

Домашнее задание 4. Дизайн исследования и сбор данных

Гладышев Виталий Владимирович

Для подготовки раздела «Методология исследования и данные» Research Proposal разработан детальный план проверки гипотезы о влиянии семантических графов на объяснимость и точность оценки этапов извлечения и генерации в RAG-конвейере. Для этого определены методы сбора и анализа данных, описан процесс эксперимента, а также обоснована связь выбранных подходов с решением поставленных исследовательских задач. В рамках исследования будут использованы смешанные методы, сочетающие количественные метрики структурного анализа графов и качественную интерпретацию их семантических связей.

1 Выбор методов исследования

Основным методом исследования станет адаптация подхода GraphRAG (Edge et al., 2024), который предполагает автоматическое построение графов знаний из неструктурированного текста с использованием больших языковых моделей (LLM). Этот метод был выбран по нескольким причинам. Во-первых, он позволяет преобразовывать текстовый контекст и ответы RAG в семантические графы, узлы которых соответствуют сущностям (например, «приказ на отпуск»), а рёбра — связям между ними (например, «оформить»). Во-вторых, GraphRAG позволяет выполнить формирование графа знаний в удобной форме (сущности, связи, комьюнити и другие элементы доступны в виде parquet-файлов) с высокой степенью автоматизации. Формирование необходимых параметров также может быть автоматизировано.

Для количественной оценки различий между графами будет использована метрика Graph Edit Distance (GED), измеряющая минимальное количество операций (добавление, удаление, замена узлов/рёбер), необходимых для преобразования одного графа в другой. Эта метрика позволяет объективно оценить, насколько структура ответа RAG соответствует эталонному контексту. Дополнительно будет проведён анализ совпадения узлов и рёбер, что поможет выявить семантические расхождения (например, пропущенные сущности).

Кроме того, для валидации результатов будет применяться метод проверки подграфа: если граф ответа является подграфом эталонного контекста, это свидетельствует о структурной согласованности. Данный подход дополняет традиционные метрики (ROUGE, BLEU), которые фокусируются на поверхностном сходстве текстов, игнорируя семантику.

Применение этих методов позволит закрыть пробел между теоретическими работами по графам и прикладными задачами RAG. Например, в отличие от исследований, где графы используются для классификации текстов, данный проект адаптирует их для оценки качества генерации.

2 Описание данных для исследования

Для исследования будут использоваться два вида данных:

1. Реальные данные из датасета Vikhrmodels/Grounded-RAG-RU-v2, содержащего вопросы на русском языке, контексты и эталонные ответы. Этот датасет охватывает разнообразные домены, что соответствует актуальности исследования для различных областей.
2. Синтетические данные, сгенерированные с помощью LLM (например, GPT-4o) для моделирования контролируемых аномалий: отсутствие ключевых сущностей, некорректные связи между узлами, избыточные элементы. Такие данные помогут проверить чувствительность метрик к структурным ошибкам.

Для обработки данных потребуются следующие ресурсы:

- Большие языковые модели (Qwen, Gemma, Saiga, Vikhr) — для извлечения сущностей и построения графов.
- Библиотеки NLP (SpaCy, Natasha) — для токенизации и лемматизации текста.
- Инструменты анализа графов (NetworkX, PyTorch Geometric) — для расчёта GED и визуализации структур.

3 Эксперимент и процесс исследования

Эксперимент будет проводиться в четыре этапа:

1. Построение эталонных графов. Исходные контексты из датасета преобразуются в графы с помощью GraphRAG. Например, текст «Сотрудник должен подать заявление за две недели до отпуска» будет представлен как граф с узлами «сотрудник», «заявление», «отпуск» и рёбро «подать».
2. Генерация тестовых графов. Ответы, полученные от RAG-системы, аналогичным образом преобразуются в графы. На этом этапе также будут созданы синтетические данные с преднамеренными ошибками для тестирования устойчивости метода.
3. Расчёт метрик. Для каждой пары «эталонный граф — тестовый граф» вычисляются GED, процент совпадения узлов/рёбер и проводится проверка на подграф.
4. Оптимизация. Графовые алгоритмы будут адаптированы для потоковой обработки данных, что повысит скорость анализа.

Инструментальная часть эксперимента включает написание Python-скриптов для автоматизации построения графов и расчёта метрик, а также использование Jupyter Notebook или Streamlit/Gradio UI для интерактивной визуализации результатов.

4 Обработка и анализ результатов

Количественный анализ будет направлен на установление корреляции между структурными метриками графов и субъективными оценками качества ответов. Например, если высокое значение GED соответствует низким оценкам точности (по данным экспертов), это подтвердит гипотезу о связи структурных различий с качеством генерации.

Качественный анализ сосредоточится на интерпретации конкретных ошибок. Визуализация графов позволит выявить паттерны: например, отсутствие узла «заявление» в тестовом графе может объяснить, почему RAG-система неверно ответила на вопрос о процедуре оформления отпуска. Такая интерпретация недоступна при использовании традиционных метрик вроде Answer Accurasy, которые лишь констатируют факт ошибки, но не объясняют её причин.

По итогам исследования будут разработаны:

1. Библиотека для анализа графов в RAG, поддерживающая расчёт GED, визуализацию и экспорт результатов в форматы JSON и CSV.
2. Аннотированные датасеты, включающие эталонные графы для контекстов и ответов RAG, что упростит воспроизведение экспериментов.
3. Методические рекомендации по интеграции графовых метрик в существующие конвейеры RAG.

5 Возможные ограничения и их нивелирование

Основное ограничение связано с зависимостью от качества LLM: ошибки в извлечении сущностей (например, объединение «заявления» и «приказа» в один узел) могут исказить графы. Для минимизации этого риска будет проведена кросс-валидация с использованием нескольких моделей (Qwen, Gemma, Saiga) и ручная проверка выборки данных.

Вычислительная сложность алгоритмов GED, которая растёт экспоненциально с увеличением размера графов.

Синтетические данные могут недостаточно точно отражать реальные сценарии. Для компенсации этого в эксперимент будут включены кейсы из медицинских и юридических датасетов, где структурная точность критически важна.

Предложенный дизайн исследования обеспечивает комплексный подход к проверке гипотезы. Сочетание GraphRAG, метрик структурного анализа и методов валидации позволяет не только оценить качество RAG-систем, но и выявить конкретные источники ошибок на этапах извлечения и генерации. Это напрямую способствует достижению цели исследования — повышению прозрачности и точности оценки конвейеров RAG через семантическое моделирование. Результаты проекта могут стать основой для новых стандартов в разработке объяснимых AI-систем для критически важных доменов.