НЕЙРОСЕТЕВЫЕ МОДЕЛИ СЕМАНТИЧЕСКОГО ПОИСКА В ПРЕДМЕТНО-ОРИЕНТИРОВАННЫХ КОЛЛЕКЦИЯХ ТЕКСТОВ

Чернявский К. Э., Марцинкевич В. А.
Отдел информационных технологий, отдел сетевых технологий, отдел информационных технологий,
Центр информатизации и инновационных разработок,
Белорусский государственный университет информатики и радиоэлектороники
Минск, Республика Беларусь
E-mail: {k.cherniavskij, vlad}@bsuir.by

В работе рассматриваются нейросетевые подходы к семантическому поиску информации в предметно-ориентированных текстовых коллекциях. Анализируются архитектуры BERT и GPT, демонстрирующие способность учитывать контекст и смысловые связи в сложных запросах. Показано, что трансформеры обеспечивают более высокую релевантность результатов по сравнению с классическими методами. Приведены примеры интеграции моделей в специализированные поисковые системы, а также результаты экспериментального сравнения на корпусе научных публикаций.

Введение

Современные предметно-ориентированные коллекции текстов – научные архивы, отраслевые базы данных, специализированные библиотеки – требуют инструментов поиска, способных учитывать сложную терминологию и контекст. Классические методы, такие как TF-IDF и BM25, основаны на частотном анализе слов и не интерпретируют смысловые связи, что ограничивает их применимость в узкоспециализированных областях. Особенно это проявляется при работе с научными публикациями, где запросы могут включать многозначные термины, синонимы и сложные грамматические конструкции.

С развитием нейросетевых технологий появились модели, способные учитывать контекст и семантику текста. Архитектуры BERT и GPT, основанные на трансформерах, демонстрируют высокую эффективность в задачах интеллектуального поиска.

Цель работы – исследовать применение трансформеров в предметно-ориентированных коллекциях, сравнить их с классическими подходами и оценить качество поиска на реальных данных.

I. Классические методы поиска

До появления нейросетевых моделей основными инструментами поиска информации были статистические методы, основанные на частотном анализе слов. Одним из наиболее распространённых подходов является TF-IDF (Term Frequency-Inverse Document Frequency), который оценивает значимость термина в документе относительно его распространённости в коллекции. Несмотря на простоту и эффективность в ряде задач, TF-IDF не учитывает порядок слов, грамматические связи и контекст, что ограничивает его применимость при обработке сложных запросов.

Другим популярным методом является BM25 – вероятностная модель, основанная на расширении TF-IDF. Она учитывает длину документа и насыщенность терминами, что позволяет более гибко ранжировать результаты. Однако, как и TF-IDF, BM25 работает на уровне отдельных слов и не способен интерпретировать смысловые связи между ними.

Классические методы широко применяются в поисковых системах, таких как Lucene и Elasticsearch, благодаря своей скорости и простоте реализации. Тем не менее, они демонстрируют ограниченную эффективность при работе с многозначными терминами, синонимами и контекстно-зависимыми запросами. Например, запрос «поиск информации с помощью нейросети» может не найти документы, содержащие фразы «глубокое обучение для извлечения данных» или «семантический анализ текста».

Несмотря на историческую значимость и широкое распространение, классические методы поиска уступают современным нейросетевым подходам в задачах, требующих понимания контекста и семантики.

II. Нейросетевые архитектуры

Ограничения классических методов стимулировали развитие моделей, способных учитывать контекст и семантику. Одним из наиболее значимых достижений в этой области стало появление трансформерных архитектур, таких как BERT (Bidirectional Encoder Representations from Transformers) и GPT (Generative Pre-trained Transformer).

Модель BERT обучается на задаче восстановления пропущенных слов и предсказания следующего предложения, что позволяет ей формировать двунаправленные контекстуальные представления текста. Значение каждого слова определяется не только его окружением слева, но и справа, что критически важно для понимания сложных языковых конструкций. GPT, в свою очередь, использует авторегрессионный подход,

при котором каждое следующее слово предсказывается на основе предыдущих, что делает модель особенно эффективной в генерации текста и диалоговых системах.

Обе архитектуры основаны на механизме внимания, который позволяет модели фокусироваться на наиболее значимых частях входного текста. Это обеспечивает высокую точность при обработке длинных и насыщенных контекстом запросов. Кроме того, трансформеры масштабируются на большие объемы данных и могут быть дообучены на специализированных корпусах, что делает их адаптируемыми к конкретным предметным областям.

Нейросетевые модели не только превосходят классические методы по качеству поиска, но и открывают возможности для семантического ранжирования, кластеризации документов, извлечения ответов и построения интеллектуальных агентов.

III. Интеграция в специализированные поисковые системы

Интеграция нейросетевых моделей в поисковые системы стала важным этапом в развитии технологий извлечения информации. Такие модели позволяют не только анализировать текстовые данные, но и интерпретировать смысл пользовательских запросов.

Одним из первых примеров масштабного внедрения трансформеров стала модель BERT, интегрированная в Google Search. Она позволила улучшить обработку длинных и сложных запросов, повысив точность ранжирования и релевантность выдачи. В отличие от классических алгоритмов, BERT учитывает контекст слов, грамматические зависимости и скрытые смысловые связи, что делает результаты поиска более осмысленными.

В научных системах, таких как Semantic Scholar и Microsoft Academic, нейросетевые модели используются для семантического поиска, автоматического аннотирования статей и построения тематических связей между публикациями. Это позволяет исследователям находить релевантные работы даже при использовании узкоспециализированной терминологии или нестандартных формулировок.

IV. Экспериментальное исследование

Для оценки эффективности нейросетевых моделей в задачах интеллектуального поиска информации было проведено экспериментальное сравнение трансформерной архитектуры BERT с классическим методом TF-IDF. В качестве тестового корпуса использовалась коллекция научных статей из области компьютерных наук, включающая более 5000 документов на русском и английском языках, отобранных из открытых научных источников. Запросы формировались вручную и охватывали как общие, так и специализированные темы, включая многозначные термины и контекстно-зависимые выражения.

Каждый запрос обрабатывался двумя системами: одна использовала TF-IDF с косинусным сходством, другая – предобученную на научных текстах модель BERT. Релевантность результатов оценивалась вручную экспертами по шкале от 0 до 3, где 3 означало полное соответствие запросу. Также рассчитывались стандартные метрики: точность, полнота и F1-мера.

Результаты показали явное преимущество нейросетевого подхода. Средняя точность BERT составила 0.87 против 0.62 у TF-IDF, полнота – 0.81 против 0.58, F1-мера – 0.84 против 0.60. Особенно заметно преимущество трансформеров при обработке запросов с вариативной терминологией и сложной синтаксической структурой. BERT демонстрировал устойчивость к вариативности формулировок и лучше справлялся с многозначностью.

Заключение

Рассмотренные в работе нейросетевые модели, основанные на трансформерных архитектурах, демонстрируют высокую эффективность в задачах интеллектуального поиска информации. В отличие от классических методов, они способны учитывать контекст, грамматические зависимости и скрытые смысловые связи, что особенно важно при обработке сложных и многозначных запросов.

Экспериментальное сравнение BERT и TF-IDF на корпусе научных публикаций показало значительное преимущество нейросетевого подхода по ключевым метрикам качества поиска. Это подтверждает целесообразность внедрения трансформеров в современные поисковые системы.

1. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // arXiv:1810.04805, 2018.
2. Brown T., Mann B., Ryder N. и др. Language Models are Few-Shot Learners // arXiv:2005.14165, 2020.
3. Lu Y., He D., Ouyang Y. и др. Neural Document Ranking with BERT // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2019.
4. Manning C. D., Raghavan P., Schütze H. Introduction to Information Retrieval. – Cambridge: Cambridge University Press, 2008. – 482 p.
5. Vaswani A., Shazeer N., Parmar N. и др. Attention Is All You Need // arXiv:1706.03762, 2017.