О построении статистических языковых моделей для систем распознавания русской речи*

А.Б. Холоденко

Естественный язык – результат многовековой параллельной работы огромного числа носителей языка – принципиально отличается от случайных комбинаций слов и от формально построенных языков. Одной из основных его особенностей является избыточность, позволяющая понимать искаженную речь. Формализация этого процесса сталкивается с трудностями больших объемов текстов (только в них можно выявить языковые особенности, а не в простом списке фраз). В таких случаях естественно исследовать различные применения частотных характеристик в текстовых базах данных.

Эта статья описывает результаты исследования различных статистических подходов к разработке языковых моделей для создания систем распознавания русской речи. Показано, что такие особенности русского языка, как большое число словоформ и свободный порядок слов в предложении, не дают возможности использовать стандартные и хорошо разработанные для других языков методы, с успехом применяющиеся на практике. Для корпуса, состоящего из текстов деловой прозы, научно-технической тематики и газетных публикаций, удалось преодолеть сложности, связанные со свободным порядком слов. Для решения проблем, связанных с большим количеством словоформ, предлагается разделить языковую модель на изменяемую (основанную на морфологической информации) и постоянную (основанную на начальных формах слов) части и решать

*Работа выполнена при поддержке фирмы Intel Corporation, USA.

задачу для каждой из них независимо. При этом сложность каждой подграмматики оказывается приемлемой для ее реализации в работающей системе. Описан разработанный пакет программ, предназначенных для построения вышеуказанных подграмматик.

1. Введение

Для многих языков, таких как английский, французский, китайский и т.п. разработаны статистические языковые модели, принципиально улучшающие распознавание слитной речи. Для русского языка различными исследователями разработаны автоматические модули, позволяющие проводить морфологический [1], синтаксический [2] и частично семантический [3] анализ предложений и текстов. Несмотря на то, что эти подходы хороши в некоторых ситуациях, связанных с обработкой текстовых данных на естественных языках, они имеют очень ограниченное применение для задачи распознавания речи. Как правило, они обеспечивают высокую точность анализа предложения, но при этом требуют значительных вычислительных ресурсов, наличия всего предложения для начала анализа и не работают с многовариантными случаями. Недостатки традиционных моделей приводят к необходимости разработки альтернативных, в том числе статистических, подходов к моделированию языка. Однако удовлетворительных статистических языковых моделей для славянской группы языков, и в том числе для русского языка, позволяющих использовать их в системах распознавания речи, до сих пор нет.

Основными отличиями славянских языков от, например, английского являются:

1) существенно большее количество словоформ для каждого слова и
2) свободный порядок слов.

Эти особенности приводят к существенным трудностям при создании языковых моделей, основанных на статистических подходах,

таких как, например, n-граммы и их вариации [4]. Свободный порядок слов «размывает регулярность» в предложениях языка, а большое количество словоформ приводит к значительному росту словаря. Кроме того, большинство словоформ одного и того же слова отличаются только в окончаниях, которые произносятся обычно не так четко, как начала слов. Это приводит к тому, что использование словоформ как отдельных независимых слов словаря (как это делается, например, во многих существующих системах распознавания для английского языка) приводит к большому количеству ошибок распознавания.

Основной целью данной работы являлась проверка пригодности существующих подходов для создания языковых моделей для построения русских систем распознавания речи (возможно, с некоторыми дополнениями) и/или разработка новых методов, пригодных для этой цели. Исследования велись в предположении, что разрабатываемые языковые модели должны быть пригодны к использованию в многоцелевых системах распознавания речи, например, для диктовки произвольных текстов и поэтому основное внимание уделялось таким моделям, которые могут быть автоматически обучены на корпусах больших размеров. Другие подходы, требующие вмешательство эксперта в процесс обучения, например, модели, основанные на семантике, в этой работе не рассматривались.

Все исследования, описываемые далее в настоящей работе, проводились на двух текстовых корпусах: части публичной электронной библиотеки Максима Машкова (расположенную по адресу http://www.lib.ru, около 70 млн. слов) и архиве «Независимой газеты» (около 30 млн. слов).

Для использования морфологической информации, была разработана оригинальная система морфологического анализа, включающая инструменты для морфологического анализа слов, то есть определения морфологических характеристик словоформ, определения начальной формы слова по любой его словоформе и построение списка всех словоформ для заданного слова. Словарь подсистемы на данный момент включает около 150 тыс. слов (около 2,5 млн. словоформ).

2. Современное состояние языковых моделей

В настоящее время основным подходом к построению языковых моделей для систем распознавания речи является использование аппарата статистических методов. При этом модель в таком понимании – просто распределение вероятности на множестве всех предложений языка. Естественно, что хранить модель в таком виде невозможно, поэтому используют более компактные способы задания. Рассмотрим вкратце, какие модели используются сегодня в коммерческих и экспериментальных системах распознавания речи с неограниченными словарями.

2.1. n-граммы

Языковые модели, основанные на n-граммах, используют явное предположение о том, что вероятность появления очередного слова в предложении зависит только от предыдущих $n-1$ слов. На практике используются модели со значениями $n = 1, 2, 3$ и 4. Наиболее удачной моделью из этого класса для английского языка оказывается триграммная модель. Все новые модели практически всегда оцениваются по отношению к триграммной модели. На сегодняшний день практически все коммерческие системы распознавания речи используют n-граммную модель в той или иной форме. При этом вероятность всего предложения вычисляется как произведение вероятности входящих в него n-грамм.

Основным достоинством данного класса моделей оказывается возможность построения модели по обучающему корпусу достаточно большого размера и высокая скорость работы. Основные недостатки – заведомо неверное предположение о независимости вероятности очередного слова от более длинной истории, что затрудняет работу и не позволяет моделировать более глубокие связи в языке и колоссальные, но все-таки недостаточные для получения достоверных оценок объемы обучающих данных. В самом деле, если словарь содержит N слов, то число возможных пар слов будет N^2. Даже если только 0,1% от них реально встречаются в языке, то ми-

нимально необходимый объем корпуса для получения статистически достоверных оценок будет иметь порядок 125 млрд. слов или около 1 терабайта при специально подобранном корпусе. Для триграмм минимальные корпуса будут достигать размеров в сотни и тысячи терабайт.

Для преодоления этого недостатка используется развитый аппарат техник сглаживания, которые позволяют производить оценку параметров модели в условиях недостаточных или вовсе отсутствующих данных. Другим подходом к решению той же проблемы является кластеризация словаря, позволяющая сократить модель.

2.2. Модели, основанные на деревьях решений

Этот класс моделей использует деревья решений для оценки распределения вероятностей очередного слова по известной истории. Под деревом решений понимается бинарное дерево, каждой листовой вершине которого приписывается распределение вероятностей на словаре, а остальным вершинам приписываются предикаты, определенные на множестве историй. В процессе работы такой модели возникает путь от корня дерева до одной из его листовых вершин, распределение в которой и объявляется результатом работы модели.

Несмотря на то, что деревья решений очень популярны в распознавании речи, их использование в задачах языкового моделирования наталкивается на существенные трудности, в первую очередь из-за колоссального объема анализируемой информации. Хотя теоретически эти модели способны показать существенное улучшение по сравнению с, например, n-граммными моделями, на практике не известно ни одной удачного примера их применения. В настоящее время этот тип моделей практически не используется.

2.3. Модели, основанные на теории формальных языков

К этому классу относятся модели, использующие развитый аппарат теории формальных языков для представления лингвистической информации. При этом подходе естественный язык описывается при

помощи систем правил. Примерами таких моделей могут служить сетевые грамматики Вудса [5] (о применении этих моделей для моделирования естественных языков смотри также [6]) и грамматики зависимостей [7]. (Следует отметить, что, как доказано в указанных работах, обе модели порождают в точности класс контекстно-свободных языков).

Обычно правила для таких языковых моделей строятся «вручную» исследователем, что сопряжено со значительными трудностями. Однако точность такой модели оказывается существенно выше, чем точность простых грамматик, типа n-грамм. Так, например, в [7] указано, что построенная таким образом грамматика для английской «деловой прозы» позволяет построить правильное дерево разбора для большей части тестового корпуса, составленного из газетных публикаций. При этом обеспечивается достаточно высокая скорость грамматического разбора: анализ предложения проводится за время порядка n^3, где n - длина предложения.

К сожалению, эти языковые модели также обладают рядом недостатков. Так, например, эти модели оказываются излишне «жесткими», то есть не пропускают предложения, не укладывающиеся в них. Для преодоления этих ограничений были разработаны вероятностные обобщения этих моделей [8]. Они сочетают в себе преимущества обеих моделей, позволяя использовать как более глубокие связи, существующие в естественном языке, так и обычный n-граммный подход, обеспечивающий «универсальность» построенной языковой модели.

Еще один возможный подход к использованию КС-грамматик в классической форме можно найти в [9].

2.4. Адаптивные модели

Этот класс моделей используется в ситуациях, когда обучающие и тестовые данные существенно различаются или не хватает данных для обучения языковой модели для какой-либо предметной области и приходится использовать данные из других предметных областей. При этом основная модель строится по обучающему корпусу, а по мере работы самого распознавателя модель корректируется «на лету»

с учетом той информации, которая извлекается из распознаваемого текста.

2.5. Анализ качества моделей

Хотя при моделировании систем распознавания речи естественно было бы использовать такой параметр, как средний уровень ошибок при распознавании, однако для его вычисления потребовалось бы проводить полный эксперимент с распознавателем, что не всегда удобно и возможно. (Более того, задача построения языковой модели может решаться и вообще в отрыве от какого-то фиксированного распознавателя и тогда вычисление среднего уровня ошибок распознавания оказывается невозможным).

Поэтому для анализа качества языковых моделей принято использовать так называемый коэффициент неопределенности (perplexity coefficient), введенный в [10], который может быть проинтерпретирован как (геометрическое) среднее ветвление в данной модели [11].

Для n-граммной модели коэффициент неопределенности задается формулой:

$$perplexity = \left(\sqrt[N]{\prod P(\omega_{i_k} \mid \omega_{i_{k-1}} \dots \omega_{i_{k-n+1}})} \right)^{-1},$$

где $\omega_{i_1}\omega_{i_2}...\omega_{i_N}$ – тестовый корпус, или

$$\log perplexity = -\frac{1}{N} \sum \log P(\omega_{i_k} \mid \omega_{i_{k-1}} \dots \omega_{i_{k-n+1}}).$$

Нетрудно видеть, что коэффициент неопределенности является функцией от построенной языковой модели и естественно языка (и текстового корпуса). Таком образом, при фиксированном языке он позволяет сравнивать различные языковые модели, а при фиксированном типе модели – оценивать сложность самих естественных языков.

2.6. Техники сглаживания

Как уже неоднократно отмечалось, на практике обучающие данные всегда неполны, то есть значительная часть теоретически воз-

можных n-грамм либо вообще отсутствуют, либо встречается слишком редко для того, чтобы можно было применить статистические методы для оценки вероятности их появления. Если такая n-грамма встретится во время работы, то правильный вариант распознавания будет отклонен или его вероятность будет существенно занижена. (Заметим также, что коэффициент неопределенности для тестового корпуса, содержащего отсутствующую в тестовом корпусе n-грамму окажется равным бесконечности).

Для преодоления этого эффекта используются техники сглаживания, которые переопределяют вероятности n-грамм таким образом, что все n-граммы получают отличные от нуля вероятности появления. При этом вероятности появления не встречавшихся или редко встречавшихся n-грамм корректируются с учетом n-грамм меньшего порядка. Подробнее о техниках сглаживания см. [11].

3. Анализ применимости существующих языковых моделей и их модификаций

Первая часть исследования состояла в анализе применимости существующих моделей к русскому языку. Были проверены следующие: стандартные n-граммы (для $n = 2$ and 3) и n-граммы со свободным порядком слов (смотри далее).

Для экспериментов с этим классом языковых моделей был разработан пакет программ, который позволяет строить полный список n-грамм по данному текстовому корпусу, использовать сглаживание (реализована техника линейного сбрасывания и торможения, linear discounting and back-off, см. [11]), вычислять вероятности предложений, коэффициент неопределенности тестового корпуса, и т.д.

3.1. Стандартные n-граммы

При проверке простейшей языковой модели (n-грамм с $n = 2$) оказалось, что число пар слов, встретившихся в полном корпусе (100 млн. слов) по одному разу, составило более 92коэффициент неопределенности превысил 500. Для $n = 3$ ситуация оказалась еще хуже. Тем

самым было подтверждено априорное утверждение о неприменимости стандартного статистического подхода для русского языка.

После этого были протестированы еще два подхода. Первый подход призван удалить из грамматической информации порядок слов, а второй – преодолеть трудности, связанные с большим количеством словоформ.

3.2. n-граммы со свободным порядком слов

Эти языковые модели вводятся следующим образом. Вероятности классических n-грамм $P(\omega_n \mid \omega_{n-1} \dots \omega_1)$ заменяются вероятностями $P(\omega_n \mid \{\omega_{n-1} \dots \omega_1\})$, где фигурные скобки обозначают множество, то есть языковая модель не учитывает порядок первых $n-1$ слов в n-грамме. Как было показано, эта модификация не привела к существенным улучшениям.

Кроме того, эксперименты показали, что такое явление, как свободный порядок слов в предложении практически не проявляется при работе с научно-техническими и деловыми текстами. Все отклонения, вызываемые свободным порядком слов, могут быть промоделированы при помощи методов сглаживания. Спонтанная речь устроена более сложно и в ней это явление будет играть серьезную роль. К сожалению, она практически не представлена в существующих текстовых корпусах и численная оценка погрешностей, которые она будет вносить в языковые модели, на данном этапе не может быть проведена.

4. Составные языковые модели

Второй подход, разработанный в ходе проведенных исследований, призван решить проблему большого количества словоформ. В русском языке связи между словами определяются не порядком слов в предложении, а морфологическими характеристиками слов. Вот простой пример. Рассмотрим следующее предложение:

«Черная собака укусила кошку»

Несмотря на то, что этот вариант является самым частотным, в русском языке можно переставить слова в предложении, например:

«Кошку укусила черная собака»

или даже

«Укусила черная кошку собака»

где между объектом (собака) и его свойством (черная) находится другой объект (кошка).

Для того чтобы иметь возможность связать между собой, например, объект и его свойства, они должны быть морфологически согласованы. Так, в этом примере, существительное и прилагательное имеют одинаковый падеж, род и число. Таким образом, очевидно, что морфологическая информация должна являться очень важной частью языковой модели.

Кроме того, как это уже упоминалось ранее, многие словоформы очень незначительно различаются с акустической точки зрения. Для того чтобы облегчить задачу распознавания этих словоформ, мы можем попытаться предварительно выделить морфологическую составляющую грамматики в отдельную модель.

Таким образом, мы можем ввести понятие категорной языковой модели, и в частности категорных n-грамм. Каждое слово в словаре имеет 15 атрибутов, определяющих грамматические свойства словоформы (см. Таблицу 1).

Каждый атрибут может иметь одно из нескольких значений (например, значениями для атрибута «часть речи» являются «существительное», «прилагательное», «глагол» и так далее). В том случае, если некоторый атрибут оказывается бессмысленным для данного слова, например, атрибут «время» для имени существительного, он получает специальное значение «не определено».

Таблица 1. Список используемых грамматических категорий и примеры разбора.

характеристики	черный		играло
	вариант 1	вариант 2	
часть речи	прилагательное	прилагательное	глагол
одушевленность	—	—	—
собственность	—	—	—
число	единственное	единственное	единственное
род	женский	женский	средний
падеж	именительный	винительный	—
краткость	нет	нет	—
вид	—	—	несовершенный
залог	—	—	—
время	—	—	прошедшее
возвратность	—	—	—
наклонение	—	—	изъявительное
лицо	—	—	—
тип числительного	—	—	—
тип местоимения	—	—	—

Множество значений определяет класс словоформы. Введенный набор атрибутов позволяет выделить в русском языке 562 морфологических класса. Таким образом, каждое слово в предложении может рассматриваться как его начальная форма плюс его морфологический класс, что позволяет разбить грамматику на две составляющие: изменяемую часть (основанную на морфологии) и постоянную часть (основанную на начальных формах слов). Например, слово «черный» будет преобразовано в начальную форму («черный») плюс два варианта списка морфологических характеристик (см. Таблицу 1).

4.1. Категорные языковые модели (изменяемая часть грамматики)

Основное внимание в данной работе было уделено разработке категорной части языковой модели. По полному корпусу (100 млн. слов) была построена категорная триграммная модель, после чего по фор-

муле (1) были вычислены коэффициенты неопределенности для возрастающей последовательности вложенных друг в друга корпусов.

Итоговый коэффициент неопределенности оказался равен 21,93. Следует дополнительно отметить, что построенная категорная грамматика позволяет нам быстро решать проблему определения окончаний, которая представляет значительные трудности в рамках традиционного подхода.

4.2. Языковая модель для начальных форм слов (постоянная часть грамматики)

Вторая составляющая грамматики строится обычным образом, это n-граммная языковая модель, построенная на начальных формах слов. В результате экспериментов удалось установить, что коэффициент неопределенности этой модели примерно в 2-2,5 раза выше, чем в случае английских (около 230 в нашей модели vs. примерно 100 в, например, [12]). К сожалению, точные значения коэффициента вычислить невозможно в связи с отсутствием текстовых корпусов достаточных размеры для обучения и тестирования. Для достижения более точных результатов необходимо создавать больших размеров.

5. Заключение

Эта работа представляет один из возможных подходов к решению проблем, препятствующих созданию промышленных систем распознавания слитной речи для русского языка. Показано, что предложенное в ней разложение общей языковой модели на две составляющие: модель, основанную на морфологии и модель, основанную на начальных формах слов, позволяет разработчикам использовать все преимущества n-граммного подхода. Кроме того, выделение морфологической информации в независимую модель позволяет справиться с проблемой акустической похожести различных словоформ одного и того же слова.

Предложенное решение проблемы может быть эффективно использовано для многих языков, где число словоформ достаточно велико, например, для языков славянской группы.

В результате проведенных теоретических изысканий был создан пакет программ для построения различных вариантов языковых моделей для русского языка, в том числе составных моделей, основанных на категорном подходе. Также было разработано программное обеспечение для осуществления морфологического анализа и синтеза со словарем 150 тыс. слов.

Автор выражает благодарность своему научному руководителю, д.ф.-м.н. Д. Н. Бабину и м.н.с. И. Л. Мазуренко за помощь в проведении исследований и написании пакета программ, а также фирме Intel Corporation (USA) за предоставленные текстовые базы данных и оборудование для расчетов.

Список литературы

[1] Соколова Е.Н. Алгоритмы лемматизации для русского языка. // Рабочий проект многоязычного автоматического словаря на 60 тыс. словарных статей. Т. 1. Лингвистическое обеспечение. М., 1984. С. 45–62.

[2] Кулагина О.С. Об автоматическом синтаксическом анализе русских текстов / Препринт. ИПМ АН СССР. М., 1987. №205.

[3] Мельчук И.А. Опыт теории лингвистических моделей «Смысл ↔ Текст». М.: Наука, 1974.

[4] Kanevsky D., Monkowsky M., Sedivy J. Large Vocabulary Speaker-Independent Continuous Speech Recognition in Russian Language // Proc. SPECOM'96. St.-Petersburg, October 28–31, 1996.

[5] Вудс В.А. Сетевые грамматики для анализа естественных языков // Кибернетический сборник. Нов. сер. М.: Мир, 1978. Вып. 13. С. 120–158.

[6] Кулагина О.С. Исследования по машинному переводу. М.: Наука, 1979.

[7] Sleator D., Temperley D. Parsing English with a link grammar. Pittsburgh, PA: Computer Science Dept., Carnegie-Mellon Univ., Oct. 1991. Tech. Rep. CMU-CS-91-196.

[8] Lafferty J.D., Sleator D., Temperley D. Grammatical trigrams: A probabilistic model of link grammar // Proc. AAAI Fall Symp. Probabilistic Approaches to Natural Language. Cambridge, MA, Oct. 1992.

[9] Холоденко А.Б. Использование лексических и синтаксических анализаторов в задачах распознавания для естественных языков // Интеллектуальные системы. Т. 4. Вып. 1–2. 1999. С. 185–193.

[10] Bahl L.R., Baker J.K., Jelinek F., Mercer R.L. Perplexity-A measure of the difficulty of speech recognition tasks // J. Acoust. Soc. Amer. Vol. 62. P. S63. 1977. Suppl. no. 1.

[11] EAGLES. Handbook of Standards and Resources for Spoken Language Systems. Mouton de Gruyter, 1997.

[12] Manhung S. and oth. Integrating a context-dependent phrase grammar in the variable n-gram framework // Proceeding of ICASSP. 2000.