

McCarthy J., Hayes P. Some philosophical problems from the standpoint of artificial intelligence. In B. Meltzer and D. Michie, editors, Machine intelligence, volume 4, Edinburgh University press. 1969. – pp.463-502.

McShane M., Zabludowski M., Nirenburg S., Beale S. OntoSem and SIMPLE: Two multi-lingual World Views. - Proceedings of the Second Workshop on Text Meaning and Interpretation: ACL 2004. 2004. – pp.25-32.

Medelyan O. Computing Lexical Chains with Graph Clustering. – Proceedings of the ACL 2007 Student Research Workshop. 2007. – pp. 85-90.

Medical Subject Headings: Annotated Alphabetic List, 1992. Bethesda, MD: National Library of Medicine, 1992.

Mihalcea R., Tarau P., Figa E. PageRank on Semantic Networks, with application to Word Sense Disambiguation. In Proceedings of The 20st International Conference on Computational Linguistics (COLING 2004), Switzerland, Geneva. 2004.

Mihalcea R., Chklovski T., Kilgarriff A. The Senseval-3 English lexical sample task In Proceedings of Senseval-3. 2004. pp.25-28.

Miller G. Nouns in WordNet. In: Fellbaum, C (ed) WordNet – An Electronic Lexical Database. – The MIT Press. 1998. – pp.23-47.

Miller G., Fellbaum C. Morphosemantic links in WordNet. – Traitement automatique de langue, 44.2. 2003. – pp. 69-80.

Miller G., Hristea F. WordNet Nouns: Classes and Instances. – Computational linguistics, Volume 32, Number 1. 2006. – pp.1-3.

Miller K. Modifiers in WordNet. In: Fellbaum, C (ed) WordNet – An Electronic Lexical Database. – The MIT Press. 1998. – pp .47 - 68.

Min-Yen Kan, Klavans J., McKeown K. Linear Segmentation and Segment Relevance. In the Proceedings of 6th International Workshop of Very Large Corpora (WVLC- 6). 1998. – pp.197-205.

Mizoguchi R., Sunagawa E., Kozaki K. and Kitamura Y. A Model of Roles within an Ontology Development Tool: Hozo. In Journal of Applied Ontology, Vol.2, No.2. 2007. – pp. 159-179.

Moens M.F. Information Extraction. Algorithms and Prospects In a Retrieval Context. Springer, 2006.

Mochizuki H., Iwayama M., Okumura M. Passage Level Document Retrieval Using Lexical Chains. RIAO 2000, Content Based MultiMedia Information Access. 2000. – pp. 491-506.

Moldovan D., Harabagiu S., Pasca M., Mihalcea R., Goodrum R., Girju R., Rus V. LASSO: A Tool for Surfing the Answer Net. - Proceedings of TREC-8. 1999. – pp. 175-184.

Moldovan D., Novischi A. Lexical Chains for Question Answering. In the Proceedings of International Conference on Computational Linguistics (COLING-2002). 2002. – pp.674-680.

Molla D., Vicedo J. Question Answering in Restricted Domains: An Overview. – Journal of Computational linguistics, v. 33, N1. 2007. – pp. 41-61.

Montejo-Ráez A., Steinberger R., López A. Adaptive selection of base classifiers in one-against-all learning for large multi-labeled collections. Advances in Natural Language Processing: 4th International Conference, EsTAL-2004, Alicante (Spain), Lectures notes in artifial intelligence, Springer, ISSN: 0302-9743. 2004. – pp. 1-12.

Morris J., Hirst G. Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of the Text. Computational Linguistics, 17(1). 1991. – pp. 21-45.

Motschnig-Pitrik R., Kaasboll J. Part-Whole Relationship Categories and their Application in Object-Oriented Analysis // IEEE TSE. – V. 11(5). 1999. – pp.779-797.

Nenadic G., Ananiadou S., McNaught J. Enhancing automatic term recognition through recognition of variation. In the Proceedings of the 20th international conference on Computational Linguistics (COLING-2004). 2004. – pp. 604-610.

Nenkova A., Louis A. Can you summarize this? Identifying correlates of input difficulty for generic multi-document summarization. Proceedings of ACL-08: HLT. 2008. – pp . 825-833.

Niles I., Pease A. Towards a Standard Upper Ontology // C.Welty and B.Smith, (Eds.) Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001), Ogunquit, Maine. 2001.

Niles I., Pease A. Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology, *Proceedings of the IEEE International Conference on Information and Knowledge Engineering.* 2003. – pp. 412-416.

Nirenburg S., McShane M., Beale S. The Rationale for Building Resources Expressly for NLP. - In Proceedings of the 4st International Conference on Language Resources and Evaluation (LREC). Lisbon, Portugal. 2004. – pp. 3-6.

Nirenburg S., Wilks Y., What's in a symbol: Ontology, representation, and language // Journal of Experimental and Theoretical Artificial Intelligence, 13(1). 2001. – pp. 9–23.

Nirenburg S., Raskin V. Ontological Semantics. MIT Press. 2004.

Noy N.F., McGuinness D. Ontology Development 101: A Guide to Creating Your First Ontology. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, March 2001. Рус. Перевод: *Разработка онтологий 101: руководство по созданию Вашей первой онтологии (http://ifets.ieee.org/russian/depository/ontology101_rus.doc).*

Noy N., Wallace E. Simple part-whole relations in OWL Ontologies. W3C Technical report. 2005. (http://www.w3.org/2001/sw/BestPractices/OEP/SimplePartWhole/index.htm)

Obrst L. Ontologies for Semantically Interoperable Systems. In Proceedings of the 12th ACM International Conference on Information and Knowledge Management (CIKM-2003). 2003. – pp.366-369.

Ogilvie, P., Callan J. Experiments using lemur toolkit. In Proceedings of 10th Text Retrieval Conference (TREC-10). 2001. – pp.103-108.

Pazienza M., Stellato A. Linguistic Enrichment of Ontologies: a Methodological Framework. In Proceedings Ontolex-2006 workshop. 2006.

Pedersen B.S., N. Sorensen. Towards Sounder Taxonomies in Wordnets. In Proceedings from Ontolex-2006, Genova, Italy. 2006. – pp. 9-15.

Pedersen, B.S., Nimb S., Asmussen J., Sɯrensen N., Trap-Jensen L., Lorentzen H. DanNet - a WordNet for Danish. In Proceedings of the Third International WordNet Conference. Jeju, Korea. 2006. – pp. 329-331.

Pedersen T. A simple approach to building ensembles of Naive Bayesian classifiers for word sense disambiguation. In Proceedings of NAACL. 2000.

Peter H., Sach H., Bechstein C. Smartindexer – Amalagamating Ontologies and Lexical Resources for document indexing. In Proceedings of OntoLex - 2006. 2006.

Peters, W., Peters, I., Vossen, P. Automatic sense clustering in EuroWordNet. In: Proc of the 1st. international conference on Language Resources and evaluations. 2000.

Petras V. GIRT and the Use of Subject Metadata for Retrieval. In: Multilingual Information Access for Text, Speech and Images. 5th workshop of the Cross-language Evaluation Forum, CLEF 2004. Lecture Notes in Computer Science, Vol. 3491. Springer-Verlag. 2004.– pp. 298-309.

Petras V. How One Word Can Make all the Difference – Using Subject Metadata for Automatic Query Expansion and Reformulation. - In: Multilingual Information Access for Text, Speech and Images. 6th workshop of the Cross-language Evaluation Forum, CLEF 2005. Lecture Notes in Computer Science, Springer-Verlag. 2005.

Pianta E., Bentivogli L., Girardi C. MultiWordNet: Developing an Aligned Multilingual Database. In Proceedings of the First International Conference on Global WordNet, Mysore, India. 2002.

Plaunt, Ch., Norgard, B. A. An Association Based Method for Automatic Indexing with a Controlled Vocabulary. Journal of the American Society for Information Science 49, (10). 1998. pp. 888-902.

Ponte J., Croft B. A Language Modeling Approach to Information Retrieval. In Proceedings of SIGIR-1998. 1998. – pp. 275-281.

Pouliquen B., Steinberger R., Ignat C. Automatic Annotation of Multilingual Text Collections. with a Conceptual Thesaurus. In: Proceedings of the International Conference *Recent Advances in Natural Language Processing,*. Borovets, Bulgaria. 2003. – pp. 401-408

Prevot L., Borgo S., Oltramari A. Interfacing Ontologies and Lexical Resources. In the proceedings of OntoLex-2005. 2005. – pp. 91-102.

Rada, R., Barlow, J., Potharst, J., Zanstra, P. and Bijstra, D. Document ranking using an enriched thesaurus. *Journal of Documentation,* 47(3). 1991. – pp. 240-253.

Radev D., Jing H., Budzikowska M. Centroid-based summarization of multiple docuemtns: sentence extraction, utility-based evaluation, and user studies. – In ANLP/NACCL Workshop on Summarization, Seattle. 2000.

Radev D., McKeown K., Hovy E. Introduction to the Special Issue on Summarization. Computational linguistics issue 4. 2002. – pp.399-408.

Reed S., Lenat D. Mapping ontologies into Cyc. In Proceedings of AAAI 2002 Conference Workshop on Ontologies for the Semantic Web, Edmonton, Canada. 2002.

Reeve L., Han H., Brooks A. BioChain: Using Lexical Chaining for Biomedical Text Summarization. – Proceedings of the ACM Symposium on Applied Computing. 2006. – pp.180-184.

Resnik P. Using information content to evaluate semantic similarity. In Proceedings of IJCAI-1995. 1995.

Robertson S., Walker S., Hancock-Beaulieu M., Gatford M. Okapi in Trec-3. Proceedings of Text Retrieval Conference TREC-3. NIST Special publication 500-225. 1994. – pp. 109-126.

Roget P. *Roget's Thesaurus*, Burnt Mill, Harlow, Essex: Longman Group Limited. 1982.

Rondeau G. Introduction at a terminologie. Quebec, 1980.

Rose T, Stevenson M., Whitehead M., The Reuters Corpus Volume 1 – from Yesterday News to tomorrow's Language. In Proceedings of the Third International Conference on Language Resources and Evaluation, Las Palmas de Gran Canaria. 2002.

Roventini A., Marinelli R. Extending the Italian WordNet with the Specialized Language of the Maritime Domain. // Proceedings of Second International WordNet Conference GWC-2004. 2004. – pp. 193-198.

Roventini A., Alonge A., Bertagna F., Calzolari N., Marinelli R., Magnini B., Speranza M., Zampolli A. ItalWordNet: a Large Semantic Database for the Automatic Treatment of the Italian Language. In Proceedings of LREC-2000. 2000.

Sag I., Baldwin T., Bond F., Copestake A., Flickinger D. Multiword expressions: A Pain in the Neck for NLP. – In proceedings of CICLING 2002, Mexico city, Mexico. 2002.

Sagri M., Tiscornia D., Bertagna F. Jur-WordNet. In Proceedings of Second International WordNet Conference GWC-2004. 2004. – pp. 305-310.

Salton G. Another look at automatic text-retrieval systems. Communications of the ACM, 29 (7). 1986. – pp.648-656

Salton G. Automatic Text Processing - The Analysis, Transformation and Retrieval of Information by Computer. Addison-Wesley, Reading, MA. 1989.

Sanderson M. Word Sense Disambiguation and information retrieval. - In Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 1994.

Scott S., Matwin S. Text classification using WordNet hypernyms, in Proceedings of the Workshop on Usage of WordNet in Natural Language Processing Systems (Coling-ACL 1998), Montreal, Canada. 1998. – pp.45-52.

Shah Ch., Croft B. Evaluating High Accuracy Retrieval Techniques. In Proceedings of SIGIR '04. 2004. – pp. 2-9.

Schott H. Thesaurus for Social Sciences. 2 vols. Vol.1. German –English. 2. English – German. Bonn: Informations-Zentrum Sozialwissenshaften. 2000.

Silber G., McCoy K. 2003. Efficiently computed lexical chains as an intermediate representation for automatic text summarization. Computational Linguistics, 29 (1), 2003.

Simons P. Parts. A study in Ontology. Oxford University Press, 1987.

Smith B. Beyond Concepts: Ontology as Reality Representation // Proceedings of International Conference on Formal Ontology and Information Systems FOIS-2004. 2004.

Smith B, Köhler J, Kumar A: On the application of formal principles to life science data: A case study in the Gene Ontology. *DILS 2004: Data Integration in the Life Sciences*. 2004. – pp.124-139.

Snyder B., Palmer M. The English all-words task. In Proceedings of SENSEVAL-3. Third International workshop on the Evaluation of Systems for the Semantic Analysis of Texts. 2004. – pp .41-43.

Soergel D., Lauser B., Liang A., Fisseha F., Keizer J., Katz S. Reengineering Thesauri for New Applications: the AGROVOC Example. - Article No. 257, 2004-03-17.

Song F., Croft B. A General Language Models for Information Retrieval. *Research and Development in Information Retrieval.* 1999. – pp. 279-280.

Soricut R., Marcu D. Sentence Level Discourse Parsing using Syntactic and Lexical Information. In Proceedings HLT/NAACL-2003. 2003.

Sowa J. Using a Lexicon of Canonical Graphs in a semantic interpreter. Relational models of lexicon / M.Evens. Cambridge University press. 1988. – pp.113-137.

Sowa J. Knowledge Representation: Logical, Philosophical, and Computational Foundations. – Brooks Cole Publishing Co., Pacific Grove, CA. 2000.

Sowa J. Building, Sharing and Merging Ontologies. Режим доступа: http://www.jfsowa.com/ontology/ontoshar.htm

Sparck Jones K. Retrieval system tests, 1958-1978. In Karen Sparck Jones, editor, Information Retrieval Experiment. Butterworths, London. 1981.

Srikanh M., Srihari R. Biterm language models for document retrieval - In Proceedings SIGIR-2002. 2002. – pp. 425-426.

Stairmand M. Textual content analysis for information retrieval. In the Proceedings of th 20th Annual ACM SIGIR Conference (SIGIR-97). 1997. pp.140-147.

Steinberger R., Hagman J. Scheer St. Using Thesauri for Automatic Indexing and Visualisation. – In Proceedings OntoLex 2000. 2000. – p. 130-141.

Steinmann F. The representation of roles in object-oriented and conceptual modelling // Data and Knowledge engineering. 35, 1. 2000. – pp. 83-106.

Stokes N., Hatch P., Carthy J. Lexical semantic relatedness and online news event detection. In the proceedings of the Annual 23rd ACM SIGIR Conference on Research and Development (SIGIR-00). 2000. pp.324-325.

Stokes N., Carthy J, Smeaton A.F. SeLeCT: A lexical Cohesion based News Story Segmentation System. – In the Journal of AI communications, 17(1). 2004. – pp. 3-12.

Swales J. Genre analysis: English in academic and research settings. Cambridge, Cambridge University Press. 1990.

Teufel S., Moens M. Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status. – Computational linguistics, vol.28, n°4. 2002. – pp. 409-445.

Tipster SUMMAC Text Summarization Evaluation. Final report. - MITRE Technical report MTR 98000138. - October, 1998. (http://www.itl.nist.gov/iaui/894.02/related_projects/tipster_summac/summac-final-report-part2.ps)

Tomlin R. S., Forrest L., Pu M. M. Discourse semantics. In T. van Dijk (Ed.), *Discourse as structure and process*. London: Sage. 1997. – pp. 63-111.

Tonta Y. Analysis of Search Failures in Document Retrieval Systems: A Review // The Public-Access Computer Systems Review 3, no. 1. 1992. – pp.4-53.

Trautwein M., Grenon P. Roles: One Dead Armadillo on WordNet's Speedway to Ontology. - In Proceedings of International Wordnet Conference (GWC – 2004). 2004. – pp. 341-346.

Tsujii J., Ananiadou, S. Thesaurus or logical ontology, which one do we need for text mining? In Language Resources and Evaluation, Springer Science and Business Media B.V., vol. 39, no 1. 2005. pp. 77-90.

Tudhope D., Alani H., Jones Cr. Augmenting Thesaurus Relationships: Possibilities for Retrieval. – Journal of Digital Libraries. Volume 1, Issue 8. 2001.

Tudhope D., Taylor C. Navigation via Similarity: automatic linking based on semantic closeness". *Information Processing and Management*, 33(2). 1997. pp. 233-242.

UNBIS Thesaurus, English Edition, Dag Hammarskjold Library of United Nations, New York, 1976.

UNBIS Guidelines for Analysis of UN Information Resources. Код доступа: http://www.un.org/Depts/dhl/unbisref_manual/indexpolicy/650.htm. Время доступа: 2009.

Varzi A. Parts, Wholes, and Part-Whole Relations: The Prospects of Mereotopology. Data and Knowledge Engineering 20. 1996. – pp. 259-286.

Varzi A. Basic Problems of Mereotopology', in N. Guarino (ed.), Formal Ontology in Information Systems, IOS Press. 1998. – pp. 29-38.

Varzi A. A Note on Transitivity of Parthood. – Applied Ontology, 1:2. 2006. – pp.141-146.

Vechtomova O., Jones R., Dias G. Report on the ACM International Workshop on Methodologies and Evaluation of Lexical Cohesion Techniques in Real-World Applications (ELECTRA 2005) held at SIGIR 2005. Sigir Forum V 39, N2. 2005. – pp. 42-45.

Voorhees E. Query expansion using lexical-semantic relations. In Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 1994. – pp. 61--69.

Voorhees E. (1998). Using WordNet for Text Retrieval. – In: WordNet – an Electronic Lexical Database. – MIT Press. – pp. 285-304.

Voorhees, E. (1999). Natural Language Processing and Information Retrieval. In M.T.Pazienza (ed.). - Information Extraction: Towards Scalable, Adaptable Systems, New York: Springer, pp. 32-48.

Voorhees E. (2004) Overview of the TREC 2004 Question Answering Track. NIST Special Publication 500-261

Vossen P.(ed.). EuroWordNet: A multilingual Database with Lexical Semantic Network. – Dodrecht. 1998.

Vossen P. Tutorial Wordnet, *EuroWordNet and Global WordNet*, International Conference RANLP – 2003 (Recent Advances in Natural Language Processing), Borovets, Bulgaria. 2003.

Vossen, P.: Extending, Trimming and Fusing WordNet for Technical Documents. In Proceedings of WordNet and Other Lexical Resources: Applications, Extensions and Customizations, Pittsburg, USA. 2001.

Vossen P., Glaser E., Gradinaru M., van Steenwijk R., van Zutphen H. Meaning-full Effects on Information Retrieval. IST-2001-34460, Deliverable 8.3. 2005.

Vossen P., Rigau G., Alegria I., Agirre E., Farwell D., Fuentes M. Meaningful results for Information Retrieval in the MEANING project. In Proceedings of Third International WordNet Conference. 2006.

Wasson M. Classification Technology at LexisNexis. SIGIR 2001 Workshop on Operational Text Classification. 2001.

Welty, C., McGuinness, D., Uschold, M., Gruninger, M., and Lehmann, F. Ontologies: Expert Systems all over again. AAAI-1999 Invited Panel Presentation. 1999.

Wielinga B., Schreiber A., Wielemaker J., Sandberg J. From Thesaurus to Ontology. - Proceedings of the 1st international conference on Knowledge capture. 2001. – pp. 194 – 201.

Wilks Y. Ontotherapy: or how to stop worrying about what there is. Invited presentation, Ontolex 2002, Workshop on Ontologies and Lexical Knowledge Bases,27th May. Held in conjunction with the Third International Conference on Language Resources and Evaluation - LREC02, 29-31 May, Las Palmas, Canary Islands. 2002.

Wilks Yorick. The Semantic Web as the apotheosis of annotation, but what are its semantics? In IEEE Intelligent Systems May/June. 2008.

Will L. Thesaurus consultancy. – The thesaurus: review, renaissance and revision / Sandra K. Roe and Alan R. Thomas, editors. - New York ; London : Haworth. 2004. - 209p.

Winston M., Chaffin R, Herrmann D. A Taxonomy of Part-Whole Relations. Cognitive Science, 11. 1987. – pp. 417-444.

Wolf F., Gibson E. Representing Discourse Coherence: A Corpus-based study. Computational linguistics, V31, N2. 2005. – pp. 249-287.

Woods W. Conceptual Indexing: A Better Way to Organize Knowledge. - Sun Microsystems, Inc., Technical Report: TR-97-61. 1997.

Wűster E., Einfurung in die Allgemeine Terminologielehre und terminologishe Lexicographie. - Vien; N.Y., 1979/Bd 1 - 2. 1979.

Yang Y., Liu X. A re-examination of text categorization methods. In Proceedings of Int. ACM Conference on Research and Development in Information Retrieval (SIGIR-99). 1999. – pp. 42-49.

Z39.19 – Guidelines for the Construction, Format and Management of Monolingual Thesauri. – NISO. 1993.

Zhai C., Lafferty J. A study of smoothing Methods for Language models applied to Information Retrieval. In Proceedings of SIGIR-2001 Conference. 2001. – pp. 334-342.

Агеев М.С., Добров Б.В., Макаров-Землянский Н.В. Метод машинного обучения, основанный на моделировании логики рубрикатора. // RCDL'2003 Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Пятая всероссийская науч. конф. — Санкт-Петербург. 2003.

Агеев М.С., Добров Б.В., Лукашевич Н.В., Поддержка системы автоматического рубрицирования для сложных задач классификации текстов // Электронные библиотеки: перспективные методы и технологии, электронные коллекции. Труды шестой Всероссийской научной конференции. Пущино, 29.09-01.10.2004 – Ин-т мат. проблем биологии, Пущино. 2004. – С .216-225

Агеев М.С., Добров Б.В., Лукашевич Н.В., Сидоров А.В. Экспериментальные алгоритмы поиска/классификации и сравнение с «basic line». // Российский семинар по Оценке Методов Информационного Поиска — Пущино, 2004. – С . 62-89

Агеев М.С., Кураленок И.Е. Официальные метрики РОМИП'2004. // Российский семинар по Оценке Методов Информационного Поиска — Пущино, 2004.

Агеев М., Добров Б., Красильников П., Лукашевич Н., Павлов А., Сидоров А., Штернов С. УИС РОССИЯ в РОМИП2007: поиск и классификация. Российский семинар по Оценке Методов Информационного Поиска. Труды РОМИП 2007-2008. (Дубна, 9 октября 2008г.) Санкт-Петербург: НУ ЦСИ, 2008, 258 с.

Агеев М.С., Добров Б.В., Лукашевич Н.В., Штернов С.В.. УИС РОССИЯ в РОМИП 2008: поиск и класификация нормативных документов. Российский семинар по Оценке Методов Информационного Поиска. Труды РОМИП 2007-2008. (Дубна, 9 октября 2008г.) Санкт-Петербург: НУ ЦСИ, 2008, 258 с.

Агеев М.С., Добров Б.В., Лукашевич Н.В. Автоматическая рубрикация текстов: методы и проблемы. - Ученые записки Казанского государственного университета. Серия Физико-математические науки. 2008. Том 150, книга 4, стр. 25-40.

Азарова И.В., Митрофанова О.А., Синопальникова А.А. Компьютерный тезаурус русского языка типа WordNet // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции Диалог'2003. М., 2003. - С . 43-50.

Азарова И.В., Синопальникова А.А., Яворская М.В. Принципы построения wordnet-тезауруса RussNet // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции Диалог'2004. М., 2004. - С. 542-547.

Азарова И.В., Синопальникова А.А., Смрж П. Представление устойчивых лексических сочетаний в компьютерном тезаурусе RussNet. Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции Диалог'2005. М., 2004. С. 11-16.

Александрова З.Е. Словарь синонимов русского языка. – М.: Русский язык, 1999.

Алексеев А.А., Лукашевич Н.В. Автоматическое порождение обновления к аннотации новостного кластера. Труды конференции RCDL-2010. 2010.

Антонов А.В., Курзинер Е.С. Автоматическое определение тематики большого необработанного текстового массива. – Труды международной конференции Диалог'2002.

Апресян Ю.Д. Лексическая семантика. Синонимические средства языка. – М.: Восточная литература. – 1995.

Апресян Ю. Д. (ред.) Языковая картина мира и системная лексикография. М.: Языки славянских культур, 2006.

Архангельская В.А., Базарнова С.В. Информационно-поисковый тезаурус по экономике и демографии. – НТИ, сер.1. Орг. и методика информ. работы. – 2001, N 7, стр. 24-32.

Белоногов Г.Г., Хорошилов Ал-р А., Хорошилов Ал-сей А. Единицы языка и речи в системах автоматической обработки текстовой информации. – НТИ, сер.2. Информационные процессы и системы, N11, 2005 – стр. 21-29.

Блюменау Д.И., Гендина Н.И. Формализованное реферирование с использованием словесных клише (маркеров) // НТИ, 2002. Сер. 2, № 5, С. 29–36.

Богомолова А.В., Дышкант Н.Ф., Юдина Т.Н. Университетская информационная система РОССИЯ: ресурсы и сервисы для поддержки общественного участия и задач государственного управления. «Труды XI Всероссийской объединенной конференции "Интернет и современное общество". Октябрь 2008 года. Санкт Петербург, стр. 196-199.

Большакова Е.И, Большаков И.А., Котляров А.П. Расширенный эксперимент по автоматическому обнаружению и исправлению русских малапропизмов. Труды Международной конференции Диалог'2006. М., 2006. – стр 78-83.

Большакова Е.И., Васильева Н.Э. (2008) Терминологическая вариантность и ее учет при автоматической обработке текстов. Труды XI национальной конференции по искусственному интеллекту.

Большаков И.А. КроссЛексика – большой электронный словарь сочетаний и смысловых связей русских слов // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной конференции «Диалог». Выпуск 8 (15) – С.45 –50.

БТС. Большой толковый словарь русского языка. – Изд-во Норинт, Санкт-Петербург, 1998.

Гаврилова Т.А., Хорошевский В.Ф., Базы знаний интеллектуальных систем. - Санкт-Петербург: Изд-во "Питер" – 2000. - 382 с.

Гаврилова Т.А. Извлечение знаний: Лингвистический аспект. Корпоративные системы. – 2001. - N10 (25), с. 24-28.

Гальперин И. О. Текст как объект лингвистического исследования / И. О. Гальперин. – М.: Наука, 1981.

Герд А.С. Прикладная лингвистика. Изд-во Санкт-Петербургского университета, 2005.

Городецкий Б.Ю. Термин как семантический феномен (в контексте переводческой литературы. Код доступа: (http://www.dialog-21.ru/dialog2006/materials/html/GrodetskiyB.htm) - 2006.

ГОСТ 7.66.-92. Индексирование документов Общие требования к систематизации и предметизации.

ГОСТ 7.74-96 Информационно-поисковые языки. Термины и определения Межгосударственный стандарт 7.74-96 – Минск: Межгосударственный совет по стандартизации, метрологии и сертификации, 1996.

ГОСТ 7.25.-2001 Тезаурус информационно-поисковый одноязычный: Правила разработки: структура, состав и форма представления: Межгосударственный стандарт. – Минск: Межгосударственный совет по стандартизации, метрологии и сертификации, 2001.

ГОСТ 7.59.-2003. Индексирование документов. Общие требования к систематизации и предметизации.

Гринев-Гриневич С.В. (2008) Терминоведение. М., Академия, 2008.

ван Дейк Т.А ., Кинч В. 1988. Стратегии понимания связного текста. // Новое в зарубежной лингвистике. Вып. 23. - М.: Прогресс. - С.153-211.

Добров Б.В., Иванов В.В., Лукашевич Н.В., Соловьев В.Д. Онтологии и тезаурусы: модели, инструменты, приложения. М.: Интуит, 2009.

Добров Б.В., Лукашевич Н.В. Тезаурус и автоматическое концептуальное индексирование в университетской информационной системе РОССИЯ // Третья

Всероссийская конференция по Электронным Библиотекам «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - Петрозаводск, 2001 – С.78-82.

Добров Б.В., Лукашевич Н.В. Автоматическая рубрикация полнотекстовых документов по классификаторам сложной структуры // Восьмая национальная конференция по искусственному интеллекту. КИИ-2002. 7-12 октября 2002, Коломна – М.: Физматлит – Т.1 – С.178-186. 2002a.

Добров Б.В., Лукашевич Н.В. Организация двуязычного поиска в Университетской системе РОССИЯ // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды Четвертой Всероссийской научной конференции RCDL'2002 (Дубна, 15-17 октября 2002 г.): В 2 т. – Дубна: ОИЯИ, 2002. – Т.2. – С.148-158. 2002b.

Добров Б.В., Лукашевич Н.В., Невзорова О.А. Технология разработки онтологий новых предметных областей // Труды Казанской шкрлы по компьютерной лингвистике TEL-2002. Выпуск 7. / Под ред. В.Г.Бухараева, В.Д.Соловьева, Д.Ш.Сулейманова - Казань: Отечество, 2002. - С.90-106.

Добров Б.В., Лукашевич Н.В., Сыромятников С.В. Формирование базы терминологических словосочетаний по текстам предметной области. - Труды пятой всероссийской научной конференции "Электронные библиотеки: Перспективные методы и технологии, электронные коллекции. – 2003, с. 201-210.

Добров Б.В., Лукашевич Н.В., Невзорова О.А., Федунов Б.Е. Методы и средства автоматизированного проектирования практической онтологии // Известия РАН. Теория и системы управления. - 2004. - N 2. - С. 58-68

Добров Б.В., Лукашевич Н.В., Синицын М.Н., Шапкин В.Н. Разработка лингвистической онтологии для автоматического индексирования текстов по естественным наукам // Электронные библиотеки: перспективные методы и технологии, электронные коллекции. Труды Седьмой Всероссийской научной конференции (RCDL'2005) г. Ярославль 4-6 октября 2005г. – Ярославль: ЯрГУ им. П.Г.Демидова, 2005. – С.70-79.

Добров Б.В., Лукашевич Н.В. Онтологии для автоматической обработки текстов: описания понятий и лексических значений. // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции "Диалог'2005 / Под ред. И.М. Кобозевой, А.С. Нариньяни, В.П. Селегея. - М.: Наука, 2005. – С.138-142.

Добров Б.В., Лукашевич Н.В. Вторичное использование лингвистических онтологий: изменение в структуре концептуализации// Восьмая Всероссийская научная конференция «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (Владимир-Суздаль, 16-18 октября 2006г.). 2006.

Добров Б.В., Лукашевич Н.В. Транзитивные нетаксономические отношения в онтологическом моделировании. // Труды симпозиума Онтологическое моделирование. Институт проблем информатики РАН, 2008. - стр.229-259.

Жинкин Н.И. Механизмы речи. М 1958.

Зализняк А. Многозначность в языке и способы ее представления. Изд. Языки Славянской Культуры, 2006.

Зализняк А. Феномен многозначности и способы его описания. Вопросы языкознания, №2, 2004, с. 20-45.

Зализняк Анна А., Шмелев А.Д. Введение в русскую аспектологию. – Языки русской культуры – Москва, 2000.

Зубов А.В., Зубова И.И. Основы искусственного интеллекта для лингвистов. – Москва: Логос, 2006.

Караулов А.В. Русский ассоциативный словарь. В 2 томах. – Изд-во АСТ, 2002.

Кибрик Андрей А. Анализ дискурса в когнитивной перспективе. Москва, РАН. 2003.

Клещев А.С., Шалфеева Е.А. Классификация свойств онтологий. Онтологии и их классификации. НТИ сер. 1, 2005 N 9, стр 16-22.

Кобозева И.М. Лингвистическая семантика. Эдиториал УРСС. Москва 2000.

Кобрицов Б.П. Методы снятия семантической многозначности // Научно-техническая информация, сер.2, 2004а, N 2.

Кронгауз М.А. Семантика. – М., РГГУ 2001.

Кураленок И., Некрестьянов И., Павлова Е. РОМИП 2003: Опыт организации. // Труды первого российского семинара по оценке методов информационного поиска. *Под ред. И.С. Некрестьянова* - Санкт-Петербург: НИИ Химии СПбГУ, 2003. – стр.1.-22.

Кустова Г.И. Типы производных значений и механизмы языкового расширения. Изд. Языки Славянской Культуры, 2004.

Леоненков А.В. Самоучитель UML. – СПб.: БХВ-Петербург, 2001. – 304 с.

Леонтьева Н.Н., Волковысская Е.В., Копылова О.Т. и др., Словарь энциклопедических функций и его роль в автоматическом индексировании // НТИ. – М., 1978. – Сер.2. – N 7. – С.23-29.

Леонтьева Н.Н. Семантика связного текста и единицы информационного анализа. – НТИ, сер. 2, 1981, N1.

Леонтьева Н.Н. Автоматическое понимание текстов: системы, модели, ресурсы. – М., Издательский центр «Академия», 2006. – 304 с.

Литвиненко А.Л. Описание структуры дискурса в рамках Теории Риторической структуры: применение на русском материале. М.: Диалог. 2001.

Лукашевич Н.В., Автоматизированное формирование информационно-поискового тезауруса по общественно-политической жизни России // НТИ. Сер.2. - 1995. - N 3. - С.21-24.

Лукашевич Н.В., Разрешение многозначности терминов в процессе автоматического индексирования // Труды международного семинара Диалог'96. - Москва, 1996. - С.142-146.

Лукашевич Н.В., Салий А.Д., Тезаурус для автоматического рубрицирования и индексирования: разработка, структура, ведение // НТИ. Сер.2. - 1996. - N 1. - С.1-6.

Лукашевич Н.В., Добров Б.В., Построение и использование тематического представления содержания документов // 5ая Национальная конференция КИИ-96. - Казань, 1996. - С. 130-134.

Лукашевич Н.В., Автоматическое рубрицирование потоков текстов по общественно-политической тематике // НТИ. Сер.2. - 1996. - N 10. - С.22-30.

Лукашевич Н.В., Автоматическое построение аннотаций на основе тематического представления текста // Труды международного семинара Диалог'97. - Москва, 1997 - С. 188-191.

Лукашевич Н.В., Салий А.Д., Представление знаний в системе автоматической обработки текстов // НТИ. Сер.2. - 1997 - N3.

Лукашевич Н.В., Добров Б.В., Построение структурной тематической аннотации текста // Труды международного семинара Диалог-98 - Том 2 – 1998 - С.795-802.

Лукашевич Н.В., От общеполитического тезауруса к тезаурусу русского языка в контексте автоматической обработки больших массивов текстов // Труды международного семинара Диалог-99, - Том 2 - 1999 - С.184 -190.

Лукашевич Н.В., Добров Б.В. Исследования тематической структуры текста на основе большого лингвистического ресурса // Труды международного семинара “Диалог 2000” – Том 2. – 2000 - С.252-258.

Лукашевич Н.В., Добров Б.В., Тезаурус для автоматического концептуального индексирования как особый вид лингвистического ресурса // Труды международного семинара Диалог-2001. - Аксаково-2001.- с.273-279.

Лукашевич Н.В., Добров Б.В. Модификаторы концептуальных отношений в тезаурусе для автоматического индексирования // НТИ, Сер.2. – 2001. – N 4. – С. 21-28. 2001.

Лукашевич Н.В., Добров Б.В., Тезаурус русского языка для автоматической обработки больших текстовых коллекций // Компьютерная лингвистика и интеллектуальные технологии: Труды Международного семинара Диалог'2002 / Под ред. А.С.Нариньяни – М.: Наука – 2002. – Т.2 - С.338-346.

Лукашевич Н.В., Добров Б.В., Двуязычный информационный поиск на основе автоматического концептуального индексирования // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции Диалог-2003. Протвино. 11-16 июня 2003г. / Под ред.И.М.Кобозевой, Н.И.Лауфер, В.П.Селегея – М.: Наука, 2003. - С.425-432.

Лукашевич Н.В., Добров Б.В., Разграничение общезначимой лексики и терминологии и автоматическая обработка больших электронных коллекций // Русский язык: исторические судьбы и современность. II Межд. конгресс исследователей русского языка. – М.: МГУ – 2004. – с.481-482. 2004a.

Лукашевич Н.В., Добров Б.В., Отношения в онтологиях для решения задач информационного поиска в больших разнородных текстовых коллекциях // Девятая национальная конференция по искусственному интеллекту с международным участием КИИ-2004 (28 сентября –2 октября 2004 г., Тверь): Труды коференции. В 3-х т. - Т2. – М.: Физматлит, 2004. – С.544-551. 2004b.

Лукашевич Н.В., Добров Б.В. Разрешение лексической многозначности на основе тезауруса предметной области. Компьютерная лингвистика и интеллектуальные технологии. // Труды международной конференции «Диалог 2007» (Бекасово, 30 мая - 3 июня 2007 г.). М. Наука – 2007. стр. 400-406.

Лукашевич Н.В., Чуйко Д.С. Автоматическое разрешение лексической многозначности на базе тезаурусных знаний. Интернет-математика 2007: Сборник работ участников конкурса. — Екатеринбург: Изд-во Урал. ун-та, 2007. Стр.108-117.

Лукашевич Н. В., Добров Б. В. Автоматическое аннотирование новостного кластера на основе тематического представления. Компьютерная лингвистика и интеллектуальные технологии по материалам ежегодной Международной конференции «Диалог 2009» Выпуск 8 (15), стр. 299-305.

Лукашевич Н.В. Моделирование отношения ЧАСТЬ-ЦЕЛОЕ в лингвистических и онтологических ресурсах. // Информационные технологии. – 2007. – N 12. 2007a.

Лукашевич Н.В. Проблемы установления родовидовых отношений в лингвистических онтологиях. – Материалы Всероссийской конференции «Знания-Онтологии-решения» (ЗОНТ-07). Стр.211-220. 2007b.

Лукашевич Н.В. Типы и роли в лингвистических онтологиях // Труды Казанской школы по компьютерной лингвистике TEL-2006. Казань: Отечество, 2007. – стр.49-64. 2007c.

Мальковский М.Г., Соловьев С.Ю. Универсальное терминологическое пространство. Труды международного семинара "Компьютерная лингвистика и интеллектуальные технологии", М: Наука, 2002, т.1, стр.266-270.

Мдивани Р.Р. О разработке серии тезаурусов по социальным и гуманитарным наукам. – НТИ, сер.2. Информ. процессы и системы. 2004. - N 7, стр. 1-9.

Мельчук И.А. Опыт теории лингвистических моделей «Смысл-текст». – Изд-во Наука, М. 1974.

Методика составления информационно-поисковых тезаурусов. - М.: ВИНИТИ, 1973.

Моисеев А.И. О языковой природе термина // Лингвистические проблемы научно-технической терминологии. – М., 1970. - С. 127-138.

Морковкин В. В. Идеографические словари. Изд-во Моск. ун-та, 1970.

Нариньяни А.С. Кентавр по имени ТЕОН: Тезаурус+Онтология // Труды Международной конференции ДИАЛОГ-2001. – М., 2001. – Т.1. – С.184-188.

Никитина С.Е. Семантический анализ языка науки. Москва, Наука, 1987.

Новиков А.И. Семантика текста и ее формализация. – М. Наука, 1983

НОСС. Новый объяснительный словарь синонимов русского языка. Третий выпуск. Под общим руководством акад. Ю.Д. Апресяна. – М.: Языки славянской культуры, 2003. – 624 с.

Осипов Г.С. Приобретение знаний интеллектуальными системами: Основы теории и технологии. М.: Физматлит, 1997.

Падучева Е.В. Когнитивные» идеи в теоретической семантике. // Тезисы конференции Русский язык: исторические судьбы и современность 2007. – с. 470-471.

Поляков В.Н. Проект WordNet и его влияние на технологии компьютерной и когнитивной лингвистики (Обзорная статья) // Труды Казанской школы по компьютерной и когнитивной лингвистике TEL-2002. – Казань, 2002. С.6-61.

Рахилина Е.В., Кобрицов Б. П., Кустова Г. И., Ляшевская О. Н., Шеманаева О. Ю. Многозначность как прикладная проблема: лексико-семантическая разметка в корпусе русского языка. // Компьютерная лингвистика и интеллектуальные технологии: Труды Международного семинара Диалог'2006 / Под ред. А.С.Нариньяни – М.: Наука – 2006. – Т.2 - С.445-450.

РОМИП. Российский Семинар по методам информационного поиска. http//www.romip.ru.

Рубашкин В.Ш., Лахути Д.Г. Семантический (концептуальный) словарь для информационных технологий Ч.1 // НТИ. – М., 1998. – Сер 2. – N 1. С. 19-24.

Рубашкин В.Ш., Лахути Д.Г. Семантический (концептуальный) словарь для информационных технологий Ч.2 // НТИ. – М., 1999. – Сер 2. – N 5. С. 1-12.

Саломатина Н.В., Гусев В.Д. *Автоматизация формирования индикаторных словарей и возможности их использования* // Труды межд. конференции Диалог-2006 "Компьютерная лингвистика и интеллектуальные технологии", Бекасово, 31мая – 4 июня 2006, Москва, "Наука", С. 121–125.

Севбо И.П., Структура связного текста и автоматизация реферирования. - Изд-во Наука, Москва, 1969.

Сегалович И., Маслов М. Яндекс на РОМИП-2004. Некоторые аспекты полнотекстового поиска и ранжирования Яндекса. – РОМИП-2004

Селезнев М.Г. Референция и номинация. //Моделирование языковой деятельности в интеллектуальных системах - М.: Наука, 1987.- С.64-77.

Степанов Ю.С. Понятие / Лингвистический энциклопедический словарь. М.: Советская энциклопедия, 1990. – С. 383-385.

Суакисян Э.Р. Школа индексирования: практ. пособие. – М.,: ЛИБЕРЕЯ-БИБИНФОРМ, 2005. – 144 С.

Суперанская А.В., Подольская Н.В., Васильева Н.В., Общая терминология: Вопросы теории / Отв. Ред. Т.Л.Канделаки. Изд. 2-е, стереотипное. – М.: Едиториал УРСС, 2003. – 248 с.

Сухоногов А.М., Яблонский С.А. Автоматизация построения англо-русского WordNet. // Компьютерная лингвистика и интеллектуальные технологии: Труды Международного семинара Диалог'2005 / Под ред. А.С.Нариньяни – М.: Наука -2005.

Тер-Минасова С.Г. Словосочетание в научно-лингвистическом и дидактическом аспектах. – М.: Изд-во ЛКИ, 2007.

Уемов А. И. Вещи, свойства и отношения. М., 1963.

(Указ, 2000) Указ Президента Российской Федерации от 15 марта 2000 N511 "О классификаторе правовых актов.

Хорошевский В.Ф. Онтологические модели и Semantic Web: откуда и куда мы идем? – Труды семинара «Онтологическое моделирование» под редакцией Калиниченко Л.А. – Москва ИППИ РАН, 2008. стр. 13-45.

Шевченко Н.В. Основы лингвистики текста. – М.: Приор-издат, 2003.

Шемакин Ю.И. Тезаурус в автоматизированных системах управления и информации. - М: Военное изд-во министерства обороны СССР, 1974. - 192 с.

Шемакин Ю.И. Тезаурус научно-технических терминов. - М: Военное изд-во министерства обороны СССР, 1972.

Языковая картина мира и системная лексикография 2006. под редакцией Ю.Д. Апресяна. – М.: Языки славянских культур, 2006.