УДК 004.8

ИССЛЕДОВАНИЕ АЛГОРИТМОВ ИЗВЛЕЧЕНИЯ ПОДГРАФА ДЛЯ ГЕНЕРАЦИИ SPARQL ЗАПРОСОВ С ИСПОЛЬЗОВАНИЕМ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

Плюхин Д.А. (ИТМО), **Радюш Д.В.** (ИТМО)
Научный руководитель – кандидат технических наук, доцент Муромцев Д.И.
(ИТМО)

Введение. Решение задачи построения вопросно-ответных систем над графами знаний с данными о результатах научных исследований включает в себя множество сложностей:

1. неоднородность способов представления данных;
2. различная степень детализации описаний исследований;
3. неполнота графов в силу эволюции научных знаний.

Ввиду указанных особенностей подходы, основанные на обработке шаблонов и использовании специализированных предобученных моделей не обеспечивают высокого качества генерируемых запросов, что обуславливает необходимость разработки методов, основанных на генеративных моделях. В частности, предлагается использование больших языковых моделей для решения поставленной задачи.

Основная часть. В данной работе представлен подход, основанный на больших языковых моделях, для генерации SPARQL-запросов в рамках корпуса Open Research Knowledge Graph (ORKG) для челленджа ISWC SciQA [5]. Данный подход предлагает несколько улучшений к ранее опубликованному методу SPARQLGEN [4], который ориентирован на генерацию SPARQL-запросов с использованием контекста, передаваемого в большие языковые модели наряду с запросом, сформулированным на естественном языке. Подобно SPARQLGEN, для генерации SPARQL запроса используются три компонента исходных данных:

1. сам вопрос;
2. RDF-подграф, необходимый для ответа на вопрос;
3. пример правильного SPARQL-запроса из обучающей выборки.

Основной акцент в исследовании сосредоточен на разработке алгоритмов извлечения подграфа, близких к реальным сценариям генерации запросов к графам знаний, а также замене случайного выбора примера вопроса-запроса на подход, использующий оценку схожести формулировки запроса на естественном языке и текстового описания запросов из обучающей выборки на основе подсчета расстояния Левенштейна.

Рассмотренные алгоритмы извлечения подграфа состоят из 4 шагов:

1. **предобработка вопроса на естественном языке** - удаление токенов, соответствующих служебным частям речи;
2. **извлечение релевантных компонентов графа** - поиск наиболее релевантных сущностей в графе на основе косинусного расстояния между векторным представлением результата предобработки запроса и текстовых меток;
3. **генерация фрагментов подграфа** - расширение сформированного множества релевантных компонентов за счет включения соседних вершин и смежных ребер, были рассмотрены два варианта реализации данной процедуры:
    a. первый подход предполагает использование описаний статей и результатов исследований в качестве основного источника информации, необходимой для генерации SPARQL запроса;
    b. второй подход предусматривает непосредственный поиск триплетов с похожими текстовыми метками независимо от того, описанию сущности какого типа в исходном графе они принадлежат;
4. **объединение компонентов подграфа** и преобразование результата в строку для

передачи в языковую модель.

Выводы. По результатам выполнения экспериментов были сформулированы следующие выводы:

1. Даже без использования графа знаний система характеризуется высоким качеством работы (значение F1-score равно 0.922), что свидетельствует о высокой однородности набора данных и большом структурном сходстве экземпляров обучающей и тестовой выборки. В данной конфигурации помимо текста вопроса на естественном языке в модель также передавался наиболее релевантный пример запроса из обучающей выборки;
2. Добавление подграфа в запрос с использованием первого подхода к извлечению релевантного фрагмента данных позволяет несколько улучшить качество работы системы (значение F1-score равно 0.935);
3. Добавление подграфа с использованием второго подхода к извлечению релевантного фрагмента данных приводит к ухудшению качества работы системы относительно базовой конфигурации (F1-score становится равным 0.916), что свидетельствует о низкой релевантности извлеченных данных.

Список использованных источников:

1. M. Färber, D. Lamprecht, J. Krause, L. Aung, P. Haase Semopenalex: The scientific landscape in 26 billion rdf triples // arXiv preprint – 2023 – arXiv:2308.03671
2. D. Dessí, F. Osborne, D. Reforgiato Recupero, D. Buscaldi, E. Motta, Cs-kg A large-scale knowledge graph of research entities and claims in computer science // International Semantic Web Conference, Springer – 2022 – C. 678–696.
3. S. Vahdati, G. Palma, R. J. Nath, C. Lange, S. Auer, M.-E. Vidal Unveiling scholarly communities over knowledge graphs // International Conference on Theory and Practice of Digital Libraries, Springer – 2018 – C. 103–115.
4. Kovriguina L., Teucher R., Radyush D., Mouromtsev D. SPARQLGEN: One-Shot Prompt-based Approach for SPARQL Query Generation // CEUR Workshop Proceedings – 2023 – Vol. 3526
5. Scholarly QALD at ISWC 2023 URL: https://www.wikidata.org (Дата обращения: 02.02.2024)