Semantic Parsing for Knowledge Graph Question Answering with Large Language Models

Debayan Banerjee[1[0000-0001-7626-8888]]

Language Technology Group, Universität Hamburg, Germany
debayan.banerjee@uni-hamburg.de

Abstract. This thesis explores the topic of Knowledge Graph Question Answering with a special emphasis on semantic parsing approaches, incorporating pre-trained text-to-text language models. We use the text generation ability of these models to convert natural language questions to logical forms. We test whether correct logical forms are being generated, and if not, how to mitigate the failure cases. As a second step, we try to make the same models generate additional information to aid the process of grounding of the logical forms to entities, relations and literals in the Knowledge Graph. In experiments conducted so far, we see encouraging results on both generation of base logical forms, and grounding them to the KG elements. At the same time, we discover failure cases prompting directions in future work. [1]

1 Introduction

A Knowledge Graph (KG) [21,32] is an information store where data is stored in the form of node-edge-node triples. Nodes represent entities and edges represent relationships between these entities. The aim of Knowledge Graph Question Answering (KGQA) [18] is to produce answers from this KG given an input question in natural language, e.g., `Where did Einstein receive his bachelor degree?`. Usually, the first steps in KGQA are to perform Entity and Relation Linking (EL, RL) where mention spans, e.g., `Einstein` representing the name of a person, place, etc., are linked to a KG node and the relationship of the entity to the potential answer in the KG is extracted, e.g., `educated at`.

Some KGQA systems [9,33] attempt to fetch the answer based on the results of the two steps above, which typically ends up being another entity (node) in the graph. However, for more complex questions, such as count queries or min/max aggregate queries (e.g.: `How many institutions did Einstein study in?`) the answer does not lie in a node or edge in the graph, but instead, a formal query or logical form must be generated. The task of generating a formal query is also known as *semantic parsing*, also the focus of this proposal.

Semantic parsing in KGQA is challenging mainly due to two factors: schema-level complexity and fact-level complexity. The schema of a modern KG is diverse. For example, Freebase [6] has over 8K schema items in total (6K relations

[1] The author considers himself a 'middle-stage' Ph.D. candidate.

and 2K types), while a relational database usually comprises dozens of schema items only (i.e., table names and column headers). Hence, learning an alignment between natural language and the schema is much more challenging in KGQA. Moreover, fact-level information (i.e., contents in the KG) plays a vital role in KGQA. Consequently, generating logical forms that can ground to non-empty answers from the KG, i.e., faithful to the KG, requires incorporating fact-level information. In addition, the graph structure of KB facts leads to an enormous search space due to combinatorial explosion, rendering generating faithful queries even more challenging.

A large body of work exists on the topics of entity and relation linking for KGs [30,13]. Similarly, significant literature exists around non-KG semantic parsing (eg: text-to-SQL) [27]. There is prior work [42,40,14] that explores the topic of KGQA semantic parsing using encoder-decoder models, that comprises of LSTMs [17], Transformers [37], and pre-trained language models (PLMs) based on the Transformer architecture [11].

However the role of text-to-text large language models (T2TLMs) like T5 [29] and BART [23] remains under-investigated [15]. *This gap motivates us to pursue further research in this direction.*

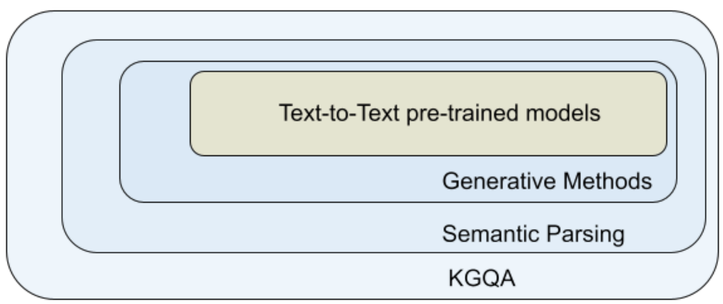

Fig. 1: Situating the topic of this proposal in the KGQA landscape.

2 States of the Art/Related Work

There are three widely followed approaches for semantic parsing approaches for KGQA: *ranking methods, coarse-to-fine methods, and generation methods.* Ranking methods [1,4,39,41] first enumerate candidate queries from the KG and semantic parsing then relies on computing the matching score for each candidate-question pair. Coarse-to-fine methods [5,10,34,42] first generate query skeletons and then ground the skeletons to the KG with admissible schema items. Generation methods [2,19,20,28] which have emerged more recently, typically use

Dataset	Top-1 F1	Family
COMPLEXWEBQ	70.0 [10]	Coarse-to-fine
LC-QUAD	75.0 [41]	Ranking
GRAILQA	74.4 [39]	Ranking
KQA PRO	90.6 [23]	Generation
WEBQSP	76.5 [8]	Generation
GRAPHQ	31.8 [16]	Generation

Table 1: Best-published results on KGQA benchmarks.

a form of Natural Language Generation (NLG) from large language models to produce token-by-token a base logical form. The logical form is often grounded to the KG via constrained decoding, and thus dynamically reduce the search space.

In Table 1, we present the F1-scores and the corresponding family of best-performing models on KGQA benchmarks. On KQA-PRO [7], WEBQSP [3], and GRAPHQ [31], the state-of-the-art models are based on generation, while ranking methods achieve the best results on LC-QUAD [36] and GRAILQA [14] and the best performance on COMPLEXWEBQ [35] is obtained by the coarse-to-fine method. No family dominates all the benchmarks, but generation methods tend to be a trending option due to the easy integration of PLMs.

3 Problem Statement

Our singular research hypothesis is as follows:

T2TLMs can be used effectively to generate logical forms for KGQA.

3.1 Research Questions

RQ 1 : Can T2TLMs generate correct logical form structure?

Given the question:
Is it true that an Olympic-size swimming pool's operating temperature is equal to 22.4?
the correct logical form structure for SPARQL would be:

```
ASK WHERE {
            ENT1 REL1 ?obj
            filter ( ?obj = LIT1 )
}
```

The query above depicts a skeleton of the correct SPARQL query, with placeholders instead of grounded entity, relations or literals. The placeholders are ENT1, REL1 and LIT1.

RQ 2 : Can T2TLMs aid in grounding of logical forms to the KG?

For the same question, a grounded query is as follows:

```
ASK WHERE { wd:Q2084454 wdt:P5066 ?obj
            filter(?obj = 22.4)
}
```

The query above is a valid and grounded SPARQL query on the Wikidata KG. `wd:Q2084454` is the entity ID for `Olympic-size swimming pool` while `wdt:P5066` is the relation ID for `operating temperature`. The placeholders have been replaced with proper entity and relation IDs, and the correct literal has been extracted from the input question and copied into the literal placeholder. Executing it provides a valid and correct response. Can T2TLMs aid in grounding of skeleton logical forms to KGs?

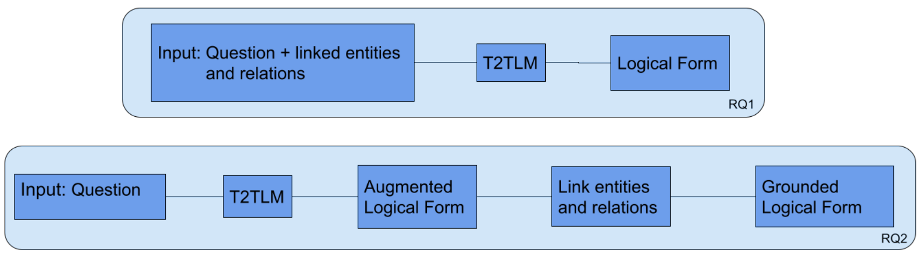

Fig. 2: RQs 1 and 2 depicted visually.

4 Research Methodology and Approach

To test our hypothesis, we require datasets that contain mappings from questions in natural language to corresponding logical forms. Since our focus is on KGQA, these logical forms must be grounded in KGs. Additionally, since our experiments involve large language models, we observed that small datasets are not suitable for our experiments. On an average, we tend to focus on datasets which have more than 5,000 questions. Apart from the datasets mentioned in Table 1, we also test on LC-QuAD 2.0[12]. For models, we primarily test on T5 and BART, which are the two most popular T2TLMs. We plan to fine-tune and prompt-tune [22] the models on the datasets of our choice, and evaluate their performance on tasks related to RQ1 and RQ2.

5 Evaluation/Evaluation Plan

For RQ1, we need to evaluate whether our models produce the correct logical forms. The two popular methods to evaluate this are the exact-match metric (whether the generated query matches the gold query exactly) and BLEU score [24]. There are several different variants of BLEU in use, however the general method is to measures how many n-grams in the reference sentence are reproduced by the candidate sentence. This value is reduced to a precision in the range of 0-1. BLEU is generally used to evaluate machine translations, however for the task of semantic parsing to logical forms, it has the obvious shortcoming that since it does not take into consideration the word ordering on the output, a logical form which is syntactically incorrect may still produce a high BLEU score. The best metric for use of evaluation of semantic parsing to logical forms remains an open question.

For RQ2, since we expect our models to produce grounded queries instead of mere logical forms, we can additionally evaluate using execution-based metrics such as F1-score.

Our approach to evaluate will be to first find existing semantic parsing and KGQA systems which have a working demo or code available. We will then evaluate our approaches using T2TLMs and compare against existing systems. If too few such systems are available, we shall resort to results as reported.

6 Results

We performed an initial set of experiments using T5 and BART on the task of KGQA semantic parsing, and paid special attention to a scenario which requires the copying of input tokens to the output logical form. Assuming that the entities and relations have been linked before-hand, one may modify the original input question to:

Is it true that an Olympic-size swimming pool's operating temperature is equal to 22.4? `[SEP] wd:Q2084454 Olympic-size swimming pool [SEP] wdt:P5066 operating temperature`

here, we append the input question with linked entities and relations and also their corresponding labels. The task for the model is to produce the correct logical form. This addresses **RQ1** and we found that T5 outperforms existing approaches on LC-QuAD and LC-QuAD 2.0 datasets, while BART lags behind [2].

In additional experiments for **RQ1**, we found that both for prompt-tuned [22] and fine-tuned T2TLMs, choosing an alternate output vocabulary for the logical-form improves semantic parsing performance. The usual logical-form vocabulary is distinct from human vocabulary. T2TLMs are pre-trained for human language tasks, and hence the logical-form vocabulary may appear alien to it. This work is currently under review.

For **RQ2**, we used T5 to generate not just the logical forms, but also entity and relation labels. As input we provided only the question. Additionally, we trained T5 jointly to produce truncated graph embeddings for each entity. In effect, we generate *augmented logical forms* (Figure 2). We use the generated augmented logical forms to ground the queries to a KG. We found that such a setup produces strong performance on LC-QuAD 2.0 and SimpleQuestions. This work is currently under review.

During our experiments we realised that T2TLMs have some common shortcomings for semantic parsing tasks. They are unable to perform compositional generalisation, handles special tokens, and have a limited ability to generate embeddings. Moreover, existing KGQA datasets are based on the three popular KGs, namely Freebase, DBpedia [21] and Wikidata [38]. T2TLMs are pre-trained on large corpus which is of similar nature as the sources from which these KGs are derived, and hence, a new dataset which belongs to different domain was required, which would remove the natural advantage for T2TLMs of having seen much of the information at a pre-training stage. As a result, we produced a new dataset consisting of 10,000 question-query pairs on a scholarly KG, which provides T2TLMs with specific challenges for KGQA and semantic parsing. This addresses **RQ1** and **RQ2**, and is currently under review.

7 Conclusions/Lessons Learned

In our initial experiments we found that T2TLMs are a potent tool for the task of KGQA semantic parsing. The natural advantage of using them is that they are easily available, well maintained and require minimal initial setup before being put to use. Through this thesis, our aim is to provide strategies to researchers on how to mitigate some of the pitfalls of using T2TLMs on the semantic parsing task, and push accuracy further when starting from a vanilla T2TLM.

At the PhD symposium, I look forward to receiving feedback from mentors on the overall structure of the thesis. In paper submissions so far, I have had trouble convincing reviewers about the quality of my evaluations. There is an evaluation crisis [26] in our field, because several older versions of KGs are not publicly available any more, neither are working versions of KGQA systems, making replication next-to-impossible. Another aspect I am concerned about is whether the current thesis contains enough substance for publication. It is, for example, possible to compare semantic parsing approaches to the SQL parsing community, or work on more popular problems like compositional generalisation [25].

8 Acknowledgements

I would like to thank my co-supervisors Chris Biemann and Ricardo Usbeck for their valuable guidance and feedback in the drafting of this document.

References

1. Abujabal, A., Yahya, M., Riedewald, M., Weikum, G.: Automated Template Generation for Question Answering over Knowledge Graphs. p. 1191–1200. WWW '17, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE (2017), https://doi.org/10.1145/3038912.3052583

2. Banerjee, D., Nair, P.A., Kaur, J.N., Usbeck, R., Biemann, C.: Modern Baselines for SPARQL Semantic Parsing. p. 2260–2265. SIGIR '22, Association for Computing Machinery, New York, NY, USA (2022), https://doi.org/10.1145/3477495.3531841
3. Berant, J., Chou, A., Frostig, R., Liang, P.: Semantic parsing on Freebase from question-answer pairs. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. pp. 1533–1544. Association for Computational Linguistics, Seattle, Washington, USA (2013), https://aclanthology.org/D13-1160
4. Berant, J., Liang, P.: Semantic Parsing via Paraphrasing. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1415–1425. Association for Computational Linguistics, Baltimore, Maryland (2014), https://aclanthology.org/P14-1133
5. Bhutani, N., Zheng, X., Jagadish, H.V.: Learning to Answer Complex Questions over Knowledge Bases with Query Composition. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management. p. 739–748. CIKM '19, Association for Computing Machinery, New York, NY, USA (2019), https://doi.org/10.1145/3357384.3358033
6. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In: Proceedings of the 2008 ACM SIGMOD international conference on Management of data. AcM (2008)
7. Cao, S., Shi, J., Pan, L., Nie, L., Xiang, Y., Hou, L., Li, J., He, B., Zhang, H.: KQA Pro: A Dataset with Explicit Compositional Programs for Complex Question Answering over Knowledge Base. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 6101–6119. Association for Computational Linguistics, Dublin, Ireland (2022), https://aclanthology.org/2022.acl-long.422
8. Cao, S., Shi, J., Yao, Z., Lv, X., Yu, J., Hou, L., Li, J., Liu, Z., Xiao, J.: Program transfer for answering complex questions over knowledge bases. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 8128–8140. Association for Computational Linguistics, Dublin, Ireland (2022), https://aclanthology.org/2022.acl-long.559
9. Christmann, P., Saha Roy, R., Weikum, G.: Beyond NED: Fast and Effective Search Space Reduction for Complex Question Answering over Knowledge Bases. In: Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining. p. 172–180. WSDM '22, Association for Computing Machinery, New York, NY, USA (2022), https://doi.org/10.1145/3488560.3498488
10. Das, R., Zaheer, M., Thai, D., Godbole, A., Perez, E., Lee, J.Y., Tan, L., Polymenakos, L., McCallum, A.: Case-based Reasoning for Natural Language Queries over Knowledge Bases. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 9594–9611. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (2021), https://aclanthology.org/2021.emnlp-main.755
11. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019), https://aclanthology.org/N19-1423

12. Dubey, M., Banerjee, D., Abdelkawi, A., Lehmann, J.: LC-QuAD 2.0: A Large Dataset for Complex Question Answering over Wikidata and DBpedia. In: The Semantic Web – ISWC 2019. Springer International Publishing (2019)
13. Dubey, M., Banerjee, D., Chaudhuri, D., Lehmann, J.: EARL: Joint Entity and Relation Linking for Question Answering over Knowledge Graphs. In: The Semantic Web – ISWC 2018. pp. 108–126. Springer International Publishing, Cham (2018)
14. Gu, Y., Kase, S., Vanni, M., Sadler, B., Liang, P., Yan, X., Su, Y.: Beyond I.I.D.: Three Levels of Generalization for Question Answering on Knowledge Bases. p. 3477–3488. WWW '21, Association for Computing Machinery, New York, NY, USA (2021), https://doi.org/10.1145/3442381.3449992
15. Gu, Y., Pahuja, V., Cheng, G., Su, Y.: Knowledge base question answering: A semantic parsing perspective, https://arxiv.org/abs/2209.04994
16. Gu, Y., Su, Y.: ArcaneQA: Dynamic Program Induction and Contextualized Encoding for Knowledge Base Question Answering. In: Proceedings of the 29th International Conference on Computational Linguistics. pp. 1718–1731. International Committee on Computational Linguistics, Gyeongju, Republic of Korea (2022), https://aclanthology.org/2022.coling-1.148
17. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
18. Lan, Y., He, G., Jiang, J., Jiang, J., Zhao, W.X., Wen, J.R.: A Survey on Complex Knowledge Base Question Answering: Methods, Challenges and Solutions. In: Zhou, Z.H. (ed.) Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21. pp. 4483–4491. International Joint Conferences on Artificial Intelligence Organization (2021), https://doi.org/10.24963/ijcai.2021/611, survey Track
19. Lan, Y., Jiang, J.: Query Graph Generation for Answering Multi-hop Complex Questions from Knowledge Bases. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 969–974. Association for Computational Linguistics, Online (2020), https://aclanthology.org/2020.acl-main.91
20. Lan, Y., Wang, S., Jiang, J.: Multi-hop Knowledge Base Question Answering with an Iterative Sequence Matching Model. In: 2019 IEEE International Conference on Data Mining (ICDM). pp. 359–368 (2019). https://doi.org/10.1109/ICDM.2019.00046
21. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer, C.: DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. Semantic Web Journal **6**(2), 167–195 (2015)
22. Lester, B., Al-Rfou, R., Constant, N.: The Power of Scale for Parameter-Efficient Prompt Tuning. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 3045–3059. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (2021), https://aclanthology.org/2021.emnlp-main.243
23. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7871–7880. Association for Computational Linguistics, Online (2020), https://aclanthology.org/2020.acl-main.703

24. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: A Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA (2002), https://aclanthology.org/P02-1040
25. Patel, A., Bhattamishra, S., Blunsom, P., Goyal, N.: Revisiting the Compositional Generalization Abilities of Neural Sequence Models. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 424–434. Association for Computational Linguistics, Dublin, Ireland (2022), https://aclanthology.org/2022.acl-short.46
26. Perevalov, A., Yan, X., Kovriguina, L., Jiang, L., Both, A., Usbeck, R.: Knowledge Graph Question Answering Leaderboard: A Community Resource to Prevent a Replication Crisis. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference. pp. 2998–3007. European Language Resources Association, Marseille, France (2022), https://aclanthology.org/2022.lrec-1.321
27. Qin, B., Hui, B., Wang, L., Yang, M., Li, J., Li, B., Geng, R., Cao, R., Sun, J., Si, L., Huang, F., Li, Y.: A Survey on Text-to-SQL Parsing: Concepts, Methods, and Future Directions, https://arxiv.org/abs/2208.13629
28. Qiu, Y., Zhang, K., Wang, Y., Jin, X., Bai, L., Guan, S., Cheng, X.: Hierarchical Query Graph Generation for Complex Question Answering over Knowledge Graph. In: Proceedings of the 29th ACM International Conference on Information and Knowledge Management. p. 1285–1294. CIKM '20, Association for Computing Machinery, New York, NY, USA (2020), https://doi.org/10.1145/3340531.3411888
29. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Journal of Machine Learning Research **21**(140), 1–67 (2020), http://jmlr.org/papers/v21/20-074.html
30. Sevgili, Ö., Shelmanov, A., Arkhipov, M., Panchenko, A., Biemann, C.: Neural Entity Linking: A survey of models based on deep learning. Semantic Web **13**, 527–570 (2022), https://doi.org/10.3233/SW-222986, 3
31. Su, Y., Sun, H., Sadler, B., Srivatsa, M., Gür, I., Yan, Z., Yan, X.: On Generating Characteristic-rich Question Sets for QA Evaluation. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 562–572. Association for Computational Linguistics, Austin, Texas (2016), https://aclanthology.org/D16-1054
32. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: A Core of Semantic Knowledge. p. 697–706. WWW '07, Association for Computing Machinery, New York, NY, USA (2007), https://doi.org/10.1145/1242572.1242667
33. Sun, H., Dhingra, B., Zaheer, M., Mazaitis, K., Salakhutdinov, R., Cohen, W.: Open Domain Question Answering Using Early Fusion of Knowledge Bases and Text. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 4231–4242. Association for Computational Linguistics, Brussels, Belgium (2018), https://aclanthology.org/D18-1455
34. Sun, Y., Zhang, L., Cheng, G., Qu, Y.: SPARQA: Skeleton-based Semantic Parsing for Complex Questions over Knowledge Bases (2020), https://arxiv.org/abs/2003.13956
35. Talmor, A., Berant, J.: The Web as a Knowledge-Base for Answering Complex Questions. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,

Volume 1 (Long Papers). pp. 641–651. Association for Computational Linguistics, New Orleans, Louisiana (2018), https://aclanthology.org/N18-1059
36. Trivedi, P., Maheshwari, G., Dubey, M., Lehmann, J.: LC-QuAD: A Corpus for Complex Question Answering over Knowledge Graphs. In: The Semantic Web – ISWC 2017. Springer International Publishing (2017)
37. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is All you Need. In: Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
38. Vrandečić, D., Krötzsch, M.: Wikidata: A Free Collaborative Knowledgebase. Commun. ACM **57**(10), 78–85 (2014), https://doi.org/10.1145/2629489
39. Ye, X., Yavuz, S., Hashimoto, K., Zhou, Y., Xiong, C.: RNG-KBQA: Generation Augmented Iterative Ranking for Knowledge Base Question Answering. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 6032–6043. Association for Computational Linguistics, Dublin, Ireland (2022), https://aclanthology.org/2022.acl-long.417
40. Yin, X., Gromann, D., Rudolph, S.: Neural Machine Translating from Natural Language to SPARQL. Future Generation Computer Systems **117**, 510–519 (2021). https://doi.org/10.1016/j.future.2020.12.013
41. Zafar, H., Napolitano, G., Lehmann, J.: Formal Query Generation for Question Answering over Knowledge Bases. In: The Semantic Web. pp. 714–728. Springer International Publishing, Cham (2018)
42. Zhang, H., Cai, J., Xu, J., Wang, J.: Complex Question Decomposition for Semantic Parsing. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 4477–4486. Association for Computational Linguistics, Florence, Italy (2019), https://aclanthology.org/P19-1440