УДК 004.89

РАЗРАБОТКА СТРУКТУРЫ УНИВЕРСАЛЬНОЙ СИСТЕМЫ АГРЕГАЦИИ И ХРАНЕНИЯ ДАННЫХ ДЛЯ СЕМАНТИКО-ОРИЕНТИРОВАННОГО ИНТЕЛЛЕКТУАЛЬНОГО ПОИСКА НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

Филатова А.А. (ИТМО), **Ковальчук М.А.** (ИТМО)
Научный руководитель – аспирант, Ковальчук М.А. (ИТМО)

Введение. В эпоху цифровизации и повсеместной интеллектуализации организации сталкиваются с вызовами, связанными с обработкой и управлением огромными объемами разнородных данных, от законодательных документов до неструктурированной аналитической информации и текстов общения. Эти массивы данных представляют собой ценный ресурс для оптимизации бизнес-процессов, обучения персонала и служат основой для создания систем поддержки принятия решений. Однако разнообразие источников и форматов данных усложняет их сбор, систематизацию и анализ для извлечения значимой информации, что подчеркивает необходимость в разработке новых методов и алгоритмов, включая технологии искусственного интеллекта, для решения данных задач.

В основе данной работы лежит разработка структуры комплексной автоматизированной системы, способной эффективно собирать, обрабатывать и анализировать большие объемы неструктурированных и полуструктурированных документов. Глобальным назначением такой системы является накопление предметных знаний с последующим использованием в доменных системах-ассистентах для ответа на экспертные пользовательские запросы на естественном языке.

Основная часть. В рамках работы была спроектирована комплексная система для решения задач, связанных с автоматизированным сбором, обработкой и анализом больших объемов текстовых данных, а также построением экспертных вопросно-ответных систем на естественном языке. Структура системы включает в себя несколько ключевых модулей: автоматизированный модуль сбора данных с семантической компонентой, модуль для парсинга и выделения семантических структур, модуль построения семантико-временного графа знаний, а также методы поиска и ранжирования релевантной информации и обработки ширококонтекстных запросов на естественном языке.

Автоматизированный модуль сбора данных с семантической компонентой (МСД-СК) будет представлять собой универсальный краулер данных из различных источников, объединенный с моделью анализа семантики и определения тематик в текстовых данных. При реализации краулера сбор данных будет производиться посредством подключения к API социальных медиа, форумов и других платформ. Анализ семантики будет основан на трансформерных моделях, таких как BERT [1]. Трансформеры, особенно BERT, предназначены для работы с контекстуальной информацией, что позволяет модели глубже понимать значение слов в различных ситуациях, что важно при семантическом анализе. Помимо этого, трансформерные модели легко масштабируются и могут обрабатывать большие объемы данных, что критически важно для работы с МСД-СК.

Модуль для автоматизированного парсинга и выделения семантических структур (МПСС) будет предназначен для выявления и извлечения текстовых блоков, объединенных единым семантическим контекстом, из собранных с помощью МСД-СК данных произвольного формата и структуры. Дополнительно на этом этапе из текстов будет выделяться информация о временной метке при ее наличии. Для уточнения полученной структуры для всех типов документов планируется дополнительно использовать методы глубокого обучения. Это поможет найти закономерности и структуры в частях документов, которые не имеют разметку. На следующем этапе после структурного разбора документов будут использоваться продвинутые модели анализа семантики, такие как методы на основе аддитивной регуляризации (ARTM) и трансформерные модели на основе BERT. ARTM [2] позволяет одновременно использовать несколько регуляризаторов, что делает модель более

гибкой, дает возможность учитывать различные аспекты текстовых данных и влиять на интерпретируемость и различимость выделяемых тематик, избегая их избыточности и смешения.

Модуль построения семантико-временного графа знаний (СВГЗ) предназначен для построения модели знаний на основе выделенных МПСС семантических структур с временной компонентой, которая позволяла бы эффективно учитывать сложные семантические зависимости между такими структурами, отслеживать изменения информации со временем и оптимизировать запросы на получение релевантных данных по заданному семантическому контексту и времени.

Для анализа и обработки пользовательских запросов на естественном языке будут разработаны метод поиска и ранжирования информации в СВГЗ и метод обработки ширококонтекстных пользовательских запросов. Метод поиска и ранжирования информации в СВГЗ (МПРИ) будет принимать на вход неструктурированный текст пользовательского запроса и выделять из него информацию о семантике и времени, по которым осуществляется поиск наиболее релевантных текстовых блоков из СВГЗ. Для выделения информации о семантике (тематиках, присутствующих в запросе) и времени будут использоваться подходы, аналогичные методам анализа семантики в МПСС – классические NLP методы и трансформерные модели. Метод обработки ширококонтекстных пользовательских запросов на естественном языке (МОПЗ) также принимает на вход неструктурированный текст пользовательского запроса. Этот текст, совместно с результатами МПРИ в виде набора релевантных запросу документов из базы знаний, представленной СВГЗ, поступает на вход алгоритму формирования запроса к большой языковой модели. Данный алгоритм будет агрегировать текст запроса и информацию из релевантных документов из СВГЗ, формируя итоговый запрос к большой языковой модели, задавая ей специализированный контекст, в рамках которого документы из СВГЗ будут представляться как энциклопедические знания для формирования ответа на профильные запросы. В качестве большой языковой модели, формирующей итоговый ответ, будет использоваться модель LLaMA [3], поскольку она обеспечивает открытый доступ и возможность донастройки и дообучения. На выходе метода будет текстовый ответ на естественном языке, который будет передаваться обратно пользователю.

Выводы. В результате данной работы спроектирована структура системы организации сбора, хранения и анализа предметных данных, которая будет использоваться для построения вопросно-ответных сервисов-ассистентов, позволяющих взаимодействовать с пользователем-экспертом, и отвечать на предметные вопросы на естественном языке. Дополнительно определен набор технологий, которые будут использоваться при реализации, а также описана структура ключевых методов.

Список использованных источников:

1. Devlin J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding // arXiv preprint arXiv:1810.04805 – 2018.
2. Konstantin Vorontsov, Oleksandr Frei, Murat Apishev, Peter Romov, Marina Dudarenko // Bigartm: Open source library for regularized multimodal topic modeling of large collections – 2015 – Analysis of Images, Social Networks and Texts: 4th International Conference, AIST 2015, Yekaterinburg, Russia, April 9–11, 2015, Revised Selected Papers 4 – C. 370-381.
3. H.Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix, B.Rozière, N. Goyal, E. Hambro, F.Azhar, et al. Llama: Open and efficient foundation language models // arXiv preprint arXiv:2302.13971 – 2023.