

УДК 004.896

Оценка семантического сходства текстов методами машинного обучения

Корлякова Мария Олеговна

mkorlyakova@yandex.ru

Васенков Павел Андреевич

pavelvasenkov@yandex.ru

КФ МГТУ им. Н.Э. Баумана, Калуга, Россия

Рассмотрены основные аспекты метода семантического поиска, проведен фантинг модели RuBERT. Изложена концепция пуллинга для уменьшения размерности векторов. Был выбран оптимальный датасет и проведена подготовка данных. Также были выбраны и предложены метрики для анализа качества обучаемой модели. Результаты были представлены в виде таблиц, был проведен анализ, выявлены недостатки предобученной модели BERT при работе с ассиметричными данными. Написан пайплайн семантического анализа текстовой идентичности. Проведенные исследования могут быть полезны в сфере обработки естественного языка.

Ключевые слова: обработка естественного языка, машинное обучение, семантический анализ, искусственный интеллект

Оценка семантического текстового сходства (англ. Semantic Textual Similarity) информации широко применяется в сфере обработки естественного языка (англ. Natural Language Processing). В рамках поставленной задачи было необходимо провести анализ алгоритма для семантического анализа текстов, а также выполнить программную реализацию и провести тесты.

Семантический поиск — это метод поиска информации, который учитывает смысл и контекст запросов. В отличие от традиционных (лексических) методов, которые опираются на точные соответствия между словами в запросе и документами, семантический поиск использует технологии обработки естественного языка (NLP) и машинного обучения для понимания намерений пользователя и контекста запроса. Модели, разработанные для этой задачи, могут оценивать степень семантической близости между текстовой информацией [1].

Основная концепция данного метода заключается в представлении текстовой информации в виде многомерных векторов, что, в свою очередь, позволяет выявить семантические связи между словами. Вектор формируется как усреднение всех векторных представлений токенов с помощью масок внимания, что позволяет исключить токены, не являющиеся значимыми.

В качестве метрики сходства используем косинусное сходство (англ. cosine similarity) [2]

$$\text{similarity}(A, B) = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n A_i^2} \times \sqrt{\sum_{i=1}^n B_i^2}}$$

Косинусная близость измеряет угол между векторами, а не их длину, что во время работы оказалось полезно при анализе текстов разной длины.

Данная метрика отличается хорошей интерпретируемостью, поскольку значения варьируются от -1 до 1 , где 1 — полная схожесть, 0 — отсутствие схожести, -1 — противоположенное направление вектора. Помимо этого, косинусная близость может быть менее чувствительна к шуму в данных по сравнению с другими метриками.

Для понижения размерности входного вектора был использован алгоритм среднего пулинга (англ. mean pooling) (рис. 1). Для проведения последующих исследований была выбрана дообученная модель трансформера [3] BERT — RuBERT [4], содержащая 180 млн параметров, 12 слоев и 768 скрытых слоев. Использование на новых данных позволило оценить смысловую идентичность текстов, модель продемонстрировала высокие результаты во время экспериментов (табл. 1). Однако при тесте на ассиметричных данных, модель показала неточные результаты (табл. 2) [5].

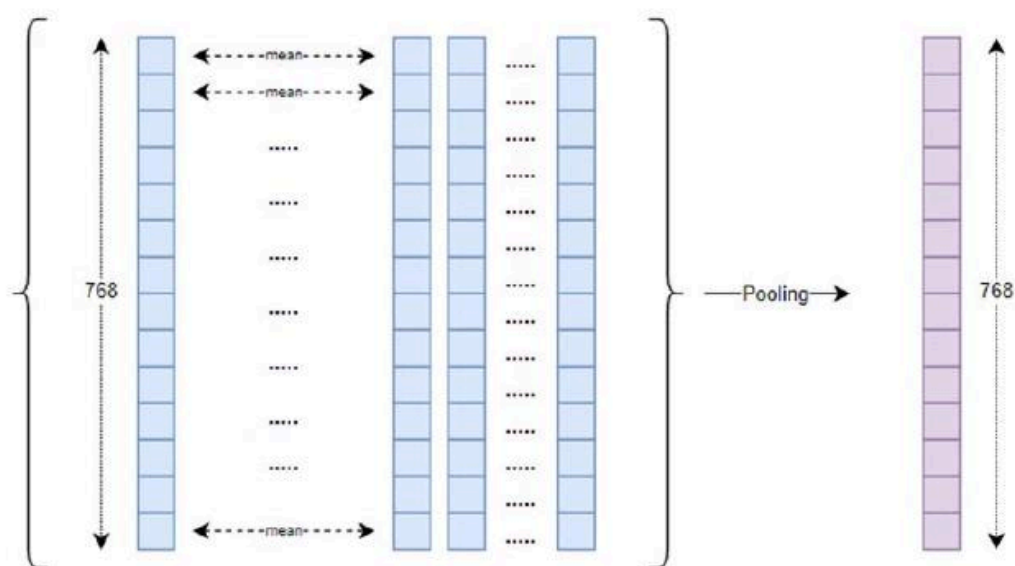


Рис. 1. Схема среднего пулинга

Таблица 1

Сравнительный анализ сходства с помощью предобученной модели

№	Текст 1	Текст 2	Вероятность сходства
1	Python — популярный язык программирования	Один из известных языков — это Python	0,896
2	Чтение — мое любимое занятие в свободное от работы время	В свободное время мне нравится читать книги	0,729
3	По выходным я часто смотрю фильмы и отдыхаю дома	Не ходите дети в Африку гулять	0,204
4	Я предпочитаю кофе утром, чтобы зарядиться энергией	Человек — самое разумное существо на планете	0,126

С целью повышения производительности модели было принято решение провести дообучение (файнтюнинг) модели на ассиметричных данных [5].

Таблица 2

Сравнительный анализ сходства с помощью файнтюненной модели

№	Текст 1	Текст 2	Вероятность сходства
1	Я люблю гулять по парку в солнечные дни	В ясную погоду, мне не нравится проводить время у экрана, я люблю проводить его на свежем воздухе, гуляя на природе и любуясь красотами окружающего мира	0,311
2	Мне нравится готовить русскую кухню	С самого детства готовка для меня — это не просто необходимость, а настоящее увлечение	0,214
3	Алексей часто слушает музыку	Музыка сопровождает Алексея на протяжении всего дня, создавая нужное настроение и настраивая на работу	0,267
4	Амурский тигр — редкое животное	Амурский тигр занесен в Красную книгу Российской Федерации, а также является подвидом, сохранившимся только на территории России	0,193

С целью повышения производительности модели было принято решение провести дообучение (файнтюнинг) модели на ассиметричных данных. Для этих целей, в качестве обучаемой архитектурой была выбрана SBERT, которая является би-энкодером (рис. 2). Данная модификация использует сиамские и триплетные сетевые структуры для получения семантически значимых вложений предложений, которые можно сравнивать с использованием косинусного сходства. Для этих целей была использована русскоязычная часть датасета Microsoft Machine Reading Comprehension. MS MARCO [6] — набор данных, предназначенный для глубокого обучения в задачах обработки и анализа естественного языка. Поскольку сборка содержит разнородные данные, полученные из разных источников, он идеально подходит для файнтюнинга [1]. Количество входных токенов в нашей задаче всегда соответствует количеству векторов. Результирующий тензор имеет размерность 1 на 768.

Обучение позволило сблизить семантически схожие вектора, а также отдалить различные вектора в гиперпространстве [7]. Для оценки качества обученной модели были использованы метрики качества ранжирования MRR, NDCG, а также MAP.

Mean Reciprocal Rank (MRR) — это статистический показатель, используемый для оценки процессов, которые генерируют список потенциальных ответов на набор запросов, отсортированных по вероятности их правильно-

сти. Метрика изменяется в диапазоне от 0 до 1, а также учитывает позицию элементов [2]:

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i}.$$

Mean Average Precision (MAP) — это среднее значение средней точности для каждого запроса. MAP учитывает точность и порядок возвращаемых результатов, благодаря чему показала высокие результаты оценки в процессе экспериментов [2]:

$$\text{MAP} = \frac{\sum_{q=1}^Q \text{Ave P}(q)}{Q}.$$

Normalized Discounted Cumulative Gain (NDCG) — нормализованная версия DCG (Discounted Cumulative Gain). Данная метрика принимает значения от 0 до 1, а также наследует от DCG учет позиции элементов в списке [2]:

$$\text{nDCG}_p = \frac{\text{DCG}_p}{\text{IDCG}_p}.$$

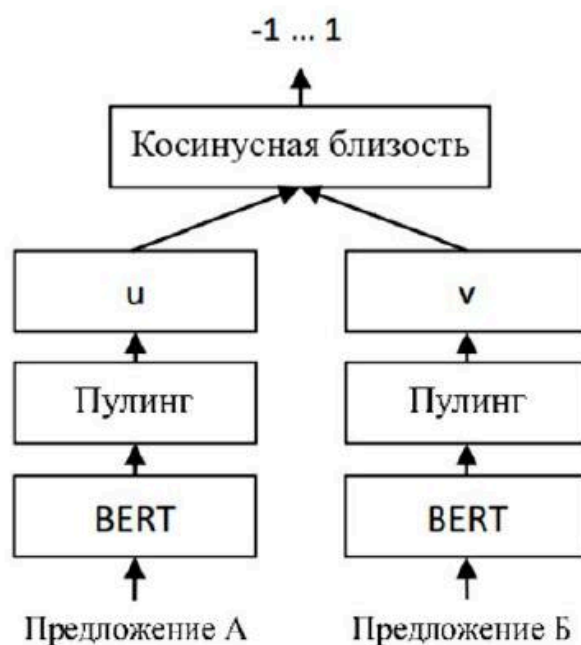


Рис. 2. Архитектура би-энкодера

В результате была обучена и была проведена валидация на 50 тысячах данных, после чего был разработан пайплайн (англ. pipeline) для решения задачи анализа семантического сходства текстов [7].

Таблица 3

Результат семантического анализа текста файнтюниной моделью

№	Текст 1	Текст 2	Вероятность сходства
1	Елена любит читать книги	Чтение книг — любимое занятие Елены, которое приносит ей радость и вдохновение	0,879
2	Он пошел в магазин	В один прекрасный день он решил отправиться в ближайший магазин, чтобы купить необходимые продукты	0,827
3	Дождь пошел	Внезапно небо затянулось тучами, и начался дождь, который наполнил улицы свежестью	0,695
4	Дорога к морю была долгой и утомительной	В году мы решили провести отпуск на море, но путь был нелегким, долгим и изнуряющим	0,783

Таким образом, был проведен анализ метода анализа семантического сходства текстов в процессе файнтюнинга модели RuBERT, рассмотрены и проанализированы алгоритмы и метрики для оценки качества данного процесса.

Литература

- [1] Pre-training of Deep Bidirectional Transformers for Language Understanding. URL: <https://arxiv.org/pdf/1810.04805> (accessed November 10, 2024).
- [2] Bishop C.M. Pattern recognition and machine learning. Springer Science+Business Media, 2006.
- [3] Transformers. URL: <https://huggingface.co/docs/transformers/main/en/index> (accessed November 10, 2024).
- [4] DeepPavlov/rubert-base-cased. URL: <https://huggingface.co/DeepPavlov/rubert-base-cased> (accessed November 10, 2024).
- [5] A Multilingual Version of the MS MARCO Passage Ranking Dataset. URL: <https://arxiv.org/pdf/2108.13897> (accessed November 10, 2024).
- [6] Unicamp-dl/mmario. URL: <https://huggingface.co/datasets/unicamp-dl/mmario> (accessed November 10, 2024).
- [7] Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. URL: <https://arxiv.org/pdf/1908.10084> (accessed November 10, 2024).

Evaluation of semantic similarity of texts using machine learning methods

Korlyakova M.O.

mkorlyakova@yandex.ru

Vasenkov P.A.

pavelvasenkov@yandex.ru

BMSTU (KB), Kaluga, Russia

The main aspects of the semantic search method are considered, the RuBERT model is fine-tuned. The concept of pooling for reducing the dimensionality of vectors is presented. The optimal dataset was selected and the data was prepared. Metrics for analyzing the quality of the trained model were also selected and proposed. The results were presented in tables, an analysis was carried out, and the shortcomings of the pre-trained BERT model when working with asymmetric data were identified. A pipeline for semantic analysis of text identity was written. The conducted research can be useful in the field of natural language processing.

Keywords: natural language processing, machine learning, semantic analysis, artificial intelligence