

ОПТИМИЗАЦИЯ ПРОЦЕССА КОНСУЛЬТИРОВАНИЯ АБИТУРИЕНТОВ С ПОМОЩЬЮ ТЕХНОЛОГИИ RETRIEVAL AUGMENTED GENERATION

Цыденов С.Б.¹

¹ Томский Политехнический Университет, ИШИТР, АЗ-39

Аннотация

В статье представлен подход к использованию технологии Retrieval Augmented Generation (RAG) для создания автоматизированной системы консультирования абитуриентов. Технология RAG комбинирует механизмы генерации ответов с поиском по большим базам данных, позволяя системе поддержки принимать на вход вопросы от пользователей и выдавать информативные, точно подобранные ответы, обогащённые актуальными данными. В качестве исходных данных используется база знаний, включающая информацию о правилах поступления. Результатом исследования является демонстрация возможности значительного повышения качества и скорости консультирования абитуриентов, снижения нагрузки на сотрудников приемных комиссий и улучшения общего пользовательского опыта на сайте университета.

Ключевые слова: Retrieval Augmented Generation, автоматизированная система поддержки, база знаний, консультирование абитуриентов, искусственный интеллект.

Введение

Современные технологии искусственного интеллекта (ИИ) находят всё новые применения в образовании, облегчая и улучшая процесс обучения и взаимодействие с абитуриентами. Одним из передовых направлений в этой области является технология Retrieval Augmented Generation (RAG) [1, 2], которая позволяет существенно расширить возможности систем автоматического ответа на вопросы за счет интеграции механизмов поиска по базам данных непосредственно в процесс генерации текста. Это позволяет не только генерировать ответы на основе фиксированного набора знаний, но и обогащать их актуальной информацией из внешних источников в реальном времени. Использование RAG для консультирования абитуриентов на сайте университета может оптимизировать процесс предоставления информации, делая его более быстрым и удобным для пользователей. В данной статье описывается разработка и внедрение системы на основе RAG, направленной на улучшение качества поддержки абитуриентов, а также анализируются полученные результаты и пути дальнейшего развития технологии в контексте высшего образования [4, 5].

Описание алгоритма

Алгоритм, представленный в данной работе, является базовой реализацией архитектуры Retrieval Augmented Generation (RAG) для системы консультирования абитуриентов университета. Алгоритм разработан таким образом, чтобы быть простым в реализации и интеграции, предоставляя при этом основу для дальнейших улучшений и расширений.

Алгоритм реализован на языке программирования Python с использованием библиотеки langchain [3] и дополнительных инструментов для работы с текстом и семантическими векторами. Информационная база собрана из официального сайта ТПУ, обработана и индексируется для ускорения поиска. В процессе сбора были извлечены данные о поступлении на бакалавриат, ответы на часто задаваемые вопросы, условия приема, сроки подачи документов, условиях проживания в общежитиях и контактные данные приемной комиссии. Эти данные были собраны в файл, автоматически разделены и проиндексированы в векторной базе данных для последующего использования в процессе извлечения информации.

Основная идея алгоритма заключается в сочетании процессов извлечения (retrieval) и генерации (generation) текста. В первую очередь, при получении запроса от пользователя, алгоритм выполняет поиск по собранной вручную базе знаний, используя вектора с семантическими значениями из запроса для извлечения наиболее релевантной информации. После этапа извлечения информации следует этап генерации ответа. На основе извлеченных данных алгоритм использует модель генерации текста «gpt-3.5-turbo-0125» для того, чтобы сформировать понятный и информативный ответ.

Важной особенностью алгоритма является его модульность. Компоненты извлечения информации и генерации текста разработаны таким образом, чтобы их можно было независимо модифицировать и улучшать. Например, методы извлечения могут быть улучшены для повышения

точности и релевантности поиска, а модель генерации текста может быть заменена на более современную версию или дообучена на более специфичном для университета наборе данных.

В целом, предложенный алгоритм является устойчивой основой для создания эффективной системы консультирования абитуриентов, предлагая пути для дальнейших исследований и разработки в области применения технологий искусственного интеллекта в высшем образовании.

Результаты

В ходе экспериментального исследования алгоритма, основанного на принципах Retrieval Augmented Generation (RAG), было выявлено его способность к адаптации под разнообразные запросы пользователей. Эмпирический анализ показал высокую точность и релевантность сгенерированных ответов.

Таблица 1

Некоторые примеры ответов алгоритма на пользовательские запросы

Запрос пользователя	Ответ сети	Комментарий
Когда начнется прием документов?	Прием документов для поступления на программы бакалавриата и специалитета начнется до подведения итогов конкурса. Документы можно отправлять по почте через операторов почтовой связи общего пользования.	Ответ верный
Какие условия в общежитиях ТПУ?	Проживающие в студенческих общежитиях ТПУ могут быть размещены одиночно, вдвоем или втроем. Общежития имеют круглосуточную пропускную систему и профессиональную охрану, а также принимают гостей с 8:00 до 22:00. Размещение студентов производится по шести квадратных метров на человека, преимущественно в трехместных и двухместных комнатах.	Ответ верный
Есть ли скидки на обучение?	Скидки на обучение предоставляются по результатам ЕГЭ или вступительных испытаний, размер скидки зависит от направления подготовки. Вход в музеи Томского политехнического университета бесплатный, но посещение осуществляется организованными группами по предварительной записи.	Ответ верный
Какие документы нужны для поступления?	Для поступления необходимы документы, удостоверяющие личность и гражданство, документ об образовании, фотографии, договор для платных мест, копия свидетельства о перемене имени и медицинская справка при необходимости. Отправляя документы по почте, следует приложить копии документов, удостоверяющих личность, военного билета (при наличии), образования, фотографии размером 3х4. Копии документов не должны быть заверены.	Ответ верный
Как дозвониться до отдела студенческих общежитий?	Для связи с отделом студенческих общежитий ТПУ можно позвонить по телефонам +7 (3822) 60-62-05 и +7 (3822) 70-18-25. Отдел студенческих общежитий находится по адресу г. Томск, ул. Вершинина, 37. В общежитии №18 на ул. Усова, 15б, размещаются до шести человек в комнатах.	Ответ верный
Могу ли я отозвать оригиналы документов после их подачи? И как мне их вернуть, если это возможно?	Да, вы можете отозвать оригиналы документов после их подачи. Оригиналы документов могут быть возвращены либо указанному лицу, либо через операторов почтовой связи общего пользования, в зависимости от указаний в заявлении об отзыве документов. Оригиналы документов будут возвращены в течение 20 рабочих дней после отзыва документов или после подведения итогов конкурса.	Ответ верный
Могу ли я отправить документы для поступления через почту?	Да, вы можете отправить документы для поступления через почту, но при этом необходимо присоединить к заявлению копии документов, удостоверяющих личность, военного билета (для лиц, проходивших военную службу), оригиналы или копии документов об образовании, другие документы при необходимости и 2 фото размером 3х4. Также возможно направление документов через личный кабинет абитуриента ТПУ или Единый портал	Ответ верный

	государственных и муниципальных услуг, где нужно будет загрузить цифровые копии необходимых документов и фотографию. После завершения приема документов через сайт подача документов недоступна.	
На сколько направлений можно подавать заявление?	На пять направлений подготовки.	Ответ верный и краткий
Есть ли дистанционное обучение?	Информация о состоянии общежитий не была предоставлена в извлеченном контексте, поэтому я не могу ответить на этот вопрос.	Алгоритм не смог найти информацию в исходных данных
Сколько стоит обучение?	Я не знаю.	Здесь был задан вопрос, для которого не было ответа в собранной базе данных. Модель не стала ничего выдумывать и ответила, что не обладает информацией

Несмотря на общую точность, в ответах периодически встречалась избыточная информация, не всегда соответствующая первоначальным запросам пользователей. Такие случаи выделяют важность последующей настройки алгоритма для оптимизации процесса извлечения и генерации данных. Примечательно, что алгоритм ограничен в создании нового содержания, опираясь исключительно на предварительно загруженные данные. Это поднимает вопрос о регулярном обновлении и расширении базы данных для сохранения актуальности информации.

Дополнительно было проведено исследование реакции системы на нестандартные запросы, которые не имели прямых ответов в базе данных. Алгоритм демонстрировал способность корректно определять границы своих знаний, указывая на отсутствие информации, что является важным аспектом в контексте пользовательского взаимодействия и удовлетворенности.

Заключение

Итоги настоящего исследования подтверждают высокий потенциал использования архитектуры Retrieval Augmented Generation для автоматизации процесса консультации абитуриентов. Демонстрируя впечатляющую способность к адаптации и генерации релевантных ответов, алгоритм открывает новые перспективы для повышения эффективности информационного обслуживания. Вместе с тем, выявленные недостатки и ограничения алгоритма подчеркивают необходимость его дальнейшего совершенствования, особенно в аспектах интеграции дополнительных данных и оптимизации механизмов извлечения информации. Дальнейшие работы планируется сосредоточить на разработке методик обновления информационной базы и на улучшении точности обработки запросов с применением методов машинного обучения для динамической оценки релевантности информации.

Список использованных источников

1. "Retrieval Augmented Generation (RAG): A Comprehensive Guide" // [www.datastax.com](https://www.datastax.com/guides/what-is-retrieval-augmented-generation). – 2023. – URL: <https://www.datastax.com/guides/what-is-retrieval-augmented-generation> (дата обращения 24.03.2024).
2. "What is Retrieval-Augmented Generation (RAG)?" // [www.analyticsvidhya.com](https://www.analyticsvidhya.com/blog/2023/09/retrieval-augmented-generation-rag-in-ai/). – 2021. – URL: <https://www.analyticsvidhya.com/blog/2023/09/retrieval-augmented-generation-rag-in-ai/> (дата обращения 26.03.2024).
3. Pandya, Keivalya; Holia, Mehfuza. Automating Customer Service using LangChain: Building custom open-source GPT Chatbot for organizations // [arxiv.org](https://arxiv.org/abs/2310.05421). – 2023. – URL: <https://arxiv.org/abs/2310.05421> (дата обращения 24.03.2024).

4. Retrieval Augmented Generation and Representative Vector Summarization for Large Unstructured Textual Data in Medical Education // ar5iv.org. – 2023. – URL: <https://ar5iv.org/html/2308.00479> (дата обращения 1.04.2024).
5. Guo, Zhicheng et al. «Retrieval-Augmented Generation for Large Language Models: A Survey» // ar5iv.org. – 2023. – URL: <https://ar5iv.org/html/2312.10997> (дата обращения: 27.03.2024).