УДК 81'25'322:811.133.1:004.4:004.738.5

Переходько И.В., Мячин Д.А.
Оренбургский государственный университет
E-mail: perehodko2008@yandex.ru

ОЦЕНКА КАЧЕСТВА КОМПЬЮТЕРНОГО ПЕРЕВОДА

Данная статья посвящена исследованию проблем оценки качества машинного перевода, что обусловлено все возрастающей Интернет-коммуникации с одной стороны и недостаточной разработанностью систем компьютерного перевода, которые в настоящее время не позволяют добиться адекватного перевода, с другой стороны. Подобное исследование позволяет выявить основные причины ошибок при машинном переводе.

В данной статье предпринята попытка исследования стратегий взаимодействия Человек – Машина при переводе, а также анализа текстов переводов Интернет-сайтов, выполненных с помощью систем компьютерного перевода, на основе лингвостилистического анализа и на основе автоматической обработки текстов с использованием метрики METEOR по методу N-грамм. Анализ извлеченных примеров позволил сделать вывод о том, что наибольшее количество ошибок связано с переводом семантических конструкций.

Практическая значимость исследования состоит в том, что разработка системы оценки качества машинного перевода позволяет выявить и систематизировать все недостатки компьютерных программ с целью их дальнейшего совершенствования, т. к. автоматизация процесса перевода сегодня приобретает ключевое значение, т. к. с ее помощью возможно выполнять большие объемы работ.

Ключевые слова: Интернет-коммуникация, компьютерный перевод, лингвостилистический анализ текста, лексические ошибки, синтаксические ошибки, стилистические ошибки.

В исследованиях проблем оценки качества машинного перевода большое внимание уделяется изучению переводов текстов Интернет-сайтов, выполненных с помощью систем компьютерного перевода. Внимание это обусловлено тем, что сегодня активное развитие получила Интернет-коммуникация. Интернет-коммуникация – это относительно новая и бурно развивающаяся речевая формация, которая отличается коммуникативным многообразием, полифункциональностью, динамизмом и не имеет себе равных по степени своего влияния на другие сферы общения. Проблемами Интернет-коммуникации занимаются такие ученые, как Колокольцева Т.Н, Лутовинова О.В. Значительная часть информации, предоставленной Интернет-коммуникациями, представлена зарубежными источниками, которые не имеют интерфейса на других языках, а следовательно, ограничивают доступ к информации, содержащейся в них, для потребителя [1]. В этих случаях пользователи сети Интернет прибегают к компьютерному (машинному) переводу.

Оксфордский словарь дает следующее определение машинного перевода: «Машинный перевод – это перевод, осуществляемый при помощи компьютера» [2]. Иными словами – это процесс, который использует двуязычные данные, включающие в себя лексику и грамматику обоих языков, а также модели фраз, используемые в данных языках. Результатом этого процесса является перекодирование текста одного естественного языка на другой.

Сегодня понятия автоматический перевод и автоматизированный перевод принято разграничивать, так как автоматический перевод – это перевод, выполненный исключительно машиной, без участия человека, а автоматизированный – это метод, при котором перевод осуществляется человеком, при использовании программного обеспечения, облегчающего этот процесс. Этот вид машинного перевода в английской терминологии также называют Machine-assisted.

Развитие систем компьютерного перевода получило свое начало в 40-х годах XX века. Впервые концепция машинного перевода была сформулирована в 1949 году Уорреном Уивером, директором отделения естественных наук Рокфеллерского Фонда, в его меморандуме, адресованном Фонду. В 1954 году в рамках Джорджтаунского эксперимента был проведен первый машинный перевод с русского языка на английский. В 1967 году в отчете наблюдательного комитета по автоматической обработке текстов национальной академии наук США была подчеркнута нецелесообразность разработки систем компьютерного перевода в качестве замены пере-

водчика, так как полностью автоматический машинный перевод удовлетворительного качества не может быть получен. В отчете также было признано, что необходимо использовать уже имеющийся опыт разработки систем машинного перевода для развития электронных программ, облегчающих работу переводчика – электронных словарей. Машина в состоянии выполнять рутинные операции не только быстрее чем человек, но и качественнее. Изменилось само отношение к электронной вычислительной машине: ее стали рассматривать как инструмент для автоматизации труда переводчика.

За последние несколько лет машинный перевод пережил значительные изменения и сегодня по типу осуществления перевода выделяют:

– системы прямого перевода: производится пословный перевод, отсутствуют модули полного семантического и синтаксического анализа;

– трансферные системы: используются переводные соответствия: эквивалентные, вариативные, трансферные, то есть преобразующие текст для правильной передачи;

– системы семантического перевода: принцип действия основан на применении семантических баз данных.

При этом машинный перевод без редактирования, выполненного человеком, не представляется возможным, так как программное обеспечение не дает возможности для подбора адекватного эквивалента некоторым семантическим конструкциям.

Следует выделить несколько стратегий взаимодействия Человек – Машина при переводе:

– компьютерный перевод с предредактированием – это преобразование текста перед его вводом в систему компьютерного перевода для воссоздания на языке оригинала конструкций языка перевода;

– компьютерный перевод с постредактированием – это преобразование «грубого» текста, выполненного системой машинного перевода, с целью привести его к нормам языка перевода;

– компьютерный перевод с интерредактированием – это взаимодействие человека и машины непосредственно во время перевода.

При этом используется простое и полное редактирование. Простое редактирование заключается в проведении как можно меньшего числа операций над текстом, с целью сделать его понятным, фактически точным и грамматически правильным и включает в себя:

– исправление наиболее очевидных опечаток и грамматических ошибок;

– изменение сложных предложений частично или полностью;

– фиксирование ошибок компьютерного переводчика;

– удаление ненужных или альтернативных вариантов перевода;

– создания глоссария, но без углубленной проработки терминов.

Этот вид редактирования применяется в тех случаях, когда необходимо передать только смысл.

Полное постредактирование – это более долгий процесс, итогом которого является текст, который читается, как будто он был написан на языке перевода. Этот процесс включает в себя:

– проверку соответствия терминологии;

– сбор информации, связанной с текстом перевода;

– синтаксические изменения в соответствии с правилами языка перевода;

– работа над стилистикой текста;

– перевод и адаптация культурологических явлений (фразеологизмы, идиомы и др.);

– попытка добиться полного соответствия с оригинальным текстом;

– выполнение форматирования, в соответствии с оригиналом текста;

– исправление всех грамматических, пунктуационных и орфографических ошибок.

Согласно данным пользователей автоматического перевода TAUS, такой вид постредактирования применяется чаще, так как позволяет сохранить инвариативность на уровне содержания [6].

В связи с тем, что машинный перевод без редактирования не способен выдавать адекватный и эквивалентный результат, возник вопрос об оценке текстов, выполненных с помощью систем компьютерного перевода.

Сегодня существует два основных направления оценки компьютерного перевода: на основе лингвостилистического анализа и на основе автоматической обработки текстов.

Макото Нагао предложил шкалу оценки машинного перевода на основе лингвостилистического анализа, от пяти до одного балла.

1 балл – смысл предложения понятен и не возникает никаких вопросов, грамматика, словоупотребление и стиль соответствуют общей структуре текста и не требуют постредактирования

2 балла – смысл предложения понятен, но возникают большие проблемы с грамматикой, словоупотреблением и стилем.

3 бала – общий смысл предложения понятен, но смысл некоторых его частей вызывает сомнение из-за неправильного грамматического строя.

4 балла – присутствуют ошибки словоупотребления и стилистики, требуется обращение к оригиналу.

5 баллов – в предложении имеется большое количество грамматических, словоупотребительных и стилистических ошибок, смысл предложения с трудом можно понять после внимательного изучения.

Недостаток этого метода оценки состоит в том, что человек не может быть объективен, так как уровень понимания текста реципиентом всегда зависит от индивидуальных, а значит, субъективных факторов, следовательно, оценка перевода по шкале Макото Нагао не может быть абсолютно точна [3].

Автоматические системы оценки машинного перевода основываются на методе N-грамм, который был введен для этой цели также Макото Нагао и Шинсуке Мори. Этот метод основан на использовании вероятности появления цепочки букв N-го порядка (N-грамм) в анализируемых текстах [7]. N-грамма – это последовательность из n-слов или знаков. По своему составу N-граммы могут делиться на униграммы (одно слово или один знак), биграммы (два), триграммы (четыре) и так далее.

Основными метриками, основанными на этом методе, являются BLEU и METEOR. Их авторами являются: Кишоре Папинени, Михаэль Денковски. Эти метрики работают по сути на одном алгоритме, но METEOR является более совершенным, так как допускает при оценке перифраз.

Алгоритм оценивания следующий: в программу загружаются два перевода одного и того же текста. Один, выполненный с помощью системы компьютерного перевода, второй, прошедший постредактирование. Оба этих текста делятся на N-граммы, после чего сравниваются друг с другом. Далее, используя методы математической статистики, высчитывается оценка компьютерного перевода. Оценка ставится от 0 до 1, где 0 – плохой результат, а 1 – отличный.

Эти метрики создали сотрудники IBM, чтобы отслеживать результат изменений в переводе в процессе разработки системы. Оценивается как меняется перевод при добавлении новых корпусов текстов для тренировки системы, при изменении программного кода и т. д.

С одной стороны, метод автоматической оценки машинного перевода более объективен, так как он позволяет оценить перевод с помощью формул, и тем самым исключить субъективность. С другой стороны, автоматические метрики оценки не могут дать полностью адекватный результат, так как при оценке с помощью N-грамм, не ставится задача понимания семантики текста, что и ведет к некоторой неточности такой оценки.

В настоящей статье освещены результаты исследования, в рамках которого были проанализированы переводы двух сайтов, один из которых имеет русский аналог, а другой не имеет такового.

Материалом нашего исследования являются технические тексты сайтов международной автомобильной компании Citroën (www.Citroën.fr) и французского Интернет-провайдера Orange (www.orange.fr).

В ходе исследования был осуществлен анализ переведенных с помощью компьютерных систем текстов с использованием метрики METEOR.

Мы перевели предложение «*Un orage, la foudre qui tombe pas loin de chez vous et votre box pourrait se retrouver hors service*» с помощью систем перевода Google translate и PROMT. Перевод от Google translate звучит так: *«Гроза, молния не выйди из дома и ваш ящик может обрываться»*, PROMT дает следующий перевод *«Буря, молния, которая падает на вас и ваш отсек мог бы вновь оказаться негодным»*. METEOR выдает переводу Google translate оценку 0,32, а переводу, выполненному с помощью системы PROMT, оценку 0,42.

В соответствии со шкалой Макото Нагао в переводе, выполненном и системой Google translate, и системой PROMT присутствуют ошибки словоупотребления и стилистики. Перевод требует обращения к оригиналу и существовать без постредактирования не может.

В связи с этим требуется лингвостилистический анализ для выявления причин ошибок в переводе. Анализ показывает, что причина неадекватного перевода в неспособности электронного переводчика понять смысл фразы. В результате постредактирования мы можем предложить следующий вариант перевода – «Гроза, молния вблизи роутера могут вывести его из рабочего состояния».

Следующий пример иллюстрирует грамматическую ошибку, допущенную машинным переводом. Так фраза *«Trois précautions valant mieux qu'une»* – «Три меры предосторожности будет лучше, чем один» оценивается по метрике METEOR 0,75. Согласно шкале Макото Нагао смысл предложения понятен, но возникают большие проблемы с грамматикой, словоупотреблением и стилем. Неверное согласование родов слов «мера» и числительного «один», возможен следующий перевод *«Три меры предосторожности лучше, чем одна».*

В процессе компьютерного перевода возможны синтаксические ошибки. Оригинал предложения – *«Fidèle à son tempérament de pionnier, André Citroën ouvre les portes de son usine du Quai de Javel au public à l'occasion du salon de l'automobile»*. Google translate выдает перевод *«Верный своему пионерской темперамент, Андре Ситроен открывает двери своего завода Кэ де Javel к публике на автошоу»*. Система PROMT *«Верный его темпераменту пионера, Андрэ Ситроен открывает двери своего завода Набережной Javel публике по случаю автомобильной выставки».* В русском аналоге сайта ситроен это предложение переведено *«Демонстрируя в очередной раз свой характер первооткрывателя, Андре Ситроен открывает двери своего завода на набережной Жавель для публики по случаю открытия парижского автомобильного салона»*. Перевод от Google получает оценку по метрике METEOR в 0,32, а PROMT – 0,39. По шкале Макото Нагао оба этих перевода получают 3 балла, т. е. общий смысл предложения понятен, но смысл некоторых его частей вызывает сомнение из-за неправильного грамматического строя.

На основе анализа текстов перевода можно сделать вывод о том, что PROMT делает меньше ошибок в синтаксисе и грамматике, в Google же, напротив, эта проблема еще не решена окончательно.

Лексические ошибки связаны чаще всего со специализированной лексикой, например, *«une rampe* ***commune d'injection alimente les injecteurs sous très haute pression»***. Эту фразу Google translate перевел как *«общий топливораспределительной рампе подает форсунки под высоким давлением»*, а PROMT *«общая эстакада инъекции снабжает инжекторы под очень высоким давлением»*. Перевод Google, и получает оценку по метрике METEOR в 0,25, а согласно шкале Макото Нагао в предложении имеется большое количество грамматических, словоупотребительных и стилистических ошибок, смысл предложения с трудом можно понять после внимательного изучения. В переводе PROMT присутствуют ошибки словоупотребления и стилистики, требуется обращение к оригиналу, и его оценка составила 0,27.

В данном случае ошибка вызвана именно наличием специализированной лексики, такой как «инжекторы» или «общая магистраль прямого впрыска». И после постредактирования в русской версии сайта Citroën это предложение выглядит *«топливо подается на инжекторы под высоким давлением по общей магистрали прямого впрыска».*

Также при сравнении переводов данных сайтов была выявлена ошибка, допущенная постпереводчиком *«Cinq ans après la Croisière Noire, la croisière Jaune.* ***Sans limites, Citroën s'attaque maintenant à la traversée du continent asiatique depuis Beyrouth jusqu'à Pékin»***. В русском варианте сайта эта фраза переведена как *«Через пять лет после «Черного рейда» организуется экспедиция «Желтый рейд». На этот раз компания Citroën не признавая никаких намеревается пересечь азиатский континент от Бейрута до Пекина».* Явно, что переводчик не перевел слово *«limites»,* что переводится как *«граница»,* и перевод, возможно, должен был звучать так: *«Через пять лет после «Черного рейда» организуется экспедиция «Желтый рейд». На этот раз компания Citroën, не признавая никаких гра-*

ниц, намеревается пересечь азиатский континент от Бейрута до Пекина».

Название одной из статей сайта Orange звучит как *«Sécuriser sa box, Quelques petits trucs à savoir»*. Электронный переводчик Google переводит это предложение как *«Безопасный ящик. Некоторые советы, чтобы знать»*, а PROMT *«Успокаивать его отсек. Несколько маленьких уловок которые надо знать»*. Этот перевод набирает 4 балла шкалы Макото Нагао, т. е. в предложении присутствуют ошибки словоупотребления и стилистики, требуется обращение к оригиналу. По метрике METEOR Google получает оценку – 0,03, а PROMT – 0,16.

Ошибка носит семантический характер, так как слово *«box»* здесь имеет значение *«роутер»,* а слово *«Sécuriser»* – это глагол, одно из значений которого – *«обеспечивать безопасность»,* т. е. такое словосочетание, по нашему мнению, может переводиться как *«несколько небольших советов по обеспечению безопасности вашего роутера»*.

Анализ текстов статей указанных сайтов, выполненных с помощью компьютерного перевода, позволил сделать вывод о том, что наибольшее количество ошибок связано с переводом семантических конструкций (50%), грамматические ошибки составили (20%), синтаксические и лексические по (15%) соответственно.

Таким образом, проведенный анализ позволяет сделать вывод о том, что наибольшие проблемы при машинном переводе связаны с семантикой, так как при переводе семантических конструкций необходимы базы данных, которые в настоящее время не разработаны. Помимо этого немалые затруднения вызывают переводы сложных грамматических, синтаксических и лексических конструкций.

Дальнейшее развитие автоматического перевода связано с возможностью целостной оценки текстов, выполненных при помощи систем компьютерного перевода. Именно адекватная и полная оценка поможет выявить и систематизировать все недостатки программы, чтобы в дальнейшем данные проблемы были решены.

08.02.2017

Список литературы:

1. Никифорова, Н.Ю. Лингвопрагматическая специфика рекламных текстов в компьютерно-опосредованном дискурсе и приемы их локализации [Электронный ресурс] / Н.Ю. Никифорова // Вестник ВолГУ. – Серия 9: Исследования молодых ученых. – 2013. – №11. – URL: http://cyberleninka.ru/article/n/lingvopragmaticheskaya-spetsifika-reklamnyh-tekstov-v-kompyuterno-oposredovannom-diskurse-i-priemy-ih-lokalizatsii (дата обращения: 27.12.2016).
2. Oxford University Press [Электронный ресурс]. – Oxford Dictionariesa, 2016. – Режим доступа: https://en.oxforddictionaries.com/
3. Бабина, О.И. Языковая личность переводчика и машинный перевод [Электронный ресурс] / О.И. Бабина // Вестник ЧелГУ. – 2011. – №24. – URL: http://cyberleninka.ru/article/n/yazykovaya-lichnost-perevodchika-i-mashinnyy-perevod (дата обращения: 27.12.2016).
4. Densmer, L. Light and Full MT Post-Editing Explained / MORAVIA [Электронный ресурс] / L. Densmer. – Режим доступа: http://info.moravia.com/blog/bid/353532/Light-and-Full-MT-Post-Editing-.
5. Чакырова, Ю.И. Постредактирование в транслатологической парадигме [Электронный ресурс] / Ю.И. Чакырова // Вестник ПНИПУ. Проблемы языкознания и педагогики. – 2013. – №8. – URL: http://cyberleninka.ru/article/n/postredaktirovanie-v-translatologicheskoy-paradigme (дата обращения: 27.12.2016).
6. Nagao, M. A New Method of N-gram Statistics for Large Number of n and Automatic Extraction of Words and Phrases from Large Text Data of Japanese / M. Nagao, S. Mori // In Proceedings of the 15th International Conference on Computational Linguistics (COLING 1994). – Kyoto, Japan, 1994.

Сведения об авторах:

Переходько Ирина Валерьевна, доцент кафедры романской филологии
и методики преподавания французского языка Оренбургского государственного университета,
кандидат педагогических наук, доцент
460018, г. Оренбург, пр-т Победы, 13, ауд. 171011, тел. (3532) 912233, e-mail: perehodko2008@yandex.ru

Мячин Дмитрий Алексеевич, студент факультета филологии и журналистики
Оренбургского государственного университета
460018, г. Оренбург, пр-т Победы, 13, e-mail: Dima.rishilie@gmail.com