УДК НОМЕР УДК 004.89

Машкин Владимир Анатольевич
доцент кафедры информатики вычислительной техники и прикладной математики
Забайкальский государственный университет
г. Чита, Россия
Замешаев Семен Игоревич
студент групп ИВТ(аи)м-24 Энергетического факультета
Забайкальский государственный университет
г. Чита, Россия

РАЗРАБОТКА ИНТЕЛЛЕКТУАЛЬНОГО АССИСТЕНТА ДЛЯ НАВИГАЦИИ И СЕМАНТИЧЕСКОГО ПОИСКА В ЭЛЕКТРОННЫХ БИБЛИОТЕЧНЫХ СИСТЕМАХ

В статье рассматривается проблема эффективного взаимодействия пользователей с крупными массивами научной информации в электронных библиотеках. Предлагается решение на основе гибридной RAG-архитектуры (Retrieval-Augmented Generation) [3] с использованием нескольких генеративных моделей и продвинутых методов эмбеддинга.

Цель разработки - создание интеллектуального ассистента, способного точно отвечать на естественно-языковые запросы студентов, извлекая релевантные фрагменты из документов и генерируя структурированные, обоснованные ответы.

Современные электронные библиотечные системы ЭБС и научные репозитории считаются частью образовательного процесса. Они предоставляют студентам доступ к огромным массивам публикаций: учебникам, монографиям, научным статьям и диссертациям [4].

Традиционные системы поиска в ЭБС часто основаны на ключевых словах и метаданных (автор, название, год издания). Такой подход имеет существенные недостатки:

- Низкая контекстуальная релевантность: Система ищет точные совпадения слов, игнорируя смысл запроса.

- Проблема лексического несоответствия: Пользователь формулирует вопрос своими словами, в то время как в документах используется иная терминология.
- Отсутствие синтеза информации: Студент получает список из десятков или сотен документов, но не готовый ответ на свой конкретный вопрос. Ему приходится самостоятельно изучать каждый источник, что требует значительного количества времени для поиска информации.

Несмотря на наличие поисковых систем, ключевые проблемы остаются нерешенными:

- Семантический разрыв: Неспособность традиционных систем понимать intent и контекст запроса пользователя.
- Формирование комплексного ответа: Существующие системы не агрегируют информацию из нескольких источников для генерации связного, краткого и точного ответа.
- Верифицируемость и доверие: Сгенерированный нейросетью ответ (как в чистых LLM, например ChatGPT) может быть непроверяемым и содержать “галлюцинации” - вымышленные или неточные факты.
- Адаптивность: Большинство систем не адаптируются под специфику конкретной дисциплины или ЭБС.

Разработка и внедрение программного ассистента на основе искусственного интеллекта, который предоставляет точные, верифицируемые ответы на естественно-языковые запросы студентов на основе документов конкретной электронной библиотеки [5].

Для достижения цели необходимо решить следующие задачи для работы:

1. Спроектировать гибридную RAG-архитектуру, объединяющую этап семантического [6] поиска и этап генерации ответа.
2. Реализовать модуль векторного поиска [7]:

- Провести препроцессинг и чанкирование документов из ЭБС.
- Подобрать или обучить модель для создания эмбеддингов, учитывающую научную и техническую терминологию.

3. Реализовать модуль генерации:

- Интегрировать мощную языковую модель для формирования ответов.
- Разработать промты, инструктирующие модель формулировать ответы исключительно на основе предоставленного контекста из базы знаний и цитировать источники.

4. Обеспечить масштабируемость и интеграцию: Разработать решение, которое можно адаптировать для работы с различными ЭБС через API.

5. Разработать механизм оценки качества работы системы [1].

Для решения задач планируется использовать встроенные методы:

RAG: Это современный архитектурный подход, который решает проблему “галлюцинаций”. Система сначала извлекает правильные фрагменты текста из векторной базы данных, а затем передает их языковой модели в качестве контекста для генерации ответа.

Векторное представление знаний (Эмбеддинги): Текстовые документы будут преобразовываться в числовые векторы с помощью предобученных моделей. Это позволит осуществлять семантический поиск по смыслу, а не по ключевым словам.

Генеративные языковые модели (LLM): Крупные модели, такие как GigaChat, будут использоваться на этапе синтеза ответа. Их ключевое преимущество - способность понимать сложные запросы и генерировать связный текст.

Гибридный поиск: Для повышения точности Retriever-модуля планируется комбинировать семантический поиск и лексемический. Это

позволит учесть как смысловую близость, так и точное совпадение ключевых терминов.

Оптимизация промтов (Prompt Engineering): Будет разработан набор строгих промтов для LLM, которые заставят модель отвечать только на основе предоставленных фрагментов текста, структурировать ответ и обязательно указывать источники.

В результате будет создано программное обеспечение, которое:

- Принимает текстовый запрос на естественном языке.
- В реальном времени находит наиболее правильные фрагменты текста в подключенной ЭБС.
- Формирует четкий, структурированный и краткий ответ, синтезированный из найденных источников.
- Сопровождает каждый ключевой факт ссылкой на источник, обеспечивая полную верифицируемость.
- Предоставляет пользователю список всех использованных документов для глубокого изучения.

Эффективность системы будет оцениваться по двум направлениям:

1. Метрики качества поиска [1]:

- Precision@K: Доля важных документов среди первых K извлеченных.
- Recall@K: Доля всех важных документов, найденных в топ-K результатах.
- MRR: Средняя величина, обратная рангу первого важного документа.

2. Метрики качества генерации (Generator):

- Экспертная оценка (Human Evaluation): Наиболее важный критерий. Эксперты будут оценивать ответы по шкалам:
- Корректность и точность фактов.
- Полнота ответа.
- Связность и понятность текста.

- Наличие и точность цитирования.
- Автоматические метрики: BLEU, ROUGE, но их использование ограничено из-за вариативности правильных ответов.

Аналоги:

Elicit.org URL: https://elicit.org: AI-ассистент для исследователей, который помогает находить и обобщать научные статьи по запросу. Использует семантический поиск и генерацию. Прямой аналог, но ориентированный на англоязычные статьи.

Функционал: Находит статьи по семантическому запросу, извлекает из них ключевые моменты, обобщает выводы, отвечает на вопросы на основе содержимого статей.

Отличие: Elicit ориентирован на англоязычные научные статьи и мета-анализ. Наше решение фокусируется на закрытых корпоративных электронных библиотеках [8] ЭБС с литературой на русском языке, включая учебники, монографии, методички, что требует иных подходов к чанкированию и эмбеддингу.

Scopus AI / Semantic Scholar AI URL: https://www.scopus.com/ai: Крупные научные базы данных начали внедрять AI-инструменты для обзора литературы и ответов на вопросы.

Функционал: Генерация кратких обзоров по теме, ответы на вопросы на основе подборки статей, определение ключевых статей.

Отличие: Это проприетарные системы, глубоко интегрированные в свою экосистему (Scopus, SciVerse). Они не предназначены для развертывания на стороне университета для работы с его внутренней, возможно, закрытой коллекцией документов. Наше решение предлагается как архитектура или платформа для интеграции с любой ЭБС.

Патент US 20240168729 A1 International Business Machines Corporation (IBM): “System and method for adaptive context management in retrieval-augmented generation”

Система динамически определяет оптимальный объем контекста, передаваемого языковой модели для генерации ответа [9]. Она оценивает сложность запроса, историю диалога и характеристики самих документов, чтобы избежать как недостаточного, так и избыточного контекста.

Позволяет значительно повысить точность и релевантность ответов ассистента, особенно на сложные, многосоставные запросы студентов, требующие синтеза информации из нескольких глав или даже разных учебников. Решает проблему "потери" ключевых фактов в большом объеме текста.

Патент US 11860972 B1 Amazon Technologies, Inc: “Domain-specific fine-tuning for retrieval-augmented generation models”

Описан метод тонкой настройки (fine-tuning) всех компонентов RAG-пайплайна (энкодера для семантического поиска, реранкера, самой языковой модели) на узкоспециализированных корпусах текстов. Это не просто настройка LLM, а комплексная совместная оптимизация всех этапов под конкретную предметную область.

Позволяет создать высокоспециализированного ассистента именно для академической среды. Система может быть дообучена на учебниках и научной литературе, что улучшит ее понимание терминологии, концепций и контекста обучения, а также повысит точность поиска и качество генерации.

СПИСОК ИСПОЛЬЗУЕМОЙ ЛИТЕРАТУРЫ

1. Основы поиска информации: Учебное пособие для вузов. - М.: Издательский дом "Вильямс", 2021. - 384 с. - ISBN 978-5-8459-2101-2.
2. Speech and Language Processing (3rd ed. draft). - 2023. - URL: https://web.stanford.edu/~jurafsky/slp3/.
3. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks // Advances in Neural Information Processing Systems 33 (NeurIPS 2020). - P. 9459-9474.
4. Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции “Диалог”. - Вып. 22. - М.: Изд-во РГГУ, 2023.
5. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of NAACL-HLT 2019. - P. 4171-4186.
6. Интеллектуальный анализ данных и семантический поиск: методы и технологии. - Новосибирск: Изд-во ИДМИ, 2020. - 210 с.
7. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP 2019). - URL: https://arxiv.org/abs/1908.10084
8. Электронные библиотеки: проблемы юзабилити и когнитивной нагрузки пользователя // Научно-техническая информация. Серия 1: Организация и методика информационной работы. - 2022. - № 8. - С. 15-23.
9. RoBERTa: A Robustly Optimized BERT Pretraining Approach // arXiv preprint arXiv:1907.11692. - 2019.