Санкт-Петербургский государственный университет
Филологический факультет
Кафедра математической лингвистики

ПРОБЛЕМЫ СОЗДАНИЯ ГИБРИДНОГО ПЕРЕВОДЧИКА С ЭСПЕРАНТО НА РУССКИЙ ЯЗЫК

Магистерская диссертация
студентки II курса
кафедры математической лингвистики
Орловой Дарьи

Научный руководитель
к.ф.н., доц.
Митренина Ольга Владимировна

САНКТ-ПЕТЕРБУРГ
2015

Введение

Системы машинного перевода приобретают всё большее значение. Люди стремятся путешествовать, узнавать мир, также становятся популярными знакомства и общение в интернете с людьми из других стран. Однако на Земле насчитывается не одна сотня различных языков, и освоить каждый из них хотя бы на базовом уровне – задача для обычного человека непосильная. Получивший в XX веке мировое распространение английский язык облегчает международную коммуникацию, однако, всё же не решает проблему языкового барьера полностью.

В ситуации, где нет возможности попросить кого-либо перевести текст, на помощь приходят системы машинного перевода. От обычного словаря они отличаются тем, что способны перевести готовую фразу целиком, тем самым не требую от пользователя знаний грамматики или лексики языка.

На данный момент в построении систем машинного перевода преуспевают крупные корпорации, такие как Яндекс, Google, PROMT и т.д. Крупные компании могут себе позволить в короткие сроки собрать большие объёмы материала и запустить на своей платформе очередную систему перевода. В основном компании концентрируются на двух подходах к машинному переводу: подходе, основанном на правилах, и статистическом подходе. Каждый из подходов обладает своими недостатками, скомпенсировать которые их объединение.

Подобное слияние двух методов перевода получило название гибридного, и именно оно представляет сейчас наибольший интерес среди компьютерных лингвистов. Несмотря на большой потенциал, разработок в этой сфере ведётся не так много.

В сложившейся ситуации чрезвычайно актуальной кажется задача улучшения систем машинного перевода с тех языков, которые до сих пор не были достаточно автоматизированы, но при этом являются популярными и распространёнными по всему миру. Одним из таких языков является

эсперанто. Однако, несмотря на то, что эсперанто считается самым успешным искусственным языком в мире, автоматических переводчиков, обслуживающих этот язык, лишь единицы. Культура эсперанто не теряет актуальности уже больше столетия, особенно на территории Европы, однако система машинного перевода с эсперанто на русский язык была разработана компанией Яндекс лишь в прошлом году. Как и в случае с другими парами языков, перевод с эсперанто основывается на статистике, что приводит к многочисленным ошибкам в согласовании. Программа, исправляющая уже готовый перевод, может не только существенно упростить понимание текста, но и продемонстрировать преимущество гибридного перевода перед другими типами. Это и определяет **практическую значимость** нашей работы.

Целью работы является выявление проблем построения гибридного компонента для статистического переводчика с эсперанто на русский.

Для достижения поставленной цели нам необходимо решить следующие **задачи**:

- изучить устройство и этапы развития систем машинного перевода;
- проанализировать лексику, морфологию и синтаксис языка эсперанто;
- разработать программу-прототип гибридного компонента переводчика;
- оценить результаты эксперимента и определить дальнейшие пути развития данного проекта.

В первой главе теоретической части рассматривается история систем машинного перевода, текущий этап их развития и основные достоинства и недостатки существующих подходов.

Вторая глава посвящена языку эсперанто: его истории, устройству и уже созданному програмному обеспечению..

Третья глава является практической и описывает основные алгоритмы и этапы работы программы, исправляющей ошибки перевода.

Глава 1. Машинный перевод

Основная задача перевода, будь то машинного или человеческого, — поставить текст на одном языке в соответствие тексту на другом языке, при этом обеспечив их смысловую эквивалентность. [Марчук 1983]. В книге «Введение в машинный перевод» («An Introduction to Machine Translation») британские лингвисты Дж. Хатчинс и Г. Сомерс замечают, что термин машинный перевод сейчас является традиционным названием для компьютеризированных систем, производящих переводы с одного естественного языка на другой с участием человека или без такового [Hutchins, Somers 1992].

Машинный перевод сегодня стал неотъемлемой частью жизни современного общества, и главная причина этому — глобализация и возможность международной коммуникации через сеть интернет. Как отмечает профессор и специалист в области машинного перевода Ю. Н. Марчук, актуальность развития машинного перевода сохраняется по следующим четырём причинам [Марчук 2007]:

- перевод с одного языка на другой – единственный эффективный способ преодоления языковых барьеров;
- растут и расширяются возможности современной компьютерной информационной технологии, поэтому всегда появляется соблазн поручить машине какую-нибудь интересную интеллектуальную задачу;
- спрос на переводы в мире увеличивается в абсолютных и относительных пропорциях соответственно тому, как всё больше естественных языков приобщается к мировой цивилизации и вступает в коммуникационную информационную сферу;
- высока научная привлекательность проблемы машинного перевода.

Профессор Л. Н. Беляева отмечает, что ежедневно пользователи сети интернет производят около миллиона запросов на перевод текстов в различных форматах:

- перевод динамических ресурсов сети;

- перевод сообщений электронной почты;

- перевод запросов к различным поисковым системам.

Таким образом, прослеживается острая необходимость в оперативном получении перевода [Беляева 2007].

В данной главе мы рассмотрим историю развития машинного перевода, основные направления, а также некоторые примеры работающих систем.

1.1 История машинного перевода

Идеи о создании механических переводчиков можно встретить ещё и XVII веке. Так, Готтфрид Вильгельм Лейбниц, философ, логик и языковед, выдвигает идею создания универсального кода, пиктографического языка для выражения мыслей, который может послужить языком-посредником для точной передачи мысли одного языка на другой [Lewis 1918]. Однако машинный перевод не мог развиваться без существовании самой «машины», поэтому историю его существования принято начинать именно с XX века. Дж. Хатчинс выделяет следующие этапы развития машинного перевода [Hutchins 2011]:

- 30-е – начало 40-х годов – докомпьютерный этап;

- 40-е – 1954 – предвестники машинного перевода;

- 1954 – 1966 – годы высоких ожиданий;

- 1967 – 1976 – период затишья;

- 70-е – 80-е годы – эпоха Возрождения машинного перевода;

- 90-е годы – появление новой ветки развития - статистической;

- 2000 – наши дни– своеобразное противостояние различных подходов.

До появления компьютерах исследователи прежде всего концентрировались на автоматизированных словарях. В 1933 году практически одновременно двум ученым из разных стран были выданы патенты на механические словари. Первый — французскому инженеру Георгу Арцруни, второй — русскому ученому Петру Смирнову-Троянскому. Машина Арцруни, которую он назвал «Механический мозг», представляла собой механическое устройство на бумажной ленте для записи и извлечения

информации (для перевода произвольного слова на другой язык), работающее на электрическом моторе. Создатель предполагал использование устройства для расписаний поездов, банковских счетов и, в частности, как механический словарь. Некоторые компании проявили живой интерес к изобретению. В 1937 году был даже продемонстрирован прототип машины. Однако начало Второй Мировой войны остановило дальнейшее развитие «Механического мозга».

Система Смирнова-Троянского опиралась на немного другие принципы. Автор выделял 3 этапа перевода как процесса, машина принимала участие только лишь во втором. На первом этапе человек, владеющий лишь языком-источником, должен был обработать входные данные следующим образом: все слова редактор приводил к начальной форме, а затем каждому слову приписывал его синтаксическую функцию в предложении с помощью специальных символов. На втором этапе машина, представленная в виде автоматического двуязычного словаря, заменяла начальные формы и символы языка-источника на последовательность начальных форм и символов для целевого языка. Третий этап требовал нового редактора, знающего целевой язык, который должен был превратить сочетания слов и символов в грамматически связанное предложение. Также, автор считал, что в будущем первый и третий этапы также должны быть автоматизированы [Hutchins 1986].

Примечательно то, что вспомогательные символы, описывающие синтаксические функции слов, Смирнов-Троянский позаимствовал из языка эсперанто, воспользовавшись буквами и буквосочетаниями грамматических показателей. Например, -j обозначало множественное число, -o указывалось для существительных в именительном падеже, а –n — в случае существительного в косвенном падеже. Выбор именно такой нотации можно объяснить чрезвычайной популярностью языка эсперанто в первой половине XX века. Многие считали эсперанто языком будущего, что, вероятно, и сподвигло Смирнова-Троянского взять эсперантские аффиксы в качестве спецсимволов. Человеку, владеющему эсперанто, не составило бы никакого

труда присвоить подобные категории словам и не пришлось бы дополнительно что-то запоминать.

«Лингвистический арифмометр» П. П. Смирнова-Троянского опередил время, но дошел до наших дней лишь в списке научных курьёзов: расширить его функциональность для работы с естественным языком так и не удалось. Идеи Смирнова-Троянского были забыты вплоть до середины 50-х годов. Впервые на достижения изобретателя обратил внимание Л. И. Жирков в своей статье «Границы применимости машинного перевода», опубликованной в 1956 году [Жирков 1956], а уже в 1959 году труды Смирнова-Троянского с комментариями различных учёных были опубликованы Академией Наук в «сборнике материалов о переводной машине для перевода с одного языка на другие, предложенной П. П. Троянским в 1933» [Бельская и др. 1959].

Началом нового этапа в развитии машинного перевода можно считать март 1947 г., когда Уоррен Уивер, директор отделения естественных наук Рокфеллеровского фонда, в письме кибернетику Норберту Винеру впервые сформулировал концепцию машинного перевода, которую несколько позже (в 1949 г.) развил в своём меморандуме, адресованном Фонду. Текст в переводе на русский звучит так: «У меня перед глазами текст, написанный по-русски, но я собираюсь сделать вид, что на самом деле он написан по-английски и закодирован при помощи довольно странных знаков. Все что мне нужно — это взломать код чтобы извлечь информацию, заключенную в тексте.». По словам О. В. Митрениной, именно дату создания письма (4 марта 1947 года) можно считать днём рождения машинного перевода как научного направления [Николаев и др. 2016].

В 1952 г. состоялась первая конференция по машинному переводу в Массачусетском технологическом университете. В 1954 г в штаб-квартире компании IBM при поддержке университета Джорджтауна была представлена первая система машинного перевода – IBM Mark II. Это событие вошло в историю как «Джорджтаунский эксперимент». Публике был представлен первый электронный переводчик — русско-английская система IBM Mark II,

содержавшая словарь из 250 единиц и 6 грамматических правил. Группа исследователей во главе с профессором Леоном Достертом решила продемонстрировать автоматический перевод на примере небольшого количества предложений из органической химии и нескольких предложений на общие темы. Поскольку машина действительно перевела предложения, пусть и на узкую тематику, результаты впечатлили как учёных, так и обывателей. Опыт получил хорошие отклики в прессе. Этот эксперимент способствовал началу исследовательских работ в этой области во многих странах, в том числе в США, СССР, Италии, Китае, Франции [Соловьёва 2008]. Казалось, что создание систем качественного автоматического перевода вполне достижимо в пределах нескольких лет.

К началу 50-х годов целый ряд исследовательских групп в США и в Европе работали в области машинного перевода. В основном они занимались разработкой двуязычных систем перевода. В исследования и разработку систем машинного перевода были вложены значительные средства, однако результаты очень скоро разочаровали инвесторов. Одной из главных причин невысокого качества машинного перевода в те годы были ограниченные возможности аппаратных средств: малый объём памяти при медленном доступе к содержащейся в ней информации, невозможность полноценного использования языков программирования высокого уровня. Другой причиной было отсутствие теоретической базы, необходимой для решения лингвистических проблем. В результате системы машинного перевода тех времён состояли в основном из объёмных двуязычных словарей, и слово языка источника порождало несколько вариантов перевода на целевом языке. [Филинов 2002].Однако, оптимизм в эти годы сохранялся на протяжении всей декады, и учёные с нетерпением ждали скорейшего прорыва в исследованиях.

В СССР первая экспериментальная система машинного перевода с французского языка на русский была создана в 1955-1956 гг. (авторы О. С. Кулагина, И. А. Мельчук). В этой системе словарь основ имел объём в 1236 слов, был введён словарь оборотов в 250 единиц. Единицей перевода являлось предложение. Алгоритм морфологического анализа на основе

работы с таблицами окончаний приписывал словам переводимой фразы необходимую грамматическую информацию, в наличии был блок снятия омонимии, в блоке синтеза русской фразы предусматривалось указание на правильный порядок слов [Беляева, Откупщикова 1996].

К середине 1960-х исследовательские группы по машинному переводу были основаны по всему миру: Европа (Венгрия, Чехословакия, Германия, Франция и т.д.), Китай, Мексика, Япония. В большинстве своем они оказались не очень долгоживущими. Главная причина такой недолговечности состояла в том, что на раннем этапе развития идей автоматического перевода недооценивалась сложность естественного языка. Первые десять лет развития данной области дали достаточно оснований для осознания этих заблуждений. [Шаляпина 1996].

В 1959 г. философ и математик Йошуа Бар-Хиллел выступил с утверждением, что высококачественный полностью автоматический перевод не может быть достигнут в принципе, потому что существует многозначность, которую нельзя снять простым обращением к энциклопедическим знаниям о мире. Поэтому полная автоматизация перевода, по мнению Й. Бар-Хиллела, является утопией [Беляева, Откупщикова 1996]. Это выступление самым неблагоприятным образом отразилось на развитии машинного перевода в США.

В 1966 г. Национальной Академией наук была специально создана комиссия ALPAC (Automatic Language Processing Advisory Committee). Основываясь, в том числе и на выводах Й. Бар-Хиллела, данная комиссия пришла к заключению, что машинный перевод медленнее, менее точен и в два раза дороже, нежели перевод, сделанный человеком, и по этой причине он нерентабелен. За докладом ALPAC последовал этап затишья [Hutchins 2011].

Несмотря на то, что доклад комиссии ALPAC был предвзят и недальновиден, он оказал существенное влияние на развитие машинного перевода как в США, так и в других странах. Однако исследования

продолжались в Канаде, Франции, Германии и других развитых странах [Hutchins 2011].

Особого упоминания заслуживает работа в этой области советских лингвистов, таких, как И. А. Мельчук и Ю. Д. Апресян (Москва), результатом которой стал лингвистический процессор ЭТАП. Основой для него стала лингвистическая теория И. А. Мельчука «Смысл<=>Текст» [Мельчук 1999], которая рассматривает язык как многоуровневую модель преобразований смысла в текст и обратно на основе использования синтаксиса зависимостей и Толково-комбинаторного словаря. В 1960 г. в составе Научно-исследовательского института математики и механики в Ленинграде была организована экспериментальная лаборатория машинного перевода, преобразованная затем в лабораторию математической лингвистики Ленинградского государственного университета.

Наибольшее развитие получили узконаправленные системы машинного перевода. Советский Союз и США сконцентрировались на русско-английских и англо-русских переводчиках научных и технических документов, которые на выходе давали грубый перевод, передающий лишь общий смысл документа. В Комиссии по атомной энергии США для получения «черновых» переводов на английский язык работ русских исследователей уже с 1964 г. стала применяться система GAT, а в Отделе зарубежных технологий ВВС США — разработанная фирмой IBM система Mark-II. В 70-х гг. обе системы были заменены созданной на базе GAT системой SYSTRAN [Шаляпина 1996].

Продолжались разработки в сфере словарного поиска, а также тех видов перевода, где он играет определяющую роль: начиная с «черновых» переводов, используемых для предварительного просмотра текста (определения его тематической принадлежности, степени его релевантности для нужд пользователя) и кончая переводами таких специфических текстов, как, например, анкетные данные или инвентарные списки. Рос спрос на системы автоматизированного перевода, т.е. перевода текстов человеком с использованием компьютерных технологий.

«Ренессанс» машинного перевода связан не только с развитием вычислительной техники в конце 70-х гг. (появление микрокомпьютеров, увеличение ресурсов памяти), но и с реалистичным взглядом на проблему. Исследователи теперь ставили целью развитие качественных черновых машинных переводов, предполагавших впоследствии участие человека. Системы машинного перевода становятся незаменимыми помощниками людей-переводчиков, способствующим экономии времени и человеческих ресурсов [Филинов 2002].

Доступность микрокомпьютеров и программного обеспечения для работы с текстом позволили появиться на рынке более дешёвым системам машинного перевода. На мировую арену выходит Япония, создающая множество систем для перевода с английского на японский и с японского на английский. Системы покупают крупные коммерческие компании, такие как Sharp, Oki, Mitsubishi, Sanyo [Hutchins 1986].

Руководитель японской государственной программы по машинному переводу профессор Макото Нагао из университета Киото опубликовал в 1984 г. статью, в которой предложил новый взгляд на системы. Он назвал свой подход «Example based translation» (перевод, основанный на примерах), а традиционный подход – как «Rule based translation» (перевод, основанный на правилах) [Белоногов 1999]. Идеи М. Нагао в дальнейшем послужат основой статистического метода перевода.

1974 г. – Москва, в институте ИНФОРМ-ЭЛЕКТРО создаются системы ЭТАП-1 (французско-русская), ЭТАП-2 (англо-русская), в ВЦП (Всесоюзный центр переводов) – АМПАР (англо-русская), НЕРПА (немецко-русская), ФРАП (французско-русская).

Система ФРАП представляет собой интересную попытку моделирования действий переводчика, для которого переводимый язык не является родным, а переводимый текст содержит элементы неизвестного (на данном этапе перевода). Научно-исследовательской целью при этом является выяснение действенности чисто грамматического подхода. В связи с этим процесс перевода разбит на большое количество блоков, каждый из которых,

соответствует определённому уровню анализа переводимой фразы, и лишь после использования всех формально-грамматических средств процедура перевода «предлагает» включение семантических уровней анализа, которых так же, как и грамматических, предусмотрено несколько [Беляева, Откупщикова 1996].

С 1975 г. в лаборатории инженерной лингвистики ЛГПИ им. А. И. Герцена разрабатывается система СИЛОД, она создаётся как многоязычная система с мощным словарным и морфологическим обеспечением и грамматикой, использующей трансфер. В этот период были созданы и экспериментально опробованы версии англо-русского и французско-русского машинного перевода, заложены основы фреймового и тезаурусного анализа, позволяющего распознавать структуру и значение отдельных синтаксических конструкций [Беляева, Откупщикова 1996].

1976 г. в Монреале (Канада) создаётся полностью автоматическая система TAUM-METEO – система перевода текстов метеосводок с английского на французский [Леонтьева, 2006].

Продолжаются исследования по поиску более продвинутых методов машинного перевода. 1980-ые проходят под знаком «непрямого» подхода: промежуточные представления, иногда по типу интерлингвы, включающие семантический, морфологический и синтаксический анализ, а иногда нелингвистические базы данных. Одними из самых выдающих проектов 80-х стали GETA-Ariane (Grenoble), SUSY (Saarbrücken), Mu (Kyoto), Rosetta (Eindhoven), а также международный многоязычный проект Eurotra.

В 90-е годы развитие систем машинного перевода перестало быть гомогенным. До 1989 года исследователи концентрировались на анализе синтаксиса и морфологии и других языковых закономерностей и правил. На смену им пришёл новый подход - подход, основанный на корпусах.

Мощным импульсом для развития статистического подхода стала публикация группы исследователей компании IBM во главе с Питером Брауном [Brown et al., 1993]. В ней разработчики представили пять моделей перевода и ввели понятие пословного выравнивания (word-to-word

alignement). Это позволило им свести к минимуму использование лингвистической информации. Успех разработанной системы задал учёным новый вектор исследований.

В 90-е годы произошло бурное развитие рынка ПК (от настольных до карманных) и информационных технологий, расширилось использование интернета, который становится всё более интернациональным и многоязычным. Всё это сделало возможным, а главное востребованным, дальнейшее развитие машинного перевода. Появление дисплеев, ПК, сканнеров, рост мощности персональных компьютеров – всё это изменило подход к теоретико-лингвистическим исследованиям и перенесло их в практическую плоскость [Беляева, Откупщикова 1996].

Работа учёных университета Южной Калифорнии, посвящённая машинному переводу, стала новым витком развития статистического подхода. В ней авторы доказали, что модели перевода, основанные на словах, уступают моделям, включающим в себя также вероятность перевода одного словосочетания другим. Такие модели получили название фразовых или основанных на сочетаниях (phrase-based translation) [Koehn et al. 2003]. На данный момент именно фразовые модели перевода являются доминирующими среди систем, основанных на статистике.

Несмотря на это. продолжал своё развитие и подход, основанный на правилах. В июле 1990 года на выставке PC Forum в Москве была представлена первая в России коммерческая система машинного перевода под названием PROMT (PROgrammer’s Machine Translation). В 1991 г. было создано ЗАО «ПРОект МТ», и уже в 1992 г. компания «ПРОМТ», будучи единственной неамериканской фирмой, выиграла конкурс NASA на поставку машинного переводчика.

В 1992 г. «ПРОМТ» выпускает целое семейство систем под новым названием STYLUS для перевода с английского, немецкого, французского, итальянского и испанского языков на русский и с русского на английский, а в 1993 г. на базе STYLUS создается первая в мире система машинного перевода для Windows. Тенденции к развитию 2 направлений машинного перевода

сохранили и в наши дни. Системы машинного перевода стали активно продаваться не только крупным компаниям и корпорациям, но и обычным пользователям персональных компьютеров. Многие системы стали доступны онлайн, т.е. больше не требуют установки специального программного обеспечения на компьютер.

В России существуют и развиваются системы с полноценным лингвистическим обеспечением семейства ЭТАП. Начато создание нескольких «молодых» систем (с татарского и турецкого языков, по которым появляются робкие заявления и публикации) [Леонтьева 2006].

Большие надежды возлагаются на мощные лексические и терминологические базы данных и базы знаний. Появляется отдельный класс систем, основанных на знаниях (knowledge-based MT, или KBMT systems). Появляются концептуальные структуры, концептуальные сети (часто мало отличающиеся от синтаксических) [Леонтьева 2006]. Однако одним из самых перспективных подходов на данный момент кажется нейронный машиннный перевод или перевод, основанный на нейронных сетях (Neural Machine Translation, NMT).

Самые первые разработки в этом направлении датируются 2013 годом [Kalchbrenner, Blunsom 2013], а уже через три года компания Google анонсирует новую технологию - Google Neural Machine Translation - и вводит её для английского, испанского, китайского, корейского, немецкого, португальского, турецкого, французского и японского языков [Wu et al.,2016]. В отличие от систем, переводящих по словам или словосочетаниям, нейронный перевод предполагает наличие так называемого кодировщика или энкодера (encoder), который превращает предложение-источник в вектор, и декодеровщика или декодера (decoder), который преобразует данный вектор в целевое предложение [Sutskever et al. 2014; Cho et al. 2014]. Как компания Google, так и другие фирмы, осваивающие новый подход к машинному переводу (например, латвийская компания Tilde, которая опробовала новую технологию на латышском и эстонском языках [Tilde neural machine translation]), утверждают, что качество перевода превосходит все предыдущие

подходы. Однако стоит оговориться, что для того, чтобы нейронная сеть обучилась действительно качественно, необходим колоссальный объём параллельных корпусов.

Гибридные технологии тоже представляют интерес в современном мире, однако многие разработки на данный момент защищены коммерческой тайной, а в академической среде гибридный перевод изучается крайне редко [Николаев и др. 2016]. В качестве примера можно упомянуть проект HyghTra (Hybrid High Quality Translation System), которым занимается университет Лидс [Costa-jussa et al. 2013]. Предполагается, что гибридный перевод объединяет сильные стороны систем, основанных на правилах, и статистических систем.

Машинный перевод является одной из десятка тем, которыми в России занимается междисциплинарный семинар ДИАЛОГ, возобновлённый в 1995 г. Это самое представительное российское мероприятие, целиком посвященное компьютерной лингвистике и ее приложениям, собирающее каждый год большое число ведущих специалистов в области интеллектуальных языковых технологий из компьютерных фирм, вузов и научных институтов со всей России и из-за рубежа [Всеволодова, 2007]. В других странах также ежегодно проходят связанные с компьютерной лингвистикой конференции, освещающие в том числе современное состояние систем машинного перевода. [Hutchins, 1986].

1.2. Виды машинного перевода

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

- По количеству языков (бинарные, расчитанные на одну пару, и многоязычные, расчитанные на работу с несколькими языками);

- По направленности (однонаправленные и многонаправленные, если целевой язык и язык-источник могут меняться местами в зависимости от требований пользователя);

- По степени автоматизации (автоматизированный перевод, где основную работу проводит человек, а система привлекается при необходимости, и машинный перевод, где процесс перевода почти полностью реализуется системой).

Основным делением систем машинного перевода, однако, считается следующее: традиционная, на основе правил (Rule-based Machine Translation RBMT), и статистическая (Statistical Machine Translation, SMT). Кроме того, есть технологии, совмещающие оба этих подхода, которые зовутся гибридными (Hybrid Machine Translation, HMT).

1.2.1. Перевод, основанный на правилах

Хронологически самым ранним является машинный перевод на основе правил. Первые RBMT системы были первыми системами машинного перевода в принципе. Разработанные еще в начале 1970-х, они оказали огромное влияние на развитие машинного перевода. Некоторые из них функционируют до сих пор (SYSTRAN), хоть и в измененном виде. Среди современных известных RBMT систем можно выделить испанскую платформу машинного перевода Apertium, датско-норвежскую Gram-Trans и некоторые системы российской компании PROMT.

RBMT системы состоят из больших двуязычных словарей и тщательно разработанных грамматик, описывающих основные морфологические, семантические и синтаксические правила входного и выходного языков. Используя всю имеющуюся информацию, текст последовательно переводится с одного языка на другой. Система построена по принципу соотнесения структуры предложения на первом языке со структурой предложения на втором.

Существует три типа систем перевода на основе правил [Щипицина 2013]:

Системы пословного перевода (Direct Systems, Dictionary-based Systems) функционируют как простой двуязычный словарь: последовательно преобразовывают слова из входного языка на выходной с использованием

самых базовых грамматических правил. Данный тип является, пожалуй, одним из самых ранних и бесхитростных подходов к машинному переводу. Он не дает удовлетворительных результатов, если применять его на длинных предложениях и текстах. Однако как бюджетный вариант может хорошо подходить для перевода списков слов, перечней товаров или простых каталогов продуктов и услуг. Также, исследования показали, что пословный перевод является самым качественным методом машинного перевода по правилам, если речь идёт об очень близких языках [Hajič et al. 2000].

Интерлингвистические системы (Interlingual Systems) трансформируют входной текст в язык-посредник - интерлингву (абстрактное представление, независимое от языка), и текст на выходе генерируется из интерлингвы. Сейчас такие системы практически не используются, однако несколько десятков лет назад на них возлагали большие надежды. Так, в качестве языка-посредника пытались использовать UNL (Universal Networkong Language, универсальный сетевой язык) - искусственный семантико-синтаксический язык, разработанный для хранения семантических данных и связей между ними, выделенных из естественного языка [Uchida 1996]. Основное преимущество такого подхода состоит в том, что с помощью интерлингвы можно работать с парами совершенно не родственных языков, а поэтому процесс добавления нового языка (многоязычный перевод) заметно проще. Недостатком является зависимость от специфической области, которую практически невозможно расширить для перевода каких-либо более общих текстов.

Трансферные системы (Transfer-based Systems) подвергают анализу предложение на входе, и применённые правила выступают как соответствие между структурой этого предложения и того, которое получится на выходе. Современные переводчики, основанные на правилах, чаще всего являются именно трансферными [Николаев и др. 2016]. Сначала из текста на входе выделяют входящие в него слова и другие элементы, которые принято называть токенами. Затем каждый токен анализируется морфологически, то есть определяется его роль в предложении и, если токен является

словоформой, то часть речи и грамматические характеристики. Далее на основе морфологических данных система анализирует синтаксическую структуру предложения. С его помощью, а также с использованием двуязычных словарей и грамматических правил, генерируется новое предложение на целевом языке. При такой технике можно получить довольно качественный текст на выходе, хотя многое зависит от языковых пар и особенностей текста. Для улучшения качества перевода часто используют базу памяти перевода, то есть некоторые заранее сохранённые сегменты текста и их перевод.

В основании трансферных и интерлингвистических систем лежит одна и та же мысль: чтобы произвести перевод, в первую очередь нужно создать промежуточное представление, отражающее смысл исходного предложения. А уже из этого представления можно генерировать перевод предложения на целевой язык. Различие заключается в том, что в интерлингвистических системах промежуточное представление должно быть независимо от языковых пар, в то время как трансферные системы как раз опираются на некоторые особенности входного и выходного языков.

1.2.2. Статистический перевод

Хотя первые идеи статистического машинного перевода были представлены Уорреном Уивером еще в середине 1949 года, серьезные разработки начались лишь в 80-90-х в исследовательском центре IBM [Koehn 2010]. Благодаря этому интерес к SMT сильно вырос в конце XX века, что сделало его самым широко изучаемым и используемым методом машинного перевода. Одной из самых известных в мире систем, использующих статистический метод, до недавнего времени являлась Google Translate, которая сейчас постепенно переходит на нейронный перевод. На российском рынке статистический подход наиболее ярко представлен компанией Яндекс.

SMT основан на сравнении больших объемов языковых пар. Работа статистического машинного перевода сводится к поиску наиболее вероятного перевода предложения с использованием данных, полученных из двуязычных

корпусов текстов. Текст переводится согласно распределению вероятностей того, что предложение *y* целевого языка является переводом предложения *x* исходного языка.

У каждого из вариантов переводов есть своя вероятность быть переводом исходного предложения. Можно записать эти вероятности как *P(y1|x), P(y2|x), P(y3|x)* и т.д. Из всех этих предложений y_n компьютер должен выбрать самое вероятное, при условии, что у нас есть предложение *x*. Чтобы посчитать эти вероятности, используется формула Байеса, которая лежит в основе статистического машинного перевода.

Совместная вероятность зависящих друг от друга событий *A* и *B* равна вероятности события *A*, умноженная на вероятность события *B* при условии события *A*, поделенная на вероятность события *B*.

«Переведём» это на язык предложений. Допустим, у нас есть предложение *x*, которое надо перевести на другой язык. Задача статистического машинного перевода сводится к нахождению такого предложения *y*, которое с наибольшей вероятностью является переводом *x*. Иными словами, ищется такой *y*, при котором вероятность *p(y|x)* принимает максимальное значение.

Стандартная современная система перевода работает следующим образом. Входной текст разбивается декодером на так называемые фразы, то есть последовательности из нескольких слов. Затем, каждая фраза независимо переводится на целевой язык. Наконец, полученные кусочки упорядочиваются, так как в языке-источнике и в целевом языке порядок слов может быть разным.

Для того, чтобы выбрать подходящий перевод, необходима фразовая таблица переводов, где каждой фразе будут сопоставлены возможные переводы и их вероятности. Чаще всего используются именно фразовые, а не пословные таблицы, так как это позволяет переводить слова, исходя из контекста, что решает проблему неоднозначности [Молчанов 2013]. Такая таблица зовётся моделью перевода (translation model). Для построения модели перевода разработчику необходим большой объём параллельных

текстов или «битекстов», то есть пар текстов на языке-источнике и целовом языке, где один из текстов является переводом другого [Harris 1988]. Параллельные тексты обычно получают путём выравнивания двух текстов с помощью специальных программ или алгоритмов, таких как, например, Hunalign [Varga et al. 2005].

На конечном этапе алгоритму необходимо определить, насколько вероятно присутствие такого предложения в языке перевода. Для этого для гипотезы перевода считается оценка по языковой модели. Наиболее общепринятый подход для моделирования языка – использование модели языка, построенной по N-граммам [Koehn 2010]. Модель языка (или языковая модель) – набор правил и связей между составляющими языка, используемых для описания строения языка. Несколько лет назад считалось, что модель языка работает тем лучше, чем больший объём тренировочных данных имеется у неё в распоряжении. Более современные модели ориентированы скорее на качество тренировочных данных, а не на их количество [Scwenk 2012].

1.2.3. Гибридный перевод

И статистическим и правиловым подходом разработчики занимаются уже не один десяток лет, однако каждый из подходов обладает существенным рядом недостатков.

Перевод по методу правил более стабилен, чем статистический перевод. Технология перевода статистической системой построена на анализе параллельных текстов, и программа может менять результат перевода в зависимости от контекста анализируемых баз. В итоге перевод одного и того же термина может быть разным, так как система вычисляет наибольшее количество совпадений перевода данного слова в определенном контексте. В отличие от статистического метода, технология перевода по правилам гарантирует одинаковый перевод одного и того же термина. Слабым местом статистических систем является отсутствие механизма анализа грамматических правил входного и выходного языков. Маловероятно, что

система, которая не анализирует текст с точки зрения грамматики, способна выдать связный перевод. Очевидно также, что причина ошибок при переводе статистическим методом заключается в недостаточном объеме параллельных баз текстов из различных областей знаний. В традиционных системах перевода, работающих на основе правил, есть возможность настройки программы для перевода текстов со специализированной терминологией. В частности, путём подключения тематических словарей, создания и редактирования собственных словарей и т. д. [Андреева 2013].

Предполагается, что именно гибридный машинный перевод должен сгладить недостатки каждого из подходов. Постепенно многие известные компании внедряют гибридные технологии для разработки собственных систем. Среди самых известных: PROMT и SYSTRAN.

Объединение двух подходов к переводу может осуществляться разными способами.

Первый способ — это добавить статистический модуль в уже готовую систему перевода, основанную на правилах. Именно так поступает компания ПРОМТ, развивая технологию DeepHybrid. RBMT-система дополнена двумя компонентами: модулем статистического постредактирования и модулем языковых моделей. Статистическое постредактирование позволяет сгладить перевод по правилам, приближая его к естественному языку и при этом сохраняя четкую структуру синтезируемого текста. Языковые модели используются для оценки гладкости и грамматической правильности вариантов перевода, порождаемых гибридной системой [Николаев и др., 2016].

Второй способ — воспользоваться правилами для улучшения статистического подхода. Применить какое-либо правило можно до осуществление перевода как такового, так и после. Например, статистические n-граммные модели плохо справлялись с переводом с японского языка на английский, так как в японском языке сказуемое обычно находится в конце предложения. Качество перевода можно заметно улучшить, если до осуществления перевода переставить сказуемое в японском предложении

сразу после подлежащего, тем самым делая порядок слов схожим с тем, что в английском языке. Другим примером может служить перевод с чешского на английский. В одном из исследований перевод улучшался, когда слова чешского предложения вначале приводили к нормальной форме [Goldwater, McClosky 2005]. С другой стороны, во многих ситуациях уместно делать именно постпроверку правилами.

На данный момент гибридные технологии ещё не стали атрибутом большинства систем машинного перевода, но именно они помогут существенно улучшить качество перевода тех пар языков, для которых сложно собрать большой объём параллельных корпусов.

1.3 Оценка машинного перевода

Несмотря на проработанность и логичность каждого из подходов, перевод, тем не менее, нередко получается с ошибками. Самый надёжный способ понять, есть ли ошибка в переводе или нет, – это оценить перевод. Для быстрой оценки большого объёма данных используют автоматические метрики, и самой известной из них является метрика BLEU.

Основную идею метрики можно выразить словами её создателей: «Чем ближе машинный перевод к профессиональному переводу человека, тем он лучше» [Papinelli et al. 2002]. Для получения результатов метрике необходимы переведённый машиной текст и один или несколько эталонных переводов, сделанных человеком. Оценка может ранжироваться от 0 до 100%, где 100% - это полное совпадение перевода с эталоном.

Несмотря на то, что метрика BLEU продолжает оставаться одной из самых распространённых для автоматической оценки перевода, она имеет ряд существенных недостатков. Метрика основана на сравнении n-грамм, поэтому для языков со свободным порядком слов она даёт оценку ниже, чем на самом деле полагается. Таким образом, даже перевод, сделанный человеком, может получить низкую оценку только из-за того, что слова в предложении он поменял местами. Также, метрика BLEU занижает результаты для языковых пар, где целевым языком является язык

флективный. Если переводчик в результате неверно согласовал слова, то оценка будет крайне низкой, несмотря на то, что смысл и даже лексемы были переданы правильно. Так, самая высокая оценка, полученная при сравнении популярных автоматических переводчиков на русский, была 15% [Браславский и др. 2013]. Лингвист Давид Коловратник в своей работе предлагает не опираться исключительно на автоматическую метрику, но и производить ручную оценку перевода: выделить основные ошибки и подсчитать, каков их процент от общего количества ошибок, высчитать процент успешно переведённых с точки зрения человека предложений и т.д. [Kolovratnik et al. 2009].

Выводы к главе 1

Машинный перевод как область науки прошел длинный путь от первых примитивных систем середины XX века до мощных, быстрых и экономичных автоматических переводчиков современности. Первые идеи использования машины как помощника в процессе перевода зародились еще в XVII веке. В 1930-х несколько ученых делали попытки по изобретению машин-переводчиков, однако их достижения остались практически незамеченными.

Середина XX века считается временем появления такой области лингвистики, как машинный перевод. Несколько удачных экспериментов стали причиной повышенного интереса к данной области, что привело к финансированию масштабных проектов по автоматизации перевода. Все системы того времени были довольно примитивными, сейчас мы могли бы отнести их к стандартному RBMT.

Десятилетие больших надежд сменилось десятилетием больших разочарований: машинный перевод не оправдал себя ни по качеству, ни по рентабельности. Во многих странах работа остановилась совсем. Интерес к машинному переводу возобновился лишь с появлением в 1980-х годах новой техники, т.е. новых возможностей.

Ближе к концу XX века разработки в области машинного перевода ведутся по всему миру. Исследователи предлагают новые подходы, появляется интерес к переводу речи, в проекты включаются все больше различных языков. Системы машинного перевода становятся популярными среди больших компаний, но одновременно и доступными для широкой аудитории. На данный момент зарождается абсолютно новый подход, основанный на глубинном обучении и нейронных сетях.

В результате более полувековых исследований и разработок современный машинный перевод представляется нам в основном в виде трех подходов: машинный перевод на основе правил (RBMT), статистический машинный перевод (SMT) и гибридный машинный перевод (HMT)

Системы с хорошим словарем, подробной грамматикой, семантическим анализатором и парсером, анализирующие предложения на входе и генерирующие их на выходе, представляют собой классические RBMT-системы. Те, что практически не включают в себя лингвистические правила, а натренировывают машину «угадывать» верный перевод на большом корпус параллельных текстов, являются более современными SMT-системами. И наконец, системы, объединяющие техники обоих подходов в том или ином сочетании, называются гибридным машинным переводом (HMT).

Каждый подход имеет свои достоинства, однако в последнее время продвигается идея о том, что HMT должен стать самым совершенным из них: объединить преимущества и устранить недостатки RBMT и SMT систем.

Для оценки качества переводов часто используют автоматичские метрики, например, BLEU, однако для флективных языков со свободным порядком слов результаты оценки метрики не всегда соответствуют действительности.

Глава 2. Язык эсперанто и его компьютерная обработка

2.1. Эсперанто в системе искусственных языков.

Традиционно языки делятся на две подгруппы: искусственные и естественные языки. Естественным языком в лингвистике называется язык, используемый для общения людей и не созданный целенаправленно. Естественный язык может существовать в письменной и бесписьменной формах.

В отличие от естественного языка, искусственный язык всегда является языком, созданным одним человеком или группой людей. Искусственные языки также называются конлангами, от английского «constructed languages» [Пиперски 2017]. Чаще всего искусственный язык имеет подробное описание лексики, произносительной нормы (если таковая требуется) и синтаксиса, поскольку освоение искусственного языка обычно происходит уже во взрослом возрасте и не может произойти само собой на интуитивном уровне. Искусственные языки создаются и применяются в тех областях, где применение естественного языка менее эффективно или невозможно.

Искусственные языки можно классифицировать по двум критериям: по назначению и по степени сходства с естественными языками. По назначению языки делятся на философские, вспомогательные и художественные, а по степени сходства — на априорные и апостериорные языки. [Trask et. al. 2007] Некоторые лингвисты также выделяют смешанные языке, т.е. языки, объединяющие априорные и апостериорные черты.

Первый тип классификации основывается на цели создания конланга. Если задача лингвиста — проверить какую-либо лингвистическую теорию посредством искусственно созданного языка, то конечный продукт его лингвоконструирования попадёт в категорию философских языков. Одним из самых известных языков такого типа является логлан, созданный Джеймсом

Куком Брауном для экспериментальной проверки теории лингвистической относительности Сэпира-Уорфа [Brown 1999], а также его «последователь» — язык ложбан. Часто подобные искусственные языки сложны для освоения и восприятия, в отличие от вспомогательных языков, задача которых — упростить общение международное общение. Одним из первых таких языков считается Lingua Ignota. Он был описан ещё в XII веке монахиней Хильдегардой Бингенской [Okrent 2010]. Если же смысл создания языка — это лишь эстетическое удовольствие или артистические нужды, то он попадает в третью категорию — категорию художественных языков. Лингвоконструированием ради эстетических целей прославился английский писатель и лингвист Джон Толкин, спроектировавший несколько языков для различных рас персонажей.

Другой тип классификации опирается на сходство с естественными языками. Если искусственный язык заимствует из одного или нескольких естественных языков грамматику, лексику и синтаксис, то он называется апостериорным языком. Яркий пример апостериорного языка — Basic English, созданный в 1925-м году британским лингвистом Чарльзом Огденом на основе английского языка [Ogden 1932]. Если же элементы искусственного языка созданы на какой-либо другой основе, то язык попадает в категорию априорных языков. Примером априорного языка является клингонский язык, придуманный лингвистом Марком Окрандом специально для американского сериала «Звёздный путь» [Okrand 1992] . Клингонский язык предназначался для расы пришельцев-клингонов, поэтому автор специально стремился сделать его непохожим ни на один уже существующий человеческий язык.

Эсперанто можно назвать вспомогательным апостериорным языком. Также, эсперанто часто называют также плановым языком, так как он широко распространился по всему миру. Так, Сергей Николаевич Кузнецов предлагает называть плановыми те искусственные международные языки, что вышли за рамки лингвопроекта и реализовались как полноценные языки в обществе (т.е. социализировались) [Кузнецов 1991].

2.2. История создания эсперанто

Автором языка эсперанто является Людвиг Заменгоф, польский окулист. Заменгоф родился в 1859 году в Белостоке в еврейской семье. В те времена Польша входила в состав Российской империи, поэтому город был населён людьми разных национальностей — русскими, поляками, немцами, евреями, — которые едва находили общий язык друг с другом. Заменгоф был убеждён, что людям в его городе мешает поладить лишь отсутствие общего языка, поэтому с самого детства он придумывал универсальный язык, но его попытки терпели неудачу.

Однажды, проходя по улице, он обратил внимание на вывески «Швейцарская» и «Кондитерская». Заменгоф осознал, что буквосочетание «ская» является продуктивным словообразовательным элементом русского языка, с помощью которого можно создавать новые слова, значения которых получаются из значения корня и значения этого самого элемента. Это навело его на мысль добавить в свой язык развитую систему аффиксации [Okrent 2010].

Заменгоф ещё в школе начал изучать английский язык. По сравнению с русским языком и его сложностями с согласованием родов, падежей и лиц английский стал для Заменгофа приятной неожиданностью. Это способствовало ещё большему упрощению будущего языка эсперанто.

В 1887-м году Заменгоф опубликовал первую книгу про его язык под названием «Международный языкъ». Заменгоф был свято уверен, что международный язык, как и любой национальный, является общественной собственностью, поэтому не считал себя в праве как-то дальше влиять или видоизменять язык. Более того, книгу он подписал псевдонимом Dr. Esperanto, что в переводе означает «надеющийся». Вскоре псевдоним автора распространился и на название нового языка.

Книга включала в себя 16 правил грамматики и около 900 слов. Поначалу книга рассказывала о корневых словах, затем их сложность возрастала, но основные принципы языка соблюдались: значение каждого

нового слово получалось из значения его корня и значения аффикса. Грамматическую категорию слово получало за счёт окончания. Также, в книге было небольшое количество двуязычных примеров и несколько текстов на эсперанто без перевода. Читателям предлагалось послать своим друзьям письмо на эсперанто и перевод основных корней, и любители загадок и головоломок разослали письма на эсперанто по всему миру.

Пиком популярности эсперанто можно назвать первую половину XX века. Первый эсперанто-конгресс, насчитывавший 688 человек из 20 стран, состоялся в 1905 году во французском городе Булонь-сюр-Мер. Именно там было решено, что основным нормативным документом языка эсперанто будет опубликованная в том же году книга Заменгофа «Основы эсперанто» [Zamenhof 1905]. Подобное решение было необходимо, так как сам создатель отказался от каких-либо авторских прав на язык эсперанто, а отсутствие закреплённых нормой правил неизбежно привело бы к различным трактовкам и разночтениям, а затем — и к расслоению языка на диалекты. После окончания Первой Мировой войны Лига Наций рассматривала вопрос о принятии эсперанто в качестве нейтрального международного языка, но из-за возражений Франции, считающей, что французский уже и так почти является универсальным языком, предложение было отклонено.

После Второй Мировой войны начал быстро набирать популярность английский язык. Эсперанто, не имевший ни страны, ни капитала, был не в состоянии конкурировать за право стать универсальным языком, и эсперанто-движение пошло на спад.

На сегодняшний день эсперанто насчитывает от 300 тысяч до 2 миллионов бегло говорящих людей и около 1000 человек, для которых этот язык является родным, что делает эсперанто уникальным среди других конлангов [Corsetti et al. 2004]. Каждый год издаются сотни книг на эсперанто и около 250 журналов, самым известным из которых считается Monato. На эсперанто пишется музыка и даже снимаются фильмы. Ежегодно Всемирная Эсперанто-Ассоциация устраивает съезды, на которые съезжаются все желающие пообщаться или выступить с лекцией.

2.3. Описание языка эсперанто

2.3.1. Алфавит и фонетика

Алфавит эсперанто построен на основе латинского. В алфавите 28 букв: A, B, C, Ĉ, D, E, F, G, Ĝ, H, Ĥ, I, J, Ĵ, K, L, M, N, O, P, R, S, Ŝ, T, U, Ŭ, V, Z, которые соответствуют 28 звукам — пяти гласным, двум полугласным и 21 согласному. Одна буква соответствует одному звуку [Заменгоф 1887].

Своего рода недостатком для современного человека являются 6 букв с диакритическими символами. 5 согласных букв имеют над собой циркумфлекс, а одна гласная — бреве. В стандартном наборе языков для операционной системы Windows эсперанто отсутствует, поэтому это может затруднять набор текста. Однако пользующиеся интернетом эсперантисты часто заменяют букву с диакритикой на её сочетание с буквой «x» (вместо ĵ пишут jx). Это не мешает прочтению, так как буквы «x» в эсперанто нет.

Также, можно использовать клавишу Alt и цифры (на цифровой клавиатуре). Сначала пишут соответствующую букву (например C для Ĉ), затем нажимают на клавишу Alt и набирают 770, и над буквой появляется циркумфлекс. Если же набрать 774, то появится бреве для ŭ.

В эсперанто каждое слово читается так, как оно пишется. Одна буква соответствует ровно одному звуку, и одному звуку соответствует ровно одна буква. Ударение всегда находится на предпоследнем слоге.

2.3.2. Морфология и словообразование

Эсперанто относится к агглютинативным языкам. Носители языка активно используют префиксацию и суффиксацию для образования новых слов и оттенков значений; все аффиксы имеют ровно одно фиксированное значение. Кроме того, язык позволяет образовывать сложные слова путём простого слияния слов.

В эсперанто отсутствуют некоторые грамматические категории, существующие в русском языке. Существительное обладает категорией числа (единственное, множественное), категорией падежа (именительный и

винительный, остальные падежи выражаются предлогами). Категории одушевлённости и грамматического рода в эсперанто нет. К существительным, о которых уже шла речь в тексте, прибавляется определённый артикль.

Прилагательные не обладают синтетической формой степени сравнения и не имеют краткого аналога. Как и существительные, они могут изменяться по числам и падежам.

Глагол в эсперанто обладает категорией наклонения (условное, изъявительное, повелительное), времени (прошедшее, настоящее, будущее), залога (активный, страдательный и средний) и переходностью. Категории вида, лица, рода и числа в языке не используются.

Наречия делятся на непроизводные и производные. Производные получаются путём добавления к корню окончания «е».

Личные местоимения в именительном падеже не имеют окончаний. В отличие от русского языка, в эсперанто присутствует неопределённо-личное местоимение (oni) и местоимение, обозначающее неодушевлённые предметы. Также, эсперанто, как и английский язык, не различает второе лицо по числам.

В эсперанто широко распространены причастия и деепричастия, каждое из которых может быть действительного или страдательного залога и относится к настоящему, прошедшему или будущему времени. Для каждого вида причастий существуют свои суффиксы.

Знаменательные части речи характеризуются особыми грамматическими окончаниями:

Окончание	Соответствующая часть речи
o	имя существительное (кроме некоторых имён собственных)
a	имя прилагательное, притяжательное местоимение, порядковое числительное, причастие

j	множественное число существительных, прилагательных, притяжательных местоимений, порядковых числительных, причастий
n	винительный падеж
e	производное наречие, деепричастие
i	инфинитив
as	настоящее время глагола
is	прошедшее время глагола
os	будущее время глагола
u	повелительное наклонение глагола
us	условное наклонение глагола

Стоит отметить, что некоторые имена собственные могут не приобретать окончания существительного и остаться неизменёнными. Кроме того, окончания существительного или артикля могут быть заменены апострофом (напр. de l’ mondo вместо de la mondo).

2.3.3. Лексика

Базовый набор словаря эсперанто состоял из 900 корней и был опубликован автором в первой книге по этому языку [Заменгоф 1887]. Сейчас словарный состав, безусловно, расширился, но в языке эсперанто содержится много продуктивных словообразовательных моделей, поэтому новые корни он заимствует менее охотно, чем, например, русский язык.

В основном корни слов являются адаптированными под фонологию эсперанто заимствованиями из романских и германских языков корней, а также интернационализмами латинского и греческого происхождения. Некоторые славянские корни также попали в словарь, английский же язык на времена создания эсперанто не был всемирно известным, поэтому лексика оттуда представлена достаточно бедно.

Благодаря тому, что эсперанто сегодня является живым языком, в нём появляется и слова, порождённые самой системой и реалиями эсперанто-

сообщества. Например, в словарях уже зафиксированы глаголы «krokodili», означающий «говорить среди эсперантистов на национальном языке, обычно — на родном для всех говорящих», и «aligatori», означающий «говорить среди эсперантистов на языке, родном лишь для части присутствующих». [Колкер 2009]

2.3.4. Синтаксис

Главная особенность языка эсперанто — свободный порядок слов, что во многом роднит его с русским языком. Это не создаёт проблем с пониманием благодаря чёткой системе окончаний и большому количеству однозначных предлогов. Сам автор пишет, что «каждый предлог имеет определённое постоянное значение; если же нужно употребить предлог, а прямой смысл не указывает, какой именно, то употребляется предлог je, который самостоятельного значения не имеет».

Порядок слов в основном SVO (субъект — предикат — объект), перестановка слов может порождать смысловые оттенки.

Благодаря наличию безличного местоимения oni, соответствующему английскому «one» или немецкому «man», в эсперанто можно строить безличные предложения.

Отрицание в эсперанто создаётся с помощью частицы «ne», которая ставится перед тем словом, которое и необходимо отрицать. При другом отрицательном слове частица опускается, т.е. язык не допускает двойного отрицания.

Вопросительные предложения строятся с помощью специальных вопросительных слов, в том числе и общие вопросительные предложения, которые снабжаются специальном частицей. В устной речи она иногда опускается, и тогда вопрос передаётся исключительно интонацией.

2.4. Компьютерные ресурсы для работы с эсперанто

Несмотря на то, что эсперанто, в отличие от национальных языков, не получает поддержки и финансирования от какого-либо государства, в сети Интернет можно найти множество программ и сайтов для работы с языком: словари, автоматические переводчики, морфологические и синтаксические анализаторы. Стоит отметить, что некоторые программы были сделаны одиночными исследователями и энтузиастами. Во многих случаях это заканчивалось тем, что создатель прекращал техническую поддержку своего продукта, поэтому работать с ним на данный момент не представляется возможным.

В этом разделе мы рассмотрим самые эффективные ресурсы для работы с языком эсперанто.

2.4.1. Словари и переводчики

На данный момент самыми значимыми системами автоматического перевода с эсперанто на русский язык являются «Google Translate» и «Яндекс Переводчик», а наиболее полноценным эсперанто-русским словарём является большой словарь Бориса Кондратьева на сайте eoru.ru.

Переводчик Google Translate.

23 февраля 2012-го года исследователь Торстен Брэнтс в официальном блоге google-переводчика заявил, что Google Translate начал поддерживать и язык эсперанто, который тем самым стал 64-м языком для машинного перевода Google. Разработчик отметил, у языка эсперанто и у переводчика общие цели: позволить людям проще понимать друг друга. Также, разработчики опасались некачественного перевода из-за небольшого количества параллельных текстов, но к своему удивлению обнаружили, что Google переводит на достаточно хорошем уровне [News about Google Translate]. Например:

Табл. 2.1. Примеры перевода Google Translate

Эсперанто	Русский язык (машинный перевод)	Русский язык (перевод человека)
Li vidas la hundon	Он видит собаку	Он видит собаку
Ŝia edzo estas franco	Ее муж-французский	Её муж — француз
Originala prozo en Esperanto	Оригинальная проза Эсперанто	оригинальная проза на эсперанто

За последние несколько лет перевод Google с эсперанто на русский многократно улучшился, в том числе и за счёт глубинного обучения. Однако система машинного перевод Google по прежнему использует английский язык в качестве языка-посредника, и убедиться в этом не составляет труда. Рассмотрим несколько примеров:

Табл. 2.2. Пример перевода Google Translate

Эсперанто	Английский язык (машинный перевод с эсперанто)	Русский язык (машинный перевод с эсперанто)	Русский язык (перевод человека)
Ili al mi respondis	They answered I	Они ответили I	Они мне ответили

Наличие I в русском переводе доказывает тот факт, что Google Translate не переводит с эсперанто напрямую. Это порождает так называемые «эффект испорченного телефона», когда ошибки при втором переводе накладываются на ошибки при первом переводе, тем самым ещё более скрывая от пользователя смысл высказывания. Данный эффект можно заметить ещё на одном примере:

Табл. 2.3. Пример перевода Google Translate

Эсперанто	Английский язык (машинный перевод с эсперанто)	Русский язык (машинный перевод с эсперанто)	Русский язык (перевод человека)

Mia kamaradino estas bela	My companion is pretty	Мой спутник красивый	Моя подруга красивая

Конечно, нельзя сказать, что слово companion полностью отражает смысл слова kamaradino, в котором содержится суффикс «in» — показатель женского рода. Однако, рассмотрев перевод на английский, понять можно вполне, о чём идёт речь. Перевод на русский точно передал лишь местоимение.

Яндекс Переводчик

В отличие от своего основного конкурента, компания Яндекс занялась языком эсперанто несколькими годами позднее. 25 июля 2016 года в официальном блоге компании появилась новость о новых 11 языках переводчики, среди которых был и эсперанто [Яндекс. Блог Переводчика].

Для всех языковых пар Яндекс переводчик использует статистический машинный перевод, однако на русский язык перевод осуществляется напрямую. В случае с эсперанто это многократно улучшает качество перевода. Несмотря на то, что для качественного статистического машинного перевода необходимы большие объёмы параллельных текстов, которых на языке эсперанто не так уж и много, схожий порядок слов и система времён позволила Яндексу обучить модель перевода на меньшем количестве данных. Перевод получается не только понятным, но и чаще всего грамотным:

Табл. 2.4. Пример перевода Яндекс.Переводчика

Эсперанто	Русский язык (машинный перевод)	Русский язык (перевод человека)
Tago de Homaj Rajtoj	День прав человека	День прав человека
Ŝia edzo estas franco	Ее муж-француз	Её муж — француз
Ili al mi respondis	Они мне ответили	Они мне ответили

К сожалению, перевод Яндекса тоже нельзя назвать совершенным. Рассмотрим некоторые примеры с грубыми ошибками:

Табл. 2.5. Пример перевода Яндекс.Переводчика

Эсперанто	Русский язык (машинный перевод)	Русский язык (перевод человека)
Vi ne molestu Tom!	Вы не беспокоить Тома	(вы) не беспокойте Тома
Li sin sentis eluzata.	Он чувствовал eluzata.	Он чувствовал себя использованным
Originala prozo en Esperanto	Оригинальная проза на английском языке	Оригинальная проза на эсперанто

В первом примере система потеряла время глагола и остановилась на инфините. Из второго примера видно, что слово не нашлось во фразовой таблице, поэтому переводчик был вынужбен оставить его как етсь и переводить все остальные распознанные слова. Третий пример, пусть и выглядит немного комично, лишний раз доказывает то, что статистический машинный перевод строится автоматически, а посему может содержать самые неожиданные ошибки и искажения.

Словарь Eoru.ru

Eoru.ru — это электронный словарь, предназначенный для перевода эсперанто на русский и с русского на эсперанто. Название происходит от общепринятых доменов — русского(.ru) и эсперанто (.eo). В отличие, например, от ABBYY Lingvo, Eoru создавался группой энтузиастов во главе с Борисом Кондратьевым. За основу был взят PIV — Plena Ilustrita Vortaro de Esperanto или Полный иллюстрированный словарь эсперанто, выпущенный в 1960-х годах и до сих пор считающийся наиболее авторитетным словарём. Добавив туда новые слова и выражения, авторы выпустили первую электронную версию Eoru в 2009 году. Словарь пополняется и по сей день и содержит в себе 76815 слов в 39927 словарных статьях.

Сайт содержит четыре раздела: предисловие от автора, описание грамматики эсперанто, словарь и раздел с информацией о команде составителей. В предисловии подробно рассказывается о том, какие слова попали в eoru, а какие — нет, и как происходил процесс создания словаря.

Грамматический раздел даёт полное лингвистическое описание языка эсперанто с примерами и спорными случаями. Раздел о составителях содержит ссылки на всех авторов, даёт возможность высказать свои замечания и получить обратную связь, а также предлагает поддержать проект материально. Словарь представляет собой строку поиска.

Пример словарной статьи:

muŝ||o
1. мýха, мýшка, мóшка;
la ~oj zumas мýхи жужжáт;
li havas ~on en la kapo, ~o zumas en lia kapo *погов.* у негó не все дóма, он чóкнутый (*дословно* у негó мýха в головé (жужжúт));
kalkuli ~ojn *погов.* считáть мух, ротозéйничать, занимáться ерундóй;
bati du ~ojn per unu bato *погов.* прихлóпнуть двух мух одни́м удáром;

2. мáленькая боро́дка в ви́де небольшóго пучкá волóс под ни́жней губóй;

3. мýшка (*искусственная родинка на щеке или губе для придания лицу пикантности*); *ср.* belgrajno;

4. *maj*; *астр.* Мýха (*созвездие*);
~a муши́ный;
~ed/o/j *энт.* настоя́щие мýхи (*семейство*);
~et/o мáленькая мýха, мýшка, мóшка.

Как видно из этого примера, словарь различает значения, предлагает пользователям примеры использования и устойчивые словосочетания, а в конце предоставляет небольшую словообразовательную справку, что достаточно актуально для такого языка, как эсперанто.

Немного поработав со словарём, мы сделали следующие наблюдения:

- Словарь не исправляет опечатки пользователя. При неправильном вводе слова он не предлагает варианты правильного написания, что затрудняет работу;
- С лексикографической точки зрения словарь ближе к бумажному варианту, чем к электронному. В словарных статья часто используются сокращения, характерные для бумажных словарей, которые стремятся

экономить бумагу. В электронных словарях такие сокращения обычно неоправданы, так как они сложнее воспринимаются на глаз. Также, для электронного словаря у Eoru не так много примеров использования;

- Словарь не даёт подробного описания устойчивых словосочетаний. Например, в статье про слово «elefanto» (слон) можно найти поговорку «fari el muso elefanton», что в пословном переводе означает «делать из мыши слона». В то же время в статье «muso» (мышь) подобный фразеологизм не упоминается. Также, стоит упомянуть, что Eoru не поддерживает поиск по фразеологизмам и примерам словоупотребления (в отличие, например, от словарей на сайте Яндекс) и позволяет искать только одно слово.

2.4.2. Морфологические и синтаксические анализаторы

Автоматической обработкой языка эсперанто чаще всего занимаются не крупные компании или корпорации, а энтузиасты. Причиной этому служит то, что успешно работающая программа, скорее всего, не принесёт никакой экономической выгоды. Несмотря на это, стоит отметить некоторые удачные проекты в это области.

Один из них — алгоритм автоматической сегментации слов языка эсперанто, разработанный институтом анализа данных Нью-Мексико [Guinard 2016]. Автор работы утверждает, что поморфемный разбор в случае с эсперанто — важный компонент обработки языка, так как сложные слова образуются, во-первых, довольно часто, во-вторых, при помощи просто агглютинации, то есть при помощи прибавления одной морфемы к другой. Для каждого слова может быть более чем одна последовательность входящих в неё морфем, однако какие-то последовательности всё же более вероятны. Разработанный алгоритм, опираясь на цепи Маркова заранее и составленный список всевозможны аффиксов, снимает неоднозначность и разбирает слово на морфемы. Алгоритм реализован на языке программирования Scala,

программа находится в открытом доступе на сайте GitHub и расчитана только на операционную систему Linux.

Синтаксический анализ представлен в большем количестве проектов. В частности, в 2009 году была представлена финальная версия парсера EspGram [Bick 2009]. Парсер основывается на грамматике ограничений [Karlsson et al., 1995] и находит в предложении структуры зависимостей. По словам автора, ему удалось достичь показателя 99,5% точности определения частей речи и 92% правильно распознанных синтаксических ролей. Основываясь на этих исследованиях, автор смог затем разработать алгоритм проверки грамматики и орфографии для эсперанто и разместил его на сайте Lingvohelpilo. Доступа к компонентам алгоритма у пользователей нет.

Более детальным описанием структуры может похвастаться безымянная работа 2006 года. В ней автор детально объясняет все этапы создания синтаксического анализатора и приводит различные примеры и сложности [Aasgaard 2006]. Однако найти работающий прототип этой программы нам не удалось.

2.5. Выводы ко второй главе

В этой главе мы рассмотрели особенности языка эсперанто, подробнее остановившись на тех их них, которые особенно актуальны для машинного перевода.

Язык эсперанто обладает простой грамматикой, не знающей исключений. Порядком слов в предложении и системой глагольных времён он напоминает русский язык, однако в эсперанто отсутствует род у существительного, согласование глагола по лицам и родам с существительным, что должно усложнять машинный перевод с русского языка на эсперанто и обратно.

Перевод с эсперанто на русский можно осуществлять как с помощью автоматического переводчика, так и с помощью словаря. Среди переводчиков наиболее значимыми являются система от компаний Google и Яндекс. Первая система существует более 5 лет и обладает обширной фразовой таблицей и

новейшими технологиями глубинного обучения, однако осуществляет перевод через английский как язык-посредник, что приводит к накоплению ошибок. Яндекс переводчик для эсперанто выпущен около года назад и демонстрирует хорошие результаты, однако допускает ошибки в согласовании и пропускает некоторые слова. Словарь Бориса Кондратьева содержит множество статей, однако не является удобным в использовании и не позволяет переводить фразы.

Утилиты для автоматической обработки эсперанто выпущены не крупными компаниями, а частными исследователями. По этой причине, несмотря на проработанные алгоритмы, анализаторы довольно сложно использовать, так как они рассчитаны на ограниченный круг операционных систем и не обладают интуитивным пользовательским интерфейсом.

Таким образом, проблема качественного машинного перевода с эсперанто на русский по сей день остаётся актуальной.

Глава 3. Создание гибридного компонента.

3.1. Общее описание эксперимента.

Как уже было сказано ранее, статистический перевод, несмотря на простоту реализации, нередко порождает ошибочные предложения в языках с развитой морфологией. Результаты предыдущих исследований на тему статистического перевода с эсперанто н арусский язык показали, что большая часть ошибок приходится на неверное согласование прилагательных и существительных по роду, числу и падежу [Orlova 2015]. Также, из-за развитой словообразовательной системы языка эсперанто в предложении может встретиться редкое слово-неологизм; статистический переводчик не найдёт слово во фразовой таблице и оставит непереведённым, что может привести к новым ошибкам.

Для того, чтобы сделать перевод более понятным для пользователя и уменьшить количество рассогласований, мы решили разработать дополнительный модуль, основанный на правилах, влияющий на качество статистического перевода. Самым целесообразным показалось сделать упор именно на реактировании уже готовго статистического перевода. В результате мы разработали программу, работающую по следующему алгоритму:

- графематический анализ предложений
- морфологический анализ слов предложений
- выравнивание предложений пословно
- первичное исправление ошибок
- поиск зависимостей в предложении на эсперанто
- перенос зависимостей на русское предложение
- вторичное исправление ошибок

Алгоритм был реализован на языке программирования Python с использованием библиотеки NLTK и морфологического анализатор pymorphy2. Именно Python на данный момент является самым подходящим языком программирования для обработки естественного языка, так как для

него существует большое еоличество библиотек с уже готовыми полезными функциями и алгоритмами. Из этого следует, что написанный в результате работы алгоритм может быть использован другими разработчиками и исследователями в своих целях.

Для анализа эффективности алгоритма мы сравнили предложения, переведённые человеком, с результатами перевода статистической системы без гибридного элемента и с ним вручную и выявили основные недостатки. Далее мы подробно опишем каждый шаг алгоритма и оценку результатов.

3.2. Графематический и морфологический анализы

Самый важный этап работы алгоритма - морфологический анализ слов, так как на основе него будет происходит последующего выравнивание и исправление ошибок. Для того, чтобы произвести морфологический анализ, в первую очередь необходимо проанализировать предложение графематически, то есть разбить входной текст на знаки препинания, словоформы и т.п. Библиотека для анализа естественного языка NLTK содержит в том числе и функцию work_tokenize, с помощью которой мы разбили предложение на токены - единицы анализа текста (рис 3.1).

```
line_eo = "Ve, mi estas esperantisto"
line_ru = "Увы, я - эсперантист"
words_eo = word_tokenize(line_eo)
words_ru = word_tokenize(line_ru)


>>>['Ve', ',', 'mi', 'estas', 'esperantisto']
>>>['Увы', ',', 'я', '-', 'эсперантист']
```

Рис. 3.1. Токенизация предложения

Для хранения морфологической информации нами был разработан специальный тип данных WordStruct, который хранит словоформу, часть речи и информацию о всех возможных характеристиках слова. Если одна из характеристик неопределена или в принципе невозможна, она получает значение None:

```
keys = [
    'word',#словоформа
    'POS',#часть речи
    'gender', #род
```

```
    'case',#падеж
    'number',#число
    'normal_form',#нормальная форма
    'mood',#наклонение
    'tense',#число
    'person'#лицо
]
WordStruct = namedtuple('WordStruct', keys)
```

Рис. 3.2. Тип WordStruct

Для анализа слов русского предложения был выбран анализатор PyMorphy2 - морфологический анализатор, написанный на языке Python [Korobov 2015]. Программа умеет выдавать графематическую информацию о слове, возвращать его к нормальной форме и ставить слово в нужную форму. Если анализатор сталкивается со словом, написанным латинскими буквами, то присваивает ему тег «LATN», что довольно важно для следующих этапов работы нашего алгоритма. Выходные данные pymorphy2 отличаются от необходимого нам вида, поэтому мы преобразуем их и переписываем в тип WordStruct:

```
def make_struct(parse):
    part = parse.tag.POS #выясняем часть речи разобранного слова
    if part == 'NOUN': #если это существительное...
        ws = WordStruct(
            word = parse.word,
            normal_form = parse.normal_form,
            POS = parse.tag.POS,
            gender = parse.tag.gender,
            case = parse.tag.case,
            number = parse.tag.number
        )#записываем необходимые характеристики
```

Рис. 3.3. Преобразование данных из pymorhy2 в Wordstruct

Для языка эсперанто мы разработали свой морфологический анализатор. Поскольку эсперанто - язык по природе своей искусственный, то в нём изначально не заложена морфологическая неоднозначность. Это позволяет создать анализатор, основанный исключительно на информации об окончаниях и списках некоторых базовых частей речи, в случае которых корень слова совпадает с самой словоформой.

На входе алгоритм получает словоформу и пытается определить её часть речи. В самом начале он сравнивает словоформу с готовыми списками слов и, если происходит совпадение, присваивает все необходимые характеристики.

Отдельным списком морфологический анализатор хранит следующие части речи: непроизводные наречия (например, «hierаŭ» - «вчера», «tre» - очень), артикли, союзы («ĉar» - поскольку, «sed» - но), количественные числительные («kvar» - четыре), междометия («ek» - айда, «hura» - ура), частицы («ne» - не), предлоги («al» - предлог дательного падежа, «kun» - с), местоимения-существительные в именительном и винительном падежах («ŝi» - она, «min» - меня), а также вопросительные слова («kiu» - кто, «kie» - где) со всеми производными вариантами. За основу готового списка мы взяли данные из морфологического анализатора [Guinard 2016], которые были изменены и дополнены данными с сайта Eoru.ru.

```
conjunctions = ("aŭ", "ĉar", "ĉu", "des", "do", "ju", "kaj", "ke", "kvankam",
"malgraŭ", "minus",
                "nek", "plus", "se", "sed", "tamen") # полный список союзов
.....
elif token.lower() in conjunctions: #если словоформа есть в списке союзов...
    ws = WordStruct(POS = "CONJ")
```

Рис. 3.4. Пример готового списка союзов и сравнение с ним переменной

Если же необходимое слово не встретилось в списке, алгоритм начинает анализировать окончания слов. Словоформу, окончание которой не является стандартным, алгоритм считает именем собственным и присваивает тег существительного.

```
elif token[-1] == "n": #если видим окончание винительного падежа...
    if token[-2] == "j": #если видим окончание множетсвенного числа...
        if token[-3] == "a": #если видим окончание прилагательного...
            ws = WordStruct(POS = "ADJF", number = "plur", case = "accs")
        elif token [-3] == "u":
            ws = WordStruct(POS = "NPRO", number = "plur", case = "accs")
        else:
            ws = WordStruct(POS = "NOUN", number = "plur", case = "accs")
    elif token[-2] == "a": #если видим окончание прилагательного...
        ws = WordStruct(POS = "ADJF", number = "sign", case = "accs")
    elif token[-2] == "e": #если видим окончание наречия...
        ws = WordStruct(POS = "ADVB")
    else:
        ws = WordStruct(POS = "NOUN", number = "sing", case = "accs")
```

Рис. 3.5. Определение части речи

После применения вначале графематического анализа, а затем - морфологического, на выходе мы получаем информацию следующего вида (Рис 3.6):

```
>>>[[Ve INTJ], [, PNCT], [mi NPRO nomn sing 1per], [estas VERB indc pres],
[esperantisto NOUN nomn sing]]
>>>[[увы INTJ], [, PNCT], [я NPRO nomn sing 1per], [- PNCT -], [эсперантист NOUN masc
nomn sing]]
```

Рис. 3.6. Морфологически проанализированные предложения

3.3. Выравнивание предложений пословно

Для выравнивания предложений пословно в компьютерной лингвистике используются различные метрики, основанные на статистике, порядке слов и т.д. Однако эти метрики чаще всего призваны работать с чистым предложением без какой-либо дополнительной информации. [Och, Ney, 2003]. В нашем же случае мы имеем дело с предложением, каждое слово которого уже имеет, в частности, информацию о части речи. Поскольку части речи в русском и в эсперанто нередко совпадают, то выравнивать предложение по частям речи кажется самым продуктивным решением.

Для того, чтобы создать элайнер, или выравниватель, на основе частей речи, мы решили взять за основу расстояние Левенштейна [Левенштейн, 1965], или редакционное расстояние. Данный алгоритм обычно применяется в биоинформатике для сравнения генов, хромосом и белков, а также в системах автоматического распознавания и исправления ошибок и опечаток в тексте. В основе алгоритма лежит идея о том, что разницу между одной строкой и другой строкой можно выразить численно, и число это складывается из штрафов за операции - удаление, вставка и несоответствие. Самая минимальная сумма штрафов называется редакционным расстоянием. Формула расстояния Левенштейна расчитывается следующим образом:

$$D(i,j) = \begin{cases} 0 & ;i = 0, j = 0 \\ i & ;j = 0, i > 0 \\ j & ;i = 0, j > 0 \\ \min\big(& \\ \quad D(i,j-1)+1, & \\ \quad D(i-1,j)+1, & ;j > 0, i > 0 \\ \quad D(i-1,j-1)+m(S_1[i],S_2[j]) & \\ \big) & \end{cases}$$

Рис. 3.7. Формула нахождения редакционного расстояния

где D - это редакционное расстояние, а m - функция, которая определяет и числинно оценивает разницу между S1[i] и S2[j]. По умолчанию она бинарна, то есть возвращает единицу, если сравниваемые элементы не идентичны.

Для вычисления редакционного расстояния часто используют двумерную матрицу:

		E	L	E	P	H	A	N	T
	0	1	2	3	4	5	6	7	8
R	1	1	2	3	4	5	6	7	8
E	2	1	2	2	3	4	5	6	7
L	3	2	1	2	3	4	5	6	7
E	4	3	2	1	2	3	4	5	6
V	5	4	3	2	2	3	4	5	6
A	6	5	4	3	3	3	3	4	5
N	7	6	5	4	4	4	4	3	4
T	8	7	6	5	5	5	5	4	3

Рис. 3.8. Матрица расчёта расстояния Левенштейна

На рисунке видно не только то, как алгоритм получил число 3 в результате, но и то, какая буква одного слова должна соответствовать букве другого слова. Для того, чтобы найти эти соответсвия алгоритмически, то есть чтобы понять, какие же операции и в каком порядке необходимо применить к строкам, чтобы из одной получить другую, необходимо совершить обратный алгоритм, то есть с помощью двумерной матрицы рассчитать самый быстрый способ свести редакционное растояние к нулю. Такой алгоритм получил название «редакционное предписание» и был предложен Р. Вагнером и М. Фишером в 1974 году [Wagner, Fischer, 1974].

Чтобы применить данные алгоритм к нашей ситуации, для начала мы ненадолго забываем на слова и рассчитываем редакционное расстояние между последовательностью частей речи в предложении. Возьмём просто предложение «я вижу собаку». Пусть массивы meo и mru - это эсперантский и русский массивы переменных типа WordStruct, каждая из которых содержит слово и морфологическую информацию о нём. Тогда расстояние Левенштейна можно высчитать с помощью следующего кода:

```
>>>meo = [[Mi NPRO nomn sing 1per], [vidas VERB indc pres], [hundon NOUN accs sing]]
>>>mru = [[я NPRO nomn sing 1per], [вижу VERB sing видеть indc pres 1per], [собаку
NOUN femn accs sing собака]]

ss1, ss2 = [], []
m, n = len(meo), len(mru)
d = [[0] * (n + 1) for _ in range (m + 1)]
for i in range (m + 1):
    d[i][0] = i
for j in range (n + 1):
    d[0][j] = j
for i in range(1, m + 1):
    for j in range (1, n + 1):
        d[i][j] = min(d[i][j-1] + 1, d[i-1][j] + 1, d[i-1][j-1] + (meo[i-1][0].POS !=
mru[j-1][0].POS))
print (d[i][j])

>>>0
```

Рис. 3.9. Алгоритм расчёт расстояния Левенштейна

В результате расстояние Левенштейна получилось равным нулю, так как все части речи совпадают. Если заменить, например, «mi» на любое существительное, то расстояние уже будет равно единице, так как нам одна часть речи перестанет совпадать. Далее, переходим к редакционному предписанию:

```
while True:
    if d[i][j] == d[i-1][j] + 1:
        print("del")
        ss1.append(meo[i-1][0])
        ss2.append("-")
        i -= 1

    elif d[i][j] == d[i][j-1] + 1:
        print("ins")
        ss2.append(mru[j-1][0])
        ss1.append("-")
        j -= 1
    else:
        print ("al/match")
        ss2.append(mru[j-1][0])
        ss1.append(meo[i-1][0])
        i -= 1
        j -= 1
    if i <= 0 and j <= 0:
        break


>>>[Mi NPRO nomn sing 1per, vidas VERB indc pres, hundon NOUN accs sing]
```

```
>>>[я NPRO nomn sing я 1per, вижу VERB sing видеть indc pres 1per, собаку NOUN femn
accs sing собака]
```

Рис. 3.10. Алгоритм выявления редакционного предписания

Более наглядно результат работы алгоритма будет выглядеть, если мы добавим в предложение на эсперанто, например, частицу «ne» - «не»:

```
[Mi NPRO nomn sing 1per, ne PRCL, vidas VERB indc pres, hundon NOUN accs sing]
[я NPRO nomn sing я 1per, '-', вижу VERB sing видеть indc pres 1per, собаку NOUN femn
accs sing собака]
```

Рис. 3.11. Пример неидентичных по частям речи предложений

Алгоритм вставил пропуск на место частицы так, что последующие глагол и существительное находятся друг под другом. Таким образом, этап выравнивания является завершённым.

3.4. Первичное исправление ошибок.

На данном этапе работы программы мы имеем предложения, выровненные пословно. Этой информации достаточно, чтобы произвести первичные исправления, а именно:

- скорректировать число существительного, прилагательного или причастия в роли определения;
- скорректировать время и/или наклонение глагола
- скорректировать личное местоимение
- приписать непереведённому слову морфологические характеристики.

Все остальные характеристики в эсперанто не выражены, поэтому просто морфологической разметки эсперантского предложения нам оказывается недостаточно. Также, стоит оговориться, что, несмотря на то, что в эсперанто у существительных есть падеж, копирование информации о падеже в русское предложение без учёта синтаксической информации может привести к ухудшению перевода, так как часть косвенных падежей русского языка в эсперанто выражается только винительным падежом, а часть - только предлогами. Морфологическая информация о непереведённом слове может быть полезна пользователю, не знакомому с языком, сориентроваться в

структуре предложения и получить информацию о нормальной форме слова, чтобы в таком виде уже посмотреть в словаре его значение.

Для исправления вышеперечисленных ошибок алгоритм пошагово сравнивает соответствующую друг другу слова и, если часть речи совпадает, обращается к характеристикам. При несовпадении характеристик алгоритм вновь анализирует слово с помощью pymorphy2, извлекает из него исправленную словоформу и записывает её в массив. На языке программирование python в случае с существительным этот алгоритм выглядит следующим образом:

```
input:
>>>[mi NPRO nomn sing 1per, vidas VERB indc pres, la ARCL, grandan ADJF accs sing,
hundon NOUN accs sing]
>>>[я NPRO nomn sing я 1per, вижу VERB sing видеть indc pres 1per, None, большую ADJF
femn accs sing большой, собак NOUN femn accs plur собака]

if s_ru[i] is not None and s_ru[i].POS != "LATN": #если слово таки переведено на
русский и существует
    if s_ru[i].POS == 'NOUN'and s_ru[i].POS == s_eo[i].POS: #если это существительное
        if s_ru[i].number != s_eo[i].number: #если по числу не совпали слова
            temp = m.parse(s_ru[i].word)[0] #анализируем слово снова, чтобы получить
весь разбор
            new = temp.inflect({s_ru[i].case, s_eo[i].number}) #просим pymorhy выдать
верную словоформу

            s_ru[i] = parse_word(new.word)[0] #и в таком новом виде уже записываем

else:
    s_ru[i] = s_eo[i]

output:
>>>[я NPRO nomn sing я 1per, вижу VERB sing видеть indc pres 1per, la ARCL, большую
ADJF femn accs sing большой, собаку NOUN femn accs sing собака]
```

Рис. 3.12. Исправление ошибок в числе

В данном примере алгоритм исправил слово «собак» на «собаку».

3.5. Поиск зависимостей

Проанализировав выходные предложения статистического переводчика [Orlova 2015], мы пришли к выводу, что самые частые ошибки, который первый этап редактирования не позволяет исправить, относятся к следующим типам:

- Нарушение согласования между существительным и глаголом (в функции подлежащего и сказуемого соответственно).
- Нарушение согласования между личным местоимением и глаголом (в функции подлежащего и сказуемого соответственно).

- Нарушение согласования между существительным в функции подлежащего и именной частью составного сказуемого.
- Нарушение согласования между местоимением в функции подлежащего и именной частью составного сказуемого.
- Нарушение согласования между существительным и прилагательным.
- Нарушение согласования между существительным и причастием.
- Нарушение согласования между глаголом и прямым дополнением.

Чтобы эти ошибки исправить, в предложении на эсперанто необходимо установить зависимости между словами, а затем перенести эти зависимости на русское предложение. В рамках данной работы мы приняли решение не присваивать синтаксические роли всем единицам предложения, так как считаем это излишним.

Итак, далее алгоритм на основе морфологической информации и порядка слов выделяет следующие группы: подлежащее-существительное + глагол, подлежащее-местоимение + глагол, прилагательное + существительное, глагол + дополнение. Для сохранения информации алгоритм записывает в новый массив данных номер того слова, которое является зависимым, в ячейку того слова, которое оказывает влияние.

Далее, алгоритм сравнивает основные характеристики главного и зависимого слов и, при необходимости, вновь обращается к уже привычной процедуре: с помощью pymorphy2 заменяет словоформу на более подходящую.

3.6. Анализ ошибок.

Для того, чтобы оценить работу системы, мы приняли решение перевести 250 предложений с помощью переводчика Яндекс, а затем обработать их с помощью нашего алгоритма.

Предложения были взяты из открытого корпуса параллельных текстов OPUS [Tiedemann et al., 2004]. OPUS – развивающийся многоязычный корпус текстов, бесплатно размещённых в сети интернет. Впервые корпус был

представлен на конференции в Лиссабоне его создателем, Йоргом Тидеманном В 2012 году OPUS содержал более чем 40 миллиардов слов на 90 языках мира [Tiedemann, 2012]. Среди всех пар текстов можно найти и пары эсперанто-русский. Среди всех возможных подкорпусов мы выбрали подкорпус Tatoeba, который существует по принципу Википедии: каждый имеет право добавить, перевести или отредактировать предложение на том языке, которым он владеет. Именно этот подкорпус наиболее всего похож на современную версию языка эсперанто и не перегружен редкими терминами, неологизмами или сложными синтаксическими конструкциями.

Переводчик от компании Яндекс был выбран как основной, так как именно он на данный момент производит самый качественный перевод с эсперанто на русский язык. Таким образом мы убедимся, что система исправляет не ошибки некачественно построенного статистического переводчика, а именно изъяны самого статистического подхода.

Для оценки качества перевода изначально мы планировали использовать автоматическую метрику BLEU , которая является на данный момент одной из самых популярных систем оценок машинного перевода. В первую очередь мы сравнили результаты перевода на русский язык с предложениями, которые уже были на русском в параллельном корпусе. Оказалось, что многие предложения, хоть и переведены стилистически верно, совсем не похожи на предложения из корпуса. Например:

Табл. 3.1. Различие переводов, сделанных машиной и человеком

Эсперанто	Русский язык (машинный перевод)	Русский язык (перевод человека)
Mi ne dubas, ke li helpos al mi.	Я не сомневаюсь, что он мне поможет.	Я не сомневаюсь, что он поможет мне.
Li kuris pli rapide, ol lia frato.	Он бежал быстрее своего брата.	Он бежал быстрее, чем его брат.
Mi ne respondis al via letero, ĉar mi estis okupita.	Я не ответил на твое письмо, потому что был занят.	Я не ответил на ваше письмо, потому что я был занят.

В данных примерах переводчик Яндекса не допустил ни смысловых, ни стилистических ошибок, однако его перевод отличается от эталонного. Это ставит под сомнение разумность использования метрики, и мы решили от неё отказаться и проанализировать количество ошибок вручную.

Среди 250 переведённых предложений абсолютно верными оказались 188 предложений, то есть 75,2% переводов были осуществлены верно. За ошибки мы не считали использование большой буквы в середине предложения, как, например, в переводе «Она часто играла в теннис в субботу Днем», так как это не нарушает восприятие текста. Среди ошибочных предложений оказались примеры с отсутствием правильного согласования, стилистически некорректные предложения, предложения с непереведёнными словами. После применения алгоритма были исправлены 8 предложений, а в 14 предложениях со словами, которые так и остались на эсперанто, была добавлена морфологическая информация о слове. В результате применения алгоритма количество правильных предложений выросло в общей сумме до 78% от общей суммы. Примеры исправлений представлены в таблице ниже:

Табл. 3.2. Скорректированный с помощью алгоритма перевод

Перевод до применения алгоритма	Перевод после применения алгоритма
Это красивый автомобиль, но не стоит цена, которую я заплатил за это.	Это красивый автомобиль, но не стоит цену, которую я заплатил за это.
Не пинать ежика	Не пинай ежика
Иностранных медитации скрыта	Иностранные медитации скрыта
Последствия болезни не был серьезным	Последствия болезни не были серьезным
Учиться, fiĉjo, русский язык, учиться	Учись, fiĉjo, русский язык, учись
Джен, Джен, скажи ему все же скажу	Джен, Джен, скажи ему все же скажи
Вы не беспокоить Тома	Вы не беспокойте Тома
Сделать это снова!	Сделай это снова!

Среди исправленных переводов 5 приходятся на исправление инфинитива глагола на повелительное наклонение, 1 - на изменение падежа существительного, 1 - на изменение падежа прилагательного, 1 - на изменение числа глагола. Часть предложений после перевода стала абсолютно верной, часть - приблизилась к грамотному виду.

3.7. Проблемы и перспективы.

На каждом из этапов разработки алгоритма мы столкнулись с некоторыми проблемами, часть из которых по-прежнему остаётся нерешёнными.

Самая первая проблема - это предложения перводчика Яндекса, где глагол «быть» заменяется на тире. По какой-то причине система перевода не ставит пробелы в таких случаях, как, например, в переводе «Он-француз». Поскольку пробелы отсутствуют, pymorphy2 трактует это как слово, пишущееся через дефис, что впоследствии не даёт алгоритмо грамотно перейти к следующему этапу. Таким образом, те случаи, где ошибки были в словах, которые оказались рядом с подобным тире, не были исправлены.

Морфологический анализатор эсперанто, основанный на окончаниях и списках основных частей речи, действительно покрывает большую часть слов, однако проблемы возникают с теми именами собственными, окончания которых совпадают с окончаниями определённых частей речи. Так, имя Daŝa (Даша) распознаётся морфологическим анализатором как имя прилагательное, что в последствии не позволяет корректно работать с этим словом.

Следующая трудность - это система местоимений языка эсперанто.

В частности, местоимение «vi» выступает аналогом английского местоимения «you», то есть одновременно может относится и к группе людей, и к одному человеку, с которым автор обращения общается неформально, и к одному человеку, к которому автор обращается

уважительно. Также, для того, чтобы подчеркнуть неформальные отношения с собеседником, некоторые авторы используют местоимение «ci», что означает «ты». По этой причине нами было принято решение всегда трактовать «vi» как «вы», несмотря на то, что в текстах и разговорной речи «ci» встречается крайне редко.

Проблемы также возникали на этапе морфологического анализа русских слов. Анализатор pymorphy2 чаще всего действительно предоставляет правильные варианты разбора слова, однако иногда результаты морфологического анализа не поддаются объяснению:

```
»> from pymorphy_proc import parse_word
»> parse_word('Я')
[я NPRO nomn sing я 1per, я NOUN masc nomn sing я, я NOUN masc gent sing я, я NOUN
masc datv sing я, я NOUN masc accs sing я, я NOUN masc ablt sing я, я NOUN masc loct
sing я, я NOUN femn nomn sing я, я NOUN femn gent sing я, я NOUN femn datv sing я, я
NOUN femn accs sing я, я NOUN femn ablt sing я, я NOUN femn loct sing я, я NOUN masc
nomn sing я, я NOUN masc gent sing я, я NOUN masc datv sing я, я NOUN masc accs sing
я, я NOUN masc ablt sing я, я NOUN masc loct sing я, я NOUN femn nomn sing я, я NOUN
femn gent sing я, я NOUN femn datv sing я, я NOUN femn accs sing я, я NOUN femn ablt
sing я, я NOUN femn loct sing я]
```

Рис. 3.14. Пример разбора PyMorphy2

На данном примере видно, что анализатор разбирает словоформу «я» не только как местоимение, но и как существительное, которое может стоять в любом падеже. Такая широкая вариативность разбора создавала трудности для последующего анализа.

С другой стороны, морфологический анализатор в некоторых случаях не видит тот вариант разбора, который, на наш взгляд, является верным. Так, слова «давай» и «прощай» он анализирует исключительно как глаголы в повелительном наклонении, несмотря на то, что чаще они встречаются в роли междометий.

Внимательно изучив предложения, мы поняли, что для исправления большего количества ошибок необходимо дополнить алгоритм дополнительными условиями. В частности, предложение «Я заплатил ему четыре долларов» с точки зрения алгоритма является анализируемым и абсолютно верным, поскольку он не учитывает особенности согласования русских числительных с последующими существительными.

Также, мы не предусмотрели то, что некоторые глаголы невозможно поставить в то или иное время, предварительно не найдя ему пару другого вида. Так, предложение «Каждый день наша мама купить фрукты» изначально содержало глагол настоящего времени, который человек перевёл бы как «покупает». Однако морфологический анализитор pymorhy2 не предусматривает такого перехода от глагола купить, поэтому предложения такого типа также остались без исправлений.

В дальнейшем мы планируем дополнить алгоритм так, чтобы он работал с разными видовыми парами и исправлял ошибки согласования с числительными, а также обратимся к нескольким вариантам морфологических анализаторов русского языка, чтобы получать более достоверную информацию.

Выводы к главе 3

Для построения гибридного компонента мы разработали и реализовали на языке Python следующий алгоритм:

- графематический анализ предложений. Для его реализации мы обратились к встроенным в библиотеку NLTK функциям;
- морфологический анализ слов предложений. Морфологический анализатор языка эсперанто мы написали без использования словарей, опираясь на готовые списки корневых слов и окончания словоформ; для русского языка мы обратились к уже готовому морфологическому анализатору pymorphy2;
- выравнивание предложений пословно. Алгоритм выравнивания предложения пословно основывается на расстоянии Левенштейна и редакционном предписании;
- первичное исправление ошибок. если число существительного или наклонение или число глагола в русском предложении не соответствовали таковым в предложении на эсперанто, алгоритм возвращал слова к нормальной форме, а затем ставил слово в нужную форму;

- поиск зависимостей в предложении на эсперанто. Для этого мы взяли основные глагольные и именные группы, в которых происходит рассогласование, и осуществили поиск по ним в предложениях;
- перенос зависимостей на русское предложение. Он происходил, если части речи русского предложения совпадали с частями речи предложения на эсперанто в конкретной группе.
- вторичное исправление ошибок.

После написания алгоритма мы проверили его эффективность на предложениях, переведённых статистической системой перевода «Яндекс Переводчик». Сравнив переводы до и после редактирования с предложениями на эсперанто мы пришли к выводу, что алгоритм справляется с исправлением рассогласования в наклонениях, числах и падежах, однако не может работать корректно, если глагол совершенного вида нужно поставить в настоящее время, а также если существительное необходимо согласовать с числительным.

Заключение

В данной работе мы проанализировали историю машинного перевода, те сложности с которыми сталкивались исследователи и разработчики прошлых лет. Также, мы сравнили различные подходы к машинному переводу (перевод по правилам, статистический перевод, гибридный перевод) и установили преимущества и недостатки каждого из подходов. Этим вопросам посвящена первая глава. Далее, мы рассмотрели язык эсперанто с исторической и лингвистической точек зрения и изучили различные технологии для автоматического анализа эсперанто во второй главе. На основании проанализированной информации мы приняли решение разработать гибридный компонент для статистического переводчика с эсперанто на русский язык, об основных элементах которого можно прочитать в главе 3.

Основная цель гибридного компонента – осуществить проверку и исправление грамматических ошибок, которые возникли в результате статистического перевода. Разработанная программа вначале анализирует предложения графематически и морфологически, затем производит пословное выравнивание по частям речи с помощью расчёта редакционного предписания и исправляет ошибки, связанные с неправильным числом или наклонением глаголов, далее ищет некоторые основные зависимости и исправляет ошибки снова, уже с использованием данных об именных и глагольных группах. Для оценки алгоритма мы загрузили в программу 250 предложений из параллельного корпуса OPUS, в результате чего количество верных предложений увеличилось на 4,25%.

В процессе работы на каждом этапе был выявлен ряд трудностей. Статистический подход Яндекс-Переводчика приводит к непредсказуемым ошибкам, с которыми не справляются дальнейшие шаги алгоритма. Также, результаты морфологического анализа pymorphy2 русских слов порой даёт

чрезмерно обширные разборы, но, несмотря на это, некоторые слова трактует однобоко и определяет части речи неверно.

Разработанный нами алгоритм тоже содержит в себе ряд недочётов. Так, он не исправляет неверное согласование существительных с числительными и на данный момент не может найти глаголу его видовую пару, чтобы её поставить в нужную форму. В дальнейшем мы планируем добавить эти опции в нашу программу. Невзирая на такие недоработки, алгоритм тем не менее смог улучшить результаты статистического перевода с эсперанто на русский язык.

В целом, задачи, поставленные в данной работе, можно считать выполенными, а цель – достигнутой.

Список литературы

1. **Андреева А. Д.** Обзор систем машинного перевода. Журнал "Молодой ученый", М., 2013.
2. **Белоногов Г. Г.** Компьютерная лингвистика и перспективные информационные технологии. М., 2004.
3. **Бельская И. К., Королев Л. Н., Панов Д. Ю.** Переводная машина П. П. Троянского: сборник материалов о переводной машине для перевода с одного языка на другие, предложенной П. П. Троянским в 1933 г. Изд. Акад. Наук СССР, Москва, 1959.
4. **Беляева Л. Н.** Лингвистические автоматы в современных гуманитарных технологиях: Учебное пособие. СПб, 2007.
5. **Беляева Л. Н., Откупщикова М. И.** Автоматический (машинный) перевод // Прикладное языкознание / под. ред. Герда А. С. СПб., 1996.
6. **Браславский П., Белобородов А., Шаров С., Халилов М.** Дорожка по оценке машинного перевода ROMIP MTEval 2013: отчет организаторов. Диалог, М., 2013.
7. **Всеволодова А. В**. Компьютерная обработка лингвистических данных. М., 2007.
8. **Жирков Л. И.** Границы применимости машинного перевода. «Вопросы языкознания», М., 1956.
9. **Заменгов Л.** Международный язык. Предисловие и полный учебник. Варшава 1887
10. **Колкер Б. Г.** Международный язык Эсперанто: полный учебник. М, 2012.
11. **Кузнецов С. Н.** Краткий словарь интерлингвистических терминов // Проблемы международного вспомогательного языка. — М.: Наука, 1991. — С. 171—228.

12. **Левенштейн В. И.** Двоичные коды с исправлением выпадений, вставок и замещений символов. Доклады Академий Наук СССР, 1965.
13. **Леонтьева Н. Н.** Автоматическое понимание текстов: системы, модели, ресурсы: учебное пособие для студентов лингвистических факультетов вузов. М., 2006.
14. **Марчук Ю. Н.** Проблемы машинного перевода. М., 1983.
15. **Марчук Ю. Н.** Компьютерная лингвистика: учебное пособие. М., 2007.
16. **Мельчук И. А.** Опыт теории лингвистических моделей "Смысл-Текст". М., 1999
17. **Молчанов А.** Статистические и гибридные методы перевода в технологиях компании ПРОМТ. CONTROL ENGINEERING Россия #4 (46), М., 2013.
18. **Николаев И. С., Митренина О. В., Ландо Т. М. (Ред.).** Прикладная и компьютерная лингвистика. URSS, Москва, 2016
19. **Пиперски А. Ч.** Конструирование языков. От эсперанто до дотракийского, М., 2017
20. **Соловьева А. В.** Профессиональный перевод с помощью компьютера. СПб, 2008.
21. **Филинов Е. Н.** 07.10.2002. ст. «История машинного перевода» //http://www.computer-museum.ru/
22. **Шаляпина З. М.** Автоматический перевод: эволюция и современные тенденции. Вопросы языкознания, М., 1996.
23. **Щипицина Л. Ю.** Информационные технологии в лингвистике: учеб. пособие. М., 2013
24. **Aasgaard B. C.** Parsing of Esperanto. Cand. Scient. Thesis, Oslo, Norway, 2006
25. **Bick E.** A Dependency Constraint Grammar for Esperanto. NODALIDA, Odendse, Denmark, 2009
26. **Brown J. C.** Loglan 1: A logical language. Gainesville, FL, 1999

27. **Brown P. F., Della Pietra S. A., Della Pietra V. J., Mercer R. L**. The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics, 19(2):263–311, Cambridge, MA, US, 1993.
28. **Cho K., van Merrienboer B., Gulcehre C., Bougares F., Schwenk H., Bengio Y.** Learning phrase representations using rnn encoder-decoder for statistical machine translation. EMNLP, Doha, Qatar, 2014
29. **Corsetti R., Pinto M. A., Tolomeo M.** Regularizing the regular: The phenomenon of overregularization in Esperanto-speaking children // Language Problems and Language Planning. 2004
30. **Costa-jussa M. R., Banchs R. E., Rapp R., Lambert P., Eberle K., Babych B.** Workshop on Hybrid Approaches to Translation: overview and developments. Second Workshop on Hybrid Approaches to Translation, Sofia, Bulgaria, 2013
31. **Goldwater S., McClosky D.** Improving statistical MT through morphological analysis. EMNLP, Vancouver, B.C., Canada, 2005
32. **Guinard T.** An Algorithm for Morphological Segmentation of Esperanto Words. PBML № 105, Prague, Czech republic, 2016
33. **Hajič J., Hric J., Kuboň V.** Machine Translation of Very Close Languages. ANLC '00, Seattle, Washington, 2000
34. **Harris B.** Bi-Text, a new concept in translation theory. Language Monthly, 54:8–10, Ann Arbor, MI, US, 1988
35. **Hutchins W. J.** Machine translation: past, present, future. Chichester, UK, 1986.
36. **Hutchins J.** Machine translation over fifty years. Histoire, Epistemologie, Langage, Tome XXII, fasc. 1, Paris, France, 2001.
37. **Kalchbrenner N., Blunsom P.** Recurrent continuous translation models. EMNLP, Seattle, USA, 2013.
38. **Karlsson F., Voutilainen A., Heikkilä J. Anttila A.** Constraint Grammar - A Language-Independent System for Parsing Unrestricted Text. Natural Language Processing №4, Berlin & New York. 1995
39. **Koehn P.** Statistical Machine Translation. Cambridge, UK, 2010.

40. **Koehn P., Och, F. J., Marcu D.** Statistical phrase-based translation. NAACL, Edmonton, Canada, 2003
41. **Kolovratník D., Klyueva N., Bojar O**. Statistical Machine Translation Between Related and Unrelated Languages. ITAT, Kralova studna, Slovakia, 2009.
42. **Korobov M.** Morphological Analyzer and Generator for Russian and Ukrainian Languages // Analysis of Images, Social Networks and Texts, pp 320-332, 2015.
43. **Och F. J., Ney H.** A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics 29(1):19–51, Cambridge, MA, US, 2003.
44. **Lewis C. I.** A survey of symbolic logic. University of California Bress, Berkley, 1918
45. **Ogden C. K.** The ABC of Basic English., Trubner, London 1932.
46. **Okrand M.** The Klingon Dictionary (paperback) (2nd: i.e., with addendum ed.). New York: Pocket Books, 1992.
47. **Okrent A**. In the Land of Invented Languages: Adventures in Linguistic Creativity, Madness, and Genius. NY, USA, 2009
48. **Orlova D.** Esperus: the First Step to Build a Statistical Machine Translation System for Esperanto and Russian Languages. AINL FRUCT, Saint Petersburg, Russia, 2015
49. **Papineni K., Roukos S., Ward T, Zhu W.** BLEU: a Method for Automatic Evaluation of Machine Translation. ACL, Philadelphia, US, 2002.
50. **Schwenk H.** Continuous space translation models for phrase-based statistical machine translation. Coling, Mumbai, India, 2012.
51. **Sutskever I., Vinyals O., Le Q. V.** Sequence to sequence learning with neural networks. NIPS, Monreal, Canada, 2014
52. **Tiedemann J.** Parallel Data, Tools and Interfaces in OPUS. LREC, Istanbul, Turkey, 2012.
53. **Tiedemann J.**, Nygaard L. The OPUS corpus - parallel & free. LREC, Lisbon, Portugal, 2004

54. **Trask R. L., Stockwell P.** Language and linguistics:The key concepts. 2nd ed. Abington, New York:Routledge, 2007
55. **Uchida H**. UNL: Universal Networking Language An Electronic Language for Communication, Understanding, and Collaboration. UNU/IAS/UNL Center, Tokyo, Japan, 1996
56. **Varga D., Nemeth L., Halacsy P., Kornai A., Tron V., Nagy V**. Parallel corpora for medium density languages. RANLP, Borovets, Bulgaria, 2005.
57. **Wagner R. A., Fischer M. J.** The string-to-string correction problem. Journal of the ACM, Vol. 21, No. 1, pp. 168-173, 1974
58. **Wu Y., Schuster M., Chen Z., Le Q. V., Norouzi M., ... & Klingner, J.** Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. Technical Report, 2016.
59. **Zamenhof L. L.** Fundamenta de esperanto. Paris, 1905

Интернет-источникики:

- Яндекс Переводчик

URL: http://translate.yandex.ru/?ncrnd=3621/

- Большие словари Бориса Кондратьева

URL: http://eoru.ru/

- Домашняя страница Moses

URL: http://www.statmt.org/moses/index.php?n=Main.HomePage

- Открытый корпус параллельных текстов OPUS

URL: http://opus.lingfil.uu.se/

- Tilde Neural Machine Translation

URL: https://www.tilde.com/products-and-services/machine-translation/neural-machine-translation