Проверка фактографической датированной информации в нарративных текстах

Ловягин Андрей Андреевич

Студент

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: andrey.lovyagin.work@gmail.com

Научный руководитель *— Добров Борис Викторович*

В современном информационном обществе актуальность и достоверность исторических данных приобретают высокую значимость, усиливая потребность в разработке эффективных систем верификации. Настоящее исследование реагирует на проблематику роста объемов дезинформации, предлагая автоматизированный метод верификации фактов. В историческом контексте такая задача традиционно требовала значительных усилий и времени. Предложенный подход основывается на применении современных алгоритмов машинного обучения и обработки естественного языка, что позволяет эффективно обрабатывать неоднозначности и контекстуальное разнообразие текстов. Это значительно увеличивает точность и сокращает время обработки, делая метод важным инструментом в академической и общественной сферах.

В рамках исследования были собраны и стандартизированы обширные наборы данных из исторических учебников, книг и научных работ, общим объемом примерно 30 гигабайтов сырого материала и около 1 гигабайта очищенного текста. Разработанные методы включают алгоритмы для выделения и стандартизации дат, а также контекстный анализ с использованием эмбеддингов. Были разработаны и апробированы три различных подхода к векторному анализу, включая учет и стандартизацию временных периодов, подход без стандартизации дат и подход, исключающий использование дат. Выделение дат осуществлялось с применением обширного набора регулярных выражений, а для векторизации предложений использовались предобученные мультиязычные модели Sentence Transformers [1]. Выделенные предложения с датами сопоставлялись с базой данных через приближенный поиск ближайших соседей, реализованный через NMSLIB [2]. Общая схема блока проверки приведена на Рис. 1.

Подход был сравнен с результатами работы алгоритма от команды DeepPavlov, который использует синтаксическое дерево для выделения связи между предложениями и датами, демонстрируя зна-

чительные улучшения в точности и скорости обработки данных. На открытых данных соревнования «ПРО//ЧТЕНИЕ» [3], направленного на разработку алгоритмов автоматической проверки сочинений ЕГЭ, качество предложенного подхода достигло 0.76 F1-меры, 0.78 Recall и 0.76 Precision, в сравнении с 0.21 F1-мерой, 0.12 Recall и 0.8 Precision у команды DeepPavlov. Благодаря быстрому выделению дат и полной векторизации предложений, время обработки алгоритмом 27.000 сочинений составило всего 3 минуты против 128 минут.

Следует отметить, что несмотря на ограничения, связанные с проверкой датированных событий, предложенный метод не требует сложной или ручной разметки базы данных и позволяет проверять тысячи текстов в сжатые сроки с приемлемым уровнем качества. Результаты подчеркивают важность развития методов автоматизированной проверки для повышения достоверности образовательных и научных материалов, а также открывают перспективы для дальнейшего развития данной области.

Иллюстрации

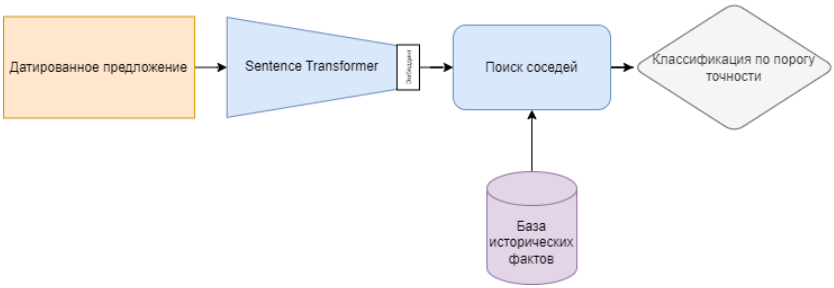

Рис. 1. Общая схема блока проверки датированного предложения.

Литература

1. Reimers N., Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks // In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, 2019, P. 3982–3993.
2. Boytsov L., Naidan B. Engineering Efficient and Effective Non-metric Space Library // In Proceedings of the Similarity Search and Applications - 6th International Conference, 2013, P. 280–293.
3. Страница конкурса «ПРО//ЧТЕНИЕ»: `https://ai.upgreat.one`