ПРИКЛАДНАЯ ЛИНГВИСТИКА

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

МАШИННАЯ ПЕРЕВОДИМОСТЬ РУССКОЯЗЫЧНЫХ НАУЧНО-ТЕХНИЧЕСКИХ ТЕКСТОВ

О.И. Бабина

В статье рассматривается проблема определения параметров текста, оказывающих негативное влияние на качество машинного перевода. Дается определение и классификация маркеров переводимости текста. Рассмотрены явления языка графического, лексического и синтагматического уровня в научно-техническом тексте на русском языке. Особое внимание уделяется проблемным с точки зрения машинного перевода особенностям текста синтагматического уровня, включающим как универсальные, так и специфические для русского языка характеристики. На основе анализа языковых особенностей составлена классификация формальных маркеров машинной переводимости русскоязычных научно-технических текстов. Выделенные классы соотнесены с проблемами перевода текстов, о которых могут свидетельствовать соответствующие маркеры. Полученные результаты могут найти применение в практике перевода и при разработке инструментария для лингвоинформационной поддержки переводческой деятельности.

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

1. Введение

Проблема сложности перевода текстов давно привлекла внимание лингвистов. Переводимость текстов определяется как проблема принципиальной возможности перевода текста на другой язык. В попытке ответить на этот вопрос лингвисты прибегают, с одной стороны, к анализу лингвистических универсалий, доказывающих возможность переводимости, с другой стороны, к изучению языкового сознания, которое специфично для представителей различных культур. В [21], выполнив исторический обзор подходов к явлению переводимости, автор заключает, что ни один перевод не возможен без потерь в силу языковых и культурных особенностей, и указывает на необходимость оценки потенциальных потерь для определения адекватности перевода.

Языковая обусловленность потерь, так или иначе, вызвана неоднозначностью и нечеткостью естественного языка. Порожденные этим проблемы возникают при переводе человеком, и, тем более, это актуально для машины. В ряде классических работ по машинному переводу детально анализируется связь между неоднозначностью языка и трудностями машинного перевода, которые она вызывает [4, 11, 27].

В общем случае, восприятие и перевод текста машиной обусловлены некоторой моделью автоматического понимания, которая – по определению модели – представляет собой упрощенное (огрубленное) представление реальной когнитивной способности человека. Это предопределяет тот факт, что машина, по сравнению с человеком, способна распознать гораздо более ограниченное подмножество текстов, которые «укладываются» в используемую модель автоматической обработки. Этот интуитивно понятный факт подтверждается также опытом работы над предписывающими языками. Например, в [23] автор указывает, что при составлении правил для предписывающего языка, ориентированного на машинный перевод, составленный экспертами (редакторами, преподавателями, авторами, переводчиками) набор правил, обеспечивающих понимание текста человеком, является лишь подмножеством набора правил для обеспечения машинной переводимости текстов (42 из 59). Таким образом, машинная переводимость текстов является частным, более строгим случаем переводимости вообще.

Использование машинного перевода для перевода текстов, написанных на понятном машине языке, могло бы обеспечить высокое качество автоматического перевода. Однако основная масса текстов создается на неограниченном естественном языке, т. е. автор такого текста не использует какой-либо предписывающий язык, который бы учитывал особенности использования автоматического перевода в целом или какой-то отдельной системы машинного перевода (при этом некоторые такие тексты, например, научно-техническую литературу, все еще целесообразно переводить с использованием средств автоматизации, хотя и с бóльшими затратами на пред- и постредактирование). В этом случае машинный переводчик сталкивается с проблемами, ухудшающими машинную переводимость текста.

2. Маркеры переводимости

Сложность текста для машинного перевода может быть заранее оценена с использованием так называемых маркеров переводимости (translatability indicator или negative translatability indicator), то есть лингвистических особенностей текста, которые негативно влияют на качество перевода.

С точки зрения системы машинного перевода в [28] авторы выделяют два вида таких маркеров:

1. Общие маркеры: потенциально вызывают сложности для всех систем машинного перевода. Очевидно, проблемы перевода текстов, характеризующиеся этими маркерами, обусловлены внутренней структурой естественного языка, многозначностью лингвистических единиц. Эти маркеры обусловлены принципиальной невозможностью (по крайне мере, в обозримом будущем) ни в какой формальной модели, дискретной по своей природе, учесть континуальность языка [1].

2. Специфичные маркеры (для определенной системы МП): вызывают проблемы для конкретной системы машинного перевода. Проблемы переводимости, связанные с этими маркерами, могут быть обусловлены двумя причинами: а) ограничениями модели естественного языка в данной системе; б) специфичными особенностями определенного языка. Так, в [10] бóльшая часть маркеров (16 из 20) специфичны для японского языка и в основном представляют собой определенные лексические единицы и конструкции, используемые в тексте на японском языке. В [8] также для некоторых маркеров указывается, что они учитываются только для немецкого языка.

Еще одно измерение для разграничения проблем переводимости можно сформулировать в зависимости от этапов выполнения перевода:

1) проблемы понимания (анализа) текста оригинала;

2) проблемы переноса (собственно перевода) текста оригинала на язык перевода;

3) проблемы синтеза текста перевода.

Таким образом, проблемы переводимости текстов можно рассматривать в нескольких ортогональных измерениях:

1) субъект перевода: человек / машина;

2) этап перевода: анализ / собственно перевод / синтез;

3) универсальность: зависимость / независимость от системы перевода и конкретного языка.

Следует отметить, что некоторые маркеры могут быть отнесены одновременно к нескольким категориям. Например, присоединение предложных групп (*John hit a girl with a skirt with white stripes with a bat with vengeance*) или разрывные лексические единицы (*Diese Technologien schliessen die Elektronik und die Büroautomation ein*) представляют сложность одновременно для человека-переводчика и машины.

Такая классификация маркеров указывает на то, что инструменты пред- и пост-редактирования должны опираться в своей работе, вообще говоря, на различные маркеры переводимости. Кроме того, разговор о максимально полной оценке переводимости уместен лишь в условиях ограничения анализа конкретной парой языков и конкретной системой машинного перевода.

3. Проблема переводимости научно-технического текста на русском языке

Мы будем ориентировать наше исследование на корпус научно-технических статей и аннотаций к ним на русском языке в предметной области «Математическое моделирование», поэтому особенности, зависящие от конкретного естественного языка, мы будем рассматривать в контексте данного подъязыка. В качестве языка перевода будем рассматривать английский язык.

Рассмотрим подробнее лингвистические особенности текстов данного подъязыка с тем, чтобы выявить проблемы, которые могут вызвать сложности при машинном переводе.

Проблемы перевода (которые могут быть идентифицированы маркерами переводимости) могут проявляться на различных лингвистических уровнях:

1) графический уровень,
2) лексический уровень,
3) синтагматический уровень.

Проблемы на ***графическом уровне*** представляют проблему на этапе анализа текста оригинала. Такие проблемы носят технический характер. Они, как правило, не представляют сложности для человека-переводчика, но могут снижать качество автоматического перевода.

Проблемы графического уровня связаны с вариативностью использования различных специальных символов и пробелов, например:

«внешние» фрагменты vs. *внешние фрагменты,*

прямое/обратное уравнение теплопроводности vs. *прямое / обратное уравнение теплопроводности,*

задачи Штурма- Лиувилля vs. *задача Штурма-Лиувилля,*

система входов/выходов vs. *система входов-выходов.*

К проблемам этого рода можно также отнести опечатки и орфографические ошибки, например:

В-третьем параграфе рассматривается задача (0.7).

Качество управления оценивается функционалом, заданым на множестве фазовых траекторий рассматриваемой системы.

Разработанные алгоритмы позволяют вычислять значения собственной функции возмущенного оператора независимо от того, известны предыдущие значения собственных функции или нет.

В первом случае автор, очевидно, объединил словосочетание *в третьем* и вводное слово *в-третьих*, что привело к образованию окказиона-

лизма *в-третьем*. Во втором – краткая и полная форма причастия *заданный* слиты в орфографически неверную форму *заданый*. В третьем случае очевидна простая опечатка в слове *независимо*.

Формальным маркером для машинной переводимости такого текста является отсутствие единицы в словаре и/или невозможность 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*, *несимметричное распределение* vs. *асимметричное распределение, доминирующее собственное значение* vs. *доминантное собственное значение* и т. д.

Другой ипостасью формальной вариативности является использование парадигматически связанных слов и выражений при обращении к одному и тому же референту (синонимов, гиперонимов), например, *вершина графа / узел графа*, *двудольный граф / биграф*, *двуместная функция /бинарная функция, теорема Гаусса-Остроградского / теорема о дивергенции, среднеквадратическое отклонение / стандартное отклонение.*

Также формальная вариативность проявляется в виде трансформаций: *биномиальное разложение* vs. *разложение бинома*, *эргодическая марковская цепь* vs. *эргодическая цепь Маркова, экстремальный анализ* vs. *анализ на экстремум* и т. д.

Проблема для человека в случае формальной вариативности вызвана тем, что он может сомневаться, являются ли данные термины морфологическими или лексическими вариантами для обозначения одного понятия или указывают на два разных понятия. 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. В системе машинного перевода такая интерпретация может эксплицироваться посредством эквивалентов на языке перевода: для вариантов одного понятия дается одинаковый перевод.

Семантическая вариативность представлена полисемией и омонимией (включая грамматическую омонимию) различных единиц. Например, *зависимости*/ед.ч., Род.п. vs. *зависимости*/мн.ч., Им.п.; *определение* (*понятия*) vs. *определение* (*параметров*). Человек, как правило, легко может восстановить из контекста лексико-грамматические и семантические характеристики терминов, в то время как для машины необходимо определить процедуру разрешения неоднозначности, обусловленной семантической вариативностью. Разрешение этой проблемы определяется способом моделирования языка в системе.

Ограничение предметной области может в значительной степени способствовать сокращению многозначности знаменательных частей речи. Однако служебные части речи, в частности, предлоги, в массе своей сходны в различных подобластях, причем многозначность, присущая служебным частям речи, в значительной степени сохраняется. Некоторые предлоги имеют различные способы перевода (*при – during, at, under, when, из – from, of, based on, по – on, along, for*), причем вряд ли можно выделить доминирующий:

при работе с системами порядка нескольких тысяч частиц (during the work)

с вырожденным оператором при производной (at the derivative)

при некоторых дополнительных условиях на оператор переопределения (under certain additional conditions)

общий результат использован при исследовании исходной обратной задачи (when investigating).

Переводимость на ***синтагматическом уровне*** определяется сложностью на уровне интерпретации зависимостей между единицами словосочетаний и предложений. Далее рассмотрим подробнее проблемы данного уровня.

4. Переводимость научно-технических текстов на синтагматическом уровне

4.1. Длина предложения

Длина (количество слов в предложении) не является собственно лингвистической характеристикой предложений. Однако, как широко отмечается исследователями в области машинной переводимости, длина предложения в значительной степени коррелирует с синтаксической сложностью: чем длиннее предложение, тем вероятнее, что оно содержит придаточные предложения, аппозитивные конструкции, предложные конструкции и другие элементы, осложняющие синтаксическую структуру. Слишком длинные предложения даже человеку бывает сложно

интерпретировать. При машинном переводе грамматические модели, заложенные в систему, не всегда способны корректно обработать такой языковой материал. При работе со сложными предложениями приходится разрабатывать техники их разбиения на части, чтобы оперировать синтаксически простыми блоками [25, 26].

В научных текстах по математическому моделированию средняя длина предложения составляет 14,84 слова, при этом предложение с максимальной длиной включает 85 слов. Из 5353 предложений корпуса, 1321 (24,68 %) имеют длину 20 и более слов. Достаточно представительная коллекций таких предложений косвенно свидетельствует о синтаксической сложности текста.

4.2. Вставные конструкции

Вставные конструкции представлены словами, словосочетаниями, предложениями, которые содержат уточнения, пояснения, поправки к сказанному, указанные в скобках или выделенные запятыми. Структурно вставные конструкции могут быть простыми (однословными) или распространенными (состоящими из нескольких слов). В текстах аннотаций к статьям по математическому моделированию такие конструкции могут быть выражены:

А. Прилагательными:

Также оставался открытым вопрос о возможности разрешения задач управления с функциональными ограничениями в более узком (классическом) множестве стратегий – позиционных стратегий.

Б. Существительными и именными группами (включая группы, распространенные причастными оборотами):

Рассматривается одна конструкция параллельной реализации метода динамического программирования для решения задачи последовательного обхода множеств (мегаполисов) с ограничениями в виде условий предшествования, именуемая обобщенной задачей курьера.

Представлены результаты математического моделирования гидрогазодинамических процессов (зависимости тяги кольцевого сопла и расхода рабочего тела от параметров двухфазной среды), протекающих при движении двухфазной среды в кольцевом сопле с укороченным центральным телом.

В. Обстоятельственными конструкциями:

При изучении прочностных свойств пластического слоя (например, при исследовании НС и несущей способности сварных соединений арматуры, в которых пластическим слоем может быть сварной шов или прослойка в ЗТВ) можно выделить три модельных случая распределения прочности по толщине слоя.

Исследование устойчивости уравнений (0.4) (в терминах дихотомии решений) было начато в [10].

Г. Предложениями (в том числе, эллиптическими):

Напряжения, направленные по нормали к поверхности оболочки, всюду внутри оболочки, считаются равными разности внешних давлений (при их равенстве или отсутствии – равными нулю).

Вставные конструкции, во-первых, могут разбивать синтаксически тесно связанные блоки, тем самым затрудняя грамматический разбор предложения и поиск контекстно-обусловленного эквивалента. Например,

Предложенный метод обращения дает возможность эффективно решить как прямую (т.е. задачу нахождения решения), так и обратную(т.е. задачу нахождения правой части уравнения по экспериментально полученному решению) задачи.

В случае многозначности единиц дистантное расположение контекста, отделенного от многозначной единицы аппозитивной вставкой, усложняет машинное «понимание» этой единицы и, как результат, перевод в общем случае оказывается неверным. Так, в приведенном примере, *обратная задача* имеет устойчивый эквивалент *inverse problem*, однако достаточно длинная аппозитивная вставка, разделяющая словосочетание, повышает вероятность того, что машина не сможет распознать слова *обратная* и *задача* как контексты друг для друга, и тогда каждая из единиц может быть переведена любой комбинацией заданных лексиконом системы эквивалентов для этих многозначных слов (например, **converse task, *reverse problem* и т. д.)

Во-вторых, проблематичным может быть перевод вставных номинативных конструкций. Синтаксически такие конструкции выступают в роли:

А. Дополнения к элементу основного текста:

Предложено расширение исходной задачи, использующее эквивалентное преобразование системы ограничений, в результате чего допустимость (маршрутов) по предшествованию заменяется допустимостью «по вычеркиванию»(заданий из списка).

Б. Парафразов, синтаксически эквивалентных именной группе из основного текста:

Параметры определяют время ретардации (запаздывания) давления.

Метод основан на детектировании концентрации стабильных радикалов в гидроксиапатите (минеральной составляющей кальцифицированных тканей).

В вершинах графа заданы условия, где через обозначено множество ребер с началом (концом) в вершине.

В. Конструкции, эквивалентной отдельному назывному предложению, синтаксически не связанному с другими элементами текста:

На основании экспериментальных данных, полученных в лабораториях Института физики металлов (Екатеринбург, Россия)…

Методика ЭПР-дозиметрии является многоступенчатой (химическое приготовление, спектрометрические измерения, анализ спектров, калибровка), и на каждом из этапов возможно привнесение дополнительных погрешностей.

В случае использования вставных конструкций, представленных существительными в родительном падеже, затруднительно по формальным критериям отнести их к первой или второй подгруппе. Например,

Получен дивергентный критерий отсутствия притяжения (аттрактора) для нелинейной системы обыкновенных дифференциальных уравнений.

Вставка «*аттрактора*» (формально) может а) выполнять функцию определения к предшествующему существительному (тогда ее следует переводить **... the lack of attraction (of the attractor) criterion…*или **... the lack of (the attractor) attraction criterion…*), б) являться уточнением, эквивалентным предшествующему существительному (перевод: *…the lack of attraction (attractor) criterion …*).

Однако следует заметить, что в корпусе в значительной степени доминируют аппозитивы в функции эквивалентных уточнений.

4.3. Падежная омонимия

Падежная омонимия в исследуемом корпусе чаще всего проявляется для родительного и творительного падежа. Существительные в родительном падеже могут выступать в функции определения к другому существительному или заполнять отдельную валентность при предикате. Например,

Разрешающая способность таких установок прямо зависит от напряжения электрического поля.

Аналогичную проблему представляет творительный падеж:

При некоторых дополнительных условиях на оператор переопределения методами теории вырожденных полугрупп операторов доказана теорема существования и единственности классического решения.

При решении задачи нахождения коэффициента гидропроводности численными методами необходимо учитывать особенности задач подземной гидромеханики.

В первом случае, именная группа заполняет самостоятельную валентность при предикате *доказана*; во втором – входит в состав более крупной фразы и определяет существительное *решении*.

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. Например,

Результаты численных расчетов собственных чисел и значений собственных функций хорошо согласуются с результатами, полученными известными методами. [инструмент]

Затем А.Л. Шестаковым и Г.А. Свиридюком впервые для решения задачи восстановления динамически искаженного сигнала было предложено использовать методы теории оптимального управления. [агент]

Она обусловлена неустойчивостью вычислительных процедур нахождения ранга матриц и решения систем линейных однородных алгебраических уравнений. [причина]

4.4. Присоединение предложной группы

Расположенная в постпозиции предложная группа может выполнять функцию определения для одного из предшествующих существительных или заполнять отдельную валентность предиката. Например, в следующем предложении предложная группа «для однородных уравнений» формально может быть подчинена нескольким кандидатам.

В статье рассматриваются аппроксимации Хилле-Уиддера-Поста для операторов разрешающей сильно непрерывной полугруппы для однородных уравнений.

При переводе на английский язык это может вызывать проблему, в случае если необходима перестановка в переводном тексте, что соответствует двум случаям:

1) предложная группа является определением к существительному, выполняющему функцию подлежащего, или определяющему его существительному, причем в русском языке имеет место обратный порядок слов. В этом случае при переводе необходимо, чтобы предикат был расположен после подлежащего, поэтому встает проблема определения границы подлежащего. В нашем корпусе текстов такая ситуация типична для предложений с возвратными глаголами в начальной позиции.

2) типичная структура предложения в английском языке для некоторого предиката требует наличия обстоятельства, выраженного предложной группой, в начале предложения. В корпусе научно-технических текстов, это, в основном, касается обстоятельственных конструкций, выражающих цель или условие. В тексте они выражаются с помощью предложных групп с предлогами *для, при*.

Следует отметить, частично 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.

4.5. Управление

Управление вызывает проблемы в случае различий в двух языках. Особенную актуальность данная проблема имеет при разрывном расположении предикатной единицы (глагола, отглагольного существительного) и предложной группы (при контактном расположении целесообразно лексикографическое решение проблемы). Например,

Выполнив необходимые преобразования, получим зависимость функционала от переменных.

4.6. Эллиптические конструкции

Чаще всего в корпусе научно-технических текстов в эллиптических конструкциях опускается существительное (восстанавливаемое из контекста текущего или предыдущего предложения) и сохраняется определение к нему. Например,

Позиции остаются при этом достаточно «простыми» и подобными (в новых условиях) используемым в [1–4].

4.7. Сочинительные конструкции

Идентификация сочинительных конструкций является одной из наиболее острых проблем при автоматическом анализе текстов и переводе. Чаще всего, существующие работы ориентированы на обработку отдельных видов сочинительных конструкций: именных групп, именных частей сложного именного сказуемого и др.

Рассмотрим виды сочинительных конструкций, встречающихся в корпусе научно-технических текстов на русском языке:

1. Сочинение прилагательных-определений к одному существительному. Например,

В качестве регуляризующего алгоритма решения при этом был использован метод проекционной регуляризации для прямого и обратного преобразований Фурье. Пространства подбираются таким образом, чтобы купировать те или иные краевые [1–4] или какие-нибудь другие [5] условия.

Вариантом сочинительной связи этого типа может выступать также конструкция со сложным союзом *как ... так и*:

Существует большое число как теоретических, так и экспериментальных исследований данной проблемы.

2. Сочинение именных частей сложного именного сказуемого:

Позиции остаются при этом достаточно «простыми» и подобными (в новых условиях) используемым в [1–4].

Это позволяет уравнения равновесия на контактной поверхности представить как систему трансцендентных (не дифференциальных) уравнений, решение которой может быть численным или приближенным аналитическим.

3. Сочинение инфинитивов, входящих в состав сложного глагольного сказуемого:

а) Сочинение собственно глагольных единиц в форме инфинитива, имеющих общий зависимый член:

Существенным является то, что подсетки являются несогласованными, что дает возможность уменьшать или увеличивать шаги подсеток в подобластях, руководствуясь лишь физическими особенностями задачи.

б) Сочинение инфинитивов, каждый из которых включает набор собственных зависимых членов:

Оказалось, что знание индексов и существенных многочленов конечной последовательности, составленной из марковских параметров системы, позволяет сразу же определить индексы наблюдаемости и управляемости системы, построить дробные факторизации передаточной матрицы-функции, решить соответствующие уравнения Безу, найти минимальную реализацию системы.

4. Сочинение простых предикатных единиц в простых предложениях:

а) Сочинение собственно предикатных единиц (для предикатов с одним обязательным актантом, синтаксически выполняющим роль подлежащего):

На этой основе получен и реализован алгоритм нахождения критической растягивающей нагрузки в зависимости от размеров и расположения дефекта, угла наклона контактной поверхности и коэффициента механической неоднородности.

б) Сочинение предикатных конструкций, включающих предикат и все зависящие от него элементы, при этом:

– предикат может быть в финитной форме:

Задача (0.2) превращается в задачу Шоуолтера – Сидорова и поэтому считается естественным обобщением последней.

Мы даем краткое введение в теорию производных в среднем, исследуем преобразование уравнений к каноническому виду и находим формулы для решений в терминах производных в среднем винеровского процесса.

– предикат может быть выражен причастием, вместе с зависимыми от него членами выполняющим функцию определения к одному из членов предложения:

Наш подход основан на концепции относительного спектра, предложенной Г.А. Свиридюком [17] и развитой его учениками [18–20].

в) Сочинение предикатных конструкций в повелительном наклонении

Зафиксируем в последнем уравнении переменные X и Y и рассмотрим его как отображение.

В отличие от предыдущей подгруппы, где предикаты зависят от одной и той же именной группы, данный вид сочинения эквивалентен сочинению простых предложений, так как формальное подлежащее в них отсутствует.

5. Сочинение предложных групп:

а) С одинаковыми предлогами:

Она редуцирована к линейной обратной задаче для дифференциального уравнения первого порядка в банаховом пространстве с вырожденным оператором при производной и с переопределением на подпространстве вырождения.

б) С различными предлогами:

Применительно к задачам электрофизики происходит локальное сгущение сетки вблизи поверхности старта заряженных частиц или в районе кроссовера пучка, то есть в подобластях зна-

чительно влияющих на поведение пучка во всей расчетной области.

Предложные группы синтаксически могут выступать в роли обстоятельств или определений к именным группам.

6. Сочинение наречий, а также наречий и обстоятельственных оборотов, выраженных именными или предложными группами:

По данным испытаний строятся кривые реагирования, которые затем обрабатываются графоаналитически или другими способами.

7. Сочинение именных групп:

а) В составе определительной предложной группы зависящие от одного и того же предлога, инициирующего определительную сочинительную конструкцию:

Абстрактные результаты проиллюстрированы конкретными начально-конечными задачами для уравнений и систем уравнений в частных производных.

б) В функции генитивного определения к существительному:

В связи с этим возникает необходимость построения адекватных математических моделей и их дальнейшего изучения

в) В функции прямого дополнения (зависят от одного и того же предиката):

Мы предполагаем, что каждая дуга имеет длину и ширину.

Возникает задача о том, как выбрать очередность посещения производственных помещений, а также конкретную траекторию перемещений, включающую «внешние» фрагменты и участки, связанные с пребыванием в самих помещениях.

Надо выбрать очередность посещения мегаполисов и конкретный вариант реализации упомянутого посещения.

г) В функции подлежащего: сочиняющиеся именные группы могут выполнять функцию подлежащего как в простом предложении, так и в различных частях сложносочиненных и сложноподчинённых предложений.

В текстах на русском языке такие сочинительные конструкции можно разделить на две подгруппы: а) подлежащее (выраженное сочиненными именными группами) расположено перед предикатом (прямой порядок слов); б) подлежащее (выраженное сочиненными именными группами) расположено после предиката (инвертированный порядок слов).

Для сочинения именных групп последний случай может представлять значительную проблему, так как перевод на английский язык требует преобразования порядка слов. Например,

В работе [2] с целью регуляризации и оценки сходимости построенных решений были применены преобразование Фурье и метод сопряженных градиентов.

Предложение на английский язык должно быть переведено с перестановкой: вначале подлежащее, включающее именные группы, соединенные сочинительным союзом, затем – предикат.

Такая конструкция может быть осложнена причастными оборотами и придаточными предложениями в функции определения, сопровождающими одно или оба элемента сочинительной конструкции:

Рассмотрен линейный дифференциальный оператор и система краевых условий, задаваемая линейными в пространстве n раз непрерывно дифференцируемых функций линейно-независимыми функционалами

В таком случае вся сочинительная конструкция, включая определительный причастный оборот, должна быть поставлена в тексте перевода перед предикатом (глаголом в пассивном залоге).

Как можно видеть из приведенных примеров, сочиняющиеся именные группы могут иметь сложную структуру и распространяться посредством генитивных определений, причастных оборотов, подчинительных предложений, что также может представлять сложность для перевода.

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. [14, 20, 24 и др.]).

Также сегодня проблему пытаются решить с применением статистических методов для решения задачи. В [9] используется статистическая модель совместной встречаемости терминов. Подобный подход, используя частоты n-грам, собранных в Интернете, применили в [15]. В [5] для определения границ сочиняющихся сложных именных групп в качестве тренировочного корпуса используется одноязычный и параллельный корпусы текстов, по данным которых выявляются ассоциации между лексическими единицами. В [19] авторы предлагают алгоритм выявления отсутствия сочинения между единицами в определенном блоке текста (на японском языке) на основе расчета расстояний между словами.

4.8. Синтаксическая асимметрия языков

В отдельных случаях «хороший» перевод текста требует синтаксического преобразования. В корпусе текстов такая ситуация наблюдается для случаев неопределенно-личных предложений. Например,

Для получения регуляризованных решений используют преобразование Фурье и стабилизирующий функционал.

Эти предложения целесообразно передавать на английский язык посредством пассивизации; тогда происходит трансформация объекта в тексте оригинала в субъект в тексте перевода.

5. Маркеры переводимости в научно-техническом тексте

Проанализировав текстовый материал, мы определили формальные маркеры, которые соответствуют выявленным лингвистическим проблемам текста и могут отрицательно влиять на переводимость текстов. Следует отметить, что для выявления формальных маркеров необходима предварительная разметка текста по частям речи с помощью лексикона, включающая для каждой единицы текста набор всех возможных меток, без снятия многозначности. Выделенные маркеры и соответствующие им проблемы при переводе включают:

1) наличие предложений с длиной более 20 слов: синтаксическая сложность;

2) количество слов с лексико-грамматическими метками из различных статей словаря: лексическая омонимия;

3) количество слов с лексико-грамматическими метками из одной статьи словари: грамматическая омонимия;

4) наличие предлогов *при, от, из*: лексическая многозначность;

5) количество скобок в предложении: разделение контекстно зависимых элементов;

6) наличие существительных с меткой творительного падежа: падежная омонимия;

7) наличие существительных с меткой родительного падежа: падежная омонимия;

8) наличие последовательности «существительное в косвенном падеже + предлог», которая стоит после предиката*, и перед предикатом нет существительного в именительном падеже*: многозначность синтаксиса предложной группы;

9) 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*): разрывное расположение управляющего предлога;

10) количество сочетаний прилагательное* (но не причастие*) + предлог: эллипсис;

11) наличие прилагательного* или причастия в конце предложения: эллипсис;

12) количество *и/или* + существительное в род.п.; *и/или* + существительное в тв.п.: падежная омонимия;

13) наличие *и/или* после предиката*, и перед предикатом нет существительного в именительном падеже*: границы подлежащего, включающего сочинение;

14) наличие активных предикатов в форме 1 л. мн.ч.: необходимость синтаксической трансформации;

15) наличие двух и более глаголов в финитной форме: синтаксическая сложность, в частности, границы сочиняющихся предикатных конструкций или сложная валентность.

Пометка звездочкой обозначает, что слово распознается как данная форма/часть речи, если в его разметке есть хотя бы одна соответствующая метка.

6. Заключение

Выделенные маркеры позволяют формализовать и автоматизировать проведение оценки «пригодности» текста для машинного перевода: наличие слишком большого количества маркеров указывает на необходимость предредактирования текстов или свидетельствует о необходимости отказаться от использования машинного переводчика. Сегодня разрабатываются различные средства автоматизации оценки перевода (ориентированные на определенные системы и языки), позволяющие дать числовую характеристику переводимости текста (оценить индекс переводимости), используя автоматическую оценку наличия маркеров в тексте (например, [8, 22]). Применяются процедуры вычисления такого индекса, которые состоят в том, что априори каждое предложение оценивается максимальным количеством баллов, а за каждый маркер – с учетом его веса – начисляются штрафные баллы, которые вычитаются из первоначальной оценки [6, 8]. При этом веса маркеров могут зависеть от пары анализируемых языков [6] или используемой системы машинного перевода [17]. Кроме веса, может также учитываться количество маркеров определенного типа (степень их «выраженности») в каждом предложении [28].

Отдельную область исследований в области машинной переводимости представляют способы подготовки текста для перевода машиной. Выделенные нами маркеры могут оказать помощь в составлении предписывающего языка для авторов аннотаций, позволяющего улучшить качество их машинного перевода. Создание текстов для машинного перевода с использованием предписывающих языков [12, 16] является одной из ключевых стратегий, обеспечивающих высокое качество автоматического перевода. Использование машинно-ориентированных предписывающих языков нацелено на то, чтобы удалить (насколько возможно) маркеры переводимости из текста оригинала с целью улучшить машинный перевод и снизить трудозатраты на постредактирование. В результате анализа маркеров на различных лингвистических уровнях определяются правила, предписывающие удалять (избегать), заменять или добавлять определенные элементы в проблемных случаях (см., например, [10, 18, 23 и др.]). Наряду с правилами предписывающего языка в подготовке текста для перевода могут участвовать автоматизированные средства предредактирования. Уникален подход, предлагающий составлять текст для перевода в интерактивном режиме [2], что позволяет избежать ряда проблем анализа на синтагматическом уровне.

Еще одно применение маркеров переводимости – создание инструментов для автоматизации постредактирования машинных переводов [3, 7, 13

и др.]. Такие инструменты позволяют сократить затраты на поиск ошибок в переводе и их исправление. Заметим также, что оценка усилий, которые необходимо затратить на постредактирование машинного перевода, является одним из подходов – наряду с отмеченным ранее взвешиванием маркеров переводимости в тексте оригинала, – используемых для оценки переводимости текста [17].

Литература/References

1. Дрейфус Х. Чего не могут вычислительные машины: Критика искусственного разума. М.: Прогресс, 1978. 334 с. [Dreyfus H.L. *Chego ne mogut vyschislitelnye machiny: Kritika iskusstvennogo razuma* (What Computers Still Can't Do: A Critique of Artificial Reason). Moscow, Progress Publ., 1978, 334 p.].
2. Шереметьева С.О. Интерактивное реферирование, ориентированной на машинный перевод. Вестник ЮУрГУ. Серия «Лингвистика». 2013. Т. 10. № 1. С. 89–92. [Sheremetyeva S.O. Interaktivnoe referirovanie, orientirovannoe na machinniy perevod (Machine Translation Oriented Interactive Summarization), *Vestnik Yuzhno-Uralskogo gosudarstvennogo universiteta, seriya Lingvistika* (The Bulletin of South Ural State University, Ser. Linguistics), 2013, vol. 10, no. 1, pp. 89–92].
3. Allen J., Hogan C. Toward the Development of a Post-Editing Module for Raw Machine Translation Output: A Controlled Language Perspective, *Proceedings of the Third International Workshop on Controlled Language Applications (Seattle, WA)*, 2000, pp. 62–71.
4. Arnold D., Balkan L., Maijer L., Humphreys R. L., Sadler L. Machine Translation: An Introductory Guide, Oxford, NCC Blackwell, 1994, 206 p.
5. Bergsma S., Yarowsky D., Church K. Using Large Monolingual and Bilingual Corpora to Improve Coordination Disambiguation, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (Portland, Oregon)*, 2011, pp. 1346–1355.
6. Bernth A., McCord M. The Effect of Source Analysis on Translation Confidence, J.S. White (ed.) *Envisioning Machine Translation in the Information Future: 4th Conference of the Association for Machine Translation in the Americas (AMTA 2000) (Cuernavaca, Mexico),* Berlin, Springer, 2000, pp. 89–99.
7. Doyon J., Doran C., Means C. D., Parr D. Automated Machine Translation Improvement through Post-editing Techniques: Analyst and Translator Experiments, *Proceedings of AMTA*, 2008, pp. 346–353.
8. Gdaniec C. The Logos Translatability Index, *Technology Partnerships for Crossing the Language Barrier: Proceedings of the First Conference of the Association for Machine Translation in the Americas AMTA*, Oct. 1994, pp. 97–105.
9. Goldberg M. An Unsupervised Model for Statistically Determining Coordinate Phrase Attachment, Proceedings of ACL, 1999, pp. 610–614.
10. Hartley A., Tatsumi M. , Isahara H., Kageura K., Miyata R. Readability and Translatability Judgments for 'Controlled Japanese', *Proceedings of the 16th EAMT Conference (Trento, Italy)*, May 2012, pp. 237–244.
11. Hutchins W. J., Somers L. An Introduction to Machine Translation, London, Academic Press, 1992, 362 p.
12. Kittredge R. Sublanguages and Controlled Language, R. Mitkov (ed.) *The Oxford Handbook of Computational Linguistics*, Oxford, OUP, 2003, pp. 430–447.
13. Knight K., Chander I. Automated Post-Editing of Documents, *Proceedings of the 12th National Conference on Artificial Intelligence (Seattle, WA),* 1994, pp. 779–784.
14. Marinčič D. Parding with Intraclausal Coordination and Clause Detection, Informatitica, 2010, no. 34, pp. 263–264.
15. Nakov P., Hearst M. Using the Web as an Implicit Training Set: Application to Structural Ambiguity Resolution, *Proceedings of HLT-EMNLP*, 2005, pp. 835–842.
16. Nyberg E., Mitamura T., Huijsen W.-O. Controlled Language for Authoring and Translation, H. Somers (ed.), *Computers and Translation*, Amsterdam, NL, Benjamins, 2003, pp. 245–281.
17. O'Brien S. Methodologies for Measuring the Correlations between Post-Editing Effort and Machine Translatability, *Machine Translation*, 2005, no. 19, pp. 37–58.
18. O'Brien S. Controlling Controlled English: An Analysis of Several Controlled Language Rule Sets, *Joint conference combining the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop (EAMT/CLAW)*(Dublin City University), May 2003, pp. 105–114.
19. Okuma H., Hara K., Shimbo M., Matsumoto Y. Bypassed Alignment Graph for Learning Coordination in Japanese Sentences, *Proceedings of the ACL-IJCNLP 2009: Conference Short Papers* (Suntec, Singapore), Aug. 2009, pp. 5–8.
20. Okumura A., Muraki K. Symmetric Pattern Matching Analysis for English Coordinate Structures, *Proceedings of the fourth conference on Applied Natural Language Processing (ANLP)*, 1994, pp. 41–46.
21. Pedro R. de. The Translatability of Texts: A Historical Overview, *Meta*, 1999, Vol. XLIV, no. 4, pp. 546–559.
22. Povlsen C., Underwood N., Music B., Neville A. Evaluating Text-type Suitability for Machine Translation a Case Study on an English-Danish MT System, A. Rubio, N. Gallardo, R. Castro & A. Tejada (eds.) *Proceedings of the First International Conference on Language Resources and Evaluation (Granada, Spain),* 1998, vol. 1, pp. 27–31.
23. Reuther U. Two in one – Can it work? Readability and Translatability by means of Controlled Language, *Proceedings of EAMT-CLAW03, Con-*

trolled Language Translation (Dublin City University, Dublin), May2003, pp. 124–132.

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

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

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

27. Trujillo A. Translation Engines: Techniques for Machine Translation, London, Sptringer-Verlag, 1999, 303 p.

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

Бабина Ольга Ивановна, кандидат филологических наук, доцент, доцент кафедры лингвистики и межкультурной коммуникации, Южно-Уральский государственный университет (Челябинск), babinaoi@susu.ac.ru

Поступила в редакцию 30 июня 2014 г.

Bulletin of the South Ural State University
Series "Linguistics"
2014, vol. 11, no. 3, pp. 5–14

MACHINE TRANSLATABILITY OF RUSSIAN SCIENTIFIC TEXTS

O.I. Babina, *South Ural State University, Chelyabinsk, Russian Federation, babinaoi@susu.ac.ru*

The problem of the identification of text parameters having negative impact on the quality of machine translation is considered in the article. Definition and classification of translatability indicators is given. Linguistic phenomena in the scientific texts in Russian are considered at the graphic, lexical and syntagmatic levels. The special attention is paid to textual features causing problems for machine translation at the syntagmatic level, including both universal features and ones specific to Russian. A classification of formal machine translatability indicators for Russian scientific texts is compiled on the basis of language features analysis. The distinguished classes are matched against text translation problems associated with the corresponding indicator. The obtained results can be applied in translation practice and when developing tools for computational support of translation.

Keywords: machine translatability, translatability indicator, machine translation, scientific text, corpus.

Olga I. Babina, Candidate of Philology (PhD), Associate Professor, Associate Professor of the Department of Linguistics and Intercultural Communication, South Ural State University (Chelyabinsk); babinaoi@susu.ac.ru

Received 30 June 2014