Разработка средства анализа ответов в конвейере RAG с применением графов

Науки о данных

Подготовил: Гладышев Виталий Владимирович

Проблема и контекст исследования

Большие языковые модели (LLM) – важный технологический прорыв современности!

Для практического применения этих моделей на практике используется технология RAG

Качественная реализация RAG на продуктовом уровне – сложная задача требующая точного представления о влиянии изменений на результат.

Метрики оценки конвейера RAG имеют ограничения:

- Часто применяется подход использования LLM как судьи. Соответственно оценки непрозрачны и сложны для интерпретации
- Традиционные метрики (ROUGE, BLEU) не оценивают структурную согласованность ответов
- Ошибки RAG-систем в критичных доменах

МФТИ

Цель исследования

Реализовать средство анализа ответов в конвейере RAG с применением графов для обеспечения интерпретируемости и прозрачности определения влияния семантики ответа на количественные характеристики

Объект исследования:

Конвейер RAG

Предмет исследования:

Семантическое моделирование с помощью графов

Исследовательский вопрос:
Как графовые представления влияют на объяснимость и точность оценки этапов извлечения и генерации в RAG?

МФТИ

Задачи исследования

1. Интеграция RAG с графовыми подходами – Изучить, как комбинированные методы, использующие графы и RAG, могут улучшить качество генерации и понимания текстов в различных доменах.
2. Применить и развить подход реализованный в библиотеке GraphRAG применительно к конвейеру RAG для численного анализа семантических признаков вопроса пользователя и контекста получаемого семантическим поиском.

Анализ источников

- Анализ структуры графов способствует более глубокому пониманию сложных взаимосвязей в данных, что критично для задач, связанных с RAG.
- Интеграция методов RAG с подходами, основанными на графах, способна повысить качество генерации и понимания текстов во множестве доменных областей.
- Определена новая методика построения семантических моделей построенных на графах – применение больших языковых моделей для получения более качественных и сложных моделей [1].

Новизна

Отсутствие унифицированных метрик для оценки RAG с использованием графов. Существуют общие метрики сравнения графов, но не для оценки качества извлечения и генерации в RAG.

Актуальность

Рост использования RAG-систем в критически важных областях (медицина, финансы, юридический анализ) требует повышения доверия к их результатам. Метрики RAG имеют сложности с интерпретируемостью

Исследование заполняет пробел между теоретическими работами по графам и прикладными задачами RAG, предлагая метод, который одновременно улучшает точность и прозрачность оценки. Графы, в отличие от векторов, сохраняют структурные и семантические связи между текстом, изображениями и знаниями

МФТИ

Гипотеза исследования

Использование семантических графов для моделирования контекста в RAG-конвейере позволяют реализовать сравнение семантической структуры вопроса пользователя и контекста с привязкой к отдельным элементам вопроса и контекста в отличии от существующих метрик таких как Faithfulness или Answer Accuracy. Соответственно можно связать количественные параметры с конкретными элементами семантической структуры.

МФТИ

Метод GraphRAG (Edge et al., 2024)

Архитектура:

Текст → LLM (извлечение сущностей) → Граф знаний → RAG

Преимущества:

- Автоматическое построение графов из неструктурированного текста
- Повышение объяснимости

Пример:

Текст → Узлы (сущности) → Ребра (связи)

Построение графа контекста

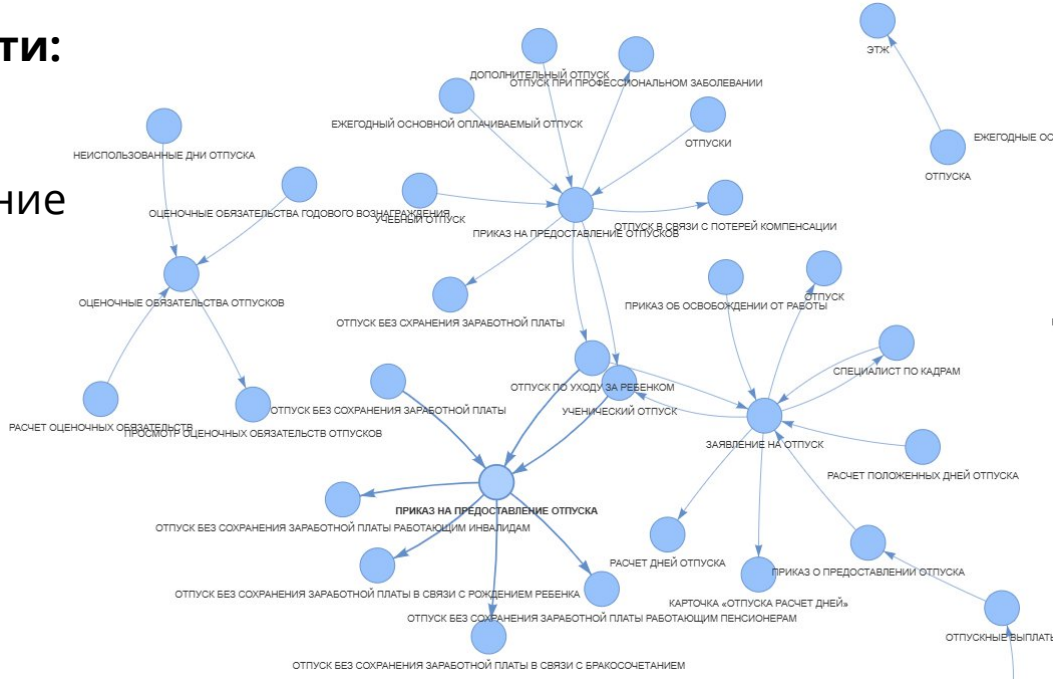
Граф:

Узлы - извлечённые сущности:

- Заявление на отпуск
- Приказ на предоставление отпуска
- Ученический отпуск

Ребра - действия:

- Подать
- Оформить
- Предоставить



Методология исследования

1. Построение графов через GraphRAG:

- Исходный контекст → Эталонный граф
- Ответ RAG → Тестовый граф

2. Расчёт метрик:

- Graph Edit Distance (GED)
- Сравнение состава узлов и связей

3. Валидация:

Подтверждение, что граф ответа является подграфом контекста

МФТИ

План реализации

1. В качестве источника данных будет использован датасет Vikhrmodels/Grounded-RAG-RU-v2, а также специально сгенерированные синтетические наборы данных

2. Реализовать редство анализа, интерпретации и визуализации результатов измерения

3. Оптимизировать GraphRAG для потоковой обработки

Список ключевых источников

1. Edge D. et al. From local to global: A graph rag approach to query-focused summarization //arXiv preprint arXiv:2404.16130. – 2024.
2. Wu L. et al. Graph neural networks for natural language processing: A survey //Foundations and Trends® in Machine Learning. – 2023. – Т. 16. – №. 2. – С. 119-328.
3. He X. et al. G-retriever: Retrieval-augmented generation for textual graph understanding and question answering //Advances in Neural Information Processing Systems. – 2024. – Т. 37. – С. 132876-132907.
4. Wills P., Meyer F. G. Metrics for graph comparison: a practitioner's guide //Plos one. – 2020. – Т. 15. – №. 2. – С. e0228728.
5. Lovász L., Plummer M. D. Matching Theory, vol. 29 //Annals of Discrete Mat North-Holland, Amsterdam. – 1986.

МФТИ

Спасибо
за внимание!

МФТИ