Санкт-Петербургский государственный университет
Кафедра математической лингвистики

АНАЛИЗ ОСОБЕННОСТЕЙ МАШИННОГО ПЕРЕВОДА
(на материале финских текстов разных функциональных стилей)

Направление: «Лингвистика»
Образовательная программа: «Прикладная и экспериментальная лингвистика»
Профиль: «Компьютерная лингвистика и интеллектуальные технологии»

Выпускная квалификационная работа

соискателя на степень
магистра филологии
Прохоровой Александры Алексеевны

Научный руководитель:
к. филол. наук, доц. М. В. Хохлова

Санкт-Петербург
2016

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Инструменты машинного перевода (МП), позволяющие работать с текстами онлайн и осуществлять быстрый перевод, служат для интернет-пользователей средством коммуникации. У таких инструментов есть ряд недостатков — ни одна из таких платформ, существующих в настоящее время, не является оптимальной с точки зрения скорости, правильности и стоимости перевода [Борисова, 2013; Аматов, 2008].

Зачастую результаты работы онлайн-инструментов требуют постредактирования [Борисова, 2014; Новожилова, 2014], и эффективно могут использоваться только теми, кто в какой-то степени владеет принимающим языком и языком-источником.

Другой проблемой является то, что не для всех малых языков существуют хорошо разработанные автоматические переводчики. Большинство систем при работе с некоторыми парами языков используют язык-посредник (обычно английский язык). Иначе говоря, перевод осуществляется не напрямую: сначала происходит трансфер текста с языка-оригинала на английский, а уже потом — на необходимый язык перевода, что во многом влияет на качество перевода [Новожилова, 2014].

В этой ситуации нам кажется актуальной задача развития систем машинного перевода для тех языков, которые до сих пор не были достаточно автоматизированы. Так, до недавнего времени для перевода в паре финский-русский язык не существовало такой системы, которая бы не использовала язык-посредник. Это определяет актуальность нашей работы. Ее **практическая значимость** обусловлена тем, что разработанные нами классификация и рекомендации могут быть использованы при создании и усовершенствовании систем машинного перевода.

Мы выбрали для работы онлайн-переводчик PROMT, во-первых, потому что на данный момент прямой перевод между русским и финским языком, без использования языка посредника, может осуществлять только этот инструмент, во-вторых, по той причине, что в нашем доступе оказались

данные реальных пользовательских запросов. Более того, компания начинает активно внедрять статистические технологии, и это один из их первых проектов, над которым началась работа в 2005 году. Для английского и русского языков компания уже сейчас ведет разработки гибридных подходов. Мы верим, что с богатым и успешным опытом компании в области классического перевода «по правилам», при дальнейшем использовании статистических методов, у PROMT есть большой потенциал также и в области гибридного перевода.

Мы считаем что, идентификация основных проблем системы перевода — это важный шаг в направлении дальнейших исследований.

Целью нашей работы является анализ особенностей перевода, связанных с работой статистического машинного переводчика PROMT в паре языков русский-финский.

Для достижения поставленной цели нами были решены следующие **задачи**:

- изучена история развития систем МП, описаны типы систем МП и рассмотрены принципы их работы;

- исследованы понятие «качества перевода» и способы оценки качества перевода;

- проанализированы типы ошибок, появляющиеся при работе системы PROMT, и дана их классификация;

- дана оценка результатам эксперимента, рассмотрены причины возникающих ошибок и определены дальнейшие пути развития.

В первой главе дается краткий обзор истории машинного перевода, приведены несколько классификаций существующих систем МП, далее подробно расписаны принципы работы перевода по правилам, статистического перевода и гибридных подходов, приведены плюсы и минусы таких систем.

Вторая глава посвящена оценке качества МП, разбираются стандартные методы оценки качества перевода (такие как экспертная оценка, метрики

BLEU, NIST и WER) ставится вопрос об интерпретации результата такой оценки, приводится стандартная классификация типов ошибок.

В третьей главе мы описываем практическую часть нашей работы, приводится классификация систематических ошибок на основании возможных причин их появления, даются рекомендации относительно улучшения работы СМТ с помощью дополнительных инструментов.

Благодарим компанию «ПРОМТ» за предоставленные нам для работы данные пользовательских запросов и возможность участвовать в разработке этого переводческого инструмента.

1. МАШИННЫЙ ПЕРЕВОД

1.1. Определение понятия перевода как вида человеческой деятельности

Перевод как один из видов языковой деятельности представляет собой процесс адекватной и полноценной передачи мыслей, высказанных на одном языке, средствами другого языка [Нелюбин, 2011, с.138]. Перевод — это сложный и многогранный вид человеческой деятельности. В процессе перевода происходит столкновение не только языков, но и культур.

Данные переводоведения используются в областях культурологи, этнографии, истории и литературоведенья. В свою очередь, в науке о переводе могут выделяться культурологические, когнитивные, психологические, литературные и прочие аспекты [Комиссаров, 2002, с.22].

Л. С. Бархударов, один из основоположников отечественной теории перевода, дает следующее определение: «Переводом называется процесс преобразования речевого произведения на одном языке в речевое произведение на другом языке при сохранении неизменного плана содержания, то есть значения» [Бархударов, 1975, с.11]. Процесс перевода — это также и процесс передачи информации, содержащейся в произведении речи, средствами другого языка [Ахманова, 1969, с.316].

По словам И. Р. Гальперина при переводе должно передаваться не только смысловое содержания текста, но и его стилистические особенности [Гальперин, 1987, с.20]. Перевод можно рассматривать, как вид коммуникативной деятельности [Колшанский, 1980, с.112], а перевести что-то на другой язык — «значит выразить верно и полно средствами одного языка то, что уже выражено ранее средствами другого языка» [Федоров, 2002, с.10]. Л. К. Латышев считает, общественное предназначение перевода заключается в том, чтобы в максимальной степени приблизить опосредованную двуязычную коммуникацию по полноте, эффективности и естественности общения к обычной одноязычной коммуникации [Латышев, 1988, с.7].

Некоторые исследователи определяют перевод через требования, предъявляемые переводчику:

1. «Он должен понимать слово в оригинале и по смыслу, и по стилю»;
2. «Он должен преодолеть различия между двумя лингвистическими структурами»;
3. «Он должен в своем переводе воссоздать стилистическую структуру оригинала» [Найда, 1978, с.121].

Другие определяют его через требования, которым должен удовлетворять сам текст перевода:

1. передавать смысл,
2. передавать дух и стиль оригинала,
3. обладать легкостью и естественностью изложения,
4. вызывать равнозначное впечатление

[там же].

1.2. Определение понятия машинного перевода

Рассмотрим, каким образом машинный перевод (МП) вписывается в наше представление о переводе. Как это ни парадоксально, но на данный момент с практической точки зрения машинный перевод остается процессом человеческой деятельности.

Термин «машинный перевод» многозначен. За долгую историю использования он приобрел множество интерпретаций. Сначала этот термин подразумевал только автоматические системы, работающие без участия человека [Sager, 1994, с.326]. Европейская ассоциация машинного перевода дала следующее определение: «использование компьютера для перевода текста с одного естественного языка на другой язык» [Сайт Европейской ассоциации машинного перевода EAMT]. А Международная ассоциация машинного перевода (IAMT) определяет машинный перевод как «единовременный ввод полного предложения и генерирование

соответствующего ему полного предложения» [Hutchins, 2000a]. Ни одно из этих определений не предполагает вмешательства человека.

Академические ученые и исследователи до сих пор расходятся во взглядах на определение машинного перевода в отношении участия человека в этом процессе. В данный момент этот термин продолжает использоваться для обозначения полностью автоматизированных систем пусть даже и с участием человека [Somers 2003: с.1—11].

Машинный перевод — это выполняемое на компьютере действие по преобразованию текста на одном естественном языке в эквивалентный по содержанию текст на другом языке, а также результат такого действия [Фролов, 2008, с.127].

Толковый переводоведческий словарь Л. Л. Нелюбина определяет машинный перевод следующим образом:

1. Автоматический перевод текста на основе заданной программы, осуществляемой ЭВМ.
2. Отрасль языкознания, разрабатывающая теорию такого перевода на основе коренного пересмотра основных положений и методов лингвистики.
3. Автоматизированная обработка информации в условиях двуязычной ситуации — передача текста с одного человеческого (естественного) языка на другой.
4. Перевод с использованием машин (ЭВМ, компьютера).
5. Общий процесс переработки информации в условиях двуязычной ситуации на любом этапе использования (и развития) технических средств.
6. Процесс перевода текста с одного языка (естественного или искусственного) на другой (естественный или искусственный), осуществляемый на электронной цифровой вычислительной машине

[Нелюбин 2011 с.107].

1.3. История развития машинного перевода

Идея машинного перевода, т.е. мысль о том, чтобы поручить машине работу по переводу с одного естественного языка на другой, насчитывает к настоящему времени уже около пятидесяти лет существования. Примерно столько же лет ведутся научно-исследовательские работы по машинному переводу во многих странах мира [Марчук, 2007, с.245].

Начиная с 40-х годов XX века, с момента создания первой ЭВМ, машинный перевод являлся одной из задач, которую ученые собирались решить в кратчайшие сроки [Baker, 2001].

Первые опыты специалистов из IBM основывались главным образом на словарном (прямом) методе и были весьма, успешными для малого (250) количества входных предложений.

Это подкрепило уверенность в том, что проблема машинного перевода — простая для решения задача. Но, после проведения дальнейших исследований, ученые обнаружили, что задача машинного перевода вовсе не является тривиальной [Кан, 2011].

Чарльз Бэббидж первым высказал мысль о возможности МП. В середине 19 века он работал над проектом цифровой аналитической машины. Это был механический прототип ЭВМ, которые потом появились только через 100 лет. Идея Бэббиджа состояла в том, что такую машину можно использовать для хранения словарей. Бэббидж привел эту идею в качестве обоснования для запроса у английского правительства средств, необходимых для физического воплощения машины, которую ему так и не удалось построить [Шаляпина, 1996, с.105].

Джон Хатчинс — один из самых активных историков машинного перевода на западе [Сайт Джона Хатчинса]. Согласно Хатчинсу пионером в области машинного перевода был Пётр Петрович Троянский, предложивший схему механического устройства перевода. Троянский был незаслуженно забыт мировым научным сообществом. Первые ЭВМ («БЭСМ» и «Стрела»)

стали использоваться для работы в СССР в 1952-53 годах уже после его ухода [Hutchins, 2000b].

Машинный перевод начали воспринимать как отдельную исследовательскую область после марта 1947 года. Тогда, Уоррен Уивер, специалист по криптографии, в своем письме Норберту Винеру сформулировал задачу машинного перевода, сравнив ее с задачей дешифровки.

В 1949 г. Уивер составил меморандум, в котором смог обосновать, каким образом возможно осуществить МП. Уивер писал: «I have a text in front of me which is written in Russian, but I am going to pretend that it is really written in English and that it has been coded in some strange symbols. All I need to do is strip off the code in order to retrieve the information contained in the text» («У меня есть текст, написанный на русском языке, но я сделаю вид, что он написан по-английски и закодирован при помощи странных символов. Тогда все, что мне нужно сделать, — это разгадать код, чтобы извлечь информацию, заключенную в тексте») [Слокум, 1989, с.56—58]. Позже идеи Уивера легли в основу подхода к МП, основанного на концепции интерлингвы (interlingva). При таком подходе стадия передачи информации разделена на два этапа. На первом этапе исходное предложение переводится на язык-посредник (созданный на базе упрощенного языка), а затем результат этого перевода представляется средствами выходного языка [Лекция о системах МП].

В 1952 г. состоялась первая конференция по МП в Массачусетском технологическом университете, а первые эксперименты по машинному переводу, подтвердившие принципиальную возможность его реализации, были проведены в 1954 г. в Джорджтаунском университете (г. Вашингтон, США). Тогда была представлена первая полноценная система машинного перевода — IBM Mark II. Это событие вошло в историю как «Джорджтаунский эксперимент». Очень ограниченная в своих возможностях система прекрасно справилась с переводом 49 специально подобранных

предложений с русского языка на английский с использованием словаря на 250 слов и шести грамматических правил [там же].

В СССР первый эксперимент по МП был осуществлен И. К. Бельской и Д. Ю. Пановым в Институте точной механики и вычислительной техники АН СССР в 1954 г. [там же].

В истории МП было несколько поворотных моментов, которые определили его развитие на долгие годы, и несколько моментов, которые привели к затишью в области на многие годы. Одной из таких поворотных точек стал доклад ALPAC, содержавший объективную оценку состояния МП в 60-е годы XX века, показавший насколько сложной задачей является машинный перевод на самом деле [Онлайн версия доклада ALPAC, 1966].

Результатом этого доклада стала идея того, что разработка систем машинного перевода нерентабельна. Это фактически привело к прекращению работ над системами машинного перевода. Однако, благодаря постоянному прогрессу вычислительной техники, исследования в этой области вновь возобновились в 70-е годы, а в конце 80-х началась разработка первых статистических систем [Молчанов, 2013].

Системы перевода разрабатывались в разных странах по всему миру: США, Германия, Франция, Россия, Япония. Из наиболее известных масштабных исследовательских проектов в области МП в Советском Союзе и России нужно отметить систему МП ЭТАП [Кан, 2011б с.71].

В СССР в 70-х годах разработку основ технологии машинного перевода продолжила группа специалистов в ВИНИТИ под руководством профессора Г. Г. Белоногова [Карасев, 2011]. В результате в 1993 г. была создана первая российская промышленная версия системы RETRANS фразеологического машинного перевода с русского языка на английский и обратно, которая применялась в министерствах обороны, путей сообщения, науки и технологий, а также во ВНТИЦ [Карасев, 2011].

Первые коммерческие продукты машинного перевода, нашедшие практическое использование, появились в середине 80-х годов. Тогда МП

стал экономически выгодным. Стоимость персональных компьютеров (ПК) понизилась, и к ним стало проще получить доступ, количество пользователей ПК увеличилось. Системы МП были реализованы на персональных компьютерах и являлись системами прямого перевода, возможности которых базировались на огромных (по сравнению с первыми системами) словарях, а не на умении анализировать и синтезировать тексты.

Одной из новых разработок 70-80-х годов стала технология ТМ (translation memory — «*память переводов*», или «переводческая память»), Такая «память», работает почти как человеческая память, по принципу накопления. Каждый раз при переводе сохраняется исходный текст и его перевод, из всех таких переводов создается лингвистическая база данных, которую затем можно использовать при последующих переводах. Инструменты ТМ сейчас активно используются большинством переводческих компаний.

Технологии МП начали развиваться еще активнее в 90-е годы. Популяризация интернета и высокий уровень возможностей персональных компьютеров обеспечили реальный спрос на МП. Так МП снова стал привлекательным для инвесторов и выгодным для разработки.

В ходе своего развития, алгоритмы МП перешли от прямого (словарного) метода к методу *трансфера*, а затем и к понятию *интерлингвы*. В итоге алгоритмы МП поделились на две группы: статистические подходы, основанные на входном корпусе данных (Data Driven Machine Translation), и классические (основанные на правилах), изучающие каждый язык во всей его лингвистической полноте (Rule Based Machine Translation) [Кан, 2011].

Сейчас основные исследования ведутся в области статистического МП [Hearne, 2011], а системы МП, основанные на правилах, считаются устаревающими. В российской литературе ситуация прямо противоположная: классические подходы привлекают значительно большее внимание специалистов [Кан, 2011].

Оба фундаментальных подхода имеют свои недостатки. Классические методы очень трудоёмки, а качество работы статистических подходов напрямую зависит от качества входного корпуса. Самые последние разработки ведуться в области гибридных систем, делающие попытку вобрать лучшие характеристики классического и статистического МП, минимизируя их недостатки [там же].

На данный момент наиболее известной и распространённой в мире системой МП является СМП компании Systran и компании Google. В России на текущий момент аналогом является СМП компании «ПРОМТ» [там же].

1.4. Типы систем МП

Классификацию систем машинного перевода можно произвести по разным основаниям [Беляева, 2007]. Например, можно выделять системы:

1. по количеству языков (бинарные, осуществляют перевод в одной паре языков, и многоязычные, работают с несколькими языками);
2. по направленности перевода (однонаправленные и многонаправленные, если целевой язык и язык-источник могут меняться местами в зависимости от требований пользователя);

В зависимости от того, какую роль играет человек в процессе МП, иными словами, по степени автоматизации, обычно выделяют три типа систем машинного перевода:

1. Полностью автоматические системы машинного перевода
2. МП-системы, машинный перевод при участии человека
3. ТМ-системы, перевод осуществляется человеком, при использовании компьютера

Полностью автоматические системы машинного перевода являются скорее несбыточной мечтой, чем реальной идей. Все системы машинного перевода (МП-системы) работают при участии человека в той или иной мере. Чтобы компьютер мог перевести текст, ему нужна помощь предредактора, который тем или иным образом предварительно обрабатывает подлежащий

переводу текст, интерредактора, который участвует в процессе перевода, и постредактора, который исправляет ошибки и недочеты в переведенном машиной тексте [Рябцева, 1986, с.167].

ТМ-системы иногда называют еще «памятью переводов». Они являются скорее просто удобным инструментом, нежели элементом автоматизации.

Другой вариант классификации систем МП, пришедший из области корпусной лингвистики, — это разделение на подходы, в которых используются параллельные корпуса и, соответственно те, в которых они не используются. Системы, использующие корпуса, далее делятся в зависимости от основной стратегии перевода — на системы, *основанные на примерах* (EBMT), и *статистические системы* (SMT) [Hearne, 2011].

Самый простой и распространенный вариант классификации - это разделение на два основных типа систем МП [Молчанов, 2013]:

- основанные на правилах (rule-based machine translation, RBMT)
- статистические

Отдельно стоят гибридные системы, которые призваны сочетать в себе лучшие черты систем, основанных на правилах, и статистических систем.

1.4.1. Память переводов (Translation Memory)

Технология памяти переводов (Translation Memory или TM) использует правила перевода и сравнивает входной документ с текстами из постоянно пополняющейся базы переводов. Находя совпадения, программа предлагает ранее одобренный вариант [Карасев, 2011].

В процессе перевода сохраняется исходный сегмент текста (предложение) и его перевод; если подобный исходному сегмент обнаруживается, он отображается вместе с переводом и указанием совпадения; затем переводчик принимает решение (редактировать, отклонить

или принять перевод), результат которого сохраняется системой [Лекция о системах МП].

1.4.2. Системы, основанные на правилах (классические системы)

Технология этого перевода состоит в применении алгоритмов, в соответствии с которыми программа анализирует текст и на основе проведенного анализа синтезирует вариант перевода.

Считается, что работа такого машинного переводчика похожа на процесс мышления человека [Новожилова, 2014].

Стандартный алгоритм действий над входным предложением в такой системе следующий: — морфологический анализ — поиск частей речи, определение входных словоформ (рода, числа, падежа, спряжения); — поиск идиом, фразеологизмов для данной предметной области и исключение их из дальнейшего анализа; — синтаксический анализ — разбор структуры, нахождение членов предложения — подлежащего, сказуемого, дополнения, обстоятельства.— лексический анализ — отделение однозначных входных слов (лексем) от многозначных (имеющих несколько переводных эквивалентов); — грамматический анализ — доопределение грамматической информации с учетом данных выходного языка; — синтез выходного предложения (перевода) [Карасев, 2011].

В системах, основанных на правилах (RBMT), можно выделить два основных подтипа: трансферные и системы-интерлингвы.

Трансферные системы машинного перевода распространены более широко, чем системы-интерлингвы. Они работают по следующим принципам: проводится морфологический, лексический и семантико-синтаксический анализ предложения на языке оригинала, создается синтактико-семантическое дерево разбора входного предложения, затем производится так называемый «трансфер», т. е. преобразование структуры входного предложения в соответствии с формальными требованиями языка перевода. На заключительном этапе синтеза формируется конечное

предложение на языке перевода. Основанная на правилах система перевода PROMT является классическим примером трансферных систем [Молчанов, 2013].

В основе систем-интерлингв лежит теория о том, что любое предложение любого языка можно преобразовать в его смысловое представление на универсальном метаязыке. Далее, используя полученное смысловое представление, можно синтезировать предложение на языке перевода. Любой текст можно преобразовать в смысл, и любой смысл в текст, используя ряд правил и семантичсекий словарь. Интерлингвы требуют очень долгой разработки и создания огромных баз знаний о языке [там же].

Системы, основанные на правилах, обладают рядом общих характеристик. Все они включают в себя словари и формальные грамматики, т. е. наборы правил морфологического, семантического и синтаксического анализа языка. С точки зрения разработки ииспользования, такие системы обладают рядом преимуществ и недостатков.

К достоинствам таких систем можно отнести высокое качество, стабильность и предсказуемость машинного перевода.

Недостатки таких систем включают высокую стоимость разработки и поддержки лингвистических алгоритмов и словарей, а также большое количество времени, необходимое для лексической настройки системы для отдельного клиента или новой предметной области. Кроме того, при высокой точности основанный на правилах перевод обладает определенным «машинным» акцентом, т. е. часто выглядит неестественно.

Существует также и проблема нарастающей сложности. Описать язык во всей его полноте — очень трудная задача, за счет того, что каждый следующий уровень языка оказывается на порядок сложнее предыдущего, и за рамками описания всегда остаются некоторые лингвистические явления.

Современные RBMT-системы обычно включают в себя общетематические словари (объемом от нескольких десятков до нескольких сотен тысяч статей) и специализированные словари по отдельным тематикам

(объемом до нескольких десятков тысяч статей). Производительность RBMT-систем машинного перевода зависит от различных параметров (среди которых количество и сложность грамматических правил, объем и количество используемых словарей) и обычно варьируется от нескольких слов до нескольких сотен слов в секунду [там же].

1.4.3. Статистический машинный перевод

Статистический М П опирается на предположение, что сказав что-то однажды, человек с некоторой вероятностью повторит это вновь [Кан, 2011].

Подход, используемый в статистическом МП, заключается в анализе колоссального массива параллельных текстов. С помощью этого двуязычного параллельного корпуса выявляются пары фраз на двух языках, которые несут один смысл. При этом использование каких-то дополнительный грамматических правил не предусматривается [Карасев, 2011].

Задача машинного перевода в этом случае на общем уровне может быть сформулирована как задача максимизации условной вероятности *P(e|f),* что обозначает условную вероятность предложения на языке *E* при заданном предложении на языке *F, e∈E, f∈F.*

Для выполнения этой задачи можно использовать теорему Байеса. Формула Байеса или теорема Байеса — одна из основных теорем элементарной теории вероятностей. Она позволяет определить вероятность чего-либо, какого-либо события при условии, что произошло другое статистически взаимозависимое с ним событие.

Тогда, применяя теорему Байеса, можно записать:

$$P(e \mid f) = \frac{P(e)P(f \mid e)}{P(f)}$$

Где

P(e) — априорная вероятность гипотезы *e*;

P(e|f) — вероятность гипотезы *e* при наступлении события *f;*

P(f|e) — вероятность наступления события *f* при истинности гипотезы *e*;

P(f) — полная вероятность наступления события *f*

Для максимизирования условной вероятности слева нужно максимизировать величину справа. Следующее уравнение называют фундаментальным уравнением машинного перевода [Кан, 2011]:

$$\max_e P(e \mid f) = \arg\max_e P(e)P(f \mid e),\ e \in E, f \in F$$

В этом уравнении вероятность *P(e|f)* называется моделью перевода, а *P(e)* — языковой моделью. Построение этих моделей является частью обучения статистической системы МП.

Для использования системы статистического МП ее нужно сначала обучить. Процесс обучения подразумевает создание двух моделей: статистической модели перевода на основании параллельного корпуса и статистической модели принимающего языка на основе (зачастую намного большего) одноязычного корпуса [Brown, 1993].

Модель перевода строится по двуязычному выровненному корпусу, то есть такому корпусу, где каждое предложение на языке F имеет перевод на языке E. Другое название такого корпуса — параллельный корпус.

Построение такого корпуса является отдельной научной задачей, а получение параллельных текстов в автоматическом режиме — также и практической (например, сканирование сети Интернет в поисках страниц, переводящих друг друга). Получение двуязычного корпуса на практике сводится к анализу форматов оцифрованных книг-переводов друг друга, а также к индексированию Интернета с целью получения параллельных страниц. В этом случае возможно применение различных эвристик с распознаванием языка и поиска шаблонов в URL адресах, подобных URL/en и URL/ni. Более качественным и, соответственно, дорогим способом получения параллельного корпуса является ручная разметка. Одним из

наиболее популярных источников параллельных корпусов для пар европейских языков является корпус Europarl [Кан, 2011].

Модель перевода составляет двуязычный словарь, где для каждого возможного перевода конкретной единицы языка-источника указана вероятность такого перевода [Hearne, 2011].

Такая модель отличается от обычного словаря, где присутствуют только правильные переводы; в этой модели будут присутствовать и маловероятные переводы. Так, лучшим переводом будет считаться самый вероятный, при этом «лучший» не означает полностью правильный [там же].

Языковая модель создает базу данных типичных цепочек слов (последовательностей словоформ) в принимающем языке (обычно от 1 до 7 слов), для каждой из которых указывается вероятность появления [там же].

После того, как система была обучена, можно начинать процесс декодирования, то есть непосредственно использовать систему для перевода. Когда система получает запрос от пользователя, переводческая модель генерирует возможные варианты перевода, а языковая модель выбирает такой перевод, который больше всего напоминает текст, написанный на естественном языке.

Когда говорят о статистическом МП, обычно имеют в виду *фразовые переводчики* (Phrase-based translation — PBT).

До появления фразовых переводчиков, стандартом считались системы *пословного перевода* (Word-based translation — WBT). В таких системах каждое слово переводится отдельно, в том порядке, в каком они встречаются в тексте, без учета синтаксических и логических связей. Появление фразовых переводчиков, позволило учитывать цепочки словоформ различной длины. В системе фразового перевода входное предложение делится на сегменты, (фразы, цепочки словоформ, n-граммы) которые переводятся отдельно. Фраза может состоять из одного и больше слов [Jehl, 2010б с.4].

Такая система позволяет легко решить проблему, когда в принимающем языке и языке источнике для некоторых слов нет точных соответствий.

В системах пословного перевода для решения этих проблем приходится водить новые сложные стратегии, такие как например нулевые слова [Jehl, 2010].

Система фразового перевода имеет следующие превосходства:

1. Позволяет разрешать лексическую неоднозначность при переводе полисемантических слов, учитывая дополнительную контекстуальную информацию
2. При увеличении количества тренировочной информации, информации в тренировочном корпусе, система может учить все более длинные фразы. Таким образом, фразовые переводчики используют тренировочные данные более эффективно [Koehn, 2010].

Многие годы фразовые системы перевода показывают лучшие результаты в области МП. В первую очередь, это связано с наличием огромных параллельных корпусов. Но необходимость использования такого большого объема данных может быть проблематична при работе с языками, для которых этих данных просто нет. Статистический перевод как подход имеет и другие внутренние ограничения [Silva, 2015, с.13].

Основанные проблемы статистического перевода связаны с использованием ограниченной лингвистической абстракции (limited linguistic abstraction), трудностям перевода определенных конструкций, как, например, получение правильного порядка слов при переводе между языками разных типов или сохранение семантического единства в выходном тексте [Silva, 2015].

К минусам статистических систем МП также можно отнести большое количество грамматических ошибок. Отдельные словосочетания при статистическом переводе получаются более точными и изящными, но грамматика хромает: иногда предложения настолько несогласованны, что невозможно понять их смысл [Карасев, 2011].

Другой проблемой является необходимость наличия представительных параллельных корпусов большого объема.

Чистая система статистического перевода без дополнительных инструментов не распознает сложные синтаксические связей, неверно определяет сказуемое, объектные, атрибутивные и другие отношения в предложении. Так, выполненные переводы могут представлять собой произвольный набор слов и словосочетаний, не объединенных смысловыми связями [Новожилова, 2014].

В целом, статистические методы перевода могут привлекать сложные алгоритмы и вероятностные модели, но главной их проблемой является то, что они не относятся к языку как к лингвистическому объекту. Они воспринимают тексты кк потоки данных, между которыми нужно найти есть соответствие.

Крупнейшие системы статистического машинного перевода, работающие с русским и финским языком — это PROMT, Google Translate и Яндекс.Перевод.

1.4.4. Гибридные системы машинного перевода

Существует мнение, что корень подавляющего большинства проблем машинного перевода лежит в несоответствии систем языков, а скромные успехи в разработке программного обеспечения для перевода текстов не связаны с плохой работой программистов или компьютерных техников, а являются результатом плохой проработки этой проблемы с лингвистической стороны [Борисова, 2014]. Результатом этой идеи, стало развитие и разработка гибридных подходов.

Гибридные подходы становятся все более популярными, так как они сочетают лучшие качества подхода, основанного на правилах, и статистического подхода [Costa-jussà, 2015].

Машинный перевод — это междисциплинарная область знаний, а к решению задач машинного перевода можно подходить с разных точек зрения,

используя данные лингвистики или статистики. Именно существование разных подходов сделало возможным создание гибридных методов. Гибридные технологии фокусируются на том, чтобы взять все лучшие качества уже существующих подходов. В настоящее время, самые распространенный вариант гибридного переводчика — это подключение правил к уже существующей системе статистического перевода (SMT). Тем не менее, проводятся также исследования, фокусирующиеся на улучшении работы систем, основанных на правилах (RBMT) с помощью дополнительной статистической информации. В настоящее время лингвисты, инженер-программисты и специалисты из области I.T. активно сотрудничают в области МП в ходе совместных семинаров, проводят эксперименты и разрабатывают архитектуру гибридных систем перевода. Например, один из таких семинаров HyTra (Workshop on Hybrid Approaches to Translation) проводится каждый год, начиная с 2012 [Costa-jussà, 2015].

Говоря о гибридных подходах, следует также упомянуть о новой уникальной технологии ABBYY Compreno, которая изначально развивалась как перевод по правилам. Сейчас она представляет собой многофункциональную лингвистическую технологию [Burukina, 2014]. Система состоит из двух основных и ряда дополнительных компонентов. Первый компонент — это универсальное дерево понятий или универсальная семантическая иерархия. Все слова в паре языков являются листьями на этом дереве, между ними задаются отношения, информация о семантической сочетаемости. Второй компонент — это синтаксический анализатор, который определяет структуру предложения и отношения между входящими в него словами. Для получения точного синтаксического анализа используются семантические данные о значении слов, которые хранятся в семантическом компоненте. Помимо этого Compreno использует статитистические методы, для снятия лексико-семантической омонимии и оценки вероятности встречаемости различных элементов лингвистического описания в текстовых корпусах.

1.5. Практическое применение систем машинного перевода

Системы МП непригодны для работы с текстами, содержащими большое количество сложносочиненных и сложноподчиненных предложений. Эти программы работают в основном на уровне словосочетания, и их можно успешно применять для перевода формализованных текстов, например технической документации, потребительских инструкций, формальных описаний и т. п., для которых характерно использование простых распространенных предложений и в которых не содержатся предложения со сложными синтаксическими конструкциями [Новожилова, 2014].

Целью использования машинного перевода может быть как получение перевода высокого качества, так и простая передача смысла исходного текста (так называемый «джистинг»). Машинный перевод применяется для перевода следующих типов текста: пользовательский контент (отзывы, комментарии и т. д.); документация (техническая, эксплуатационная, юридическая и т. д.); новостной контент; каталоги интернет-магазинов; личная и деловая переписка. К основным сферам применения машинного перевода относятся: локализация (ускорение и удешевление перевода больших объемов текста, например документации к ПО); оптимизация работы переводчиков и переводческих бюро (результат машинного перевода редактируется переводчиками); Интернет (электронная торговля, новостные и образовательные сайты) [Молчанов, 2013].

МП особенно востребован в коммерческой сфере продаж или рекламы товаров [Красных, 2011].

Системы МП являются хорошим подспорьем для специалистов различных профилей, нуждающихся в оперативных переводах иноязычной информации [Карасев, 2011].

Другое применение систем — облегчение работы профессиональных переводчиков, выполнение такого перевода, который можно было бы

исправить с помощью постредактирования. Такая технология позволяет увеличить количество переводимых в день слов с 2000 до 3500 единиц [Koponen, 2015].

Машинный перевод активно используется в области языковой локализации, где количество информации, которую требуется перевести, уже превышает реальные возможности людей переводчиков. Хотя процесс языковой локализации - это больше, чем просто перевод, тем не менее, качественный перевод является ключом к качественной локализации [Zhechev, 2010].

В настоящее время основные усилия прилагаются к уменьшению количества информации, которую переводчикам приходится переводить «с нуля». Таким образом, исследования ведутся в основном в области развития систем статистического машинного перевода (SMT) в качестве дополнения к уже сложившейся технологии памяти переводов (TM) [Zhechev, 2010].

На данный момент, профессиональные переводчики, использующие инструменты CAT, все еще с недоверием относятся к результатам работы SMT, по той причине, что в некоторых случаях постредактирование может занять больше времени, чем перевод «с нуля» [Zhechev, 2010].

В целом, SMT используется только в тех случаях, когда в результате работы TM не был получен качественный перевод [Heyn, 1996].

К основным факторам, затрудняющим машинный перевод, исследователи относят:

1. языковую неоднозначность, которая может быть как лексического, так и грамматического характера;
2. наличие сложных синтаксических структур, которые могут значительно различаться в языке оригинала и в языке перевода;
3. различия в порядке слов в предложении (прямой / обратный, строгий / свободный); наличие анафорических связей в тексте;

4. наличие идиом, смысл которых невозможно передавать посредством пословного перевода;

[Шевчук, 2013, с.222].

1.6. Перспективы развития систем машинного перевода

В настоящее время потребность в переводах растет. Увеличивается также необходимость выполнять переводы быстрее и уменьшить возможные затраты [Koponen, 2015].

Так, пользователи сети интернет ежедневно производят около миллиона запросов на перевод текстов в различных форматах [Беляева, 2007].

Если предположить, что дальнейшей направление развития систем МП будет двигаться в сторону гибридных подходов, использующих лингвистическую информацию, то это означает привлечение к работе над такими системами лингвистов и филологов. Это означает также, что необходимо обеспечить лингвистов необходимой информацией о том, как они могут повлиять на развитие этой области. Подобные руководства уже создаются [Hearne, 2011], и дальнейшее сотрудничество экспертов в области языка и специалистов в области информационных технологий может быть весьма продуктивно.

К наиболее обещающим направлениям развития статистического МП и его оценки можно отнести использование структурированной лингвистической информации (синтаксиса, иерархических структур и семантических ролей) при создании системы перевода, и разработки в области систем, которые могут выйти за пределы уровня предложений, и работать на уровне документов. Эти вопросы активно рассматриваются в области дискурсивного анализа [Guzmán, 2014].

В идеале, система перевода должна подстраиваться под формат того, то именно она переводит. Можно предположить, что для перевода различных сайтов, например, потребуется разная стратегия перевода, относительно того,

что нужно, а что не следует переводить, и какие лингвистические характеристике нужно сохранять [Jehl, 2010, с.21].

Признавая существующие недостатки, производители систем МП подчеркивают, что их программы не ориентированы на создание художественного текста. И заменить человека они не смогут даже в долгосрочной перспективе — пока не будет создан полноценный искусственный интеллект [Карасев, 2011].

1.7. Выводы

Перевод — это один из важнейших видов коммуникативной деятельности, передача информации, смыслового содержания и стилистических особенностей высказывания на одном языке средствами другого языка.

Изначально термин «машинный перевод» подразумевал только автоматические системы, работающие без участия человека. Но на данный момент с практической точки зрения машинный перевод остается процессом человеческой деятельности.

Машинный перевод — это выполняемое на компьютере действие по преобразованию текста на одном естественном языке в эквивалентный по содержанию текст на другом языке.

В ходе долгого развития алгоритмы машинного перевода (МП) поделились на две группы: статистические подходы, основанные на входном корпусе данных (Data Driven Machine Translation), и классические (основанные на правилах), изучающие каждый язык во всей его лингвистической полноте (Rule Based Machine Translation).

В настоящее время все большую популярность приобретают гибридные подходы, призванные соединить в себе плюсы классических и статистических подходов.

На данный момент системы МП непригодны для работы с текстами, содержащими большое количество сложносочиненных и

сложноподчиненных предложений и качественно работают в основном на уровне словосочетания.

При этом системы МП являются хорошим подспорьем для специалистов различных профилей, нуждающихся в оперативных переводах иноязычной информации.

Системы также применяют для облегчения работы профессиональных переводчиков, для выполнения такого перевода, который можно было бы исправить с помощью постредактирования.

2. ОЦЕНКА КАЧЕСТВА МАШИННОГО ПЕРЕВОДА

2.1. Качество перевода

Когда мы говорим о качестве переводе вообще, важно понимать, что к переводу, выполненному человеком, будут предъявляться значительно более высокие требования. Так, при экспертной оценке перевода, выполненного человеком, рассматриваются такие детали как прагматика, соответствие перевода историческому и культурному контексту, стилистика и другие моменты, касающиеся создания правильного впечатления у читателя. При переводе определенной лексики могут рассматриваться даже оттенки значений некоторых слов. Рассматривать в таком же ключе машинный перевод представляется невозможным, хотя бы по той причине, что для осознания метаинформации, которую учитывает переводчик при своей работе, программа машинного перевода должна обладать искусственным интеллектом.

Оценка качества МП является сложной задачей, уже хотя бы потому, что для исходного текста может существовать множество различных правильных переводов.

Для оценки работы систем МП используются следующие методы:

- Экспертная оценка
- Автоматические методы
- Оценка с точки зрения конкретной задачи

Мы не станем подробно разбирать оценку с точки зрения конкретной задачи, по той причине, что она полностью зависит от целей исследования. В этом случае, могут рассматриваться такие вопросы, как, например, сколько времени уходит на постредактирование текста или насколько точно передается информация при переводе [Koehn, 2010].

2.1.1. Экспертная оценка

Иногда для оценки качества перевода, используется текст из узкой специальной области, и исследователь сам осуществляет оценку качества

перевода, сопоставляя результаты работы нескольких систем [Новожилова, 2014; Борисова, 2014; Максименко, 2014].

Стандартная процедура оценки подразумевает больше одного эксперта.

Эксперты проводят субъективную оценку работы системы по двум параметрам: *адекватность* (adequacy) и *гладкость* текста (fluency). Для этого им предоставляют результаты работы системы МП, исходный текст и/или эталон перевода. Эталон перевода часто присутствует в том случае, если эксперт не владеет принимающим или исходным языком. Адекватность и гладкость текста оцениваются по шкале от одного до пяти. Адекватность в данном случае означает правильную передачу смысла исходного текста, а гладкость текста демонстрирует соответствие перевода нормам принимающего языка, правильность с точки зрения грамматики [Koehn, 2010].

Одна из проблем, возникающих при такой системе оценивания, это несогласие между экспертами. Эта проблема разрешается при использовании коэффициента каппа:

$$K = \frac{p(A) - p(E)}{1 - p(E)}$$

,

где p(A) — доля случаев, когда эксперты дали одинаковую оценку, а p(E) — вероятность того, что эксперты случайно дадут одинаковую оценку. Коэффициент Каппа равный единице будет означать полное согласие экспертов [Viera, 2005].

Существует также ранговая система оценки, когда перевод системы МП попарно сравнивается с переводами других систем в терминах «лучше» (один из переводов явно превосходит другой по качеству), «хуже» и «эквивалентно» (переводы принципиально не отличаются по качеству). В этом случае эксперты, как правило, более последовательны в своих оценках. Для достижения непредвзятости эксперты обычно не знают, результаты работы какой системы они оценивают.

2.1.2. Автоматическая оценка

Инструменты, использующиеся для автоматической оценки МП, в идеале, должны соответствовать следующим критериям: низкая стоимость работы, интуитивно понятные и значимые результаты, постоянство результатов при повторном использовании и, наконец, правильность оценки систем, которые работают лучше. Учитывается также скорость работы, возможность индивидуальной настройки под интересы пользователя и объем памяти, который требуется системе [Koehn, 2010].

Задача таких инструментов это, при наличии эталонного перевода и перевода, осуществленного МП, сравнить их и вычислить, насколько они похожи.

Для автоматической оценки работы машинных переводчиков зачастую используются показатель **Word Error Rate** или **WER**, метрики **BLEU** и **NIST**. Эти инструменты позволяют успешно сравнивать работу разных систем МП и оценивать улучшения в работе конкретной системы [Vilar, 2006]. Используются также метрики **точность (precision), полнота (recall)** и **F-мера** [Koehn, 2010].

Рассмотрим подробнее принципы их работы.

Word Error Rate, или взвешенное расстояние Левенштейна, позволяет измерять расстояние между машинным и образцовым переводом так же, как мы измеряем расстояние между словарным словом и словом с опечаткой (считая символами не буквы, а целые слова) [МП: обзор методов]. По сути WER измеряет минимальное количество изменений, которые необходимо сделать, чтобы из результата работы МП получить эталонный перевод [Koehn, 2010]. При этом WER может учитывать различные варианты эталонного перевода с разным порядком слов [Zhang, 2004].

По формуле взвешенного расстояния Левенштейна:

$$\mathrm{WER} = \frac{substitutions + insertions + deletions}{reference\text{-}length}$$

,

где

замена (substitutions): необходимость замены одного слова другим;

вставка (insertions): необходимость добавления слова;

удаление (deletions): необходимость удаления слова;

длина эталонного перевода (reference-length).

В случае с WER, чем меньше расстояние Левенштейна, тем лучше оценивается работа системы.

Метрика **BLEU** (Bilingual Evaluation Understudy) на данный момент самая популярная в современной оценке МП. Позволяет учитывать не только точность перевода отдельных слов, но и цепочек слов (n-граммы) [МП: обзор методов].

Метрика BLEU была разработана сотрудниками компании IBM и является одной из самых простых в использовании метрик оценки машинного перевода. Алгоритм BLEU оценивает качество перевода по шкале от 0 до 100 на основании сравнения машинного перевода с человеческим и поиска общих слов и фраз. Основная идея разработчиков метрики состоит в том, что чем лучше машинный перевод, тем больше он должен быть похож на человеческий [Молчанов, 2013].

Вариант метрики BLUE с ограничением до 4-грамм выглядит следующим образом:

$$\mathrm{Bleu} - 4 = min(1, \frac{\mathrm{output} - \mathrm{length}}{\mathrm{reference} - \mathrm{length}}) \sqrt[4]{\prod_{i=1}^{4} precision_i}$$

,

где

precision — отношение количества корректных i-грамм к общему количеству i-грамм в переводе;

output — lenght — длинна перевода, который оценивает метрика;

reference — length — длинна эталлоного перевода;

Лучше всего такая метрика работает не на уровне предложений, а на уровне большого текста. На маленьком объеме текста метрика зачастую обнуляется из-за отсутствия совпадающих 4-грамм и работает некорректно. Существуют также доработанные варианты метрики, которые подходят для сравнения на уровне предложения.

Метрика **NIST** была разработана на основе BLEU, но имеет одно фундаментальное отличие. Если для получения высокой оценки BLEU важнее правильный порядок слов, то NIST выше оценивает правильный выбор лексики [Zhang, 2004].

Для использования метрик BLUE и NIST требуется корпус предложений на исходном языке и различные эталонные переводы этих предложений, выполненные человеком [Zhang, 2004].

Очевидно, что ни метрика BLUE, ни NIST не работают так же как экспертная оценка. Эксперты выше оценивают грамматически верные переводы, которые напоминают тексты на естественном языке, тогда как метрики оценивают тексты в пределе 5-грамм [Zhang, 2004], а значит, не могут оценить то, например, как связаны между собой предложения в переведенном тексте.

Хотя использование метрик BLEU и NIST, становится все более популярным, мы не до конца понимаем, как именно они работают [Zhang, 2004], Часто результаты их работы сложно интерпретировать, а выяснить причины ошибок, появляющихся в конкретной системе, с помощью только этих мер невозможно [Vilar, 2006].

То же касается и всех остальных инструментов автоматической оценки работы МП в целом. Сама по себе оценка работы системы без каких-то дополнительных исследований, не предоставляет полезной информации, которую можно было бы использовать для дальнейшего развития системы

МП. Одним из вариантов такого исследования является подробный анализ результатов работы системы и появляющихся при работе ошибок.

2.2. Типология ошибок машинного перевода

Для поиска и анализа ошибок удобно использовать один или несколько вариантов эталонного перевода, чтобы можно было противопоставить результат работы системы МП и правильный текст [там же]. В случае, когда таких эталонных переводов нет, то эту задачу должен выполнять эксперт, владеющий принимающим языком и языком-источником.

Классификация ошибок, появляющихся при работе системы МП, ни в коем случае не является однозначной. Не существует общепринятой, единой классификации, а исследователи могут создавать собственные в зависимости от преследуемых целей. В нашей работе мы используем две разные классификации ошибок: одна из них уже использовалась ранее другими исследователями [Vilar, 2006; Llitj´os, 2005] и полезна с точки зрения статистической информации о количестве и типах ошибок. Вторая классификация — наша собственная, используется с другой целью, и будет рассмотрена подробнее в третьей главе.

Первая классификация имеет иерархическую структуру. Ошибки делятся на пять больших классов: пропущенные слова, неправильный порядок слов, неверные слова, неизвестные слова и пунктуационные ошибки. Классы в свою очередь делятся на меньшие категории.

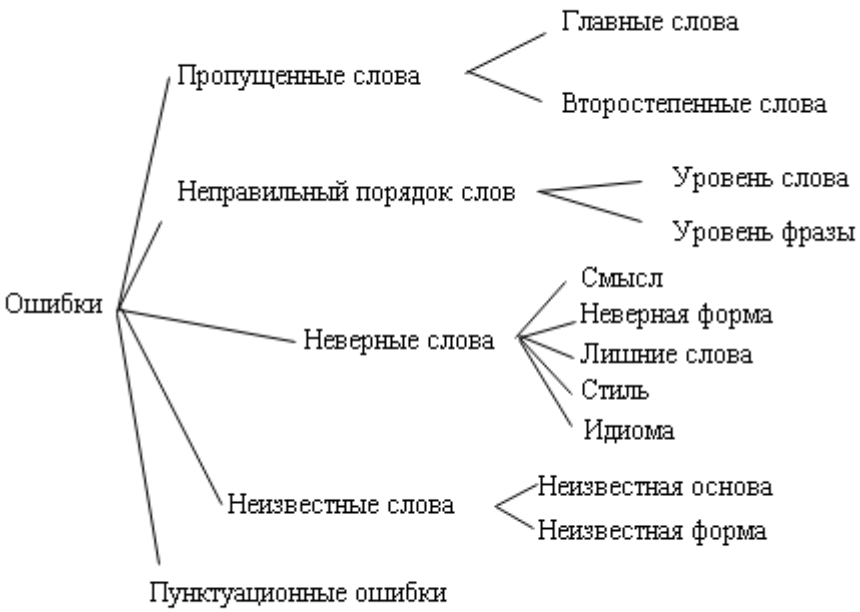

Рис. 1. Классификация ошибок, предложенная Vilar и Llitj´os

Следует помнить, что все типы ошибок взаимосвязаны и ошибка одного типа может быть причиной появления ошибки другого. Так, например, неправильный перевод одного слова может привести к неправильному порядку слов в предложении.

Все примеры, которые мы используем для иллюстрации классификации на рис.1, мы взяли из корпуса пользовательских запросов, более подробно описанного в третьей главе.

2.2.1. Пропущенные слова

Ошибка относится к этой категории, когда в результате работы системы в выходном тексте пропущены слова. Слова по важности можно разделить на «главные»” и «второстепенные». К главным словам относятся существительные и глаголы, которые несут в себе основной смысл предложения. К второстепенным относятся слова, которые требуются для составления грамматически правильного предложения. Так, выделяются два

подкласса ошибок — первый, более важный случай, когда в результате отсутствия некоторого слова оказывается непонятен смысл предложения, и второй случай, когда смысл сохраняется, но предложения оказывается некорректным с точки зрения грамматики.

Первый тип:

прикольный малый	rento

Rento ‘прикольный’. В результате перевода отсутствует существительное, соответственно, не понятно, кто именно «прикольный».

Второй тип:

Kokouskutsu ja esityslista liitteenä.	Приглашение на встречу и повестка дня вложения.
jossa hallitus aikoo rajoittaa maahanmuuttajien oikeutta sosiaalitukiin.	где правительство намеревается ограничить права иммигрантов социальная помощь.

В обоих случаях отсутствуют необходимые для построения грамматически правильной фразы предлоги. В первом примере отсутствует предлог «во». При этом «вложение» стоит в некорректном падеже. Во втором примере отсутствует предлог «на». «Социальная помощь» стоит в некорректном падеже.

2.2.2. Неправильный порядок слов

При неправильном порядке слов могут быть два случая: в первом случае для получения правильного перевода нужно перенести слово в другую часть текста. Во втором случае, когда ошибка допущена на уровне фразы, для получения правильного перевода нужно перенести блоки идущих последовательно слов.

Первый тип:

-Teimme kahden vuoden sopimuksen ja pidämme siitä kiinni.	- Мы сделали на два года и соглашение мы придерживаемся его.

Если мы перенесем слова «и» и «соглашение», то получим правильный перевод: «Мы сделали соглашение на два года, и мы придерживаемся его».

Второй тип:

Mahdollisuus talvella moottorikelkan vuokraamiseen	Возможность зимой арендной платы для аренды снегохода.

Если перенести словосочетания «арендной платы» и то можно получить правильный перевод: «Возможность арендной платы зимой для аренды снегохода».

2.2.3. Неверные слова

Это самый широкий класс ошибок. Внутри этого класса можно выделить пять подкатегорий.

Первая подкатегория это смысловые ошибки, когда в результате неправильного перевода слова меняется смысл предложения. Такие ошибки могут появляться в двух случаях — когда не было найдено правильного варианта перевода, или когда была неверно снята лексическая неоднозначность. Следующий пример иллюстрирует неправильное снятие лексической неоднозначности:

anna vaikuttaa hetki ja huuhdo vedella, kostealla liinalla .	Анна влияет момент и прополоскать с водой, влажной тканью.

Anna это не только имя, в данном контексте это императив (повелительное наклонение) глагола antaa *‘дать’*. Дословный перевод в этом случае был бы «дайте (чему-то) подействовать и сполосните водой, влажной тканью».

Вторая подкатегория — слово стоит в неверной с точки зрения грамматики форме, но основа слова переведена правильно. Такие ошибки особенно часто встречаются при работе с языками, в которых есть

словоизменение, в том числе, в финском и русском языках. В основном, такие ошибки связаны с неправильным согласованием по числу и роду.

Tilauksenne on käsitelty ja postitetaan viimeistään seuraavana arkipäivänä. setin avulla voit suunnitella ja värittää oman olkalaukkusi.	Ваш заказ обработаны и отправлены самое позднее в следующий рабочий день. набор позволяет планировать и вышить свою сумка.

Есть также случаи, когда вместо одной части речи в оригинале уместно использовать родственное слово другой части речи. Например:

Vuotuinen polttoainetarve on enimmillään lähes 3 miljoonaa kuutiota.	Ежегодная потребность в топливе максимум почти 3 миллионов кубических метров.

Смысл предложения в целом понятен, но в идеале нужно было бы перенести слово «максимум» в начало и образовать от него прилагательное: «Максимальная ежегодная потребность в топливе составляет почти 3 миллиона кубических метров»

Третья подкатегория — лишние слова. Ошибки такого типа появляются в основном при работе с инструментами автоматического распознавания речи.

Четвертая и пятая категории менее важные для работы МП. К четвертой категории относятся стилистические ошибки (например, повторное использование одного и того же слова в узком контексте вместо замены этого слова синонимом). Сюда относится пример, который мы рассматривали до этого, где слово «мы» повторяется два раза.

-Teimme kahden vuoden sopimuksen ja pidämme siitä kiinni.	- Мы сделали на два года и соглашение мы придерживаемся его.

Пятая категория — это ошибки, связанные с переводом идиоматических выражений, которые система не распознает как таковые и переводит как обычный текст.

- Ну, ты талант! Человек пришел за удочкой, а уехал на мотоцикле.	- No, sinä rekvisiittani! Mies tuli / onkia ja lähti moottoripyöräkolari.

В данном примере «ты талант» является устойчивым выражением и не может быть переведено дословно.

2.2.4. Неизвестные слова

Неизвестные системе слова тоже приводят к двум разным типам ошибок. В первом случае, системе неизвестна само слово или его основа (стем), во втором — система не знает определенной формы слова. Такие ошибки связаны с недостаточным объемом корпуса. В случае такой ошибки слово не переводиться, а остается в том виде, в котором оно было в пользовательском запросе. Такие ошибки часто появляются при переводе именованных сущностей, слов, принадлежащих узкой тематике и разговорной лексике.

Продам благоустроенный коттедж на участке 8 соток в п.Сотниково	Myy hyvinhoidettu oli mökki on 8 aarin п.Сотниково
В продолжение нашего разговора о балансе прошу Вас уточнить у бухгалтера Юккостарвике следующие вопросы: Магистратура, туризмовед, экскурсовод.	Keskustelumme jatkoa tasapainosta kirjanpitäjän pyydän tarkentaa joillekin Юккостарвике seuraavat kysymykset: Maisteriohjelmissa, туризмовед, opas.
Pakkolainsäädäntö ei tämän mukaan rajoita sopimusvapautta.	Согласно этому не pakkolainsäädäntö ограничить свободу контракта.

2.2.5. Пунктуация

Пунктуационные ошибки представляют некоторую сложность при работе с языками без фиксированных правил пунктуации. Например, правила расстановки запятых в финском языке значительно мягче, чем в русском.

Kyllä olen töissä. asunto on pieni ja siksi me muutamme elokuussa	Да я работаю. Небольшая квартира и поэтому мы переезжаем в августе.

2.3. Выводы

Оценка качества МП представляет собой сложную задачу, в первую очередь по той причине, что оценить перевод объективно невозможно. При экспертной оценке мнения асессоров могут расходиться, а при оценке с помощью метрик все равно требуется наличие эталонного перевода, выполненного вручную. Проблему также представляет интерпретация результатов такой оценки для дальнейшей работы, так как просто оценка не способствует устранению причин ошибок. При этом можно эффективно использовать экспертную оценку и метрики для сравнения различных систем перевода, или отслеживания улучшений в какой-то одной системе.

Чтобы приблизиться к пониманию того, что именно препятствует улучшению работы системы, можно по разному классифицировать ошибки.

Некоторые ошибки могут провоцировать появление других. К таким ошибкам, в первую очередь, относятся неизвестные слова и неправильный порядок слов.

3. АНАЛИЗ РАБОТЫ СТАТИСТИЧЕСКОЙ СИСТЕМЫ МП

3.1. Корпус и процентное соотношение ошибок

В ходе нашего эксперимента, нами были проанализированы 15043 реальных пользовательских запроса к финско-русскому онлайн-переводчику PROMT на момент 02.11.2015 года. 6804 из них это перевод с финского на русский, 8239 с русского на финский. При этом в нашем списке оказался достаточно большой процент запросов, выполненных на других языках (эстонском, французском, украинском и т.д.), которые были ошибочно распознаны как русский или финский инструментом автоматического распознавания языка.

Мы использовали классификацию типов ошибок, приведенную во второй главе, выбрали 300 первых запросов из обоих списков (2% от общего числа запросов), посчитали количество ошибок того или иного класса, которые мы кратко описали в пункте 2.2., и получили для нашего корпуса следующие приблизительные статистические данные:

Рис.2. Диаграмма процентного соотношения ошибок согласно классификации, приведенной в пункте 2.2.

Мы обнаружили, что больше всего ошибок относятся к категории «неверный перевод слова». Примерно треть из всех ошибок можно отнести к классу «неизвестные слова». Из этого мы можем сделать вывод о недостаточной представительности корпуса.

3.2. Причины возникновения ошибок

Как уже было сказано, нашей основной задачей является не столько перечисление типов ошибок, сколько анализ возможных причин их возникновения. Для того чтобы это выяснить, мы, учитывая полученные ранее теоретические данные о принципах работы статистического МП, подробно рассмотрели наш корпус и создали свою собственную классификацию ошибок.

В целом, ошибки, возникающие при работе системы PROMT, попадают в четыре большие категории.

- Ошибки, вызванные отсутствием или некорректной предварительной обработкой запроса.
- Ошибки, связанные с содержанием параллельного корпуса.
- Ошибки, связанные с особенностями языков.
- Ошибки, связанные с работой алгоритма системы перевода.

Необходимо упомянуть, что встречаются и некоторые пограничные случаи, когда ошибка вызвана более чем одной причиной. Далее мы рассмотрим упомянутые категории и выделим в них подкатегории.

3.2.1. Ошибки, вызванные отсутствием или некорректной предварительной обработкой запроса

Как мы уже выяснили, сейчас инструменты онлайн перевода активно используются обычными пользователями, которые зачастую либо не знают принимающего языка, либо языка источника. Пользователям свойственно допускать опечатки, нарушать правила пунктуации и использовать разговорную лексику, которую невозможно найти ни в одном словаре. В связи

с этим, для корректной работы системы МП, должна проводиться предварительная обработка запроса и его нормализация.

3.2.1.1. Некорректное распознавание языка

Если предположить, что переводчиком будут пользоваться обычные пользователи, не имеющие базовых знаний о языках, которые не могут сами выбрать необходимый язык, то им необходим инструмент автоматического распознавания.

Рассмотрим следующие примеры работы программы.

take off the gun powder	take off the gun порошок
I went to Mexico in the calendar year before this one.	I вентти to Мехико in the calendar year before this one.
I want be with you	I want be with вы

В этих примерах язык определяется по одному слову (powder, went, you), для которых есть перевод в параллельном корпусе. В финском языке часто встречаются английские слова в неизмененной форме, особенно в названиях товаров, заведений. Если такое слово попадет в финно-язычный корпус, который использовался при обучении инструмента распознавания языка, то в дальнейшем программа будет определять его как принадлежащее финскому языку.

В случае с финским языком распознавание языка идет с использованием стемминга, учитываются не слова целиком, а их основы. Такой вывод можно сделать, посмотрев на следующие примеры.

Suimming kool	Suimming kool
immensily	immensily
herrasment	herrasment

Sui — основа множественного числа слова *‘рот’*, herra *‘господин’*, immen — генетив единственного числа от слова impi *‘дева’*. Все три запроса написаны с опечатками, и при правильном написании, проблем бы не возникло.

Но бывают такие случаи, когда в двух различных языках присутствуют одинаковые лексемы.

He has never traveled by train.	Они у never путешествовал by train.
Kutsu Teile mässajad pärast ta tapeti. Me olime abielus üksteist aastat	Приглашение в закладки m ä s s a j a d p ä r a s t Тэ убиваемые. Мы инцидентами abielus üksteist совещание

Не — англ. *‘он’*, финск. *‘они’*. Во втором примере за счет стэмминга появляется основа teile, которая распознается как teile — эстон. *‘вам’*, teille — финск. *‘вам’*.

В результате в список попадают запросы на французском, латыни, эстонском и немецком. Такая стратегия как стэмминг не может использоваться в случае с финским языком, по той причине, что в финском и, относящемуся к той же финно-угорской языковой семье, эстонском языке присутствует большое количество родственных лексем [Härmävaara, 2015].

3.2.1.2. Запросы, оформленные некорректно с точки зрения синтаксиса

Предположим, что пользователь не знает языка, с которым работает в программе переводчике. Он просто копирует текст и вставляет его в окошко программы. Следующий пример текст с сайта интернет-магазина.

N y t j o u s t a v a t maksuehdot.Käsiraha alk.0e.Hinta 4490e-syysale 1000.Tarjoushinta 3490e.Hyvä vaunu ympärivuotiseen käyttöön.Siisti, kuiva ja hajuton vaunu.Kaikki laitteet toimii.Pitkä etuteltta kuluu kauppaan.Arvo 1000e.Kts kuva., Parivuode	Сейчас гибкие maksuehdot.Käsiraha alk.0e.Hinta падение продажа 4 4 9 0 e 1000.Tarjoushinta 3490e.Hyvä käyttöön.Siisti вагон весь год, сухая и без запаха vaunu.Kaikki устройства t o i m i i . P i t k ä тент проходит kauppaan.Arvo 1000e.Kts нашем номере двуспальная

кровать, рисунок.

В тексте источнике нет орфографических ошибок, но благодаря отсутствию пробелов, текст был неправильно разбит на токены, и половина слов оказалась не переведена. Знакомые с алгоритмами МП пользователи догадаются, что для корректной работы нужно расставить пробелы, но обычный пользователь, скорее всего, нет. Так, обычный пользователь может совершить ошибку уже на стадии запроса, а значит, и получить неудовлетворительный результат работы программы.

При переводе с русского на финский язык многие пользователи оформляют запросы без знаков пунктуации, а благодаря свободному порядку слов в русском языке, вопросительную конструкцию невозможно отличить от утвердительной. Например,

а ты нет	Sinä et.
так что, можешь приезжать	Joten voit tulla sinne kun olet
туда после того как освободишься	vapaa
ты будешь потом еще одну	
катку	Olet sitten yhden jäähalli

Пользователь, естественно, не знает о том, что в финском языке существует фиксированная конструкция для оформления вопросительного предложения. В такой конструкции недопустим свободный порядок слов. Если бы в финском можно было задать вопрос с помощью той же конструкции, что и в русском, то проблемы разрешения неоднозначности бы не возникло.

-Знаешь, я ведь просто хотел	- - Tiedätkö, minä vain halusin
уберечь тебя от плохих вестей.	suojella sinua huonoista uutisia.

Таким образом, даже если в корпусе будет содержаться правильный перевод для следующего предложения, при неправильном оформлении запроса, перевод будет неадекватным. В идеале, при оформлении запроса программа должны учитывать необходимость использования пунктуации и напоминать об этом пользователю.

3.2.1.3. Некорректное распознавание именованных сущностей

Rela Colic Drops vähentää tutkitusti vauvojen koliikki-itkua	Rela) colic ронявшая уменьшить плач колика доказанные быть младенцев
Gustaf Mannerheim syntyi Louhisaaren kartanossa, Turun lähellä.	Густав Маннергейм родился заминировали остров в особняке, Турку близко.

Отдельную проблему при работе системы представляет перевод именованных сущностей как обычных слов. Представить, что может существовать такой корпус, где будут содержаться все возможные имена собственные и их переводы — немыслимо. Тогда перед нами встает вопрос: как с ними поступать? Очевидно, что необходим дополнительный инструмент для выделения или распознавания таких сущностей. Один из самых простых способов сделать это — подключить словарь, опирающий на традиции русской и финской терминологии. Другим возможным решением может быть написание отдельной грамматики для выделения имен (например, учитывать наличие кавычек, написание с большой буквы и так далее).

3.2.2. Ошибки, связанные с содержанием параллельного корпуса

Корпус параллельных текстов, использующийся для построения переводческой модели должен соответствовать ряду критериев, таких как отсутствие опечаток, изначально неверных переводов, качественно выполненное и, по возможности, проверенное экспертом выравнивание.

К сожалению, при работе с малыми языками, такими как финский, найти уже готовый корпус параллельных текстов широкой тематики, который мог бы использоваться для обучения онлайн-переводчика сложно. Учитывая небольшое количество параллельных ресурсов для финского и русского языка создавать такой корпус придется с нуля.

Следует также учитывать, что не все тексты требуют единой стратегии перевода. В каждом языковом коллективе существуют нормы особенности

расположения и структурирования информации, различнык способы её подачи и представления, нормы языкового оформления в рамках различных функциональных контекстов. Тексты могут различаться по жанру и стилю, и знание жанровых и стилистических особенностей играет значительную роль при создании качественного перевода. Существует ряд формальных признаков, которые присущи текстам того или иного стиля. Так, функциональный стиль текста будет влиять на выбор лексики, которую следует использовать при переводе. В первую очередь это касается текстов, принадлежащих узкой предметной области. Существует также вопрос соблюдения норм, которые существуют для соответствующей разновидности текстов в принимающем языке. Все это в значительной степени усложняет задачу построения параллельного корпуса. Может ли быть эффективным использование переводческого инструмента, обученного на текстах художественной литературы, для перевода текстов из области медицины? Рассмотрим следующий пример:

Kennoja teillä jo on. Lähettäisittekö tilausvahvistuksen?	Клетки у вас уже есть. Будете ли вы отправить подтверждение заказа?

Kenno – многозначное слово. Оно действительно переводится как *‘клетка’*.

Но во множественном числе это слово означает *‘теплообменник для отопления салона автомобиля’*. А в составе сложного термина, например, valoherkkä kenno, оно переводится как *‘ светочувствительная матрица’*.

Можно, конечно, предположить, что кто-то покупает клетки, как например, клетки для содержания животных, но в финском языке для этого используется совсем другое слово – häkki. При ручном переводе, если переводчик знает о том, что он работает с текстом из узкой предметной области, он может проверить точное значение этого термина. При статистическом переводе это возможно осуществить, только подключив отдельные терминологические словари. После этого придется либо

предложить пользователю самому выбрать предметную область, либо создать автоматический анализатор текстов.

Отдельной проблемой является нелитературная (разговорная) лексика. Невозможно однозначно ответить на вопрос, следует ли включать ее в параллельный корпус. Её отсутсвие в параллельном корпусе приводит к тому, что такие слова не переводяться.

ПРОШУ ВАС ЛЮДИ НЕ ИГНОРЬТЕ ЭТО ИНФО!	pyydän ihmiset eivät ИГНОРЬТЕ se tietoa!
Ахахахахахаха Ангелин тебя тоже ждёт такая же судьба, эта шибанулась и ты с ней за одно	Ахахахахахаха Ангелин sinua odottaa sama kohtalo, tämä шибанулась ja olet hänelle yhden

Еще хуже ситуация обстоит с финским языком. Финский разговорный язык настолько не похож на литературный, что для его изучения его переводчики проходят отдельные курсы. Рассмотрим следующий пример:

Sun meikit levii ku kyyneleet niin valuvat	Ваш косметика Ливай текут слезы, тогда за

Исходный текст написан на разговорном языке, его литературная версия будет выглядеть вот так: «Sinun meikki leviää kun kyyneleet valuvat».

Попробуем перевести этот вариант:

Sinun meikki leviää, kun kyyneleet valuvat	Вы косметика распространяется, когда слезы текут

Как мы видим, результат уже значительно лучше.

В свою очередь, присутствие разговорной лексики в корпусе может приводить к стилистическим ошибкам:

ja juuri nyt hymyilet ilman epäilyttä.	и прямо сейчас лыбишься без сомнения.

Hymyllä *‘улыбаться’* не отновится к разговорной лексике и стилистически не окрашен, а значит перевод его глаголом «лыбиться» неадекватен.

3.2.2.1. Недостаточный объем корпуса

Финский язык относится к агглютинативным языкам, а значит, словоизменение в финском языке происходит с помощью агглютинации, то

есть присоединения формантов (суффиксов или префиксов), каждый из которых несет определенное значение. В финском языке всего пятнадцать падежей и также есть формы единственного и множественного числа. На практике это означает, что, например, каждое изменяемое по падежам слово в корпусе теоретически должно встречаться в своих пятнадцати формах в единственном числе, и еще пятнадцати формах множественного числа.

Эта ситуация осложняется тем, что русский язык, в свою очередь, относится к флективным языкам, где словоизменение происходит при помощи флексий, то есть формантов, сочетающих сразу несколько значений. Это означает, что в русскоязычной части параллельного корпуса каждое изменяемое по падежам слово должно встречаться в минимум шести формах основных падежей единственного числа и еще шести множественного.

Нужно помнить, что в крайнем случае, если какое-то слово появляется в тренировочном корпусе в какой-то одной форме, а в другой нет, то другая форма того же самого слова будет рассматриваться, как если бы слово не содержалось в словаре. Следующие примеры демонстрируют нам, что некоторые слова содержаться в корпусе только в одной форме. Во всех примерах инфинитив слова переводится другим падежом или формой множественного числа.

Kertakäyttögrilli	Одноразовые мангалы
зять	vävypoikasi
удача	Onnea
Вариант:	Vaihtoehdot:

3.2.2.2. Иноязычные слова в корпусе

Thx for your order from us .Your item has been shipped by China post International Mail Service. It is estimated to arrive in 15-40 days in normal conditions.	Thx for your order from США . Y o u r i t e m имеет Би поставляется b y C h i n a p o s t International Mail Service. It is расчетное to приходим в 15-40 days

	in normal условиях.
He has never traveled by train.	Они у never путешествовал by train.

Данные примеры были ошибочно определены как тексты, написанные на финском языке. Тем не менее, часть слов переведена, а значит, что они содержатся в корпусе, использовавшемся для тренировки переводческой модели.

Невозможно полностью избавиться от иноязычных вкраплений в корпусе, из-за большого количества названий брендов и товаров на английском языке. Для решения этой проблемы на уровне морфологической разметки можно определять иностранные слова и помечать их как не требующие перевода.

3.2.2.3. Неправильный перевод и опечатки

Очевидно, что если обучать модель на текстах с изначально неправильным переводом, то ошибка будет повторяться.

Tarjolla on aina suihkumahdollisuus ja puhtaat pyyhkeet.	На рану всегда вероятность ливней и чистые полотенца.

У слова suihkumahdollisuus есть только один возможный перевод — *‘возможность помыться’*. “Вероятность ливней” — это либо неверный перевод, осуществленный переводчиком, который решил положиться на интуицию и предположил, что suihku *‘душ’* может принимать значение *‘дождь’*, либо результат машинного перевода с помощью английского как посредника. Если обратиться к Google-translate, то в результате перевода на английский у нас появится «chance of showers», что при переводе на русский даст «вероятность ливней».

Следующие примеры демонстрируют наличие опечаток в параллельном корпусе:

Anna puhdistaa itse.	Анна вънистить сам.

Arvoisa asiakkaamme,	Г-н наши клиенты, тагшта
Maritim-verkkokauppamme uudistuu	интернет-магазин теперь
nyt kokonaan.	полностью обновляется.
Hymyillä	Улыбяться
Ihoni on läpinäkyvä ja kylmä	Моя кожа посмотрите-thr и холодная

3.2.3. Ошибки, связанные с особенностями языков

При этом общепризнанно, что системе МП легче переводить текст, где порядок слов в предложении жестко фиксируется. Русский язык (как и финский) поддерживает свободный порядок слов в предложении, что значительно усложняет процесс его формализации [Карасев, 2011].

Сходства и различия между языками изучаются и эти данные активно применяются в том числе в области перевода. Структурные различия между двумя языками могут оказывать значительное влияние на качество перевода между ними [Koppel, 2011].

Нельзя не упомянуть также о том, что работая с переводами, мы, по сути, имеем дело с текстами, написанными на переводческом языке (translationese).

Так, исследования в области теории перевода показывают, что тексты, полученные в результате перевода, принципиально отличаются от текстов, изначально написанных на принимающем языке [Twitto-Shmuel, 2015]. Идея существования специального переводческого языка (translationese) и его различных диалектов, зависящих от комбинации языка-источника и принимающего языка [Koppel, 2011], может в значительной степени изменить наше представление о работе SMT.

Например, работа переводческой модели может напрямую зависеть от того, на каком языке были изначально написаны тексты в параллельном корпусе. Так, в ходе одного из проведенных экспериментов [Kurokawa, 2009], было создано два корпуса текстов, специально переведенных экспертами на английском и французском языках. Модель, которую обучали на корпусе параллельных текстов, переведенных с французского языка на английский,

работала значительно лучше, чем модель, обученная на текстах, переведенных с английского на французский. Эти данные подтвердились в ходе дальнейших экспериментов с переводческим языком [Lembersky, 2013].

Все это доказывает, что особенности языков должны учитываться и при составлении параллельного корпуса.

3.2.3.1. Прагматические адаптации

Одна из проблем, с которой сталкиваются переводчики, это вопрос прагматических адаптаций. Мы уже упоминали, о том, что невозможно заставить программу принимать решения с учетом внешней метаинформации о тексте. Тем не менее, при работе с онлайн-переводчиком, проблемы, связанные с прагматикой все равно появляются. Одной из таких проблем является проблема перевода местоимений и глаголов в форме второго лица множественного и единственного числа.

В финской культуре общения совершенно нормально обращаться к большинству людей «на ты», при этом «на вы» обращаются к посетителям или покупателям. Существует даже мнение, что обращение «на вы» является индикатором создания некоторой дистанции между собеседниками и вообще может быть воспринято как грубость. Тогда как в русской культуре общения обращение «на ты» к случайному прохожему или начальнику будет воспринято как фамильярность. В связи с этим, рассмотрим следующие примеры.

Missä olit kesällä?	Где вы были летом?
Kysy lisää.	Спроси еще.
Terve igor! Olitko jatkamassa datanomin opintoja nyt päivälinjan puolella vai oletko saanut kaikki näytöt suoritettua?	Здравствуй Игорь! Вы были выполняются данные самостоятельно теперь исследований на день линия или вы позвонили, все мониторы выполненную?

Данные примеры демонстрируют непоследовательность в переводе местоимений и глаголов во втором лице. Во всех трех примерах в исходном тексте использована форма глагола второго лица единственного числа.

Возможное решение этой проблемы – использовать единую стратегию перевода текстов, которые содержаться в параллельном корпусе.

3.2.3.2. Тире в русском

Проблема перевода предложений с тире в том, что в финском языке тире редко используется, а эллиптических конструкций нет. Поэтому при переводе с русского на финский в предложениях с тире отсутствует глагол. А в финском языке это является грамматической ошибкой.

Слабые-мстят, сильные-прощают, счастливые-забывают....	Heikot kostavat, vahvat - anteeksi antamusta, onnellisen sydän unohtaa.

3.2.3.3. Вопросительная форма глагола в финском языке

Как мы уже упоминали в предыдущем пункте, в финском языке существует фиксированная конструкция для оформления вопросительного предложения.

На первое место выходит вопросительное слово или глагол с вопросительной частицей **ko/kö .** В данном примере, видимо вследствие отсутствия вопросительной формы глагола **työskennellä** *‘работать’* в параллельном корпусе, вместо вопросительной конструкции получилась утвердительная, даже не смотря на то, что в запросе присутствует вопросительный знак.

Сколько это км от Хельсинки ? Где ты работаешь ? Ты работаешь ночью ?	Paljonko se km Helsingistä? Missä olet töissä? **Työskentelet** yöllä?

Как только в корпусе появится эта форма глагола, ошибка исчезнет, но её можно было бы решить иначе, благодаря наличию фиксированной конструкции в финском языке. Достаточно прибавить **ko/kö** к нужной личной форме глагола и перенести его на первое место — так можно создать вопросительные формы для любого глагола. Если использовать

3.2.3.4. Обобщенно-личные предложения по смыслу, но не по форме

В финском языке есть тип предложений, который является калькой с английского. Это обобщенно-личные предложения по смыслу, в которых глагол стоит во втором лице. Например:

Tilaukset teet helposti: Заказы делаеш легко:

Когда мы переводим это предложение на русский язык дословно, то есть оставляем глагол во втором лице, то теряется смысл оригинала.

В данном случае предложение следовало перевести, как *‘Заказы делать легко!’* или *‘Вы сможете легко сделать заказ!’*. Правильно перевести такие конструкции, можно только используя синтаксический анализ. С помощью синтаксического анализа можно определить, что в предложении объект стоит на первом месте, а глагол во втором лице единственного числа и предположить, что мы имеет дело с обобщенно-личным предложением. Это не решит проблему полностью, так как такие конструкции определяются скорее по контексту, чем по формальным признакам.

3.2.3.5. Предложения с глаголом olla

К таким предложениям относятся прежде всего экзистенциальные предложения. В финском языке они имеют следующий вид: слово в инессиве/аллативе + глагол существования olla «есть, существует» + подлежащее в инфинитиве/партитиве. Такие конструкции выражают местонахождение чего-то или кого-то где-то. Похожие на них конструкции — конструкции обладания: местоимение или имя собственное в аллативе + глагол существования olla «есть, существует» + подлежащее в инфинитиве/партитиве. Основная проблема при переводе на русский таких конструкций в том, что в большинстве случаев глагол olla переводится на русский либо эллиптической конструкцией, либо заменяется подходящим по

смыслу глаголом. Рассмотрим самые простые примеры, которые демонстрируют использование таких конструкций.

Minulla on siniset silmät. *‘У меня голубые глаза’.*

Poydällä on kuppi. *‘На столе стоит кружка’.*

Poydällä on kissa. *‘На столе сидит кошка’.*

Tampereella on paljon turisteja. *’В Тампере много туристов’.*

Huoneessa on paljon ihmisiä. *’В комнате много людей’.*

В трех случаях глагол **olla** при переводе заменялся эллиптической конструкцией, а в еще двух случаях был заменен подходящим по контексту глаголом, *‘сидеть’* и *‘стоять’*. В случае с переводом экзистенциальных конструкций на русский — это два самых распространеннх варанта перевода. Тогда перевод **olla** можно свести к принятию решения между эллиптической конструкцией, и глаголами *‘сидеть’* и *‘стоять’*.

Эту проблему можно попробовать решить с помощью соответствующего грамматического показателя одушевленности в морфологической разметке. Можно использовать эту дополнительную лингвистическую информацию для принятия решения между глаголами *‘сидеть’* и *‘стоять’*. Но проблема будет возникать с глаголом *‘лежать’,* так как с некоторыми предметами будет использоваться только этот глагол. Самый эффективный способ — распознавать такие конструкции и создать правило, которое позволяет заменять глагол **olla** в таких предложениях на тире.

Перевод других типов предложений с **olla** тоже представляет собой сложную задачу.

Мне кажется, что Немецкий популярный язык. Поэтому я теперь буду писать по-Фински	Minusta tuntuu, että Saksan suosittu kieli. Siksi olen nyt kirjoitan suomea

Полученное в результате предложение грамматически неверно, в нем отсутствует глагол olla в форме третьего лице единственного числа. Правильный перевод звучит как «Saksa on suosittu kieli» При работе

статистического переводчика каждому слову должно соответствовать одно слово. Мы не может перевести пустое место каким-нибудь словом. Когда мы переводим русскую эллиптическую конструкцию на финский язык, мы получаем неправильный перевод, так как система МП не может просто так добавить в него глагол olla, который ничему не соответствует в исходном тексте.

3.2.4. Ошибки, связанные с работой алгоритма системы перевода

Чтобы с уверенностью рассуждать об ошибках работы алгоритма перевода, нужно знать архитектуру конкретной системы перевода. Тем не менее, имея даже общее представление о работе статистических алгоритмов, мы можем выделить случаи, которые относятся к некорректной работе алгоритма. Следующие примеры демонстрируют систематическую ошибку, которая появляется регулярно. В столбце справа дается правильный перевод.

Ohi kulkeneen	Прохождение	Вместе с мимо
	магазина	проходившими
Koulunkäynnin	Школу	Посещение
		школы
Käsitellään	Дело	Рассмотрим
		(императив)
noutopisteessä.	Взяв.	На пункте
		выдачи
kanssa	врача	Вместе с
tuossa	сидит	Там

Во всех приведенных примерах можно заметить повторяющийся шаблон, где слово переводится неправильно. В качестве перевода ошибочно выбирается слово, которое стояло за данным словом в корпусе, использовавшемся для обучения переводческой модели.

Например, **noutopisteessä** *‘на пункте выдачи’* переведено как *‘взяв’*. Можно с уверенностью сказать, что в параллельном корпусе содержалось предложение **ottaessa noutopisteessä** *‘взяв на пункте выдачи’*.

Käsitellään *‘рассмотрим’* переведено как *‘дело’*. Вероятно, что в корпусе было предложение **käsitellään asiaa** *‘рассмотрим дело’*.

Kanssa *‘вместе с’* переведено как *‘врача'*. Вероятно, в корпусе содержалось словосочетание **lääkarin kanssa** *‘вместе с врачом’*.

Эта ошибка также может быть связана с относительно свободным порядком слов в финском языке и свободным порядком слов в русском языке.

Существует также ряд ошибок, который связан с несовершенством алгоритма перевода, в том плане, что инструмент не «осознает», что он переводит. Такие ошибки самые сложные и исправить их без использования семантического анализа невозможно.

Minulla on myynnissä punainen pentu	У меня есть в продаже красный парень

Pentu ‘щенок’,’парень’. В данном примере неправильно проводится снятие лексической неоднозначности. Система МП не может «знать», что в интернете не принято публиковать объявления о продаже людей. Для того чтобы правильно перевести такой текст система должна иметь представление о семантических связях между понятиями в реальном мире. Для таких целей могут использоваться иерархические семантические деревья (сети).

3.3. Выводы

С помощью подробного анализа ошибок, учитывая теоретические значения о лингвистических особенностях языков и принципах работы статистического МП, можно обнаружить причины ошибок и предложить возможности их исправления.

Основные проблемы в работе инструмента PROMT вызваны недостаточной представительностью корпуса, который использовался для создания переводческой модели. Также в корпусе содержаться некоторые изначально неверные переводы и опечатки, что приводит к повторению той же ошибки при работе системы.

Работу в некоторой степени нарушает присутствие иностранных слов в корпусе, использовавшемся для построения модели языка.

Отсутствие предварительной обработки запроса приводит к тому, что пользователи оформляют запросы неверно и на выходе получают неудовлетворительный результат.

Система не всегда справляется с переводом имен собственных, либо оставляя их в изначальной форме, либо переводит их как обычные слова.

В работе алгоритма перевода присутствует повторяющаяся систематическая ошибка, связанная со слишком большим окном перевода.

Пользователи, не знакомые с принципами работы статистического МП, часто некорректно оформляют запросы, что приводит к получению неудовлетворительных результатов.

ЗАКЛЮЧЕНИЕ

В данной работе мы определили понятие машинного перевода, описали основные типы систем и методы оценки МП. На основании изученной нами теоретических данных, описанных в первой и второй главах, мы проанализировали работу статистического онлайн-переводчика PROMT, подробно разобрали ошибки, появляющиеся при работе этой системы, привели нашу собственную классификацию ошибок и предложили способы их устранения.

В первой главе мы описали историю развития систем МП, современное состояние этой области, и рассмотрели три основных современных подхода к МП: основанного на правилах, статистического и гибридного. Далее, во второй главе мы описали популярные способы оценки МП, экспертную оценку и различные метрики. Мы также привели одну из возможных классификаций ошибок, появляющихся в ходе работы систем МП.

В третьей главе для анализа работы статистического переводчика PROMT мы использовали корпус из 15043 реальных пользовательских запросов (295 тысяч токенов). Мы привели статистические данные типов ошибок и дали свою собственную классификацию ошибок, на основе причин их появления.

Исследовательские работы, проводящиеся в области МП, часто можно разделить на две категории: написанные с точки зрения лингвистики, и написанные с точки зрения точных вычислительных наук. Так, работы, в которых дается оценка качества перевода, часто полностью опускают или не учитывают принципы работы программ, которые используются для этого перевода. Исследования, которые не учитывают данных лингвистики, предоставляют статистические данные о количестве и типах ошибок, оценки BLEU или NIST, которые сложно интерпретировать. В итоге, это приводит к тому, что причины появления ошибок остаются за пределами исследования. Для улучшения результатов таких исследований, специалисты из разных областей должны больше взаимодействовать.

Для дальнейшего развития систем перевода, которые будут использоваться реальными пользователями, нужно понимать, как и кто в конечном итоге будет ими пользоваться. Нужно учитывать потребности пользователя. Так для профессиональных переводчиков будет полезна возможность выбора между несколькими вариантами перевода, а для обычного пользователя потребуются различные инструменты, осуществляющие предварительную обработку запроса.

Наше собственное исследование запросов продемонстрировало, что пользователи переводят тексты различных функциональных стилей, и ограничить тематику или стилистику текстов практически невозможно. Тем не менее, можно утверждать, что значительная часть запросов относится к области коммерции и развлечений. Эти данные можно учитывать в дальнейшем при составлении корпуса для переводческой модели.

Мы считаем задачи, поставленные в данной работе, выполненными, а цель — достигнутой.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Аматов А. М. К вопросу машинного перевода: энтропия языковой системы и способы ее преодоления // Вестник ЛГУ им. А.С. Пушкина. 2008. №2 (13) С.71-90.
2. Ахманова О. С. Словарь лингвистических терминов. М., 1969.
3. Бархударов Л. С. Язык и перевод. М.,1975.
4. Беляева Л. Н. Лингвистические автоматы в современных гуманитарных технологиях: Учебное пособие. СПб, 2007.
5. Борисова И. А. К опыту постредактирования на материале англо-русского перевода с помощью автоматических систем Google translate и Prompt // Вестник МГЛУ. 2014. №13 (699) С.53-59.
6. Борисова И. А. Коммуникация между интернет-пользователями — носителями различных языков // Вестник МГЛУ. 2013. №13 (673) С.28-34.
7. Гальперин И. Р. Введение. // Большой англо-русский словарь. М., 1987.
8. Кан, Д. А. Применение теории компьютерной семантики русского языка и статистических методов к построению системы машинного перевода: диссертация кандидата физико-математических наук. Место защиты: Федеральное государственное образовательное учреждение высшего профессионального образования Санкт-Петербургский государственный университет. Санкт-Петербург, 2011.
9. Карасев И. В., Артюшина Е. А. Системы машинного перевода // Успехи современного естествознания. 2011, №7, С.117-118.
10. Колшанский Г. В. Контекстная семантика. М., 1980.
11. Комиссаров В. Н. Современное переводоведение. Учебное пособие. М., 2002.
12. Красных В. В., Изотов А. И. Язык, сознание, коммуникация: Сборник статей. М., 2011.
13. Латышев Л. К. Перевод: проблемы теории, практики и методики преподавания. М., 1988.

14. Максименко О. И., Чинина Д. С. Обзор системы машинного перевода «Google Переводчик» (на примере финского языка). // Science Time, 2014, №5 (5), С.133-139.
15. Марчук Ю. Н. Компьютерная лингвистика: учебное пособие. М., 2007.
16. Молчанов А. Статистические и гибридные методы перевода в технологиях компании ПРОМТ. М., 2013.
17. Найда Ю. К науке переводить // Вопросы теории перевода в зарубежной лингвистике. М.,1978.
18. Нелюбин Л. Л. Толковый переводческий словарь. М., 2011.
19. Новожилова А. А. Машинные системы перевода: качество и возможности использования // Вестник ВолГУ. Серия 2: Языкознание. 2014. №3 С.67-73.
20. Рябцева Н. К. Информационные процессы и машинный перевод. Лингвистический аспект. М., 1986.
21. Слокум Дж. Обзор разработок по машинному переводу. Новое в зарубежной лингвистике. М., 1989.
22. Федоров А. В. Основы общей теории перевода (лингвистические проблемы). М., 2002.
23. Фролов С. В., Паньков Д. А. Проблемы построения машинного перевода. Тамбов, 2008.
24. Шаляпина З. М. Автоматический перевод: Эволюция и современные тенденции // Вопросы языкознания, 1996, №2, С. 105—117.
25. Шевчук, В. Н. Информационные технологии в переводе. Электронные ресурсы переводчика. М., 2013.
26. Baker M. Routledge Encyclopedia of Translation Studies. London & New York, 2001.
27. Brown P. F., Delia Pietra V. J., Delia Pietra S. A., Mercer R. L. The mathematics of statistical machine translation: Parameter estimation // Computational Linguistics, 1993, Vol. 19, №2, P. 263—311.

28. Burukina, I. Translating implicit elements in RBMT. // Translating and the Computer 36, 2014, Asling, P. 182—193.
29. Costa-jussà, M., Fonollosa, J. Latest trends in hybrid machine translation and its applications. // Computer Speech & Language, 2015, №32(1), P. 3-10.
30. Guzmán F., Joty S., Marquez L., Nakov P. Using Discourse Structure Improves Machine Translation Evaluation. // ACL (1), 2014, P. 687-698.
31. Härmävaara H. Trouble sources in Finnish-Estonian RM interaction. Helsinki, 2015.
32. Hearne M., Way A. Statistical Machine Translation: A Guide for Linguists and Translators // Language and Linguistics Compass, 2011, №5, P. 205-226.
33. Heyn M. Integrating MachineTranslation into Translation Memory Systems.// Proceedings of the EAMT Machine TranslationWorkshop, Vienna, Austria, 1996, P. 113—126.
34. Hutchins, 2000a — John Hutchins. Hutchins J. The IAMT Certification Initiative and Defining Translation System Categories // Proceedings of 5th EAMT Workshop, Slovenia, 2000.
35. Hutchins, 2000b — John Hutchins. Petr Petrovich Troyanskii (1894-1950): A forgotten pioneer of mechanical translation. // Machine Translation, vol. 15 no. 3, 2000. P. 187—221.
36. Jehl L. Machine translation for Twitter. Master's thesis. The University of Edinburgh, 2010.
37. Koehn, P. Statistical Machine Translation. Cambridge, UK, 2010.
38. Koponen M., Salmi L. On the correctness of machine translation: A machine translation post-editing task. // The Journal of Specialised Translation, 2015, №23, P. 118—136.
39. Koppel M., Ordan N. Translationese and its dialects. // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Techologies, Portland, Oregon, 2011, P. 1318—1326.

40. Kurokawa D., Goutte C., Isabelle P. Automatic detection of translated text and its impact on machine translation. // Proceedings of MT-Summit XII, 2009, P. 81—88.
41. Lembersky G., Ordan N., Wintner S. Improving statistical machine translation by adapting translation models to translationese. // Computational Linguistics, 2013, №39(4), P. 999—1023.
42. Llitj´os A., Carbonell J., Lavie A. A framework for interactive and automatic refinement of transfer-based machine translation. // Proceedings of the 10th Annual Conference of the European Association for Machine Translation (EAMT), Budapest, Hungary, 2005.
43. Sager J. C. Language Engineering and Translation: Consequenses of Automation. Amsterdam, 1994.
44. Silva J., Rodrigues J., Gomes L., Branco A. Bootstrapping a hybrid deep MT system. Lisbon, 2015.
45. Somers H. L. Intoduction // Computers and Translation: A Translator's Guide. Amsterdam, 2003.
46. Twitto-Shmuel, N., Ordan, N., Wintner, S. Statistical machine translation with automatic identification of translationese. // Proceedings of WMT-2015, 2015
47. Viera A., Garrett J. Understanding interobserver agreement: The Kappa Statistic. // Family Medicine, 2005, №37, P. 360-363.
48. Vilar D., Jia Xu, D'Haro L., Ney H. Error Analysis of Machine Translation Output. In International Conference on Language Resources and Evaluation, pages 697—702, Genoa, Italy, 2006.
49. Zhang Y., Vogel S., Waibel A. Interpreting BLEU/NIST Scores: How Much Improvement do We Need to Have a Better System? // Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC-2004), Lisbon, Portugal, 2004.
50. Zhechev V., Genabith V. Seeding Statistical Machine Translation with Translation Memory Output through Tree-Based Structural Alignment. //

Proceedings of the 4th Workshop on Syntax and Structure in Statistical Translation, Beijing, China, 2010, P. 43—51.

ИНТЕРНЕТ-ИСТОЧНИКИ

51. Сайт Европейской ассоциации машинного перевода EAMT. European Association for Machine Translation EAMT.
URL: http://www.eamt.org/mt.html (дата обращения: 6.01.2016)
52. Сайт Джона Хатчинса.
URL: http://www.hutchinsweb.me.uk/history.htm (дата обращения: 6.01.2016)
53. Лекция о системах МП - Системы автоматического (машинного) перевода текста. История, основные сведения, описание. Лекция №13.
URL: http://itclaim.ru/Education/Course/Lingvistika/Lecture/Lecture13.pdf (дата обращения: 9.11.2015).
54. Онлайн версия доклада ALPAC.
URL: http://www.nap.edu/openbook.php?isbn=ARC000005 (дата обращения: 6.12.2015)
55. МП: обзор методов - Презентация: Математические модели в лингвистике 7. Машинный перевод: обзор методов и оценка качества.
URL: http://lpcs.math.msu.su/~pentus/mfk2015/Lecture07_20151021.pdf (дата обращения: 9.11.2015).