Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ ИУ «Информатика и системы управления»

КАФЕДРА ИУ-7 «Программное обеспечение ЭВМ и информационные технологии»

РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА
К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ
НА ТЕМУ:
«Методы машинного перевода»

Студент ИУ7-53Б _____ **А. А. Светличная**
(Группа) (Подпись, дата) (И.О.Фамилия)

Руководитель _____ **Л. Л. Волкова**
(Подпись, дата) (И.О.Фамилия)

2022 г.

СОДЕРЖАНИЕ

РЕФЕРАТ

Научно-исследовательская работа 29с., 1 табл., 8 ист., 1 прил.

КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА, МАШИННЫЙ ПЕРЕВОД, МЕТОДЫ МАШИННОГО ПЕРЕВОДА, СТАТИСТИЧЕСКИЙ МАШИННЫЙ ПЕРЕВОД, МАШИННЫЙ ПЕРЕВОД НА ОСНОВЕ ПРАВИЛ

Цель — провести обзор методов машинного перевода.

В процессе работы проводилась классификация существующих методов машинного перевода, а также сравнительный анализ данных методов по сформулированным критериям оценки.

Результаты — статистический метод машинного перевода имеет большее количество достоинств, чем метод на основе правил, однако для некоторых задач его недостатки являются критичными, по этой причине оба метода до сих пор активно используются.

ВВЕДЕНИЕ

На сегодняшний день существует явная тенденция развития международных коммуникаций как в рабочих сферах, так и за их пределами. Для качественного взаимодействия между людьми, компаниями крайне необходима возможность выстраивать коммуникацию с наибольшей точностью обработки передаваемой или получаемой информации. Данную возможность могут представлять профессиональные переводчики, однако этот способ не является доступным для большинства, по этой причине актуальны методы машинного перевода и, в частности, важной является проблема качества машинного перевода.

Цель данной научно-исследовательской работы — провести обзор методов машинного перевода.

Для достижения поставленной цели необходимо решить следующие задачи:

— описать существующие методы машинного перевода;

— классифицировать рассмотренные методы;

— сформулировать критерии сравнения методов машинного перевода;

— сравнить методы на основании выделенных критериев;

— описать результаты сравнения рассмотренных алгоритмов.

1 Анализ предметной области

1.1 Основные направления компьютерной лингвистики

К компьютерной лингвистике относят все, что связано с использованием компьютеров в языкознании.

Основные направления, рассматриваемые в компьютерной лингвистике:

— информационный поиск;

— машинный перевод;

— выделение терминов;

— компьютерная лексикография;

— распознавание и синтез речи;

— проблемы обучения языку с помощью компьютера.

К прикладным направлениям компьютерной лингвистики относятся следующие:

— машинный перевод;

— распознавание и синтез речи;

— разработка и использование языков программирования, языков информационных систем;

— компьютерная лексикография и терминография;

— лингвистические основы информационного поиска;

— автоматическое индексирование, реферирование и классификация текстов;

— автоматический контент-анализ и установление авторства текстов;

— гипертекстовые технологии представления текста;

— корпусная лингвистика;

— компьютерная лингводидактика [1].

В данной работе будет рассмотрена область машинного перевода, а именно используемые методы машинного перевода.

1.2 Определения машинного перевода

Термин «машинный перевод» понимается по крайней мере в двух смыслах.

Машинный перевод в широком смысле — это область научных исследований, находящаяся на стыке лингвистики, математики, кибернетики, и имеющая целью построение систем, реализующих машинный перевод в узком смысле.

Машинный перевод в узком смысле — это процесс перевода некоторого текста с одного естественного языка на другой, реализуемый компьютером с возможным участием человека. В ходе данного процесса на вход машины подается текст, словесная часть которого не сопровождается никакими дополнительными указаниями, а на выходе получается текст на другом языке, являющийся переводом входного [2].

Также **машинным переводом** называется результат работы машинного перевода в узком смысле, то есть текст, переведенный на другой естественный язык посредством машины.

Однако далее под машинным переводом будет пониматься определение в узком смысле.

1.3 Этапы анализа и синтеза текста

Машинный анализ текстов на естественном языке содержит несколько этапов, рассмотренных подробно ниже.

1) **Графематический анализ** представляет собой самый нижний уровень обработки текстов и, соответственно, первый этап машинного анализа. Главными задачами данного этапа являются определение границ предложений (сегментация) и разделение предложений на слова (токенизация). Под токенами понимаются такие минимальные неделимые единицы текста, как слова, знаки препинания, даты, числа, сокращения и т.д.

2) **Морфологический анализ** представляет собой второй этап в обработке текста. Он позволяет определить морфологические характеристики для каждого из выделенных слов текста. При этом для каждого слова определяется его лемма – нормальная форма. В русском языке нормальными считаются следующие морфологические формы слов:

   — существительные — именительный падеж, единственное число;

   — прилагательные — именительный падеж, единственное число, мужской род;

   — глаголы, причастия, деепричастия — глагол в инфинитиве.

   Процессу лемматизации (приведения слов к нормальным формам) присуща лексическая и морфологическая неоднозначность, которая выражается в том, что одному слову могут соответствовать несколько лемм.

3) **Синтаксический анализ** позволяет снять неоднозначность, которая возникает на морфологическом уровне, то есть дополняет результаты морфологического. На этом этапе анализа формируется дерево разбора предложения — структура, элементы которой связаны синтаксическими связями в соответствии с синтаксическими правилами. Существует два основных способа построения дерева синтаксического разбора: в виде дерева зависимостей и в виде дерева составляющих. В некоторых задачах используют частичный синтаксический анализ.

4) **Семантический анализ** является переходом от структуры синтаксических связей к ее смысловой интерпретации. Таким образом, на

вход данного этапа подается синтаксическая структура текста, представленная в виде деревьев разбора. На выходе формируется множество семантических структур, построенных в соответствии с принятой формальной нотацией (семантической моделью) [2, 3].

Задача синтеза является обратной анализу текста задачей, поэтому ее решение проходит вышеописанные этапы в противоположном порядке. Эта задача все же обладает собсвтенными особенностями, поэтому рассматривается в работе также отдельно поэтапно.

1) **Семантический синтез** осуществляет переход от смысловой записи фразы к её синтаксической структуре, то есть дереву, описывающему связи между словами в предложении.

2) **Синтакстический синтез** превращает дерево, описывающее связи между словами в предложении, в линейный порядок слов. При этом осуществляется согласование параметров слов между собой.

3) **Морфологический синтез** по нормальной форме слова и его параметрам находит соответствующую словоформу.

4) **Графематический синтез** составляет из слов в заданном порядке единый текст, следит за соответствием фрагментов входного текста фрагментам выходного [4].

Вывод

В данном разделе были рассмотрены основные направления применения методов компьютерной лингвистики, определен термин «машинный перевод», а также описаны этапы автоматической обработки и синтеза текстов, тем самым проанализирована предменая область.

2 Классификация методов машинного перевода

2.1 Виды переводов

Существуют три основные классификации видов перевода, приведенные ниже.

1. По характеру переводимых текстов:
   - художественный;
   - общественно-политический;
   - специальный.
2. По характеру речевых действий переводчика в процессе перевода:
   - письменный;
   - устный [5].
3. По степени точности перевода:
   - подстрочный;
   - художественный.
4. По степени вовлеченности человека-редактора.
   1) **С постредактированием** — человек-редактор вносит правки в полученный машинный перевод для формирования конечного продукта надлежащего качества.
   2) **С предредактированием** — редактирование текста осуществляется непосредственно перед началом машинного перевода документа. Данная форма взаимодействия человека и компьютера подразумевает внесение правок в исходный документ, заранее облегчающих работу системы, а также устранение любых

потенциальных ошибок, которые могут возникнуть при машинном переводе. Данный процесс включает в себя сокращение длины предложений, унификацию терминологии, исправление орфографических и пунктуационных ошибок, упрощение грамматических структур и устранение двусмысленностей.

3) **С интерредактированием** — подразумевается процесс вовлечения человека-редактора непосредственно во время работы системы машинного перевода с целью исправления или удаления неоднозначностей, возникающих при машинном переводе многозначных слов, а также с целью уточнения формулировок и различных грамматических структур [6].

4) **Без редактирования** — редактирование текста человеком не осуществляется, то есть машина принимает произвольные необработанные данные и самостоятельно выдает перевденный текст. Такая степень вовлеченности человека является наиболее желанной и проблемной, решением данной задачи и занимаются методы машинного перевода.

2.2 Методы машинного перевода

Для начала приводятся некоторые определения, необходимые для рассмотрения методов машинного перевода.

Исходный язык – это язык, на котором текст написан изначально.

Целевой язык – это язык, на который данный текст должен быть переведен.

2.2.1 Машинный перевод на основе правил

Машинный перевод на основе правил (Rule-based Machine Translation) опирается на бесчисленные встроенные лингвистические правила и множество двуязычных словарей для каждой языковой пары. Программа анализирует исходный текст и превращает его в так называемый переходный текст, из которого впоследствии создается текст на целевом языке.

Данный процесс невозможен без объемной словарной базы, содержащей наборы морфологической, синтаксической и семантической информации, а также объемные наборы правил. Специальное программное обеспечение использует эти сложные наборы правил, а затем трансформирует грамматическую структуру исходного языка в структуру целевого языка. Процесс перевода в данном случае основан на использовании огромного числа словарей и сложных лингвистических правил. Качество перевода, предоставляемого системой машинного перевода, основанного на правилах, в большинстве случаев можно повысить двумя способами: осуществить первоначальные вложения, которые значительно повысят качество при ограниченных затратах; осуществлять постоянные вложения для постепенного повышения качества перевода. Несмотря на то, что процесс обучения систем машинного перевода, основанных на правилах, неминуемо приводит к повышению качества перевода текста, данный процесс может быть довольно долгим.

2.2.2 Статистический машинный перевод

Статистический машинный перевод (Statistical Machine Translation) анализирует базу данных существующих переводов, сделанных ранее человеком-переводчиком (известны как двуязычные текстовые корпуса). Большинство современных систем, основанных на статистическом машинном переводе, используют не слова, а фразы и собирают переводы с помощью перекрывающейся фразы [7]. При фразовом переводе цель состоит в том, чтобы уменьшить ограничения перевода на основе слов путем перевода целых последовательностей слов, длина которых может отличаться. Последовательности слов называются фразами, но обычно это не лингвистические фразы, а фразы, найденные с помощью статистических методов из двуязычных текстовых корпусов.

Анализ двуязычных текстовых корпусов (исходный и целевой языки) и одноязычных (целевой язык) создает статистические модели, которые преобразуют текст с одного языка на другой с использованием статистических весов для определения наиболее вероятного перевода [8].

Качество полученного перевода во многом зависит от того, насколько

полны и репрезентативны базы данных необходимых языковых пар и насколько согласуются друг с другом эти базы данных.

2.3 Критерии оценки методов машинного перевода

Сравнение описанных методов машинного перевода будет проводиться по следующим критериям:

1) возможность использования с любой языковой парой;
2) возможность выхода текстов за предметную области;
3) возможность настройки перевода под заданную предметную область;
4) получение перевода для любого исходного текста;
5) отсутсвие акцента (здесь и далее под акцентом понимается непохожесть переведенного текста на работу человека-переводчика);
6) простота разработки ситем, основанных на определенном подходе.
7) необходимость привлечения экспертов для разработки систем.

2.4 Сравнение методов машинного перевода

Обозначим введенные критерии оценки методов следующим образом:

— К1 — возможность использования с любой языковой парой;

— К2 — возможность выхода текстов за предметную области;

— К3 — возможность настройки перевода под заданную предметную область;

— К4 — получение перевода для любого исходного текста;

— К5 — отсутствие акцента;

— К6 — отсутствие необходимости привлечения экспертов для разработки систем.

Результаты сравнения методов машинного перевода представлены в таблице 2.1.

Таблица 2.1 – Сравнение методов машинного перевода

Метод	К1	К2	К3	К4	К5	К6
Статистический	+	-	+	-	+	-
На основе правил	-	+	+	+	-	-

По результатам сравнения методом с наибольшим количеством достоинств в общем случае оказался статистический метод (группа методов) машинного перевода. Однако в силу наличия недостатков у каждой группы методов для точной оценки необходимо рассмотреть отдельно задачи, в которых недостакти или достоинства крнкретных методов оказывают критическое влияние на решение задачи.

Рассмотрим отдельно несколько конкретных критериев.

Возможность использования с любой языковой парой — в данном случае применять метод перевода на основе правил практически невозможно, так как существуют языки, правила в которых не являются явно заданными, так например в русском языке практически произвольный порядок слов, который к тому же иногда тонко отличает смысл предложений. Таким образом, метод на основе правил будет иметь слишком сильный машинный акцент или даже неточности передачи информации.

Получение перевода при любом исходном тексте — если тексты содержат нестандартные речевые обороты, например, редко используемые фразеологизмы или терминологию, то результат систем, основанных на статистическом методе, будет не только не соответствовать предметной области, но даже не передавать даже части смысла исходного текста. В то время как метод, основанный на правилах, гарантировано вернет хотя бы подстрочный результат.

Таким образом, можно сделать вывод о том, что несмотря на кажущееся большее соответсвие статистической группы методов выделенным критериям, говорить о том, что данная группа методов является наилуч-

шей для решения любой задачи, некорректно, так как существуют такие задачи, решение которых гарантированно предоставляет только один из рассматриваемых методов.

Вывод

В данном разделе были рассмотрены существующие методы машинного перевода, сформулированы критерии их оценки, а также проведено сравнение по сформулированным критериям.

ЗАКЛЮЧЕНИЕ

В ходе выполнения данной работы были классифицированы методы машинного перевода.

На основании результатов сравнения методов можно сделать вывод о том, что в общем случае наиболее эффективным методом является статистический, так как имеет возможность использования с любой языковой парой без машинного акцента, однако также существуют такие задачи, которые могут быть гарантировано решены исключительно методом, основанном на правилах. По этой причине оба метода все еще используются.

Цель, поставленная в начале работы, была достигнута — проведен обзор методов машинного перевода. В ходе ее выполнения были решены все задачи:

— описаны существующие методы машинного перевода;

— классифицированы рассмотренные методы;

— сформулированы критерии сравнения методов машинного перевода;

— проведено сравнение методов на основании выделенных критериев;

— описаны результаты сравнения рассмотренных алгоритмов.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Умаров Абдурасул Абдурахманович ОСНОВНЫЕ НАПРАВЛЕНИЯ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ // Наука и образование сегодня. 2021. №2 (61). URL: https://cyberleninka.ru/article/n/osnovnye-napravleniya-kompyuternoy-lingvistiki (дата обращения: 27.11.2022).

2. Воронович В. В. Машинный перевод : учеб. пособие. — Минск: Белорусский государственный университет, 2017. — С. 5.

3. Дунаев Л. Л. Исследовательская система для анализа текстов на естественном языке // Проблемы интеллектуализации и качества систем информатики :сборник. Вып. 13. Новосибирск : Изд-во ИСИ СО РАН, 2006. – С. 56.

4. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : учеб. пособие / Е.И. Большакова, Э.С. Клышинский, Д.В. Ландэ, А.А. Носков, О.В. Пескова, Е.В. Ягунова. — Москва: МИЭМ, 2011. — С. 106–108.

5. «Введение в переводоведение» — М.: Изд-во РУДН, 2006. Валеева Наиля Гарифовна — к.п.н. профессор, зав. кафедрой иностранных языков №2 Института иностранных языков Российского университета дружбы народов.

6. Батуев А. А. Системы машинного перевода: сравнение качества перевода и возможностей их использования (на примере технической документации в металлургической отрасли) : учеб. пособие — Екатеринбург: Уральский гуманитарный институт, 2021. — С. 10.

7. What is Statistical Machine Translation (SMT)? [Электронный ресурс]. URL: `https://omniscien.com/faq/what-is-statistical-machine-translation/` (дата обращения: 28.11.2022).

8. Раренко М.Б. МАШИННЫЙ ПЕРЕВОД : ОТ ПЕРЕВОДА «ПО ПРАВИЛАМ» К НЕЙРОННОМУ ПЕРЕВОДУ // Социальные

и гуманитарные науки. Отечественная и зарубежная литература. Сер. 6, Языкознание: Реферативный журнал. 2021. №3. URL: https://cyberleninka.ru/article/n/mashinnyy-perevod-ot-perevoda-po-pravilam-k-neyronnomu-perevodu (дата обращения: 27.11.2022).

ПРИЛОЖЕНИЕ А

(обязательное)

Презентация

Методы машинного перевода

Студент: Светличная Алина Алексеевна ИУ7-53

Научный руководитель: Волкова Лилия Леонидовна

Цель – провести обзор методов машинного перевода

Задачи:

- описать существующие методы машинного перевода
- классифицировать рассмотренные методы
- сформулировать критерии сравнения методов машинного перевода
- сравнить методы на основании выделенных критериев
- описать результаты сравнения рассмотренных алгоритмов

Основные направления компьютерной лингвистике

- информационный поиск
- машинный перевод
- выделение терминов
- компьютерная лексикография
- распознавание и синтез речи
- проблемы обучения языку с помощью компьютера

Определения машинного перевода

Машинный перевод в широком смысле – это область научных исследований, находящаяся на стыке лингвистики, математики, кибернетики, и имеющая целью построение систем, реализующих машинный перевод в узком смысле.

Машинный перевод в узком смысле – это процесс перевода некоторого текста с одного естественного языка на другой, реализуемый компьютером с возможным участием человека.

Автоматическая обработка текста

Этапы анализа

1) Графематический анализ
2) Морфологический анализ
3) Синтаксический анализ
4) Семантический анализ

Этапы синтеза

1) Семантический синтез
2) Синтаксический синтез
3) Морфологический синтез
4) Графематический синтез

Виды переводов

По характеру переводимых текстов:
- художественный
- общественно-политический
- специальный

По характеру речевых действий переводчика в процессе перевода:
- письменный
- устный

По степени точности перевода:
- подстрочный
- художественный

По степени вовлеченности человека-редактора:
- с постредактированием
- с предредактированием
- с интерредактированием
- без редактирования

Классификация методов машинного перевода

Метод на основе правил

опирается на бесчисленные встроенные лингвистические правила и множество двуязычных словарей для каждой языковой пары

Статистический метод

анализирует базу данных существующих переводов, сделанных ранее человеком-переводчиком (известны как двуязычные текстовые корпуса)

Сравнение методов

Метод	К1	К2	К3	К4	К5	К6
Статистический	+	-	+	-	+	-
На основе правил	-	+	+	+	-	-

- К1 – возможность использования с любой языковой парой
- К2 – возможность выхода текстов за предметную области
- К3 – возможность настройки перевода под заданную предметную область
- К4 – получение перевода для любого исходного текста
- К5 – отсутствие акцента
- К6 – отсутствие необходимости привлечения экспертов для разработки систем

Заключение

Цель, поставленная в начале работы, была достигнута — проведен обзор методов машинного перевода.

В ходе ее выполнения были решены все задачи:

- описаны существующие методы машинного перевода
- классифицированы рассмотренные методы
- сформулированы критерии сравнения методов машинного перевода
- проведено сравнение методов на основании выделенных критериев
- описаны результаты сравнения рассмотренных алгоритмов

Спасибо за внимание