УДК 004.89

ИНСТРУМЕНТ ДЛЯ КОНТРОЛЯ КАЧЕСТВА ПАРСИНГА ТЕКСТОВЫХ ДОКУМЕНТОВ НА ОСНОВЕ LLM

Мурашов Никита Александрович (ИТМО), Филатова Анастасия Алексеевна (ИТМО), Тимощак Егор Владимирович (ИТМО), Сандалов Александр Дмитриевич (ИТМО)

Научный руководитель — кандидат технических наук, доцент Насонов Денис Александрович (ИТМО)

Мотивация и актуальность

Современные технологии обработки данных позволяют извлекать информацию из неструктурированных источников, таких как PDF-документы, что открывает возможности для создания более точных и информативных моделей в таких областях, как информационный поиск и извлечение знаний. Однако, несмотря на достижения в области оптического распознавания символов (OCR) и парсинга, PDF-документы, полученные через оцифровку, часто содержат ошибки, такие как потеря форматирования, неверная интерпретация таблиц и диаграмм, а также неправильная кодировка текста. Эти дефекты, в свою очередь, оказывают негативное влияние на последующие этапы обработки данных, например, в системах на базе больших языковых моделей (LLM), где качество исходных данных критически важно для корректной работы RAG-систем (retrieval-augmented generation) и других информационно-аналитических пайплайнов.

Описание входных данных

В качестве входных данных для разработанного алгоритма используются текстовые фрагменты, называемые чанками, полученные путем парсинга PDF-документов. Каждый чанк соответствует одному параграфу исходного документа и включает метаинформацию, которая описывает структуру текста, в частности, иерархию заголовков. Эта иерархия определяет, к какой главе или разделу относится данный чанк, что важно для правильной интерпретации контекста текста. В процессе валидации осуществляется оценка не только качества самого текста чанка, но и корректности извлеченной иерархии заголовков, а также проверки на правильность разделения текста на отдельные фрагменты. Это позволяет убедиться, что каждый чанк адекватно представляет части исходного документа, а структура текста сохранена.

Описание алгоритма

Предложенная система основывается на методах машинного обучения (ML), которые используют неэвристические подходы для оценки качества парсинга. Система направлена на проверку трех ключевых аспектов текста: целостности, связности и логической последовательности фрагментов. Алгоритм включает следующие важнейшие компоненты:

1) **Использование эмбеддеров для анализа структуры текста.**
   В рамках оценки структуры текста применяется модель-эмбеддер e5-large-multilingual, которая используется для анализа иерархии заголовков и проверки связности текстовых фрагментов (чанков) в пределах одного раздела документа. Эмбеддеры создают высококачественные векторные представления каждого чанка, что позволяет эффективно измерять схожесть между текстами. Это позволяет выявить потенциальные несоответствия в разбиении текста, такие как ситуации, когда текстовый чанк был ошибочно выведен за

пределы своей логической части (например, заголовок оказался в другом разделе или чанк был разорван). Такой подход помогает гарантировать, что структура документа сохраняется и каждый чанк логически связан с другими частями.

2) **Применение больших языковых моделей (LLM) для оценки логических признаков текста.**
   Для более глубокого анализа содержания чанков применяется языковая модель Qwen-2.5-32b, которая выполняет оценку смысловой полноты и логической связности каждого чанка текста. На этом этапе система проверяет, насколько хорошо сформулированы мысли внутри чанка, насколько грамматически и семантически правильно построены предложения, а также оценивает их последовательность. Модель анализирует, нет ли пробелов в логике текста или нарушений в его структуре, таких как отсутствие необходимых связок между предложениями или неполные, обрывочные мысли. Дополнительно проводится проверка на наличие ошибок в кодировке текста (например, некорректно отображаемые символы) и шумов, которые могут возникнуть в процессе парсинга.

Методы оценки

Для автоматической оценки результатов используется модель Geval, специально обученная для анализа качества текста по множеству различных критериев. Среди этих критериев – грамматическая корректность, структурная целостность и семантическая последовательность. Модель Geval [1] позволяет оценить, насколько хорошо структурированы текстовые чанки и насколько они соответствуют стандартам грамотности и логической последовательности. Дополнительно используется модель DeepEval [2], которая помогает проводить более сложный анализ текста, фокусируясь на таких аспектах, как читабельность, контекстуальная адекватность и соответствие смысловых связей. DeepEval позволяет более глубоко оценить семантические и стилистические особенности текста, что важно для выявления сложных ошибок, которые могут быть незаметны при более поверхностной проверке.

Список использованных источников:

1. Yang Liu et al. "G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment", arXiv preprint arXiv:2303.16634, 2023

2. Confident AI. DeepEval: A framework for deep learning-based evaluation of natural language generation [Электронный ресурс] / Confident AI. — GitHub репозиторий. — URL: https://github.com/confident-ai/deepeval (дата обращения: 23.01.2025).

Автор Мурашов Н. А.

Научный руководитель Насонов Д. А.