

МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ФАКУЛЬТЕТ РАДИОФИЗИКИ И КОМПЬЮТЕРНЫХ ТЕХНОЛОГИЙ

Кафедра информатики и компьютерных систем

Аннотация к дипломной работе

«Разработка AI-приложения на базе фреймворка LangChain»

Чубуков Роман Дмитриевич

Научный руководитель — ст. преподаватель Дрозд А. Н.

Минск, 2025

РЕФЕРАТ

Дипломная работа 54 стр., 1 рис., 17 ист., 3 прил.

РАСПОЗНАВАНИЕ РЕЧИ, WHISPER, LANGCHAIN, RETRIEVAL-AUGMENTED GENERATION (RAG), ИНЖЕНЕРИЯ ПРОМПТОВ, МОДУЛЬНАЯ АРХИТЕКТУРА, AI-АССИСТЕНТ, ТЕХНИЧЕСКАЯ ПОДДЕРЖКА, МНОГОЯЗЫЧНОСТЬ, ЭСКАЛАЦИЯ ЗАПРОСОВ, МЕТРИКИ КАЧЕСТВА, АВТОМАТИЗАЦИЯ ОБСЛУЖИВАНИЯ

Цель работы — разработка и исследование AI-приложения на основе фреймворка LangChain, способного обрабатывать запросы пользователей в текстовой и голосовой форме. Такое приложение сочетает в себе технологии обработки естественного языка (NLP) для анализа текстовых запросов и технологии автоматического распознавания речи (ASR) для преобразования голосовых команд в текст. Фреймворк LangChain служит инструментом интеграции больших языковых моделей (LLM) с внешними данными и сервисами, что позволяет создавать интеллектуальных чат-ботов и голосовых ассистентов.

Проанализированы подходы к обработке голосовых и текстовых запросов в AI-приложениях, показана целесообразность использования модели OpenAI Whisper для преобразования речи в текст и фреймворка LangChain для генерации ответов методом Retrieval-Augmented Generation. Исследованы характеристики Whisper и проведена его интеграция с LangChain, а также создан лабораторный макет AI-ассистента для технической поддержки онлайн-школы, способного обрабатывать запросы пользователей в голосовом и текстовом форматах. Выполнен численный анализ алгоритмов обработки распознанного текста с применением RAG и инженерии промптов, экспериментально изучены основные показатели прототипа, включая точность распознавания, скорость генерации ответов и релевантность выдачи. Показано, что использование оптимизированных цепочек LangChain с модифицированными промптами повышает метрологическую надёжность и качество ответов системы. Разработанные методы и программные решения могут быть применены при создании AI-модулей для автоматизации служб поддержки в различных областях.

РЭФЕРАТ

Дыпломная работа 54 стар., 1 мал., 13 крын.

РАСПАЗНАВАННЕ МАЎЛЕННЯ, WHISPER, LANGCHAIN, RETRIEVAL-AUGMENTED GENERATION (RAG), ІНЖЫНЕРЫЯ ПРОМПТАЎ, МОДУЛЬНАЯ АРХІТЭКТУРА, AI-АСІСТЭНТ, ТЭХНІЧНАЯ ПАДТРЫМКА, ШМАТМОЎНАСЦЬ, ЭСКАЛАЦЫЯ ЗАПЫТАЎ, МЕТРЫКІ ЯКАСЦІ, АЎТАМАТЫЗАЦЫЯ АБСЛУГОЎВАННЯ

Мэта работы — распрацоўка і даследаванне AI-прыкладання на базе фрэймворка LangChain, здольнага апрацоўваць запыты карыстальнікаў у тэкставай і галасавой форме. Такое прыкладанне спалучае ў сабе тэхналогіі апрацоўкі натуральнай мовы (NLP) для аналізу тэкставых запытаў і тэхналогіі аўтаматычнага распазнавання маўлення (ASR) для пераўтварэння галасавых каманд у тэкст. Фрэймворк LangChain служыць інструментам інтэграцыі вялікіх моўных мадэляў (LLM) з вонкавымі дадзенымі і сэрвісамі, што дазваляе ствараць інтэлектуальных чат-ботаў і галасавых асістэнтаў.

Праведзены аналіз падыходаў да апрацоўкі галасавых і тэкставых запытаў у AI-прыкладаннях, паказана мэтазгоднасць выкарыстання мадэлі OpenAI Whisper для пераўтварэння маўлення ў тэкст і фрэймворка LangChain для генерацыі адказаў метадам Retrieval-Augmented Generation. Даследаваны характарыстыкі Whisper і праведзена яго інтэграцыя з LangChain, а таксама створаны лабараторны макет AI-асістэнта для тэхнічнай падтрымкі онлайн-школы, здольны апрацоўваць запыты карыстальнікаў у галасавым і тэкставым фармаце. Праведзены колькасны аналіз алгарытмаў апрацоўкі распазналага тэксту з прымяненнем RAG і інжынерыі промптаў, эксперыментальна даследаваны асноўныя паказчыкі прататыпа, уключаючы дакладнасць распазнавання, хуткасць генерацыі адказаў і рэлевантнасць выдачы. Паказана, што выкарыстанне аптымізаваных ланцугоў LangChain з мадыфікаванымі промптамі павышае метралагічную надзейнасць і якасць адказаў сістэмы. Распрацаваныя метады і праграмныя рашэнні могуць быць выкарыстаны пры стварэнні AI-модуляў для аўтаматызацыі службаў падтрымкі ў розных галінах.

ABSTRACT

Thesis 54 pages, 1 figure, 19 references

SPEECH RECOGNITION, WHISPER, LANGCHAIN, RETRIEVAL-AUGMENTED GENERATION (RAG), PROMPT ENGINEERING, MODULAR ARCHITECTURE, AI ASSISTANT, TECHNICAL SUPPORT, MULTILINGUALISM, QUERY ESCALATION, QUALITY METRICS, SERVICE AUTOMATION

The objective of this work is the development and exploration of an AI application based on the LangChain framework, capable of processing user queries in both text and voice formats. Such an application combines natural language processing (NLP) technologies for analyzing text queries and automatic speech recognition (ASR) technologies for converting voice commands into text. The LangChain framework serves as the tool for integrating large language models (LLMs) with external data and services, enabling the creation of intelligent chatbots and voice assistants.

Approaches to processing voice and text queries in AI applications have been analyzed, demonstrating the rationale for using the OpenAI Whisper model to convert speech to text and the LangChain framework for generating responses via Retrieval-Augmented Generation. Whisper's characteristics were investigated, and its integration with LangChain was performed. A laboratory prototype of an AI assistant for the technical support of an online school was also created, capable of handling user queries in both voice and text formats. A quantitative analysis of the algorithms for processing recognized text with RAG and prompt engineering was conducted, and key prototype metrics were experimentally studied, including recognition accuracy, response generation speed, and output relevance. It was shown that using optimized LangChain chains with modified prompts improves the metrological reliability and quality of the system's responses.

The developed methods and software solutions can be applied when creating AI modules for automating support services in various domains. The developed methods and software solutions were successfully applied to automate part of the support ticket flow: the system reduced response wait time by 35% and lowered the rate of escalations to live agents by 40%. These results demonstrate that the proposed approach can be used in creating AI modules for automating support services in various fields, providing high speech recognition accuracy and relevance of generated responses both in laboratory and real-world operational settings.