СРАВНЕНИЕ ПОДХОДОВ К ИЗВЛЕЧЕНИЮ ДОПОЛНИТЕЛЬНЫХ ДАННЫХ ДЛЯ ГЕНЕРАЦИИ В RAG-СИСТЕМАХ

Ксенофонтов А. А., Михайлов А. А.

Ключевые слова: семантический поиск, генерация, векторизация

В настоящее время RAG-системы являются одним из актуальных направлений развития больших языковых моделей. Несмотря на высокие способности современных больших языковых моделей к генерации текста, их применение при работе с документами ограничено длиной контекстного окна, ухудшением точности по мере увеличения размера контекста и вероятностью галлюцинаций. RAG позволяет объединить генеративные возможности языковых моделей с механизмами поиска релевантной информации.

Одним из ключевых этапов работы любой RAG-системы является извлечения данных из исходных документов. Именно качество извлеченного текста и визуальных элементов напрямую влияет на дальнейший поиск релевантных данных и итоговое качество ответа модели. В настоящее время активно разрабатывают и изучают различные подходы для извлечения данных из документов, начиная от оптического распознавания текста и заканчивая использованием продвинутых мультимодальных языковых моделей, которые способны одновременно распознавать текст и визуальные элементы.

В рамках данной работы сравниваются различные подходы к извлечению дополнительных данных в RAG-системе. Рассматривается сценарий, в котором извлеченный текст документа и пользовательский вопрос преобразуются в векторные представления, после чего выполняется поиск релевантных фрагментов по косинусному сходству. Основное внимание уделяется влиянию методов извлечения и предварительной обработки текста на итоговое качество получаемых данных.

Для проведения экспериментов использовалось подмножество набора данных LongDocURL[1], предназначенного для оценки качества языковых моделей при работе с длинными документами. Данный набор данных включает документы большого объема и вопросы различных типов. Для тестирования были выбраны вопросы на понимание текста и обнаружение, при этом в качестве элементов с ответом были взяты текст и разметка.

В работе были протестированы пять различных подходов к подготовке данных и поиску контекста.

Первый подход представляет собой базовый сценарий без дополнительных данных, при котором модель получает только вопрос. Этот вариант используется как контрольный и позволяет оценить вклад внешнего контекста в итоговую точность ответов.

Второй подход заключается в том, что из исходного PDF-документа выбираются 30 последовательных страниц, содержащих данные для ответа. Затем весь извлеченный текст из этих страниц подается модели.

Третий подход основан на разделении документа на части. После извлечения текста документ разбивается на отдельные фрагменты (размер 500 с перекрытием 100 символов), каждый из которых превращается в вектор и индексируется отдельно.

Четвертый и пятый подходы основаны на использовании PDF-парсеров MinerU[2] и PageR[3], которые разбивают документы на определенные блоки.

Результат представлены в следующей таблице.

Подход	1	2	3	4	5
Общая точность	14.09%	40.66%	34.80%	30.39%	31.90%
Понимание	18.46%	45.85%	42.57%	36.83%	40.11%
Обнаружение	6.20%	31.31%	20.79%	18.75%	17.09%

Как можно видеть, прямое сопоставление методов, основанных на PDF-парсинге (MinerU и PageR), с более простым подходом разбиения текста на части показывает, что парсеры в текущей постановке задачи демонстрируют более низкую точность.

Это может свидетельствовать о том, что структурная информация, извлекаемая PDF-парсерами, в рамках тестируемого подхода не конвертируется напрямую в прирост качества. Иными словами, преимущества более сложного парсинга не реализуются без дополнительных механизмов обработки

Список литературы

1. LongDocURL [Электронный ресурс]. – URL: https://github.com/dengc2023/LongDocURL (дата обращения: 30.04.2026).

2. MinerU [Электронный ресурс]. – URL: https://github.com/opendatalab/mineru (дата обращения: 30.04.2026).

3. PageR [Электронный ресурс]. – URL: https://github.com/YRL-AIDA/PageR (дата обращения: 30.04.2026).