A Structure and Content Prompt-based Method for Knowledge Graph Question Answering over Scholarly Data

Longquan Jiang[1,2], Xi Yan[1,2] and Ricardo Usbeck[3]

[1]*Semantic Systems Group, Universität Hamburg, Vogt-Kölln-Str. 30, F-434, 22527, Hamburg, Germany*
[2]*Hamburger Informatik Technologie-Center e.V. c/o, Vogt-Kölln-Str. 30, 22527, Hamburg Germany*
[3]*Leuphana Universität Lüneburg, Universitätsallee 1, C 4.314, 21335 Lüneburg, Germany*

Abstract

Answering scholarly questions is challenging without the help of query-based systems. Thus, we develop a divide-and-conquer approach based on a Large Language Model (LLM) for scholarly Knowledge Graph (KG) Question Answering (QA). Our system integrates the KG ontology into the LLM prompts and leverages a hybrid prompt learning strategy with both query structure and content. Our experiments suggest that given an ontology of a specific KG, LLMs are capable of automatically choosing the corresponding classes or predicates required to generate a target SPARQL query from a natural language question. Our approach shows state-of-the-art results over one scholarly KGQA dataset, namely sciQA [1].

Keywords

Question Answering, KGQA, Scholarly KGQA, Large Language Model

1. Introduction

Motivation. Over the last decades, Knowledge Graph Question Answering (KGQA), which translates natural language questions into SPARQL queries, has gained much attention from both the industry and academia as it eases access to Knowledge Graphs for non-experts. While substantial progress has been made in this domain, there remains a pressing need for more versatile and adaptable KGQA systems.

One of the key challenges in advancing KGQA technology is the heterogeneity of KG ontologies across various domains. Each knowledge domain often comes with its own unique set of classes, properties, and relationships, which poses a significant hurdle in building generalizable systems. Subsequently, the majority of the existing systems are developed and trained against specific KG ontologies, such as [2], [3]. KGQA algorithms based on LLMs heavily rely on large amounts of training data. This poses a threat to the domain-specific KGQA problems where often only limited amounts of data are available such as in the scholarly or biomedical domain. This necessitates the development of KGQA solutions that are not bound to a specific KG ontology, but rather possess the ability to seamlessly adapt unseen KGs.

Scholarly QALD at ISWC 2023
longquan.jiang@uni-hamburg.de (L. Jiang); xi.yan@uni-hamburg.de (X. Yan); ricardo.usbeck@leuphana.de (R. Usbeck)
0000-0002-7333-2589 (L. Jiang); 0009-0004-7379-4080 (X. Yan); 0000-0002-0191-7211 (R. Usbeck)

Therefore, we want to create a KGQA system that overcomes the limitations of ontology-specific models. To achieve this, we propose leveraging LLMs through powerful and efficient prompt tuning [4]. By integrating the KG ontology into the LLM prompts and applying a judicious prompt tuning strategy, we aim to empower our system to understand and navigate the intricacies of different KGs, irrespective of their specific ontological configurations.

State-of-the-Art and Challenges. Research Knowledge Graphs, such as DBLP [5] and ORKG [6], have no existing natural language interface hindering users from posing complex queries such as *How many papers in conference X talk about Y?*. Due to the lack of such an interface, there are few studies on scientific KGQA, with only 3 datasets, namely ORKG-QA benchmark [7], DBLP-QuAD [8] and SciQA [1]. Among the former works, no research has been done on SciQA and ORKG-QA benchmarks in terms of KGQA system development. As for DBLP-QuAD, the state-of-the-art system is a T5-based question answering (aka Semantic Parsing) approach [8] which achieves 86.8% F1 on the testing set. However, their model is built on the premise of perfectly linked entities and relations, which limits the system to a specific KG ontology and heavily depends on the separate linking module.

Approach. We develop a KGQA system based on a Large Language Model, namely T5 [9], using prompt tuning techniques. T5 is a pioneering pre-trained language model in converting a wide range of NLP (natural language processing) tasks into a unified text-to-text format. It greatly pushes forward the prompt tuning technique, a pivotal technique that empowers the model to adapt to specific tasks through finely crafted prompts. Our system's idea follows the experience of a hybrid prompt tuning [10], which generates the structure and the content of queries in a two-step manner. An overview can be seen in Figure 1. More details can be found in Section 3

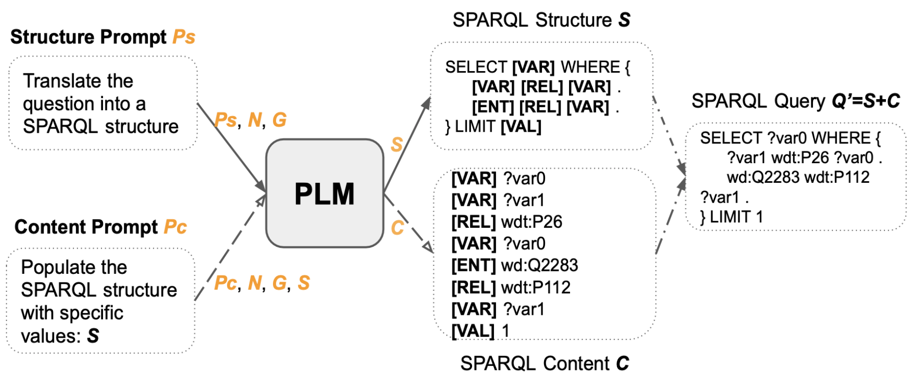

Figure 1: The framework with two-stage prompt learning

Evaluation. We evaluate our system over two KGQA datasets, namely SciQA and DBLP-QuAD. They serve as two tracks of the scholarly QALD challenge at ISWC 2023[1]. The organizers host separate benchmark pages for SciQA [2] and DBLP-QuAD [3], where the leaderboard of

[1]https://kgqa.github.io/scholarly-QALD-challenge/2023/
[2]https://codalab.lisn.upsaclay.fr/competitions/14759
[3]https://codalab.lisn.upsaclay.fr/competitions/14264

the challenge and dataset descriptions are illustrated. Our evaluation result is uploaded and published under the "Public Submissions" section, see user `longquanj`. As we see from the leaderboards, our system is not only competitive but also adaptable to various knowledge graphs with different topologies.

Contribution. We develop the first, scholarly KG-agnostic KGQA system based on an LLM. Our source code can be found at our github repository[4].

2. Tasks and Datasets

This paper describes our solution submitted for both Task 1 and Task 2 of the Scholarly QALD 2023 challenge both focusing on simple and complex question answering over scientific knowledge graphs.

Task 1 of this challenge is using the DBLP-QuAD dataset [8] to perform Entity Linking (EL) and Question Answering (QA) respectively over the DBLP KG, a well-known repository for computer science bibliography. DBLP-QALD consists of 10,000 questions paired with their corresponding SPARQL queries that were generated using a variety of human-written templates, covering both simple and complex questions with some constraints. In addition, it shows the potential of evaluating the compositional generalization of KGQA systems.

Task 2 uses the SciQA [1] dataset, a novel scientific QA benchmark for the Open Research Knowledge Graph (ORKG), to answer complex questions about bibliographic metadata and scientific elements, e.g., ideas, theories, and approaches. The formation process of scientific questions follows a bottom-up method in which a small set of complex questions is manually developed that is answerable over ORKG, and a larger set of questions is automatically generated using delicate eight templates.

We conduct an analysis of the structure of the SPARQL queries of both datasets. The numbers of the unique SPARQL structures in both datasets are displayed in Table 2. By structure, we refer to SPARQL keywords, operators, aggregators, and placeholders for KG schema items (e.g., relations, entities and classes). For SciQA, there are 7 specific structures that are in the validation set but not in the training set and 18 in the test set but not in training. For DBLP-QuAD, 9 structures from the validation set are not found in the training set and 9 from the test set are missing in the training set.

	train	valid	test
DBLP-QuAD	56	64	65
SciQA	64	19	28

Table 1
The numbers of the unique query structures

[4] https://github.com/semantic-systems/ScholarlyQuAD-QA-Solution

3. System Description

Due to the poor generalization of LLMs in low-resource scenarios, the ”pre-train and fine-tune” paradigm tends to be ineffective in few-shot settings. To alleviate such issues, we resort to prompting, where LLMs are explicitly guided by special prompts to reason about the downstream tasks, e.g., ”translate the sentence into English”.

In this paper, we employ a two-stage framework that divides the SPARQL query generation task into two sub-tasks, following the experience of Gu et al. [10]. An illustration of our model pipeline is shown in Figure 1.

3.0.1. Training Data Preprocessing

Our two-stage learning framework requires the availability of ground truth structural representations of SPARQL queries. To this end, we replace the literal values and schema items within a SPARQL query with placeholders. We design four special tokens as placeholders: [var], [ent], [rel] and [val], representing a variable, an entity, a relation, and a literal value, respectively. To ensure the better generalization of such structural representation as well as higher coverage of complex questions with constraints, we design another special token, namely [con], to represent a constraint or condition in FILTER keywords.

```
SELECT DISTINCT ?answer
WHERE {
   ?answer dblp:authoredBy pid:61_5165 .
   ?answer dblp:yearOfPublication ?y
   FILTER ( ?y > YEAR ( NOW() ) - 9 )
}
```

Listing 1: An example in the DBLP-QALD dataset.

Taking Listing 1 as an example, the SPARQL query is first preprocessed via the following procedure: prefix removal, variable name standardization, lowercase, redundant whitespace removal, etc. and then converted into its corresponding structural representation ”select distinct [var] where { [var] [rel] [ent] . [var] [rel] [var] filter [con] }”. Here, we treat the constraint ” (?y YEAR(NOW())-9)” as a whole and replace it with the placeholder [con].

3.0.2. Structure and Content Prompts:

We divide the task of semantic parsing by two, structure and content prompting.

For a question N, we feed the structure prompt P, N, and the KG ontology G to the LLM, which outputs a SPARQL structure S. In the second stage, this structure is sent to the LLM combined with content prompt Pc, N, and G to generate a complete SPARQL filled with the content. The content is based on known schema items.

In the semantic web community, ontology is a data model that defines and represents the relations and connections between different entities and concepts within a specific domain in a structured and standardized way. In this work, we assume that prompting LLMs with the ontology of a specific knowledge graph would facilitate the understanding and reasoning of

classes: Creator, Publication, Inproceedings;

properties: authoredBy, doi, isbn, publishedBy, title, yearOfPublication, wikidata, orcid, publishersAddress, bibtexType, monthOfPublication, numberOfCreators, editedBy, orderedCreators, thesisAcceptedBySchool, yearOfPublication, publishedIn, publishedInSeries, publishedInSeriesVolume, publishedInJournal, primaryAffiliation publishedInJournalVolumeIssue, publishedInJournalVolume, publishedInBook;

entities: pid:61_5165 (Dan O. Popa).

Figure 2: An example prompt for the question "Which papers did Dan O. Popa publish in the last 9 years?" in DBLP-QALD.

LLMs over structured data in terms of the SPARQL Query generation task. Due to the possibly large size of the ontology of a specific knowledge graph, in this work, the ontology is limited to the schema items (e.g., relations, entities, classes) used within a specific dataset.

An example prompt for verbalizing the ontology of a specific KG (DBLP or ORKG here) is shown in Figure 2. For the task 1 of this challenge, our framework does not perform Entity Linking as it requires the availability of linked entities, which are missing in the final stage. To alleviate such issue, we resort to a simple but effective method where a text similarity model is used to sample top K candidates semantically similar to these questions so that our model learns to select suitable entities for SPARQL query generation. For the task 2 of this challenge, the size of the set of schema items within the SciQA dataset is relatively small, so we extract and use all entities or resources, except for classes and relations, from ground truth SPARQL queries in the training set to construct the prompts.

Therefore, the model is capable of dealing with complex SPARQL queries, which might include multiple entities, relations, and operators. As for the LLM, we use mT5-base [11], which is trained massively on a multilingual corpus.

4. Results and Discussion

Our model achieves state-of-the-art results on SciQA with an F1 score of 99.19% and second in DBLP-QuAD's QA task with an F1 score of 66.19%. In Tables 2 and 3, you can find the leaderboards based on the challenge page. Our model is illustrated under the name of "longquanj" both in the below tables and the challenge website. The best results are marked in bold.

User	Date	F1
zeio	Oct 07 2023	0.9358
tilahun	Oct 06 2023	0.9904
longquanj	Oct 04 2023	**0.9919**
tilahun	Oct 03 2023	0.9602

Table 2
KGQA evaluation result on SciQA dataset

User	Date	F1-EL	F1-QA
Shreyaar12	Oct 07 2023	**0.8353**	0.0000
ruijie.wang	Oct 07 2023	0.7961	**0.8488**
ruijie.wang	Oct 06 2023	0.6235	0.2175
nsteinmetz	Sep 17 2023	0.8320	0.0000
nsteinmetz	Sep 16 2023	0.8283	0.0000
longquanj	Sep 20 2023	0.0000	0.6619
hannaabiakl-dsti	Sep 18 2023	0.7100	0.0018

Table 3
KGQA and EL evaluation result on DBLP-QuAD dataset

As can be seen from the tables, our system not only achieves high performance on different KGQA tasks but it also shows its strong generalization capacity through KGs of different topologies.

However, our hybrid prompt tuning methodology presents several weaknesses, each of which poses unique challenges. First, the unordered nature of triple patterns introduce ambiguity and hinder pattern recognition. This can result in difficulties in capturing the intended semantic relationships. Secondly, the sheer quantity of target KG schema terms, especially when dealing with fine-grained relations and entities, can lead to computational complexity and resource constraints. Additionally, the selection of the top-k relations, linked entities, n-hop neighbors, or subgraphs may inadvertently exclude valuable information, potentially limiting the system's scope. Furthermore, our reliance on an entity linking module may introduce propagation errors, undermining the overall accuracy of the system. Lastly, the expansive structure space, given the granularity at the level of relations and entities, can intensify the challenge of navigating and modeling the KG.

5. Summary

We presented our solution for the KGQA task on the DBLP-QALD and SciQA datasets, respectively for the Scholarly QALD 2023 challenge [5]. We found that integration of the KG ontology into LLM prompts can guide LLMs better in order to generate the corresponding correct SPARQL query for a given natural language question. Additionally, we analyze the results of our model and point out limitations that shall be improved in our further work.

Acknowledgments

This work has been partially supported by grants for the DFG project NFDI4DataScience project (DFG project no. 460234259) and by the Federal Ministry for Economics and Climate Action in the project CoyPu (project number 01MK21007G).

[5] https://kgqa.github.io/scholarly-QALD-challenge/2023/

References

[1] S. Auer, D. A. C. Barone, C. Bartz, E. G. Cortes, M. Y. Jaradeh, O. Karras, M. Koubarakis, D. Mouromtsev, D. Pliukhin, D. Radyush, I. Shilin, M. Stocker, E. Tsalapati, The sciqa scientific question answering benchmark for scholarly knowledge, Scientific Reports 13 (2023) 7240. URL: https://doi.org/10.1038/s41598-023-33607-z. doi:10.1038/s41598-023-33607-z.

[2] D. Banerjee, P. A. Nair, R. Usbeck, C. Biemann, Gett-qa: Graph embedding based t2t transformer for knowledge graph question answering, in: C. Pesquita, E. Jimenez-Ruiz, J. McCusker, D. Faria, M. Dragoni, A. Dimou, R. Troncy, S. Hertling (Eds.), The Semantic Web, Springer Nature Switzerland, Cham, 2023, pp. 279–297.

[3] X. Hu, X. Wu, Y. Shu, Y. Qu, Logical form generation via multi-task learning for complex question answering over knowledge bases, in: Proceedings of the 29th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 1687–1696. URL: https://aclanthology.org/2022.coling-1.145.

[4] B. Lester, R. Al-Rfou, N. Constant, The power of scale for parameter-efficient prompt tuning, arXiv preprint arXiv:2104.08691 (2021).

[5] M. Ley, The dblp computer science bibliography: Evolution, research issues, perspectives, in: International symposium on string processing and information retrieval, Springer, 2002, pp. 1–10.

[6] M. Y. Jaradeh, A. Oelen, M. Prinz, M. Stocker, S. Auer, Open research knowledge graph: a system walkthrough, in: Digital Libraries for Open Knowledge: 23rd International Conference on Theory and Practice of Digital Libraries, TPDL 2019, Oslo, Norway, September 9-12, 2019, Proceedings 23, Springer, 2019, pp. 348–351.

[7] M. Y. Jaradeh, M. Stocker, S. Auer, Question answering on scholarly knowledge graphs, in: International Conference on Theory and Practice of Digital Libraries, Springer, 2020, pp. 19–32.

[8] D. Banerjee, S. Awale, R. Usbeck, C. Biemann, Dblp-quad: A question answering dataset over the dblp scholarly knowledge graph, arXiv preprint arXiv:2303.13351 (2023).

[9] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, The Journal of Machine Learning Research 21 (2020) 5485–5551.

[10] Z. Gu, J. Fan, N. Tang, L. Cao, B. Jia, S. Madden, X. Du, Few-shot text-to-sql translation using structure and content prompt learning, Proc. ACM Manag. Data 1 (2023) 147:1–147:28. URL: https://doi.org/10.1145/3589292. doi:10.1145/3589292.

[11] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mt5: A massively multilingual pre-trained text-to-text transformer, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tür, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021, Association for Computational Linguistics, 2021, pp. 483–498. URL: https://doi.org/10.18653/v1/2021.naacl-main.41. doi:10.18653/v1/2021.naacl-main.41.