ПРОЕКТИРОВАНИЕ СЕРВИСА МАШИННОГО ПЕРЕВОДА С ИСПОЛЬЗОВАНИЕМ МЕХАНИЗМА ВНИМАНИЯ

А. С. Соломевич

Белорусский государственный университет, пр. Независимости, 4, 220030, г. Минск, Беларусь, alexandr.solomevich@gmail.com
Научный руководитель – В. И. Малюгин, доктор экономических наук, доцент

Данная работа посвящена экспериментальному сравнению различных моделей и методов работы с текстами в рамках задачи машинного перевода, оценке эффективности и выявлению их преимуществ и недостатков. По результатам экспериментов отобрана лучшая из версий моделей и интегрирована в веб-сервис, который предоставит возможность каждому пользователю удобно и быстро получать переводы.

Ключевые слова: обработка искусственного языка; машинный перевод текста; рекуррентная нейронная сеть; долгая краткосрочная память; механизм внимания; трансформер; PyTorch.

1. Введение в машинный перевод, цели исследования

Машинный перевод представляет собой важную область исследований, цель которой – автоматизировать процесс перевода текстов и речи с одного языка на другой. Этот процесс играет ключевую роль в устранении языковых барьеров и способствует улучшению глобальной коммуникации.

Целями данной работы являются практическое сравнение различных методологий и подходов для задачи машинного перевода, а также получение готового сервиса для перевода, которым смогут использовать для общения разные языковые группы.

Для достижения цели исследования решались следующие задачи:

1) формулировка актуальных технологий подготовки и обработки текстовых данных для решения задачи машинного перевода;

2) сбор и подготовка данных для обучения моделей, а также проведение сравнительного анализа различных архитектур нейронных сетей;

3) проектирование веб-сервиса, для доступа к разработанным моделям перевода.

2. Представление и обработка текста

Первым этапом решения задачи машинного перевода является получение так называемых эмбеддингов из токенов текста, то есть их векторных представлений. **Word2Vec** – это метод, используемый для преобразования слов в числовые векторы в многомерном пространстве. Основная идея заключается в том, что семантически близкие слова будут иметь близкие векторы. Continuous Bag of Words (CBOW) и Skip-Gram [1] – две основные модели, используемые

в Word2Vec. Целевая функция CBOW определяет вероятность появления целевого слова при условии контекстных слов, а Skip-Gram отличается тем, что предназначен для предсказания окна контекстных слов по заданному целевому слову.

Как и в любой задаче машинного обучения, важным этапом является подготовка данных [2]. Ключевыми компонентами данного этапа можно выделить токенизацию, управление стоп-словами и фильтрацию шума.

Токенизация представляет собой процесс разбиения текста на отдельные единицы, называемые токенами. Токенизацию можно рассматривать как процесс разбиения предложений в тексте на осмысленные единицы. На сегодняшний день существует множество способов токенизации, в сравнительном анализе были использованы токенизация по отдельным словам и BPE (Byte Pair Encoding) токенизация.

Стоп-слова подразумевают часто встречающиеся слова, которые, как правило, не несут значительной смысловой нагрузки и могут быть удалены из текста без потери его глобального смысла. Исключение стоп-слов позволяет моделям сосредоточиться на более значимых терминах, что улучшает точность перевода.

Фильтрация шума включает удаление ненужных символов, специальных знаков, цифр и других элементов, которые могут вызвать путаницу в процессе обучения. Это особенно актуально при работе с текстом, полученным из интернета, где часто встречается множество шумовых элементов.

3. Архитектуры нейронных сетей для задач Seq2Seq и их сравнительный анализ

Задача машинного перевода подразумевает работу с последовательными данными. Для данного типа задач существует множество вариантов архитектур нейросетей, таких как рекуррентные нейронные сети [3], LSTM (Long Short-Term Memory) сети [4], GRU (Gated Recurrent Unit). Помимо данных классических архитектур особое внимание было уделено механизму внимания и использующей данный механизм архитектуре трансформер [5].

Для проведения обучения моделей [6] и их сравнительного анализа был собран датасет, представляющий параллельные корпуса текстов на русском и английском языках, в размере 2-х миллионов записей. Далее данные были подвергнуты обработке, которая помогла снизить размерность данных и улучшить качество перевода, так как модели стали обрабатывать более чистый и структурированный текст. После тщательной фильтрации, размер датасета уменьшился до 505-ти тысяч пар, которые оказались наиболее пригодными для обучения и валидации моделей.

Для оценки качества перевода использовалась метрика BLEU. BLEU вычисляет сходство машинного перевода с эталонными переводами на основе подсчета n-грамм (последовательностей слов).

По результатам проведенных исследований была составлена таблица, демонстрирующая финальные характеристики для пула лучших моделей на основе каждого из видов архитектур.

Результаты различных моделей на тестовой выборке

Model	Tokenization	BLEU Score	Epoch Training Time (sec)
RNN	Words	0,09	951
RNN	BPE	0,12	964
GRU	Words	0,14	937
GRU	BPE	0,175	996
LSTM	Words	0,18	989
LSTM	BPE	0,19	1045
Transformer	Words	0,24	1337
Transformer	BPE	0,28	1396

В результате сравнения метрик качества видно, что трансформеры превосходят рекуррентные архитектуры в большинстве случаев. Их способность эффективно обрабатывать длинные последовательности и улавливать долгосрочные зависимости в тексте привела к значительному улучшению качества перевода [7].

Немаловажно отметить влияние механизма внимания. Благодаря данной технологии модель фокусируется на ключевых элементах исходного текста при формировании соответствующего перевода. Пример работы механизма внимания показан на Рисунке 1:

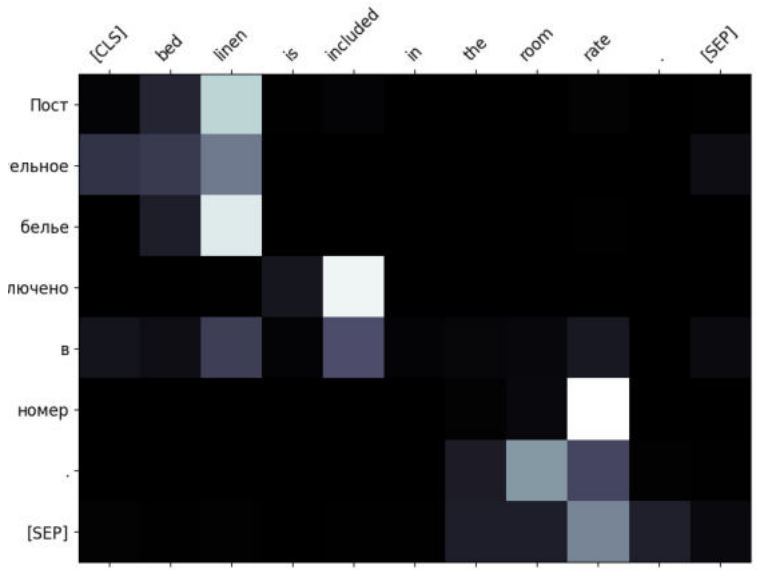

Рис. 1. Влияние механизма внимания

4. Проектирование веб-сервиса

В качестве веб-сервиса было принято решение реализовать Telegram-бот, который предоставит пользователям удобный интерфейс и позволит им получать переведенные результаты напрямую в мессенджере. Пример работы бота, а также результаты простейших запросов на разных языках показаны на Рисунке 2:

Рис. 2. Интерфейс разработанного бота

Одним из преимуществ разработанного бота является то, что его можно добавить в группы. Так участники могут отправлять текст для перевода прямо в чат, и бот будет предоставлять переводы, которые сразу же доступны всем членам группы. Это особенно полезно для международных команд, образовательных групп или любых других сообществ, где важно быстро переводить информацию между различными языками.

5. Заключение

В работе рассмотрены различные аспекты задачи машинного перевода. Проведён сравнительный анализ различных архитектур и методов, а также их производительности в разных сценариях машинного перевода. В результате этих экспериментов, трансформер с BPE токенизацией показал наилучшие результаты, достигнув 28.46 по метрике BLEU на тестовой выборке. Этот результат продемонстрировал превосходство трансформерной архитектуры в задачах машинного перевода. Финальная архитектура была внедрена в веб-сервис в виде телеграм-бота, который предоставляет удобный интерфейс для взаимодействия пользователей с моделью.

Таким образом, в работе получены следующие основные результаты:

1) проведен анализ теоретических аспектов и современных подходов в области машинного перевода;

2) на основе экспериментальных исследований продемонстрирована работа различных архитектур, а также выявлена лучшая из них;

3) разработаны алгоритмические и программные средства для возможности обучения, тестирования и эксплуатации моделей машинного перевода.

Библиографические ссылки

1. *Brown P. F.* The Mathematics of Statistical Machine Translation: Parameter Estimation / P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, R. L. Mercer. – Cambridge, MA, 1993. – P. 263-311.

2. Автоматическая обработка текстов на естественном языке и анализ данных: учеб. пособие / Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С. – М.: Изд-во НИУ ВШЭ, 2017. – 269 с.

3. *Осовский С.* Нейронные сети для обработки информации / пер. с польск. М.: Финансы и статистика, 2002. 344 с.

4. Long Short-Term Memory [Electronic resource]. URL: https://en.wikipedia.org/wiki/Long_short-term_memory (date of access: 17.03.2024).

5. *Vaswani A.* Attention Is All You Need / N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N Gomez, Ł. Kaiser, I. Polosukhin // In Advances in Neural Information Processing Systems (NIPS'17): Proceedings of the 30st International Conference, December 4–9, 2017, Long Beach, CA, USA. – P. 6000-6010.

6. *Sutskever I.* Sequence to Sequence Learning with Neural Networks / O. Vinyals, Q. V. Le // In Advances in Neural Information Processing Systems (NIPS'14): Proceedings of the 27th International Conference, December 8–13, 2014, Montreal Convention Center, Montreal, Canada. – Vol. 2. – P. 3104-3112.

7. G. Foster // In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, Copenhagen, Denmark. – P. 2486-2496.