**МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ**

**э**

**Федеральное государственное бюджетное образовательное учреждение**

**высшего образования «Кабардино-Балкарский государственный университет**

**им. Х.М. Бербекова» (КБГУ)**

**ИНСТИТУТ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА И ЦИФРОВЫХ ТЕХНОЛОГИЙ**

**КАФЕДРА КОМПЬЮТЕРНЫХ ТЕХНОЛОГИЙ И ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ**

**КУРСОВАЯ РАБОТА**

по дисциплине

Разработка программных приложений

на тему:

**«**Разработка переводчика**»**

|  |
| --- |
| Выполнила:  студент 3 курса направления  09.03.03 Прикладная информатика  (профиль «Прикладная информатика в экономике») ОФО |
| Гунукбаев М.Э. |
|  |
| Руководитель:  доцент кафедры КТиИБ  Елеев И.З \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |

**Нальчик 2023**

Содержание

[ВВЕДЕНИЕ 3](#_Toc154697702)

**Глава 1: Основы автоматического перевода………………………4**

**1.1 История развития автоматического перевода…………………………4**

**1.2 Принципы и подходы автоматического перевода………………………5**

**1.3 Принцип работы современного машинного перевода………………….6**

**Введение**

В современном мире, где глобализация и международное сотрудничество становятся неотъемлемыми частями нашей повседневной жизни, языковые барьеры становятся все более преодолимыми. Возникает необходимость в разработке и использовании автоматических переводчиков, которые могут обеспечить перевод текстов и устной речи между различными языками. Развитие современных технологий, таких как искусственный интеллект и машинное обучение, открывает новые перспективы в создании эффективных и точных переводчиков.

Целью данной курсовой работы является изучение процесса создания переводчика и исследование различных подходов, методов и алгоритмов, применяемых в этом процессе. Будет проведен анализ существующих технологий и исследований в области машинного перевода, а также рассмотрены основные принципы и подходы, используемые при разработке переводчиков.

В работе также будет исследовано использование нейронных сетей и глубокого обучения для повышения качества автоматического перевода. Будут рассмотрены различные архитектуры нейронных сетей, методы их тренировки и оптимизации для достижения наилучших результатов в переводе текстов и речи.

Однако, несмотря на постоянное развитие технологий в области машинного перевода, существуют определенные ограничения и проблемы, с которыми сталкиваются автоматические переводчики. В работе будет проведен анализ этих ограничений и поисков возможных путей и улучшений для повышения качества перевода.

В итоге, главной целью данной работы является разработка и реализация прототипа переводчика, а также проведение экспериментов для оценки его качества и эффективности. Результаты и выводы исследования позволят лучше понять возможности и ограничения автоматического перевода и найти пути для его улучшения.

**Глава 1: Основы автоматического перевода**

**1.1 История развития автоматического перевода**

История развития автоматического перевода простирается на десятилетия, начиная с первых исследований в этой области в середине XX века. В то время, когда компьютеры только начинали свое появление, ученые уже задумывались о возможности создания машин, способных переводить тексты с одного языка на другой.

Одной из ранних концепций автоматического перевода была идея использования словарей и грамматических правил для замены слов и фраз из исходного языка на эквивалентные в целевом языке. Эта подход называется "языковым переводчиком из языка в язык" (Language-to-Language Translator) и был назван Эллойсом (ELLIOT) в 1952 году.

Однако, такие ранние системы столкнулись с большим количеством проблем и ограничений. Во-первых, они не учитывали контекст при переводе, поэтому выводили дословные и нелогичные переводы. Во-вторых, они были ограничены в словарном запасе и грамматических правилах, что делало их неэффективными для перевода широкого спектра текстов.

С появлением компьютерной лингвистики и статистического подхода к автоматическому переводу в 1980-х и 1990-х годах, принципы и методы перевода начали меняться. Вместо основного использования словарей и грамматических правил, появилась идея использования больших параллельных корпусов для обучения модели перевода. Этот подход называется "статистическим машинным переводом" (Statistical Machine Translation, SMT).

С использованием статистических моделей и методов машинного обучения, системы автоматического перевода стали гораздо более эффективными и гибкими. Они могут учитывать контекст, лексические и грамматические особенности языка, а также семантический контекст для более точного перевода.

В последние годы нейронные сети и глубокое обучение (Deep Learning) стали основными инструментами в области автоматического перевода. Нейронные сети позволяют моделировать сложные зависимости между словами и фразами в тексте и обеспечивают превосходные результаты в переводе. Это открыло новые возможности для улучшения качества и точности автоматического перевода.

Сегодня автоматический перевод активно развивается, и исследователи продолжают искать новые подходы и методы для улучшения его качества и эффективности. С развитием искусственного интеллекта и новых технологий ожидается, что автоматический перевод станет все более точным и надежным, открывая новые горизонты для межъязыковой коммуникации.

**1.2 Принципы и подходы автоматического перевода**

Автоматический перевод - это процесс перевода текста с одного языка на другой при помощи компьютерных программ и алгоритмов. Вот некоторые принципы и подходы, используемые в автоматическом переводе:

1. Статистический машинный перевод (SMT): Этот подход основывается на статистическом анализе больших параллельных корпусов текстов на разных языках. Система SMT использует вероятностные модели и статистические алгоритмы для определения наиболее вероятного перевода для данного предложения на основе уже существующих параллельных данных.

2. Нейронные сети и глубокое обучение: В последние годы нейронные сети и глубокое обучение стали основными инструментами в области автоматического перевода. Этот подход позволяет строить мощные модели, которые учитывают сложные зависимости между словами и фразами в тексте. Благодаря своей способности извлекать семантическую информацию из данных, нейронные сети достигают хороших результатов в переводе.

3. Гибридные системы: Это подход, который объединяет различные методы и подходы к автоматическому переводу. Гибридные системы могут использовать как статистический машинный перевод, так и нейронные сети, чтобы достичь лучшей производительности и точности. Например, гибридные системы могут использовать статистический машинный перевод для определения наиболее вероятного перевода и нейронные сети для его уточнения.

4. Алгоритмы выравнивания: При переводе предложений необходимо определить соответствие между словами и фразами в исходном и целевом языках. Алгоритмы выравнивания помогают определить, какие слова и фразы соответствуют друг другу. Выравнивание может быть выполнено на основе статистического анализа параллельных корпусов или с использованием нейронных сетей.

5. Контекстуальный перевод: Перевод без учета контекста может привести к неправильному пониманию и неверному переводу. Поэтому многие современные системы автоматического перевода учитывают контекст, включая предыдущие предложения или даже целые документы, для более точного перевода.

6. Постредактирование: Одним из важных аспектов автоматического перевода является возможность постредактирования перевода человеком. Машина может предложить начальный вариант перевода, который затем может быть отредактирован и улучшен людьми, чтобы достичь желаемого качества.

Это лишь некоторые из основных принципов и подходов автоматического перевода. С каждым годом развитие технологий продолжает усовершенствоваться и вносить новые инновации в эту область.

**1.3 Принцип работы современного машинного перевода**

В основе современных систем машинного перевода лежит алгоритм перевода, использующий формальную грамматику языков и статистические данные. Чтобы выучить язык, система сравнивает тысячи параллельных текстов, содержащих одну и ту же информацию, но на разных языках. Для каждого изученного текста система строит список уникальных признаков. Например, редко используемые слова и специальные знаки, которые встречаются в тексте с определенной частотой.

В системах машинного перевода, как правило, три основные части: модель перевода, модель языка и декодер. Модель перевода — это таблица, в которой для всех слов и фраз на одном языке перечислены возможные переводы на другой язык с указанием вероятности этих переводов. Система сравнивает не только отдельные слова, но и словосочетания из нескольких слов, идущих подряд. Модели перевода для каждой пары языков содержат миллионы пар слов и словосочетаний. Что касается модели языка, то она создается системой на этапе изучения текстов.

Переводом занимается декодер. Он проводит морфологический и синтаксический анализ текста и для каждого предложения подбирает все варианты перевода с сортировкой по убыванию вероятности. Затем все полученные варианты декодер оценивает с помощью модели языка на частоту употребления и выбирает предложение с наилучшим сочетанием вероятности и частоты.

Использование статистических данных позволяет системам машинного перевода меняться вместе с языком. Если люди начинают писать какое-то слово по-другому, система видит это, как только к ней попадают новые тексты. Чтобы улучшать качество перевода, систему регулярно обновляют и проводят проверки. Впрочем, высококачественный машинный перевод текстов по-прежнему недостижим. Однако он значительно облегчает и ускоряет работу переводчикам.

Вместо «машинный» иногда употребляется слово **автоматический**, что не влияет на смысл. Однако термин **автоматизированный перевод** имеет совсем другое значение — при нѐм программа просто помогает человеку переводить тексты.

**Автоматизированный перевод** – перевод текстов с одного языка на другой человеком с использованием специализированных программ, приложений.

**Машинный перевод** – процесс перевода текстов с одного языка на другой машиной посредством специальной компьютерной программы. Между ними есть три главных различия:

• Трудозатраты переводчика.

•Специализированное программное обеспечение.

•Качество.

Так, при автоматизированном переводе основную работу выполняет переводчик, а специальные программы выступают только в качестве вспомогательного инструмента. Его главной целью является сокращение времени процесса, обеспечение единообразия терминов и общего соответствия (качества).

Машинный же перевод производит сама программа, а человек только редактирует полученный результат. И здесь уже цель – заменить труд человека, получая при этом быстрый перевод низкого качества.

Системы **автоматизированного перевода** – это собирательное определение для специализированных программ и интернет-сервисов, которые используют переводчики в процессе работы. Они хорошо подходят для работы над художественными, юридическими и техническими переводами. К ним относятся:

• Отдельные либо встроенные редакторы, которые помогают автоматически проверить грамматику текста и правописание слов.

• ПО, таблицы, текстовые редакторы, обеспечивающие управление терминологией (MultiTerm, Termex и т.д.).

• ПО, с помощью которого осуществляется менеджмент переводческих проектов.

• CAT-инструменты, использующие TM (TranslationMemory – Память переводов), которые включают в себя образцы ранее переведѐнных текстов или предложений. Яркими их представителями стали программы Trados, DéjàVu, MemoQ, MemSource, Wordfast.

• Корпусы, представляющие собой большой ряд документов с использованием одного или нескольких языков. С их помощью составляется сжатое описание употребления слов и выражений в общих случаях или с учѐтом какой-либо определѐнной предметной темы.

**TranslationMemory** – «Память перевода»

Сегодня письменный перевод нужно делать не только качественно, но и быстро. На помощь специалисту приходят программные разработки, такие как MemoQ. Их общее распространѐнное название «Память перевода», принцип – «не переводи одно и то же дважды». Память перевода не следует путать с программами машинного перевода. Фактически TranslationMemory – это наработанная база переведѐнных слов, выражений, синтаксических целых, которая создаѐтся самим переводчиком. В каждом новом тексте программа определяет уже «знакомые» ей понятия и предлагает специалисту перевод из базы. Таким образом, достигается единообразие в письменном переводе документов и, разумеется, значительная экономия времени.

**Машинный перевод** также представлен набором программ и интернетсервисов. Наиболее популярные из них: PROMT, Google translate, Яндекс.Переводчик.

**Google translate**

Переводчик Google Translate использует SMT (статистический уникальный машинный перевод текста - Statistical Machine Translation) Особенностью Google Translate является метод перевода: он не основан на анализе правил грамматики, а основан на поиске соответствий языка между переводимым текстом и гигантским массивом сервиса, который состоит из слов, вносимых пользователем ранее во время их перевода. Данные блоки текста составляются из всех возможных достоверных источников. В качестве примера можно привести документы государственных организаций. Они являются одними из наиболее проверенных источников для базы данных. Преимущество заключается в том, что такие документы содержат информацию, доступную на многих языках.

Также для развития Google Translate важны книги. В настоящий момент известно, что «корпус» Google Translate обладает более чем триллионом слов. Так сервис активно использует пользовательский уникальный перевод для повышения качественных характеристик машинного текстового перевода путем занесения пользовательских вариаций фраз в базу данных. Таким образом, такое сочетание способствует увеличению уровня качества генерации уникального текста.

Статистическому современному переводу необходим анализ параллельных языковых пар для осуществления генерации текста. Он самостоятельно приспосабливается к новой лексике. Система самостоятельно заносит в свою совершенную базу данных новые вариации перевода слов или словосочетаний, если они отличаются. Она хранит в памяти новые варианты возможного перевода и может использовать эти версии в будущем. Одним из главных достоинств SMT является то, что он обновляется и развивается параллельно с языком. То есть, если появляется новое слово или меняется его лексическое значение, то система распознает это и обновляет свою базу данных. И, как следствие, ускоряется «обучение» системы и совершенствуется качество генерации текста.

Но есть и значительный недостаток SMT. Этой системе необходимо высокопроизводительное аппаратное обеспечение. Для совершенствования программы требуется огромное число вычислений. Также для перевода SMT характерно то, что качество генерации текста полностью зависит от количества данных в корпусе сервиса. Чаще генерируемые сервисом вариации перевода текста оправдывает возложенные ожидания современного человека. Хотя предсказать заранее текстовое содержание, которое сгенерирует Google Translate, сложно.