Петрова А. Н., Фролов Д. О.
A. N. Petrova, D. O. Frolov

ИСПОЛЬЗОВАНИЕ ПОИСКА ФОНОВЫХ ССЫЛОК ПО СМЫСЛУ В СИСТЕМАХ БОЛЬШИХ ДАННЫХ

USE OF MEANING-BASED BACKGROUND REFERENCE SEARCH IN BIG DATA SYSTEMS

Петрова Анна Николаевна – кандидат технических наук, заведующая кафедрой «Проектирование, управление и развитие информационных систем» Комсомольского-на-Амуре государственного университета (Россия, Комсомольск-на-Амуре). E-mail: PetrovaAN2006@yandex.ru.
Anna N. Petrova – PhD in Engineering, Head of the Department «Design, Management and Development of Information Systems», Komsomolsk-na-Amure State University (Russia, Komsomolsk-on-Amur). E-mail: PetrovaAN2006@yandex.ru.

Фролов Дмитрий Олегович – аспирант Комсомольского-на-Амуре государственного университета (Россия, Комсомольск-на-Амуре). E-mail: optcompanys@mail.ru.
Dmitriy O. Frolov – Graduate Student, Komsomolsk-na-Amure State University (Russia, Komsomolsk-on-Amur). E-mail: optcompanys@mail.ru.

Аннотация. Задача фонового связывания направлена на рекомендацию новостных статей для читателя, которые наиболее релевантны для предоставления контекста и предыстории для статьи запроса. Для выполнения этой задачи предложен двухэтапный подход IR-BERT, который сочетает в себе поисковую мощность BM25 с контекстным пониманием получения с помощью модели на основе BERT. Далее предложили использовать меры разнообразия для оценки эффективности подходов к связыванию фона при извлечении разнообразного набора документов. Приведено сравнение IR-BERT с другими подходами в TREC 2023.

Summary. The background linking task aims to recommend news articles to the reader that are most relevant to provide context and background for the query article. To fulfill this task, a two-stage IR-BERT approach is proposed that combines the search power of BM25 with the contextual understanding of retrieval using a BERT-based model. The authors also proposed the use of diversity measures to evaluate the performance of background linking approaches in retrieving a diverse set of documents. A comparison of IR-BERT with other approaches in TREC 2023 is given.

Ключевые слова: фоновое связывание, двухэтапный подход, семантический поиск.

Key words: background linking, two-step approach, semantic search.

УДК 517.95

Введение. Интернет-службы новостей стали ключевыми источниками информации и повлияли на то, как мы потребляем и делимся новостями. При составлении новостной статьи часто предполагается, что читатель имеет достаточную информацию о случившихся событиях. Это не всегда так, что требует необходимости предоставить читателю ссылки на полезную информацию, которые могут установить полную картину событий. Новостная информация может быть написана разными авторами до или после выпущенной статьи и служить для предоставления дополнительной информации о статье, что даёт возможность познакомить читателя с ключевыми идеями. Однако определение того, что можно отнести к статье, предоставление фонового контекста и получение таких документов не является простым. В связи с этой проблемой была решена задача фонового связывания. Целью является получение списка публикаций, которые можно включить в поле «Пояснение» рядом с текущей статьёй, чтобы помочь читателю понять или узнать больше об истории или основных вопросах. В данной работе предложен двухэтапный подход IR-BERT для решения проблемы фоновых ссылок. Первый этап фильтрует список для идентификации набора

документов-кандидатов, которые имеют отношение к рассматриваемой статье. Это достигается путём объединения взвешенных ключевых слов, извлечённых из запроса документа в эффективный поисковый запрос, и использования BM25 для поиска в корпусе. На втором этапе используется Sentence-BERT для изучения контекстных представлений запроса в проведении семантического поиска по кандидатам, включённым в список. Мы предполагаем, что использование языковой модели может быть полезным для понимания контекста статьи-запроса и поможет найти публикации, которые предоставляют полезную информацию для текущей статьи.

Инструменты для работы. BM25 – одна из самых популярных функций ранжирования, используемых поисковыми системами для оценки релевантности документов по заданному поисковому запросу. BM25 основан на функции поиска «bag-of-words», которая ранжирует набор документов на основании терминов запроса, встречающихся в каждом документе независимо от их близости внутри документа. Несколько предыдущих подходов для поиска фоновых ссылок построено с использованием BM25.

Данная задача очень актуальна для специального поиска, в котором запросы написаны на естественных языках. Но есть основные проблемы с использованием BERT для поиска семантического сходства между текстовыми документами. Во-первых, чтобы сравнить пару документов, оба должны быть введены в модель, что приводит к значительным вычислительным затратам во время вывода. Во-вторых, для решения задачи семантического поиска широко используется подход, суть которого заключается в отображении документов в векторное пространство, где подобные документы ближе. Общие практики, такие как усреднение выходного слоя BERT или использование выходных данных [CLS] токена, дают плохие вложения. BERT выводит семантически значимые вложения предложений, которые можно сравнить, используя косинусное подобие. В этом докладе используем Sentence-BERT как часть архитектуры.

Методология. Задачу фонового связывания можно сформулировать следующим образом: учитывая новостную статью S и набор новостных статей A, получить другие новостные статьи из A, которые содержат важный контекст или справочную информацию о S. Эту задачу разумно рассматривать как частный случай новостей, рекомендацию, направленную на получение соответствующих статей из набора новостных статей A для запроса, созданного на основе статьи S. Мы предполагаем, что большинство статей, которые могут предоставить контекстную информацию о статье-запросе, вероятно, были опубликованы перед этим. С этой целью мы отфильтровываем прямые ссылки из полученных результатов, т. е. статьи, опубликованные после статьи запроса, не рассматриваются.

IR-BERT пытается решить проблему фонового связывания в два этапа. На первом этапе строим взвешенный запрос Q_i из статьи S_i и используем BM25 для получения p набора кандидатов документа. Пусть этот набор документов будет $R_b^i = \{d_1, d_2, d_p\}$, где $|R_b^i| = p$. На втором этапе мы проводим смысловой поиск Q_i по множеству найденных документов R_b^i, чтобы получить окончательный комплект документов $R_f^i = \{d_1, d_2, d_t\}$, где $|R_f^i| = t$. На рис. 1 показаны два этапа IR-BERT.

Взвешенный поисковый запрос и BM25. Сначала мы создаём эффективный поисковый запрос, который лучше всего отражает соответствующие темы статьи-запроса. Проблема сформулирована как извлечение основных ключевых слов из статьи запроса с присвоением им веса в соответствии с их релевантностью и объединением их для формирования запроса. Этот запрос затем используется для выполнения поиска по корпусу с помощью BM25, посредством которого ранжированный список генерируется.

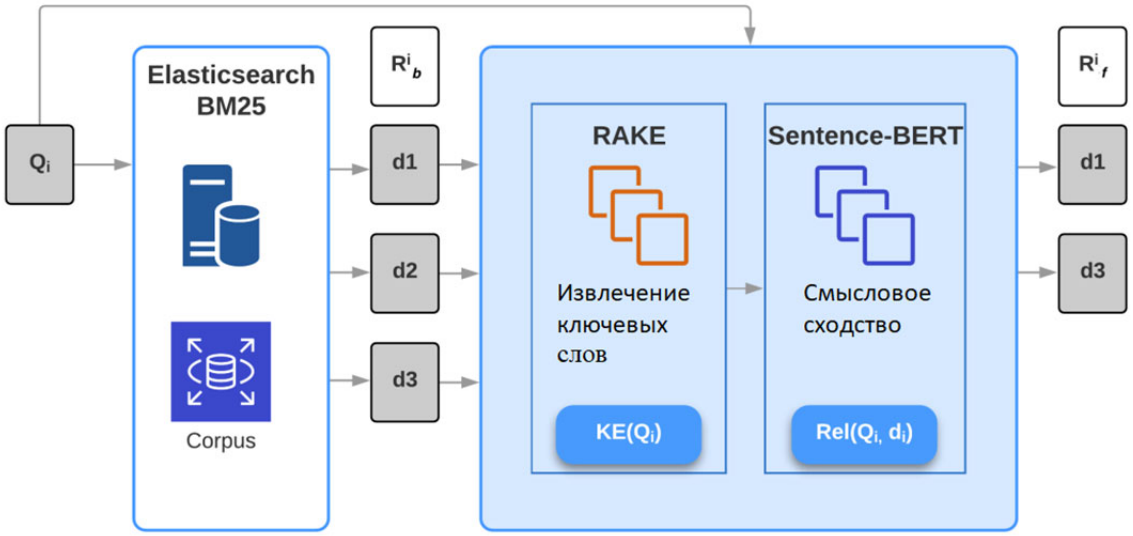

Рис. 1. Этапы поиска в конвейере IR-BERT

Чтобы найти ключевые слова $\{k_1,..., k_n\}$ из документа запроса S, мы сортируем все слова в S в порядке убывания их TF-IDF. Чтобы присвоить ключевым словам разные оценки релевантности, мы определяем вес w_j для каждого ключевого слова k_j следующим образом:

$$w_j = nint\left(\frac{s_j}{\sum_{k=1}^{n} s_k} * n\right),$$

$$s_j = \mathrm{TF}(k_j, S) * \mathrm{IDF}(k_j, A),$$

где n – количество ключевых слов, а TF и IDF – это две статистики: частота терминов и обратная частота документов. Чтобы применить вес для каждого ключевого слова, мы округляем его значение до ближайшего целого числа w_j и повторяем j-е ключевое слово k_j w_j количество раз в запросе. Мы также назначаем разные веса во вкладе ключевых слов в заголовке и теле статьи. Этот взвешенный запрос передаётся в BM25 и верхний p найденных статей (R_b^i), выбраных в качестве документов-кандидатов.

Семантический поиск с использованием BERT. На первом этапе для получения документов кандидатов используется BM25, полностью основанный на терминах частот слова, встречающихся в статье запроса. Чтобы понять контекст статьи запроса, важно принять семантику слова во внимание, потому что фон статьи необязательно может содержать те же ключевые слова, что и поисковый запрос, созданный на основе статьи запроса.

RAKE. Прежде чем проводить семантический поиск по множеству документов R_b^i, важно вводить только те слова Sentence-BERT, смысловое значение которых могло бы принести нам пользу. Таким образом, каждый документ в R_b^i передаётся через Rapid automatic Keyword Extraction algorithm (RAKE).

RAKE принимает список стоп-слов и запрос в качестве входных данных и извлекает их ключевые слова из запроса за один проход.

В основе RAKE лежит идея о том, что совпадения слов имеют значение при определении, являются ли они ключевыми словами. Отношения между словами автоматически адаптируются к стилю и содержанию текста. Это позволяет RAKE иметь адаптивное измерение совпадения слов, которые используются для оценки ключевых слов кандидата.

Sentense-BERT. Sentence-BERT (SBERT) представляет собой модификацию предварительно обученной сети BERT, которая добавляет объединение, работает поверх последнего уровня BERT и точно настроена на получение предложения фиксированного размера.

Siamese Network работает с вложениями предложений, которые специально обучены для работы с мерой сходства, такой как косинус-подобие. Архитектура Sentence-BERT во время вывода представлена на рис. 2.

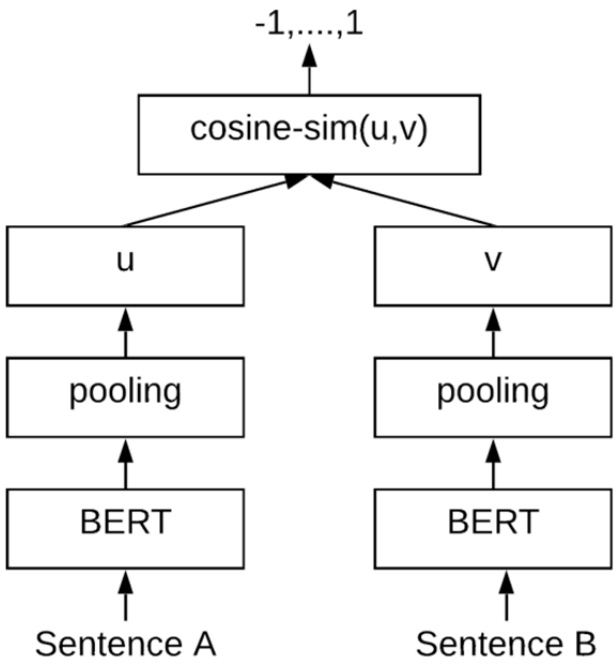

Рис. 2. Архитектура Sentence-BERT при вычислении сходства баллов

SBERT используется для получения вложений для документа запроса Q_i и каждого из документов R_b^i, извлечённых RAKE. Документы в R_b^i затем сортируются в соответствии с их косинусным сходством с запросом Q_i при помощи уравнения

$$\text{CosineSim}(e_1, e_2) = \frac{e_1 . e_2}{||e_1||. |e_2||},$$

где e_1 и e_2 – вложения двух сравниваемых документов.

Алгоритм 1 описывает шаги, необходимые для формирования окончательного списка документов R_f^i через SBERT-вложения.

Алгоритм 1. Изменение ранга кандидатов (Q_i, R_b^i):

1: p ← Количество документов, полученных BM25
2: t ← Требуемое количество итоговых документов
3: q_i ← SBERT(Q_i)
4: for j = 1, . . . p do
5: E_j = SBERT ($R^i_{b,j}$)
6: $f_j = \text{CosineSim}(E_j, q_i)$
7: end for
8: F ← R^i_b отсортировано по убыванию f_j
9: R^i_f ← топ t документов в F
10: вернуть R^i_f

В качестве набора данных мы использовали выпуски российской газеты «Вечерняя Москва». Сборник «Вечерняя Москва» содержит 630 новостных статей и сообщений в блогах за 11 месяцев 2023 года.

Набор данных был предварительно обработан. Шаги обработки данных газеты «Вечерняя Москва» показаны на рис. 3.

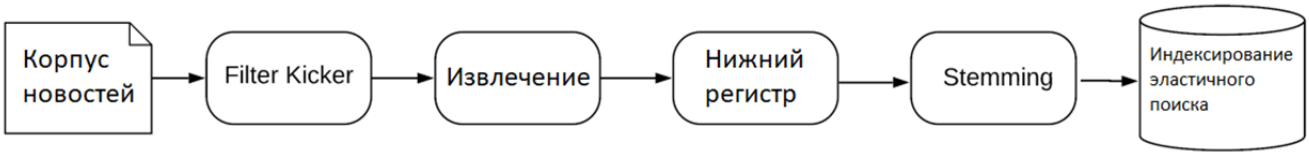

Рис. 3. Шаги обработки данных российской газеты «Вечерняя Москва»

Статьи представлены в формате JSON и включают поля для названия, даты публикации, кикера (заголовка раздела), текста статьи, а также ссылок на встроенные изображения и мультимедиа. Наш метод основан на Elasticsearch в качестве платформы индексации. Во время индексирования мы извлекли информацию из различных полей и проиндексировали их как отдельные поля Elasticsearch. Мы также создали новое поле для хранения текста статьи. Для этого мы сначала извлекли текстовое содержимое HTML из полей, отмеченных типом «sanitized_html» и подтипом «paragraph», а затем объединили их после использования регулярных выражений для извлечения необработанного текста из текста HTML. Далее мы выполнили нижний регистр, удаление стопслов и создание исходного текста. Окончательный предварительно обработанный текст затем индексировался как отдельное текстовое поле в Elasticsearch, представляющее тело статьи.

Мы использовали метод оценки по умолчанию в Elasticsearch, чтобы установить относительные веса для заголовка и тела статьи в поисковом запросе. Мы экспериментировали с различными сочетаниями ряда параметров, конечные значения для которых приведены в табл. 1.

Таблица 1

Значения параметров

# слов в построенном запросе Q	100
# отфильтрованных результатов из BM25	180
# ключевых слов, созданных на основе RAKE	100
% ключевых слов в Q из заголовка	70
% ключевых слов в Q из тела	30

Метрики оценки. Основной метрикой, используемой TREC для фоновой связи задачи, является nDCG@5 со значением усиления 2^{r-1}, где r – уровень релевантности в диапазоне от 0 (даёт мало или вообще не даёт полезной информации) до 4 (обеспечивает критический контекст).

Мера разнообразия. Задача фонового связывания – использовать полученный список разнообразных статей. Идея разнообразия может показаться субъективной, но можно вывести формулу

$$Diversity = \frac{1}{|Q|}\sum_{Q_i}\frac{1}{|R_f^i|}\sum_{a\in R_f^i}\sum_{b\in R_f^i, b\neq a} dist(d_a, d_b), \quad (1)$$

где для каждого полученного списка документов R_f^i вычисляем сумму расстояний между всеми возможными парами документов d_a и d_b. Расстояния между представлением документов могут быть зафиксированы с помощью таких показателей, как косинусное сходство. Эти расстояния затем суммируются по всем запросам/темам Q, чтобы получить разнообразные счета.

Полученные результаты. Мы рассматриваем относительную эффективность IR-BERT по сравнению с некоторыми другими методами участия в TREC 2021 в табл. 2.

Результаты показывают, что IR-BERT работает лучше, чем TKB48, который использует повторное ранжирование Doc2Query на основе преобразователя. IR-BERT также превосходит такие методы, как FUH-N и QU, которые используют матричные операции индексации и трансферное обучение из подтем соответственно.

Таблица 2

Оценки nDCG@5 IR-BERT и других участвующих методов в наборе данных российской газеты «Вечерняя Москва» за 2023 год

Методы	**nDCG@5**
KWVec	**0.4620**
IR-Cologne	0.4423
TKB48-DTQ	0.2925
FUH-N	0.2655
IR-BERT	0.3613

С другой стороны, такие методы, как KWVec и IR-Cologne, получили более высокие оценки nDCG@5, чем IR-BERT. KWVec похож на IR-BERT в использовании Sentence-BERT и Elasticsearch. IR-Cologne использует извлечённые объекты и отношения для реранжирования.

При исследовании влияния использования языковой модели на фоновую задачу связывания мы сравниваем производительность альтернативных архитектур в российской газете «Вечерняя Москва» за 2023 г. Мы перечисляем оценки nDCG@5 и nDCG@10 для каждого из этих подходов (см. табл. 3). Первые два подхода используют только первый этап нашей архитектуры, т. е. они просто создают поисковый запрос и используют BM25 для поиска. В то время как wBT+BM25 при построении запроса использует только взвешенное тело и заголовок, wQ+BM25 использует также веса для всех присутствующих слов в документе запроса. Мы видим, что wQ+BM25 даёт лучший показатель nDCG@10, что означает, что статьи, содержащие полезную справочную информацию, скорее всего, будут содержать ключевые слова аналогично статье запроса. Кроме того, IR-BERT достигает наивысшего балла nDCG@5, что позволяет предположить, что контекстуальное понимание истории статьи может принести пользу. Также интересно отметить, что использование IR-RoBERTa поверх BM25 из-за повторного ранжирования ухудшает производительность по сравнению с использованием стандартной модели BERT.

Таблица 3

Сравнение оценок nDCG для альтернативных методов в наборе данных российской газеты «Вечерняя Москва» за 2023 год

Методы	nDCG@5	nDCG@10
wBT+BM25	0.4088	0.4155
wQ+BM25	0.3942	**0.4315**
IR-RoBERTa	0.394	0.3918
IR-BERT	**0.4199**	0.4104

В нашей последней серии экспериментов мы сравнивали разнообразные документы, полученные всеми нашими подходами к российской газете «Вечерняя Москва» за 2023 год с использованием уравнения (1) (см. табл. 4). Мы наблюдаем тот IR-RoBERTa, который относительно хуже работает на nDCG@10, извлекает самый разнообразный список справочных статей по заданному запросу.

Таблица 4

Сравнение разнообразия найденных документов из разных методов в наборе данных российской газеты «Вечерняя Москва» за 2023 год

Методы	Оценка разноообразия
wBT+BM25	0.9067
wQ+BM25	0.912
IR-RoBERTa	**0.921**
IR-BERT	0.9084

Заключение. В этой статье мы описали двухэтапный подход к решению проблемы. Задача фоновой связи новостного трека TREC 2023 – извлечь репрезентативные ключевые слова из запрашиваемой статьи и использовать их для получения набора кандидатов фоновых ссылок. Далее используется контекстуальное понимание, полученное от BERT, для выполнения семантического поиска. Наша модель IR-BERT достигла оценки nDCG@5 0,3613 по версии TREC российской газеты «Вечерняя Москва» за 2023 год. В целом, участвующие модели и их эффективность демонстрируют эффективность повторного ранжирования за счёт использования контекстуального понимания моделей, основанных на трансформаторах.

ЛИТЕРАТУРА

1. Shchadnaya, M. A. Neural networks as a factor / M. A. Shchadnaya, M. A. Sparing. – M.: Aeterna, 2018. – 163 p.
2. Esin, R. V. Structural diagram of the organization in an electronic learning environment / R. V. Esin. – M.: Piter, 2017. – 321 p.
3. Grishaeva, K. E. Functionality of various types and forms of aspects of their use in the process / H. E. Grishaev. – M.: Scientific Almanac, 2016. – 103 p.
4. Buzdova, A. A. Models of neural network / A. A. Buzdova, I. A. Semenov. – M.: AST-CENTER, 2013. – 84 p.
5. Nazarov, S. V. Architecture and design of software systems: monograph / S. V. Nazarov. – 2nd ed., revised. – M.: INFRA-M, 2018. – 374 p. // ZNANIUM.COM: electronic library system. – Access mode: http://znanium.eom/catalog.php#, restricted. – Zagl. from the screen.
6. Wainer, H. (Ed.). (2000). Computerized adaptive testing. A, 2nd Edn. Hilsdale, NJ: Lawrence Erlbaum Associates. doi: 10.4324/9781410605931.
7. G. Gage Kingsbury, Steven L. Three Measures of neural network based on Optimal Information. Journal of Computerized Adaptive Testing. Vol 8, No 1 (2020).
8. Eggen Theo J. H. M. Multi-Searching for information in a neural network. Front. Educ., 11 December 2018, 3: 111. doi: 10.3389/feduc.2018.00111.
9. Lord, F. M. (1970). «Relevant search using a neural network» in Computer-Assisted Instruction, ed W.H. Holtzma (New York, NY: Harper and Row), 139-183.