Enhancing QA over Scholarly Knowledge Graphs: Addressing Semantic and Structural Challenges

Xueli Pan$^{1[0000-0002-3736-7047]}$, Victor de Boer$^{1[0000-0001-9079-039X]}$, and Jacco van Ossenbruggen$^{1[0000-0002-7748-4715]}$

Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, Netherlands
{x.pan2, v.de.boer, jacco.van.ossenbruggen}@vu.nl

Scholarly knowledge graphs (SKGs) represent the bibliographic metadata and scientific elements such as research problems, theories, approaches, evaluations. Question answering (QA) over SKGs demonstrates significant challenges due to the intricate nature of scholarly data and the complex structure of SKGs. The task of QA over SKGs usually takes a natural language question (NLQ) as the input and generates a corresponding SPARQL query to determine its correctness [3, 6].

The emergence of large language models (LLMs) has inspired a growing body of research exploring their potential to address the challenges of QA over SKGs [3, 5, 6]. Lehmann et al.[4] presented an in-depth examination of the performance of LLMs on the SciQA benchmark, focusing on different optimizing techniques such as zero-shot learning, few-shot learning and fine-tuning. Despite the improvements of LLMs on QA tasks over SKGs, LLMs face limitations when handling KG-specific parsing due to their lack of direct access to entities within the knowledge graph and insufficient understanding of the ontological schema, particularly for low-resource SKGs like the Open Research Knowledge Graph (ORKG)[1]. The experimental results conducted by Sören Auer et al.[1] showed that only 63 out of 100 handcrafted questions could be answered by ChatGPT, of which only 14 answers were correct.

To better understand why LLMs fails to generate the correct SPARQL query to a NLQ, we conduct a pilot experiment on using ChatGPT(GPT-4) to generate SPARQL queries for 30 handcrafted questions in the SciQA benchmark datasets. Insights from this pilot experiment revealed two major categories of errors LLMs tend to make in this task: semantic inaccuracies and structural inconsistencies.

Semantic inaccuracies occur when LLMs fail to link the correct properties and entities in ORKG, despite generating SPARQL queries with correct structure. Our observations reveal that LLMs tend to rely on examples provided in the few-shot learning process to generate the correct structure for a certain type of questions, but often struggle with linking the correct properties and entities because LLMs do not learn the content of the ORKG. We propose a RAG approach to generate the top k candidate properties or entities from ORKG based on the properties and entities mentioned in the NLQs, for LLMs to use as a context while generating the SPARQL queries.

Structural inconsistencies arise due to LLMs' lack of ontological schema of the ORKG, leading to errors in query structure, such as missing or abundant links (triples), despite correctly linking to the mentioned entities or properties.

We suggest that fine-tuning LLMs with ontological information from the ORKG can help address these structural issues, allowing the model to generate more accurate queries with appropriate multi-hop relations. We proposed to address these problems by fine-tuning LLMs with two different datasets: 1) the NL-SPARQL pairs in SciQA benchmark dataset and 2) the triples in ORKG.

To make sure the final generated SPARQL queries are syntax correct, we also add a LLM-based SPARQL corrector component to our proposed framework named Natural Language Question to SPARQL with RAG and Fine-Tuning (NLQ2SPARQL-RAGFT), as shown in Figure 1.

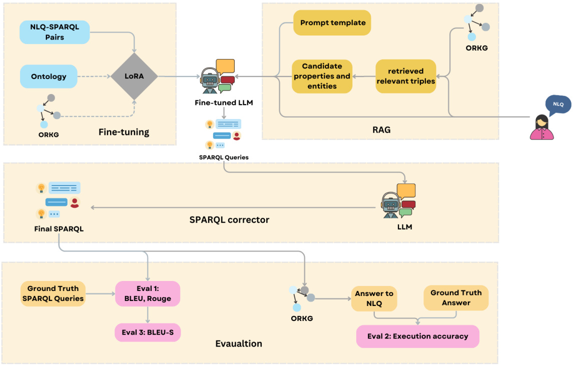

Fig. 1. NLQ2SPARQL-RAGFT Framework

Additionally, we highlight the limitations of traditional machine translation evaluation metrics like BLEU and ROUGE, which rely on n-gram token overlap and fail to detect semantic issues in generated queries. These evaluation metrics lead to high scores despite low execution accuracy when queries contain incorrect properties or entities. To address this, we propose a more nuanced metric named BLUE-S, considering both structural correctness and semantic accuracy.

$$\text{BLEU-S} = \lambda B + (1 - \lambda) S$$

where λ is a weighting factor (e.g., 0.5 for equal contribution), B is the BLEU score and S is the cosine similarity between generated and ground-truth SPARQL query embeddings.

We conducted experiments on the SciQA Benchmark dataset using a fine-tuned Llama 3.2-3B with LoRA[2], excluding the RAG component. The model achieved a BLEU-4 score of 0.49, a ROUGE-1 score of 0.76, and a ROUGE-2 score of 0.71. More details could be found in this GitHub repository [1]

[1] https://github.com/sherry-pan/QAoverSKGs

References

1. Auer, S., Barone, D.A.C., Bartz, C., Cortes, E., Jaradeh, M.Y., Karras, O., Koubarakis, M., Mouromtsev, D.I., Pliukhin, D., Radyush, D., Shilin, I., Stocker, M., Tsalapati, E.: The sciqa scientific question answering benchmark for scholarly knowledge. Scientific Reports **13** (2023), https://api.semanticscholar.org/CorpusID:258507546
2. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
3. Jiang, L., Yan, X., Usbeck, R.: A structure and content prompt-based method for knowledge graph question answering over scholarly data. In: QALD/SemREC@ ISWC (2023)
4. Lehmann, J., Meloni, A., Motta, E., Osborne, F., Recupero, D.R., Salatino, A.A., Vahdati, S.: Large language models for scientific question answering: An extensive analysis of the sciqa benchmark. In: The Semantic Web: 21st International Conference, ESWC 2024, Hersonissos, Crete, Greece, May 26–30, 2024, Proceedings, Part I. p. 199–217. Springer-Verlag (2024). https://doi.org/10.1007/978-3-031-60626-7_11
5. Pliukhin, D., Radyush, D., Kovriguina, L., Mouromtsev, D.: Improving subgraph extraction algorihtms for one-shot sparql query generation with large language models. In: QALD/SemREC@ ISWC (2023)
6. Taffa, T.A., Usbeck, R.: Leveraging llms in scholarly knowledge graph question answering. In: QALD/SemREC@ ISWC (2023)