УДК 004.93'12

Разработка метода извлечения семантической структуры документов для ИИ-консультанта

Верзаков А. Ю. (ИТМО)
Научный руководитель – кандидат технических наук, доцент Быковский С. В. (ИТМО)

Введение

Активное развитие алгоритмов машинного обучения, в частности нейронных сетей и больших языковых моделей [1] открывают огромное количество новых возможностей для создания ИИ-консультантов на основе RAG-систем [2]. Однако качество работы таких систем напрямую зависит от качества представленных знаний в виде векторов, которые формируются из различных видов источников – текста, изображений, видеоматериалов.

Ключевая проблема при формировании баз знаний является обработка документов визуального формата (PDF, сканированные изображения). Традиционные методы часто теряют семантическую структуру текста (многоколоночную верстку, таблицы, заголовки), что значительно снижает качество сформированных векторов и, как следствие, релевантность ответов больших языковых моделей. Целью данной работы является разработка гибридного метода извлечения текста и его семантической структуры для повышения качества работы RAG-систем.

Основная часть

В ходе исследования был проведен сравнительный анализ существующих готовых решений и подходов к извлечению информации. Рассмотрены два основных направления:

1. Эвристический программный анализ (чтение PDF-тегов и координат символов). Метод обладает высокой скоростью, но низкой устойчивостью к сложной верстке.
2. OCR-подходы (Optical Character Recognition) [3]. Использование нейронных сетей, таких как YOLO и TableFormatter обеспечивает высокую точность извлечения, но требует значительных вычислительных ресурсов.

На основе анализа предлагается гибридный алгоритм, объединяющий преимущества обоих подходов (Text + Visual Recognition). Были выделены пять основных этапов:

1. Классификация и детекция: преобразование страниц документа в изображения с последующим обнаружением и классификацией каждого отдельного блока текста на странице.
2. Фильтрация и упорядочивание: очистка блоков от шумных данных (колонтитулы, нумерация страниц) по меткам классов и сортировка блоков текста в порядке чтения (XY-sort).
3. Определение читаемости: валидация возможности программного извлечения текста (проверка кодировки и целостности символов) для минимизации использования OCR.
4. Адаптивное извлечение: выбор метода обработки (программное чтение или визуальное распознавание через OCR) на основе предыдущего этапа.
5. Семантическое связывание: восстановление логической целостности абзацев, разорванных переносами страниц и колоночной верстской.

Выводы

Разработан гибридный метод извлечения семантической структуры текста из визуальных форматов документов (PDF, сканированные документы). Метод обеспечивает баланс между скоростью обработки документов и вычислительной эффективностью.

Литература

1. Vaswani A. et al. Attention Is All You Need // Advances in Neural Information Processing Systems (NeurIPS). — 2017. — Vol. 30.
2. Patrick Lewis Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks // Patrick Lewis // Facebook AI Research – 2020.
3. Xu Y. et al. LayoutLM: Pre-training of Text and Layout for Document Image Understanding // Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '20). — 2020. — P. 1192–1200.