Лабораторная работа 13 (!индивидуальный проект)

Тема: «Математические модели в лингвистике: Теория и практика»

* Основные математические модели, применяемые в лингвистике
* Применение математических моделей для автоматического перевод
* Моделирование лексической и синтаксической структуры языка
* Оценка и тестирование математических моделей в NLP
* Применение математических подходов для анализа больших текстовых данных

**1.Введение**

Лингвистика и математика тесно связаны: формальные модели помогают анализировать язык, улучшать машинный перевод, распознавание речи и обработку естественного языка (NLP). В данном проекте рассматриваются ключевые математические модели, применяемые в лингвистике, их практическое использование и оценка эффективности.

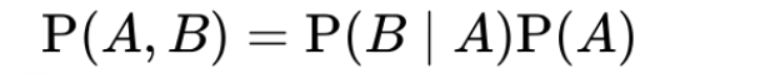
**2.Основные математические модели в лингвистике**

*Статистические модели (N-граммы, скрытые марковские модели)*

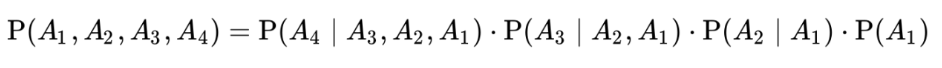
N-граммы – это статистические модели, которые предсказывают следующее слово после N-1 слов на основе вероятности их сочетания. Например, сочетание I want to в английском языке имеет высокую вероятностью, а want I to – низкую. Говоря простым языком, N-грамма – это последовательность n слов. Например, биграммы – это последовательности из двух слов (I want, want to, to, go, go to, to the…), триграммы – последовательности из трех слов (I want to, want to go, to go to…) и так далее.

Такие распределения вероятностей имеют широкое применение в машинном переводе, автоматической проверке орфографии, распознавании речи и умном вводе. Например, при распознавании речи, по сравнению с фразой eyes awe of an, последовательность I saw a van будет иметь большую вероятность. Во всех этих случаях мы подсчитываем вероятность следующего слова или последовательности слов. Такие подсчеты  называются языковыми моделями.

Как же рассчитать P(w)? Например, вероятность предложения P(I, found, two, pounds, in, the, library). Для этого нам понадобится цепное правило, которое определяется так:



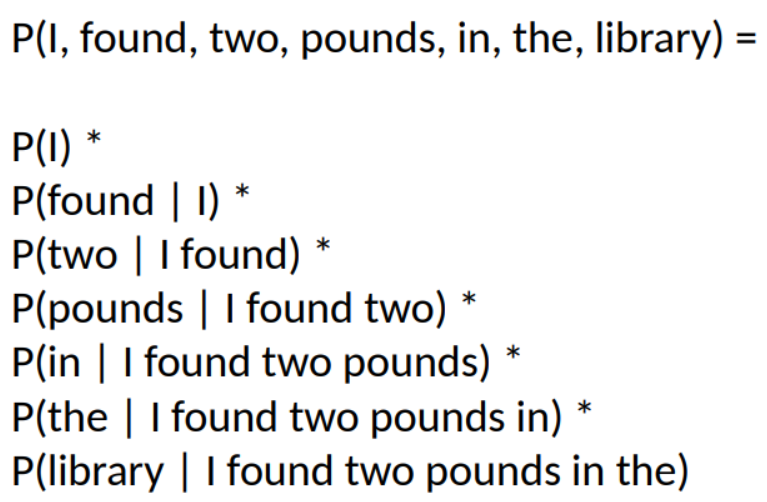
А если переменных больше, то вот так:



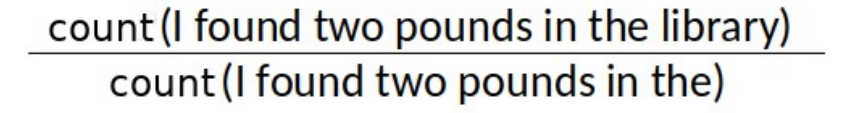
В общем виде правило выглядит следующим образом:



Вероятность нашего предложения будет рассчитываться так:



Таким образом, мы можем оценить совместную вероятность всей цепочки, перемножив условные вероятности. Однако мы не можем рассчитать точную вероятность слова при условии длинной последовательности предшествующих слов, как в этом примере:



Это потому, что возможных последовательностей очень много, а в наших данных просто может не оказаться этих выражений. Поэтому, вместо того, чтобы рассчитывать вероятность слова с учетом всех предшествующих слов, мы можем аппроксимировать вероятность, упростив ее. В этом заключается смысл цепей Маркова, с помощью которых мы можем предсказать вероятность элемента последовательности, не учитывая слишком широкий контекст.

Например, марковская модель первого порядка:

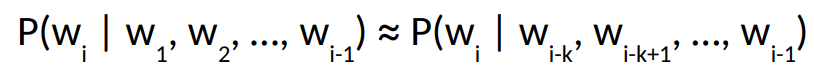


И марковская модель второго порядка:

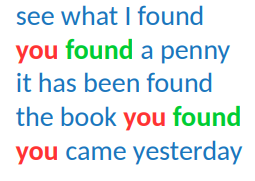


Ведь мы встретим в наших данных the library или in the library с гораздо большей вероятностью, чем I found two pounds in the.

В общем виде марковская цепь k-ого порядка (когда мы учитываем контекст только последних k слов) будет выглядеть так:



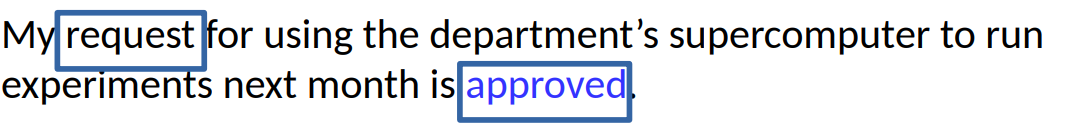
Допустим, у нас есть следующие предложения для обучения моделей:



Используя марковскую цепь первого порядка можно легко посчитать вероятность выражения you found:



Таким образом мы можем рассчитывать биграммы, триграммы, квадрограммы и т.д., причем чем длиннее цепь, тем более детализированной является наша модель, т.е. более длинные предложения включают больше грамматики, чем короткие. В то же время, чем длиннее цепь, тем меньше у нас случаев употребления, а это значит что многие наблюдения будут попадаться всего один раз.Стоит отметить, что языковые модели – не всегда идеальное решение. Возьмем, к примеру, следующее предложение:



В этом случае маловероятно, что approved относится к month, хоть они и расположены близко, и более вероятно – к request, однако n-граммы эту связь не обнаруживают.

*Векторные представления слов*

Векторные представления слов позволяют нам преобразовывать слова в числовые векторы. Это существенно улучшает качество моделей NLP.

1. Word2Vec
2. GloVe
3. FastText
4. Word2Vec

Word2Vec - это группа моделей, которую разработала команда Google. Word2Vec преобразует слова в векторы определённой размерности, где схожие по смыслу слова имеют близкие векторные представления.

**Архитектуры Word2Vec**

Word2Vec представляет два основных метода обучения:

* Continuous Bag-of-Words (CBOW): предсказывает текущее слово по контексту (окружающим словам).
* Skip-Gram: предсказывает окружающие слова по текущему слову.

**Особенности**

* Эффективность: Быстро обучается на больших объёмах текстах.
* Качество: Учитывает семантические отношения между словами.

GloVe

Glove - это модель, разработанная в Стэнфордском университете. В отличие от Word2Vec, который фокусируется на локальных связях между словами, а GloVe фокусируется на глобальных связях, проверяя, насколько часто два слова встречаются вместе в тексте.

**Особенности**

* Глобальная матрица соотношений - это способ увидеть, как слова связаны между собой в целом по всем текстам, проверяя, как часто они появляются вместе.
* Преимущества: GloVe лучше понимает общую структуру и тематику текста, что особенно полезно, когда важно понять глобальные связи и темы в большом количестве данных.

2.3 FastText

FastText — модель, разработанная Facebook AI Research. Она расширяет Word2Vec, учитывая морфологию слов.

**Особенности**

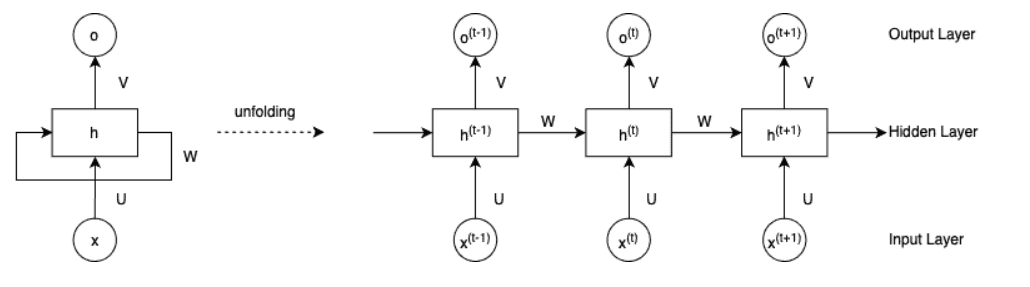
* Символьные n-граммы: Представляет слова как группу буквенных n-грамм.
* Преимущества: Лучше справляется с редкими словами и опечатками.

*Нейронные сети (RNN, LSTM, Transformer)*

Что такое RNN и LSTM?

[Рекуррентные нейронные сети](https://translated.turbopages.org/proxy_u/en-ru.ru.e9fb3e98-681f1ea7-6e962f7a-74722d776562/https/www.baeldung.com/cs/nlp-encoder-decoder-models#2-basic-rnn-units) — это класс нейронных сетей, которые способны обрабатывать последовательные данные, используя информацию из предыдущих входных данных. Рекуррентные нейронные сети — это нейронные сети, которые содержат скрытое состояние, позволяющее им сохранять информацию из предыдущих временных отрезков. Они работают с последовательными данными, обрабатывая по одному входному значению за раз и обновляя своё скрытое состояние на основе текущего и предыдущих скрытых состояний.

На рисунке ниже показана базовая структура RNN:

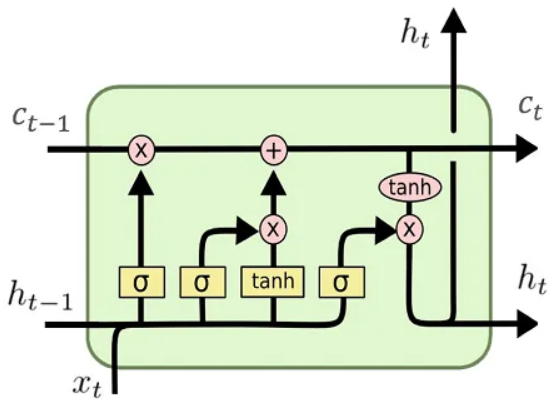


Где x, h, o — входная последовательность, скрытое состояние и выходная последовательность соответственно. U, V и W — весовые коэффициенты.

Рекуррентные нейронные сети особенно эффективны при выявлении краткосрочных зависимостей в последовательностях. Однако они страдают от проблемы исчезающего градиента, когда влияние предыдущих входных данных экспоненциально уменьшается по мере продвижения по последовательности, что затрудняет выявление долгосрочных зависимостей.

[LSTM](https://translated.turbopages.org/proxy_u/en-ru.ru.e9fb3e98-681f1ea7-6e962f7a-74722d776562/https/www.baeldung.com/cs/nlp-encoder-decoder-models#3-lstm) — это особый тип архитектуры RNN, который решает проблему исчезающего градиента, возникающую при обучении глубоких нейронных сетей. LSTM используют ячейки памяти и шлюзы для выборочного хранения и извлечения информации из длинных последовательностей, что делает их эффективными для выявления долгосрочных зависимостей.

На [рисунке](https://commons.wikimedia.org/wiki/File:Long_Short_Term_Memory.png) ниже показана ячейка LSTM:



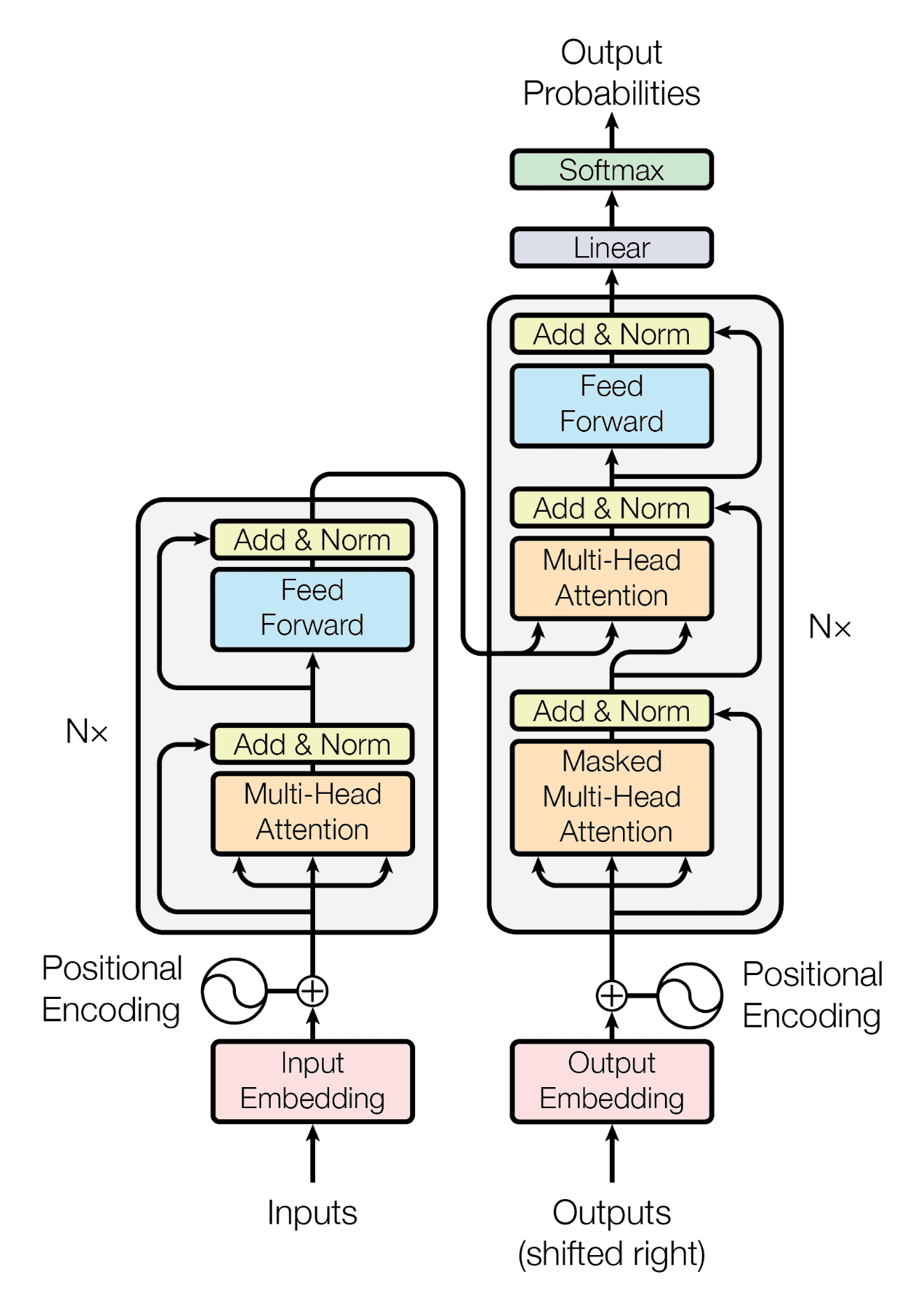
Где ht — скрытое состояние на временном шаге t, а c — состояние ячейки.

Что такое трансформаторы?

[Трансформер](https://translated.turbopages.org/proxy_u/en-ru.ru.e9fb3e98-681f1ea7-6e962f7a-74722d776562/https/www.baeldung.com/cs/large-language-models#5-arrival-of-transformer-model) — это архитектура нейронной сети, предложенная в основополагающей статье [«Всё, что вам нужно, — это внимание»](https://translated.turbopages.org/proxy_u/en-ru.ru.e9fb3e98-681f1ea7-6e962f7a-74722d776562/https/arxiv.org/abs/1706.03762) Вавани и др. В отличие от рекуррентных нейронных сетей, трансформеры не полагаются на рекурсию, а вместо этого используют самовнимание.

Самосознание позволяет модели оценивать важность различных входных токенов при прогнозировании, что даёт ей возможность выявлять долгосрочные зависимости без необходимости последовательной обработки. Трансформеры состоят из кодирующего и декодирующего слоёв, в которых используются механизмы многоголосного самосознания и нейронные сети прямого распространения.

На [рисунке](https://translated.turbopages.org/proxy_u/en-ru.ru.e9fb3e98-681f1ea7-6e962f7a-74722d776562/https/arxiv.org/abs/1706.03762) ниже показана архитектура трансформаторной сети:



Существует множество [типов трансформеров](https://translated.turbopages.org/proxy_u/en-ru.ru.e9fb3e98-681f1ea7-6e962f7a-74722d776562/https/www.baeldung.com/cs/chatgpt-model#1-types-of-transformers). В области НЛП появились две влиятельные модели: [BERT](https://translated.turbopages.org/proxy_u/en-ru.ru.e9fb3e98-681f1ea7-6e962f7a-74722d776562/https/www.baeldung.com/cs/transformer-text-embeddings#bert) (двунаправленные представления кодировщика из трансформеров) и [GPT](https://translated.turbopages.org/proxy_u/en-ru.ru.e9fb3e98-681f1ea7-6e962f7a-74722d776562/https/www.baeldung.com/cs/chatgpt-model) (генеративный предварительно обученный трансформер). BERT использует кодировщик трансформера. Как и трансформер, GPT состоит из кодировщика и декодировщика. Эти модели расширили границы понимания и генерации языка, продемонстрировав возможности крупномасштабного предварительного обучения и трансферного обучения.

**3. Применение математических моделей для автоматического перевод**

*Статистический машинный перевод*

Статистический машинный перевод (англ. Statistical machine translation, SMT) — это метод машинного перевода, при котором перевод генерируется на основе статистических моделей, основанных на анализе двуязычных корпусов текста.

История возникновения статистического машинного перевода

В 1949 году Уоррен Уивер, один из пионеров информатики, опубликовал работу, в которой предложил использовать принципы теории информации Клода Шеннона для автоматического перевода текстов. Уивер представил идею, что перевод можно рассматривать как статистическую задачу - определение наиболее вероятных соответствий между словами и фразами двух языков на основе анализа больших коллекций параллельных текстов.  
  
Однако в то время вычислительные мощности и доступные наборы данных были недостаточными для реализации этого подхода. Поэтому идеи Уивера не получили широкого развития вплоть до конца 1980-х годов. В этот период исследователи из Исследовательского центра Томаса Дж. Уотсона IBM вновь обратились к концепции статистического машинного перевода.  
  
Таким образом, первоначальные концептуальные идеи и их дальнейшее развитие в 1980–90-х годах привели к практической реализации и широкому распространению статистического машинного перевода.

Как работает статистический машинный перевод?

Статистический машинный перевод рассчитывает наиболее вероятное значение каждого слова или фразы в предложении, используя статистические методы. Чем чаще встречается вариант перевода, тем выше вероятность, что он верный.  
  
Статистический машинный перевод работает следующим образом:  
  
1. Сбор параллельных корпусов данных. Собираются большие коллекции текстов на двух языках, где каждый текст на одном языке имеет соответствующий перевод на другом. Эти параллельные тексты используются в качестве обучающих данных для системы перевода.  
  
2. Обучение моделей. На основе параллельного корпуса обучаются вероятностные модели перевода. Модели учатся предсказывать, какой перевод слова или фразы является наиболее вероятным.  
  
3. Перевод. Когда нужно перевести новый текст, статистическая модель анализирует его и генерирует наиболее вероятный перевод каждого слова и фразы. Таким образом, весь текст переводится путем поиска наиболее вероятных переводческих соответствий.  
  
Рассмотрим на примере фразы «Завтра я полечу в Нью-Йорк на конференцию». Сначала система разбивает текст на слова или фразы. В данном случае: «Завтра», «я», «полечу», «в Нью-Йорк», «на конференцию». В нашем случае «завтра» может переводиться как «tomorrow», «я» — как «I», «полечу» может переводиться как «will fly» / «am flying» / «fly», «в Нью-Йорк» — «to New York», «на конференцию» может переводиться как «for a conference» или «to a conference». Для каждой комбинации переводов оценивается вероятность на основе частоты встречаемости в параллельных текстах. Например: Tomorrow I will fly to New York for a conference / Tomorrow I am flying to New York for a conference и т.п. Система выбирает наиболее вероятную комбинацию переводов. В данном случае, это I'm flying to New York tomorrow for a conference.

Этапы тренировки статистических моделей машинного перевода

Сбор обучающих данных  
  
Первый шаг в создании машины SMT — сбор большого количества двуязычных текстов. Их берут из переводной литературы, многоязычных веб-сайтов и документов. От качества и разнообразия учебного материала зависит точность перевода.  
  
Фразовый или n-граммный перевод  
  
На данном этапе исходные предложения разбиваются на более мелкие части — фразы. Их длина варьируется от отдельных выражений до более длинных последовательностей. Машина распознает и сопоставляет единицы языка на исходном и целевом языках.  
  
Подсчет вероятностей  
  
После сопоставления система вычисляет вероятность того, что фраза на одном языке будет переведена определенным образом на другом. Делается это на основе переводов в обучающих текстах. Чем чаще встречается вариант перевода, тем выше его вероятность.  
  
Языковое моделирование  
  
Затем происходит языковое моделирование получившегося контента — оценка вероятности последовательности слов в исходном и целевом языках. Это улучшает грамматическую корректность и связность сгенерированных переводов, помогает учитывать контекст и лингвистические особенности текста.  
  
Декодирование  
  
На этом этапе автоматический переводчик ищет наиболее подходящий перевод для каждой исходной фразы. Чтобы оценить несколько вариантов учитываются статистические оценки и такие алгоритмы, как лучевой поиск или динамическое программирование.  
  
Переупорядочивание и постобработка  
  
На этом этапе программа использует статистические данные из огромного количества текстов чтобы учесть особенности языков для получения нового связного текста. В разных, даже родственных языках, грамматика и порядок слов в предложении может значительно различаться.  
  
Например, в русском языке порядок слов довольно гибкий, в немецких придаточных предложениях глагол часто ставится в конец, в английском языке порядок слов строго фиксирован: подлежащее, сказуемое, дополнение. В испанском языке, например, обычно прилагательное следует за существительным, но в некоторых случаях оно может стоять перед ним, это слегка меняет заложенный смысл. Часто личные местоимения просто опускаются, поскольку лицо можно определить по форме глагола.  
  
Оценка  
  
На финальном этапе человек оценивает качество переводов статистической системы с помощью метрик, как например BLEU (Bilingual Evaluation Understudy). Результаты оценки используются для улучшения статистических моделей, обучающих данных и конфигурации системы.

Методы оценки качества машинного перевода

Для оценки общей эффективности результатов машинного перевода было разработано несколько метрик. Вот несколько часто используемых:  
  
BLEU (Bilingual Evaluation Understudy)  
BLEU является наиболее популярной метрикой, которая измеряет сходство между сгенерированным переводом и одним или несколькими эталонными переводами. Она рассчитывает точность путем оценки n-грамм (непрерывных последовательностей фраз) в переводе-кандидате и эталонных переводах. Чем выше показатели BLEU, тем лучше качество перевода.  
  
COMET (оценка машинного перевода на основе консенсуса)  
COMET — это метрика, которая принимает во внимание не только машинный перевод и эталон, но и переводимый текст-первоисточник. Оценивает адекватность и эквивалентность.  
  
METEOR (Метрика для оценки перевода с явным упорядочиванием)  
METEOR - метрика чувствительная к порядку слов и синонимам.  
  
TER (Translation Edit Rate)  
TER показывает, сколько правок нужно сделать человеку в полученном переводе, чтобы он полностью соответствовал эталонному. Эти правки могут включать добавление, удаление и замену слов, а также изменение порядка слов в предложении. Чем ниже показатели TER, тем лучше перевод.  
  
NIST (Национальный институт стандартов и технологий)  
NIST — это метрика оценки, разработанная Национальным институтом стандартов и технологий США для оценки качества автоматического машинного перевода. Она использует точность и запоминание n-грамм для оценки результатов машинного перевода по сравнению с эталонными переводами.  
  
ROUGE (Recall-Oriented Understudy for Gisting Evaluation)  
ROUGE измеряет совпадение n-грамм между переводом устройства и эталонным переводом.  
  
HTER (Human-focused Translation Edit Rate)  
HTER — метрика, сочетающая компьютерную оценку с человеческими суждениями. Она, как и TER, измеряет дистанцию редактирования между переводом, созданным системой, и эталонными переводами. Однако при этом правки сравниваются с вариантами, отредактированными переводчиками-людьми, чтобы оценить влияние на усилия по редактированию.  
  
Важно отметить, что ни одна метрика не может охватить все аспекты. Выбор метрики зависит от конкретных целей. Для получения более полной информации о качестве перевода наряду с метриками часто используются человеческий субъективный анализ.

*Нейронный машинный перевод (NMT)*

Нейронный машинный перевод (NMT) – метод машинного перевода, который завоевал популярность благодаря своей способности обучаться на больших объемах параллельных данных и использовать глубокие нейронные сети для прямого преобразования текста с одного языка на другой. Важной особенностью NMT является единая модель, отличающаяся от традиционных систем машинного перевода, таких как статистический машинный перевод (SMT), где различные компоненты модели (например, фразовые таблицы, языковые модели) работают изолированно.

В модели нейронного машинного перевода (NMT) весь процесс перевода обрабатывается единым модулем, что позволяет учитывать контекст и зависимости между словами при переводе. Это способствует более естественному и качественному переводу.

Одной из важных архитектурных концепций в NMT является энкодер-декодер. В этой архитектуре энкодер преобразует исходное предложение на одном языке в векторное представление, а декодер использует этот вектор для генерации соответствующего предложения на целевом языке. Такой подход позволяет учитывать семантические и синтаксические особенности предложений при переводе.

Для обучения нейронных машинных переводчиков (NMT) используются параллельные корпуса данных, где каждое предложение имеет свой перевод на другой язык. Эти данные играют ключевую роль в обучении моделей, позволяя им установить соответствия между предложениями на разных языках и повысить точность перевода.

Современные системы NMT часто базируются на архитектуре трансформеров, которая является эффективным и мощным подходом к обработке последовательностей данных. Трансформеры способны обрабатывать длинные текстовые последовательности и учитывать сложные взаимосвязи между словами, что приводит к улучшению качества перевода.

Одним из ключевых преимуществ Нейронной сети машинного перевода (NMT) является ее способность к адаптации и переобучению на специфические данные, а также к приспособлению к различным контекстам или областям. Это позволяет существенно улучшить качество перевода в рамках конкретных задач или сфер деятельности.

В общем, NMT представляет собой мощный и эффективный метод машинного перевода, который обеспечивает высокое качество и естественность перевода за счет применения современных архитектур нейронных сетей и методов обучения на больших объемах данных.

Плюсы и минусы

Плюсы нейронного машинного перевода (NMT):

* Высокое качество перевода: NMT обеспечивает более точные и естественные переводы благодаря использованию глубоких нейронных сетей, которые способны улавливать сложные лингвистические закономерности.
* Единая модель: в отличие от традиционных методов, в NMT весь процесс перевода обрабатывается единым модулем, что позволяет учитывать контекст и зависимости между словами, повышая качество перевода.
* Адаптивность: NMT может быть адаптирован к различным задачам и контекстам, что позволяет улучшить качество перевода для конкретных областей знаний или доменов деятельности.
* Обучение на больших данных: NMT может использовать огромные параллельные корпуса данных для обучения, что способствует более эффективному захвату разнообразия языковых особенностей и лучшему качеству перевода.
* Современные архитектуры: Использование современных архитектур нейронных сетей, таких как трансформеры, позволяет учитывать долгосрочные зависимости в тексте, что способствует улучшению качества перевода.

Минусы нейронного машинного перевода (NMT):

* Требовательность к вычислительным ресурсам: Обучение и применение NMT требует значительных вычислительных ресурсов, включая высокопроизводительные графические процессоры (GPU) или тензорные процессоры (TPU).
* Зависимость от данных: Качество перевода NMT напрямую зависит от качества и объема обучающих данных. Недостаточное количество данных или несбалансированность корпуса данных может привести к недостаточно точным результатам.
* Не всегда сохраняет структуру и стилистику: в некоторых случаях NMT может изменять структуру или стилистику оригинального текста, особенно при переводе специфических жанров или идиоматических выражений.
* Не всегда учитывает контекст: несмотря на преимущества единой модели, некоторые аспекты контекста или локальных языковых особенностей могут быть недостаточно учтены при переводе.
* Неэффективность на редких языках: для редких языков может быть сложно найти достаточное количество параллельных данных для обучения надежной модели NMT, что снижает качество перевода.

Заключение

Нейронный машинный перевод (NMT) – передовой метод автоматического перевода текста, обеспечивающий высокое качество и естественность перевода. Основанный на глубоких нейронных сетях, этот подход способен улавливать сложные лингвистические закономерности, делая перевод более точным и приближенным к человеческому.

Преимущества NMT включают высокое качество перевода, единую модель, адаптивность к различным контекстам, возможность обучения на больших объемах данных и использование современных архитектур нейронных сетей. Однако существуют некоторые недостатки: требовательность к вычислительным ресурсам, зависимость от данных, недостаточная сохранность структуры и стилистики, а также не всегда полное учет контекста.

Несмотря на некоторые ограничения, нейронные машинные переводы (NMT) остаются мощным инструментом для автоматического перевода, находя применение в различных сферах, включая перевод веб-страниц, документов, разговоров и многих других. С постоянным развитием технологий и улучшением методов обучения ожидается, что качество и эффективность NMT будут продолжать расти, делая автоматический перевод более эффективным и точным для пользователей по всему миру.

**4. Моделирование лексической и синтаксической структуры языка**

*Лексическое моделирование*

Лексическое моделирование структуры языка — это процесс анализа и описания языковых единиц (слов, фраз и их значений) с целью выявления закономерностей и структурных особенностей языка. Такие модели помогают понять, как слова взаимодействуют друг с другом и как формируется смысл.  
  
Основные аспекты:  
1. Лексическая запись — создание словарей, где фиксируются значения и употребление слов.  
2. Связи между словами — изучение синонимов, антонимов, гипонимов и гиперонимов.  
3. Контекст — анализ, как контекст влияет на значение слова.  
4. Семантические сети — графическое представление связей между словами и их значениями.  
5. Корпусный анализ — использование больших объемов текстов для выявления частотности употребления слов и устойчивых выражений.  
Такое моделирование способствует пониманию не только лексики, но и грамматических структур языка.

*Синтаксическое моделирование*

Синтаксическое моделирование структуры языка — это изучение способов, которыми слова объединяются в более крупные единицы, такие как фразы и предложения. Этот процесс помогает выявить грамматические правила и структуры, управляющие формированием предложений.  
Основные аспекты синтаксического моделирования:  
1. Грамматические категории — изучение частей речи (существительные, глаголы и т.д.) и их ролей в предложении.  
2. Структура предложений — анализ простых и сложных предложений, включая составные части (подлежащее, сказуемое).  
3. Синтаксические деревья — графическое представление иерархической структуры предложений.  
4. Синтаксические правила — формулирование правил, регулирующих порядок слов и согласование.  
5. Контекстно-свободные грамматики — использование формальных систем для описания синтаксиса языка.  
Синтаксическое моделирование важно для автоматического анализа текста и разработки языковых технологий.

**5. Оценка и тестирование математических моделей в NLP**

Оценка и тестирование математических моделей в NLP являются критическими этапами для обеспечения их эффективности и надежности. Основные методы и подходы включают:  
1. Метрики оценки  
   Используются различные метрики, такие как точность, полнота, F-мера и BLEU для оценки качества моделей. Выбор метрики зависит от конкретной задачи (например, перевод, классификация).  
2. Кросс-валидация  
   Применяется для проверки устойчивости модели и предотвращения переобучения. Данные делятся на несколько частей, и модели тестируются на разных подмножествах.  
3. Тестовые наборы  
   Создание и использование стандартизированных тестовых наборов данных позволяет объективно сравнивать качество разных моделей.  
4. Аппаратное тестирование  
   Оценка производительности модели на различных устройствах и условиях, чтобы гарантировать ее работоспособность в реальных приложениях.  
5. Анализ ошибок  
   Изучение и анализ ошибок, допущенных моделью, помогают понять ее слабые места и направляют дальнейшую оптимизацию.  
Эти методы способствуют улучшению качества и надежности моделей в области обработки естественного языка.

**6. Применение математических подходов для анализа больших текстовых данных**

Математические подходы играют ключевую роль в анализе больших текстовых данных. Вот некоторые из них:  
Векторизация текста  
   Текст преобразуется в числовые векторы с помощью методов, таких как TF-IDF или word embeddings (например, Word2Vec, GloVe). Это позволяет алгоритмам обрабатывать текстовые данные.  
 Статистические методы  
   Применение статистики для извлечения информации, выявления паттернов и тенденций. Например, использование регрессионного анализа для прогнозирования на основе текстовых данных.  
 Классификация и кластеризация  
   Алгоритмы машинного обучения (например, SVM, Random Forest) применяются для классификации текстов или их кластеризации, что помогает в сегментации пользователей или тем.  
Тематика и частотный анализ  
   Основывается на анализе частоты слов и их сочетаний для выявления основных тем и рекомендаций.  
 Сети и графы  
   Математическое моделирование связанных текстовых данных с помощью графов позволяет анализировать структуру данных и выявлять связи между понятиями.  
Эти подходы позволяют получать ценные инсайды из больших объемов текстовой информации и поддерживать принятие обоснованных решений.

**Практическая часть**

**Источники:** <https://lingvanex.com/ru/blog/statistical-machine-translation-what-is-it/>

<https://translated.turbopages.org/proxy_u/en-ru.ru.03c9fbfc-681f1e4a-ba4ef991-74722d776562/https/www.baeldung.com/cs/rnns-transformers-nlp> <https://na-journal.ru/5-2024-informacionnye-tekhnologii/11496-neironnyi-mashinnyi-perevod?t>