Построение системы поиска похожих статей

Д.С. Баканов

Самарский национальный исследовательский университет имени академика С.П. Королева, Самара, Россия

Обоснование. В наши дни новостных статей становится все больше. Количество тематик таких статей увеличивается экспоненциально, поэтому тяжело отслеживать интересующие новости и искать их вручную. Таким образом, возникла нужда в автоматизации данного процесса. Однако сложность в решении данной задачи заключается в том, что данные статьи написаны на естественном языке.

Цель — спроектировать и разработать систему, которая по просмотренной пользователем новостной статье выдавала бы релевантные.

Методы. Для построения системы семантического поиска необходимо обучить модель, которая бы переводила все документы в векторное представление. В качестве моделей для векторизации будут выступать языковые модели BERT (модель кодирующего трансформера) [1], RoBERTa (улучшенная BERT модель) [2] и MBart (модель трансформера кодировщик-декодировщик, которая хорошо себя показала на задачах отчистке текста и автореферирования) [3]. Для обучения модели векторизации был собран корпус новостных статей из разных источников: Lenta.ru, Russia Today и РИА Новости. Всего было собрано более 831 тыс. статей. Основной особенностью новостных статей является то, что при их заведении авторы и редакторы определяют тематику, рубрику и порой составляют список ключевых слов. Используя данную информацию, можно сделать автоматическую маркировку. Для обучения моделей была использована архитектура сиамской сети, которая имеет два идентичных подмодуля и единый агрегирующий вывод [4]. Модели с данной архитектурой будут тренироваться на трех задачах определения семантической близости:

1. Близость текстов на основе того, что каждый из них является частью одного большого. Тексты статей могут составлять части одного большого цикла, тем самым они дополняют идеи и мысли друг друга в рамках одной доменной области. К примеру, цикл статей, посвященных Великой Отечественной войне, в каждой из которых затрагивается определенный год. В результате все тексты разбиваются на части. Положительные примеры — части, взятые из одного текста, негативные — из разных.

2. Семантическая близость текстов к конкретному ключевому слову. Чаще всего для поиска статей пользователь вводит ключевое слово и хочет получить релевантные статьи. Поэтому набор для обучения формируется следующим образом: 1 помечаются пары — статья и ключевое слово из нее, 0 — статья и ключевое слово из другой.

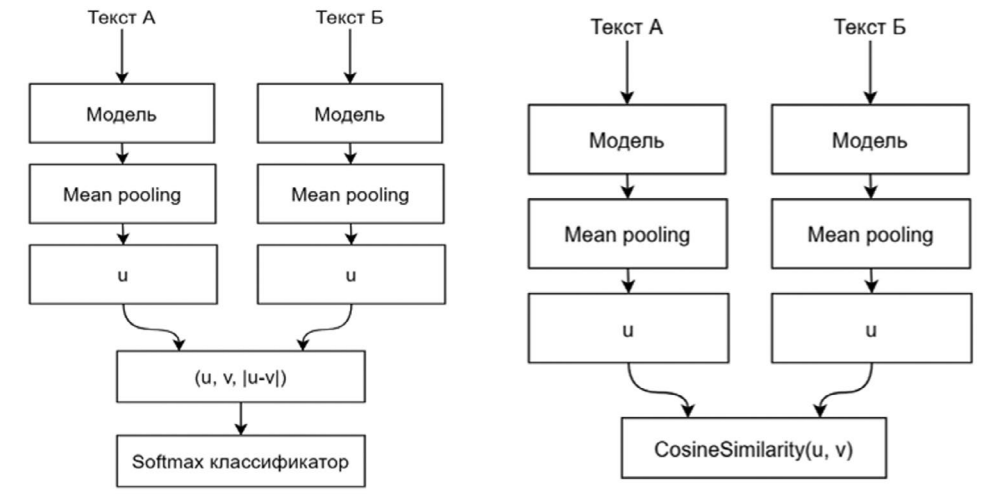
3. Близость текстов на основе пересечения множеств ключевых слов. В качестве меры близости используется следующая величина: $\mathrm{sim}(i, j) = |K_i \cap K_j| \,/\, |K_i \cup K_j|$, где K_i — множество ключевых слов каждой из статей.

Поскольку 1-я и 2-я задачи — это задачи бинарной классификации, а 3-я — регрессия, в ходе которой вычисляется значение в диапазоне от 0 до 1, то для них требуются специфичные архитектуры сетей, которые представлены на рисунке 1.

Результаты. Модели были обучены на каждой из задач определения семантической близости статей. Результаты показаны в таблице 1.

Таблица 1. Точность языковых моделей на задачах

Задача / Модель	Близость на основе встречаемости текстов в одной статье (средняя F1-мера макро)	Близость на основе сопоставления с ключевым словом (средняя F1-мера макро)	Близость на основе пересечения множества ключевых слов (RMSE)
BERT	0,95	0,5	0,1
RoBERTa	0,67	0,33	0,16
MBart	0,67	0,48	0,14



а) архитектура сиамской сети для определения близости текстов как частей целого и сопоставления с ключевым словом

б) архитектура сиамской сети для определения схожести текстов на основе пересечения множеств ключевых слов

Рис. 1. Архитектуры сиамской сети для каждой задачи

Можно видеть, что модель BERT для векторизации статей справляется лучше всех, поэтому для дальнейшего построения системы была использована она.

Для хранения векторов и проведения операции над ними была применена специальная векторная база данных. В качестве таковой выступила ChromaDB — быстроразвивающийся проект векторной базы данных с открытым исходным кодом. Поиск документов по векторам осуществляется при помощи алгоритма приближенного поиска ближайших соседей. Векторное пространство документов в векторной базе данных образует граф, используя алгоритм HNSW [5].

Сам сервис является веб-приложением, которое в качестве запроса от пользователя принимает статью: тему (заголовок) и основной текст. Данная информация векторизуется обученной моделью BERT, и по полученному вектору из базы данных получаем список похожих статей и значение близости. Список, отранжированный по близости, выдается пользователю. Данный процесс показан на рисунке 2.

Выводы. В результате были обучены модели векторизации документов, которые с помощью использования технологии трансформеров показали достаточно хорошую точность, и построена система, которая может быть использована для получения персональной выборки по просмотренной статье.

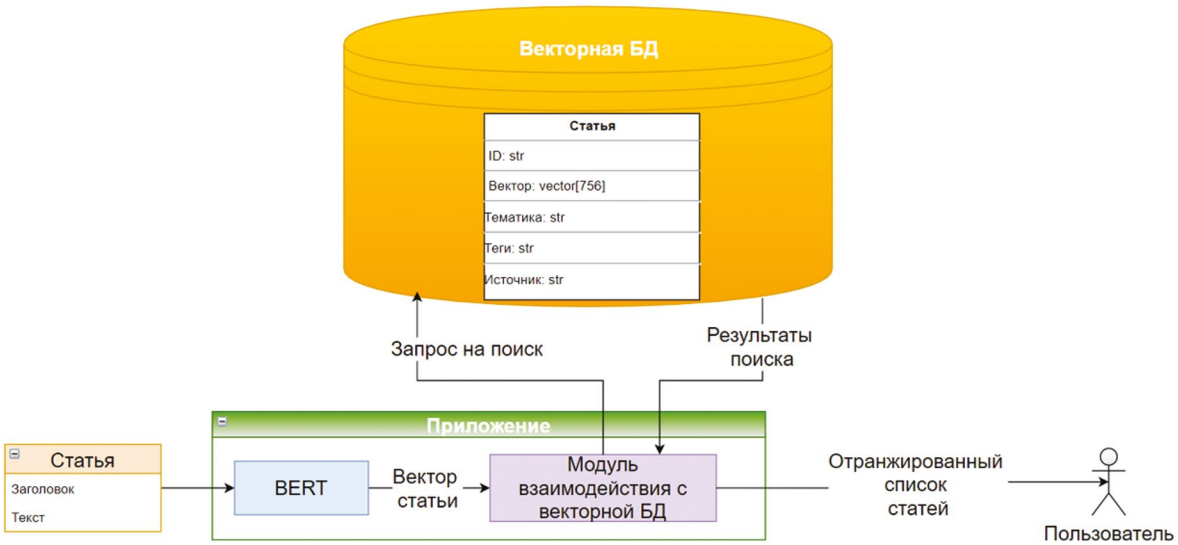


Рис. 2. Схематичное описание работы системы

Ключевые слова: машинное обучение; информационный поиск; сиамская модель; языковая модель; векторная база данных.

Список литературы

1. Devlin J., Chang M.-W., Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding // arXiv. 2018. ID 04805. doi: 10.48550/arXiv.1810.04805
2. Liu Y., Ott M., Goyal N., et al. RoBERTa: A robustly optimized BERT pretraining approach // arXiv. 2019. ID 11692. doi: 10.48550/arXiv.1907.11692
3. Liu Y., Gu J., Goyal N., et al. Multilingual denoising pre-training for neural machine translation // arXiv. 2020. ID 08210. doi: 10.48550/arXiv.2001.08210
4. Reimers N., Gurevych I. Sentence-BERT: Sentence embeddings using siamese BERT-networks // arXiv. 2019. ID 10084. doi: 10.48550/arXiv.1908.10084
5. docs.trychroma.com [Электронный ресурс]. Chroma docs [дата обращения: 05.04.2024]. Режим доступа: https://docs.trychroma.com/

Сведения об авторе:

Дмитрий Сергеевич Баканов — аспирант, кафедра «Технической кибернетики»; Самарский национальный исследовательский университет им. С.П. Королева, Самара, Россия. E-mail: dima.bakanov.1999@mail.ru

Сведения о научном руководителе:

Александр Викторович Куприянов — доктор технических наук, доцент; заведующий кафедрой технической кибернетики; Самарский национальный исследовательский университет им. С.П. Королева, Самара, Россия. E-mail: akupr@ssau.ru