ББК Ш141.2 + Ш11
УДК 821.161.1 + 81’322.4

ОЦЕНКА МАШИННОЙ ПЕРЕВОДИМОСТИ ТЕКСТОВ

О.И. Бабина

Статья посвящена проблеме определения параметров, затрудняющих адекватный машинный перевод текстов. Рассмотрены особенности графического, лексического и синтагматического уровня в научно-техническом тексте на русском языке. На основе анализа языковых особенностей составлена классификация формальных маркеров машинной переводимости русскоязычных научно-технических текстов. Полученные результаты могут найти применение в практике перевода и при разработке лингвистического программного инструментария для оценки машинной переводимости текстов.

Ключевые слова: корпус текстов, машинный перевод, машинная переводимость, маркер переводимости.

Ни один перевод не возможен без потерь в силу языковых и культурных особенностей. Поэтому для оценки адекватности перевода необходимо оценить, насколько принципиально возможно сохранить семантику исходного сообщения, выполнив адекватный перевод. Однако такая задача слишком обширна, и все корректные случаи охватить довольно сложно. С другой стороны, обратная задача – оценить потенциальные потери, которые могут возникнуть при переводе в силу различий в языках или сложности восприятия отдельных языковых конструкций, – представляется более осязаемой, и именно такая интерпретация переводимости зачастую используется [4]. Тогда задача сводится к проверке текста на предмет наличия заранее исчисленных проблемных случаев. Причем некоторые проблемы универсальны; другие – зависят от языков оригинала и перевода.

Сужая задачу определения доступности текста для перевода к оценке сложности перевода текста машиной (машинной переводимости текста), следует добавить еще одно «измерение», связанное с системой машинного перевода и обусловленное языковой моделью, которую данная система использует. Так, по нашим наблюдениям, статистические системы (например, использующие двух-, трехграммные модели), как правило, хорошо справляются с оценкой сочетаемости отдельных единиц, но в пределах весьма ограниченных языковых блоков (2–3 слов); с увеличением расстояния между связанными блоками качество правильной идентификации синтагм снижается – синтаксически такие системы довольно слабы. В системах, основанных на правилах, наоборот, точнее могут идентифицироваться

синтаксические единицы, что благотворно влияет на порядок слов в переводе, падежно-ролевую идентификацию и т.п., но при переводе, как правило, в недостаточной степени учитывается лексическая сочетаемость единиц. (Такое противопоставление статистических и детерминистких систем перевода не претендует на универсальность, это лишь общее наблюдение – для каждой индивидуальной системы могут в различной степени учитываться отдельные аспекты языка (в том числе, соотношение случаев, вызывающих проблемы при различных способах моделирования языка, может для отдельных систем быть прямо противоположным представленному), что обусловливает разнообразие систем, в том числе и точки зрения их качественной оценки). Вместе с тем, это не противоречит нашему общему тезису о том, что показатель сложности текста для машинного перевода – в связи с разнообразием систем – неизбежно зависит от того, перевод какой системы оценивается.

Стандартным подходом к оценке сложности текста для машинного перевода является использование так называемых маркеров переводимости – (negative) translatability indicator, – лингвистических особенностей текста, которые негативно влияют на качество перевода.

В зависимости от области распространения влияния такие маркеры можно классифицировать по различным критериям. Учитывая обозначенную ранее вариативность в реализации языковой модели в различных системах машинного перевода, в [9] авторы разграничивают общие и специфичные маркеры. К общим относят такие особенности, которые потенциально вызывают сложности для всех систем машинного перевода в силу континуальности языка, которую принципиально невозможно учесть ни в какой формальной модели, дискретной по своей природе. Специфичные маркеры характеризуют определенную систему машинного перевода с точки зрения ее способности успешно справляться с лингвистическими проблемами при автоматическом переводе.

Также маркеры можно разграничить по тому, какой этап процесса перевода они осложняют: понимание текста (маркеры, характеризующие текст на языке оригинала) или процесс переноса текста на другой язык (маркеры, обусловленные структурными различиями в рассматриваемой паре языков).

Проблемы машинного перевода (которые могут быть идентифицированы посредством маркеров переводимости) могут проявляться на различных лингвистических уровнях. Проблемы на **графическом уровне** представляют проблему на этапе анализа текста оригинала. Они носят скорее технический характер. Проблемы графического уровня связаны с вариативностью использования различных специальных символов и пробелов (например, система входов/выходов vs. система входов-выходов), опечатки и орфографические ошибки. Формальным маркером для машинной переводимости такого текста является отсутствие единицы в словаре и/или не-

возможность вывести единицу по предусмотренным системой правилам (например, правилам генерации форм морфологической парадигмы слова).

Проблемы переводимости на **лексическом уровне** связаны с: а) недостаточным объемом вокабуляра (покрываемость лексикона); б) вариативностью лексических единиц; в) асимметричностью языков оригинала и перевода.

Недостаточность вокабуляра является проблемой как для человека-переводчика, так и для системы машинного перевода. Однако если человек в ряде случаев может догадаться о значении лексической единицы на основе контекста, для системы машинного перевода отсутствие лексической единицы в словаре является критичным.

Вариативность лексических единиц может иметь следующие формы:

1) формальная вариативность (вариативность в плане выражения);

2) семантическая вариативность (вариативность в плане содержания).

Проблемы, связанные с формальной вариативностью могут быть представлены использованием морфологических дериватов и трансформаций словосочетаний для обозначения одного и того же понятия (например, симметричная задача коммивояжера vs. симметрическая задача коммивояжера). Другой ипостасью формальной вариативности является использование парадигматически связанных слов и выражений при обращении к одному и тому же референту (синонимов, гиперонимов) (например, вершина графа vs. узел графа, двуместная функция vs. бинарная функция). Также формальная вариативность проявляется в виде трансформаций (например, биномиальное разложение vs. разложение бинома). В машинном переводе проблема морфологической вариативности, фактически, сводится к ранее упомянутой проблеме покрываемости лексикона: если оба варианта указаны в словаре, они имеют заданную языковой моделью системы интерпретацию.

Семантическая вариативность представлена полисемией и омонимией (включая грамматическую омонимию) различных единиц (например, зависимости/ед.ч., Род.п. vs. зависимости/мн.ч., Им.п.). Проблема разрешения неоднозначности является универсальной для всех систем автоматической обработки текстов. Ограничение предметной области может в значительной степени способствовать сокращению многозначности знаменательных частей речи.

Переводимость на **синтагматическом уровне** определяется сложностью на уровне интерпретации зависимостей между единицами словосочетаний и предложений. Эта группа проблемных случаев наиболее существенна.

Идентификацию общих маркеров синтагматического уровня мы проводили на основе анализа корпуса научно-технических статей и аннотаций к ним на русском языке в предметной области «Математическое моделирование». В качестве языка перевода мы ориентировались на английский язык.

Маркер **длины предложения** (в словах) выделяют многие исследователи в области машинной переводимости (например, [7, 8]), так как, не-

смотря на то что он не является собственно лингвистической характеристикой предложения, этот показатель коррелирует с синтаксической сложностью предложения, что, в свою очередь, сказывается на автоматическом «понимании» текста. В рассматриваемом нами корпусе средняя длина предложения составляет 14,84 слова; максимальная длина слов в предложении составляет 85 слов; 24,68 % предложений имеют длину 20 и более слов. Эти показатели свидетельствует о синтаксической сложности текста.

Вставные (разрывные) конструкции затрудняют автоматическое установление границ синтагматических блоков и типа синтаксической связи между этими блоками. Неверное распознавание зависимостей между разрывно расположенными элементами текста также приводит к невозможности правильного учета лексической сочетаемости текстовых единиц. Все это в общем случае приводит к неверному переводу синтагматически тесно связанных составляющих, разделенных аппозитивной вставкой.

Кроме того, сама аппозитивная вставка иногда может вызывать трудности интерпретации, вызванные синтаксической омонимией. В рассматриваемых текстах аппозитивные вставки встречались в функции: а) дополнения: например, допустимость (маршрутов); б) парафразов-субститутов, эквивалентных по синтаксической функции уточняемому слову: например, время ретардации (запаздывания) давления; в) синтаксически не связанные со структурой предложения вставки: например, Методика ЭПР-дозиметрии является многоступенчатой (химическое приготовление, спектрометрические измерения, анализ спектров, калибровка), и на каждом из этапов возможно привнесение дополнительных погрешностей.

Согласно данным анализа корпуса, проблемные случаи с точки зрения автоматической идентификации синтаксической функции практически ограничиваются омонимией родительного падежа существительных, выступающих в качестве ядра аппозитивной вставки, так как родительный падеж допускает трактовку вставки как случай а) или б). При этом использование существительных в функции эквивалентных уточнений (случай б) значительно превосходит по частотности остальные случаи, что дает основания придавать небольшой вес данному признаку при определении степени его влияния на качество машинного перевода.

Падежная омонимия в исследуемом корпусе проявляется, чаще всего, для родительного и творительного падежа. Существительные в родительном падеже могут выступать в функции определения к другому существительному или заполнять отдельную валентность при предикате. Аналогичную проблему представляет творительный падеж. Кроме того, омонимия творительного падежа наблюдается на уровне семантических ролей: он может выражать инструмент (требуя предлога with при переводе), агент (требуя предлог by) или выполнять другие роли, не требующие при переводе специальной маркировки.

Сходная проблема возникает при идентификации синтаксических характеристик **предложных групп**: расположенная в постпозиции предлож-

ная группа может выполнять функцию определения для одного из предшествующих существительных или заполнять отдельную валентность предиката. Отчасти эта проблема может быть решена на словарном уровне – исчислением терминологических словосочетаний с предложными определениями в словаре. Например, целесообразно включение в словарь таких устойчивых терминов как производные в среднем, уравнение в частных производных и т.д.

Управление вызывает проблемы в случае различий в двух языках (например, сочетания зависимость от, влияет на и т.п. при переводе на английский язык будут требовать нетипичных для используемых предлогов эквивалентов). При контактном расположении единиц в тексте целесообразно лексикографическое решение проблемы. Однако, в тексте нередки случаи разрывного расположения предлогов.

Проблему при переводе могут вызывать **эллиптические конструкции** на русском языке, так как английский язык имеет более строгие требования к эксплицитности отдельных элементов предложения. Так, допустимое в русском языке опущение существительного в предложении Позиции остаются подобными используемым, что представляется неприемлемым при передаче на английский язык.

В рассматриваемых текстах достаточно распространены **сочинительные конструкции**, идентификация которых является одной из наиболее острых проблем при автоматическом анализе текстов и переводе. Сочинительная связь в корпусе зафиксирована между всеми группами знаменательных частей речи (прилагательными, именными и предложными группами, наречиями, глаголами в различных формах: финитных и нефинитных), а также между отдельными клаузами.

Одной из наиболее сложных проблем автоматического перевода является сочинение между именными группами, выполняющими функцию подлежащего в предложении. Учитывая, что в русскоязычных текстах аннотаций превалируют неопределенно-личные предложения, в которых подлежащее расположено в конце предложения, при переводе на английский язык – язык с фиксированным порядком слов – требуется позиционное изменение подлежащего, и корректность определения границ подлежащего, заданного сочинительной конструкцией (то есть границ сегмента текста, положение которого в линейной последовательности в языке перевода необходимо изменить), при такой постановке весьма существенно.

На сегодня существуют многочисленные работы, посвященные идентификации границ сочиняющихся конструкций. Общая стратегия решения проблемы в рамках рационалистского подхода состоит в определении аналогий в грамматической форме, семантике и структуре сочиняющихся членов предложения (главных элементов соответствующих фраз), см. [2, 3, 6 и др.).

Проанализировав случаи употребления конструкций, затрудняющих понимание и перевод, были выявлены формальные признаки, позволяю-

щие с высокой степенью достоверности отождествить наличие этого признака с проблемным (с точки зрения машинной переводимости) случаем в тексте. Такие признаки, объединенные в группы по типам сложностей, которые они вызывают, включают:

1. Синтаксическая сложность:

a) наличие предложений с длиной более 20 слов;

b) наличие двух и более глаголов в финитной форме;

c) количество скобок в предложении;

d) наличие последовательности «существительное в косвенном падеже + предлог», которая стоит после предиката, и перед предикатом нет существительного в именительном падеже – такое сочетание указывает на многозначность синтаксиса предложной группы;

e) наличие глаголов и существительных, для которых в словаре имеется статья, включающая предлог (например, в тексте распознано *зависит*, а в словаре имеется два вхождения – *зависит* и *зависит от*). Здесь мы исходим из гипотезы о существовании лексикографического решения проблемы различий в управлении между языками;

f) эллипсис:

i) количество сочетаний прилагательное (но не причастие) + предлог;

i.i) наличие прилагательного или причастия в конце предложения;

g) наличие *и/или* после предиката, и перед предикатом нет существительного в именительном падеже, что в большинстве случаев свидетельствует о наличии подлежащего, включающего сочинение.

2. Лексическая омонимия:

a) количество слов с лексико-грамматическими метками из различных статей словаря;

b) наличие предлогов *при, от, из.*

3. Лексико-грамматическая омонимия:

a) количество слов с несколькими лексико-грамматическими метками;

b) падежная омонимия:

i) наличие существительных с меткой творительного падежа;

ii. наличие существительных с меткой родительного падежа;

iii. количество *и/или* + существительное в род. п.; *и/или* + существительное в тв. п.

Выявленные признаки позволяют формализовать и автоматизировать проведение оценки машинной переводимости текстов: наличие слишком большого количества маркеров указывает на необходимость предредактирования текстов или должно свидетельствовать о необходимости отказаться от использования машинного переводчика. Сегодня разрабатываются различные средства автоматизации оценки перевода, позволяющие дать числовую характеристику переводимости текста, используя автоматиче-

скую оценку наличия маркеров в тексте (например, [1, 5]). Данное исследование позволило выявить ряд формальных критериев, которые должны лечь в основу подобного средства по оценке переводимости научных текстов для пары языков русский-английский.

Выявленные маркеры могут также оказаться полезными при составлении контроллируемого языка, целью которого является определение лингвистических ограничений, которым должен соответствовать текст, предназначенный для автоматического перевода.

Кроме того, чтобы составить основу контроллируемого языка, который должен использоваться для создания текста на языке оригинала, маркеры могут также послужить задаче разработки базы знаний для программного инструментария постредактирования переводов, выполненных с помощью автоматического переводчика. Следует заметить, что, фактически, сложность перевода и оправданность использования машинного переводчика для минимизации усилий предопределяется не столько сложностью исходного текста (хотя большинство систем автоматической оценки машинной переводимости и ориентированы именно на идентификацию маркеров в исходном тексте), сколько временем, потраченным на редактирование перевода, выполненного машиной. В этом смысле, оценка усилий на постредактирование представляется более адекватной, чем оценка синтаксической сложности исходного текста (хотя последняя, в целом, коррелирует со временем, затраченным на постредактирования). Вместе с тем, возникает противоречие, связанное с невозможностью использовать этот критерий на практике, так как учет времени в процессе реального постредактирования определяет значение параметра затраченных усилий уже постфактум – после того, как усилия уже затрачены и их оценка более не актуальна.

Как нам представляется, дальнейшие исследовательские усилия должны быть сосредоточены, с одной стороны, на разработке механизма учета влияния выделенных признаков на корректность машинного перевода (и, следовательно, опосредованно на объем усилий, которые потребуется затратить на постредактирование), с другой стороны, на определении путей использования извлеченной информации в автоматизированной процедуре постредактирования текстов, переведенных машиной.

Библиографический список

1. Gdaniec, C. The Logos Translatability Index / C. Gdaniec // Technology Partnerships for Crossing the Language Barrier: Proceedings of the First Conference of the Association for Machine Translation in the Americas AMTA. – Oct. 1994. – P.97–105.

2. Marinčič, D. Parding with Intraclausal Coordination and Clause Detection / D. Marinčič // Informatitica. – 2010. – No. 34. – P. 263–264.

3. Okumura, A. Symmetric Pattern Matching Analysis for English Coordinate Structures / A. Okumura, K. Muraki // Proceedings of the fourth conference on Applied Natural Language Processing (ANLP). – 1994. – P. 41–46.

4. Pedro, R. de. The Translatability of Texts: A Historical Overview / R. de Pedro // Meta. – 1999 – Vol.XLIV, No. 4. – P. 546–559.

5. Povlsen, C. Evaluating Text-type Suitability for Machine Translation a Case Study on an English-Danish MT System / C. Povlsen, N. Underwood, B. Music, A. Neville // Proceedings of the First International Conference on Language Resources and Evaluation (Granada, Spain) / ed. by A. Rubio, N. Gallardo, R. Castro & A. Tejada. – Vol. 1. – 1998. – P. 27–31.

6. Roh, Y.-H. Recognizing Coordinate Structures for Machine Translation of English Patent Documents / Y.-H. Roh, K.-Y. Lee, S.-K. Choi, Oh-W. Kwon, Y.-G. Kim // Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation *(De La Salle University, Manila, Philippines). – Nov. 2008*. – P. 460–466.

7. Roh, Y.-H. Long Sentence Partitioning using Structure Analysis for Machine Translation / Y.-H. Roh, Y.-A. Seo, K.-Y. Lee, S.-K. Choi // Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium(Hitotsubashi Memorial Hall, National Center of Sciences, Tokyo, Japan). – Nov. 2001. – P. 646–652.

8. Sheremetyeva, S. Handling Low Translatability in Machine Translation / S. Sheremetyeva // Proceedings of the Eleventh Conference of European Association of Machine Translation (EAMT) (Oslo, Norway). –June 2006. – P. 105–114.

9. Underwood, N.L. Translatability Checker: A Tool to Help Decide Whether to Use MT / N.L. Underwood, B. Jongejan // Proceedings of MT Summit VIII (Santiago de Compostela, Galicia, Spain) / ed. by B. Maegaard. – Sept. 2001. – P. 363–368.

К содержанию