Question Answering over Knowledge Graphs

Sareh Aghaei [0000-1111-2222-3333]

Semantic Technology Institute Innsbruck, Department of Computer Science, University of Innsbruck, Austria
sareh.aghaei@sti2.at

Abstract. With the increasing maturity of large-scale knowledge graphs, question answering over knowledge graphs has become a crucial topic and attracted massive attention. A knowledge graph-based question answering system targets to leverage facts in knowledge graphs to answer natural language questions and assist users to access the meaningful and pertinent knowledge, without knowing data structures. This research intends to propose an approach to answer natural language questions over knowledge graphs in three main steps including identification of optimal subgraphs, creation of candidates and answer selection. The proposed approach leverages the state-of-the-art techniques including graph alignment, neural networks and natural language process to generate more accurate answers for questions, either simple questions or multi-hop questions. Experiments are to be conducted over different knowledge graphs to demonstrate the effectiveness of the approach, which can outperform novel existing approaches.

Keywords: Question Answering, Knowledge Graphs, Graph Alignment, Neural Networks.

1 Introduction

With the rapid progress of the data web, a large amount of structured data has become available on the web in the form of knowledge graphs (KGs). A knowledge graph is huge semantic net which integrates various, inconsistent and heterogeneous information resources to represent knowledge about different domains [1]. Basically, a KG is a directed graph where its nodes are entities with different types and attributes and its edges are relations of entities. In KGs, each directed edge, along with its head entity and tail entity, is considered as a triple which is also named a fact. Numerous real-world KGs such as DBPedia [2], Freebase [3] contain millions or billions of facts. The increasing volume and complexity of the data structures make it difficult for end users to access the substantial and profitable knowledge in the KGs. In order to bridge the gap, question answering over KGs has been proposed and attracted massive attention [4].

Making the facts of KGs accessible and beneficial for end users is one of the primary goals of question answering (QA) over linked data [4]. QA over knowledge graph provides a way for artificial intelligence systems to incorporate knowledge graphs as a key ingredient to answer human questions, which can benefit a variety of applications, such as search engine design, automatic customer service, smart home devices, chatbots and search engine optimization (SEO) [5]. Research projects such as WordLiftNG [6] and KI-NET [7] show that QA over KGs is essential for construction of the most SEO-friendly websites and chatbots in industrial areas, respectively, and knowledge graphs and their supporting systems are already being practically deployed, for example, in the domain of tourism marketing on the Web [8].

Considering the number of KG triples required to obtain answers, natural questions can be summarized into two groups: simple questions and complex questions. A simple question which is named single-hop question, requires only one triple to fetch the answer whereas a complex question which is called multi-hop question needs two or more triples [9, 10]. The research's key motivation is proposing an approach to improve answering multi hop questions over knowledge graphs.

2 State of the art

QA over KGs has attracted wide attention from researchers and the research progress can be generally categorized into three groups: rule-based techniques, information retrieval-based techniques and semantic parsing techniques [10].

2.1 Rule-based techniques

Most of the early research in this field leverage on predefined rules or templates to parse questions and provide logical forms. Defining templates leads to limited scalability and necessity for researchers to be familiar with the linguistic knowledge [10]. Although, there are some techniques which try to automatically or semi-automatically generate templates [10]. In [11], a question answering system has been proposed to automatically learn utterance-query templates with alignments between the constituents of the question utterance and the KG query through integer linear programming. These templates are generated in an offline step by distant supervision [12] at training step. Then, in an online step, the templates are used to answer structurally questions of users.

2.2 Information retrieval-based techniques

For each given natural language question, these techniques extract the entities of interest and determine the links between the extracted entities and the KG. Then, topic-entity-centric sub-graphs are extracted and the nodes of the sub-graphs are assumed as candidate answers. Based on the features extracted from the questions and candidate answers, the matching scores between the encoded answers and questions are calculated and the final answer is selected. Although, these techniques overcome manually defined templates and rules, but suffer from model interpretability and lack of training data [10].

H. Sun et al. [13] has proposed an open domain QA, namely PullNet which uses an iterative process to construct a question-specific subgraph that contains nodes relevant to the question. The initial subgraph is constructed only based on the question and is expanded iteratively. In each iteration, a graph convolutional neural network (CNN) is used to detect nodes that should be expanded on the KG. Then, another graph CNN is employed to detect the answer from the constructed subgraph.

Stepwise reasoning network (SRN) [9] is a neural method based on reinforcement learning which considers multi-hop QA as a sequential decision problem. SRN leverages path search and beam search in order to fetch answer and reduce the number of candidates. To enhance the unique impact of different parts of a question, the attention mechanism and neural networks are used to determine the parts that need more focus. Moreover, a potential-based reward shaping strategy is applied to address the delayed and sparse reward problem.

2.3 Semantic parsing techniques

These methods usually convert natural language questions into executable queries or intermediate query forms such as query graphs based on neural semantic parsing with high scalability and capability.

Zhu et. al. [14] proposed a knowledge-based QA by tree-to-sequence learning. The basic idea behind the system can be summarized in five steps. In the first step, it determines the information (entities, types and numbers) in the KG which is referred by the question, so it leads to constraining the construction of queries. In the next step, the constraints are used to construct candidate queries for the given question. Each candidate query is encoded into a set of hidden states which are decoded into the given question in the fourth step. Finally, using the decoding probabilities, the best query is chosen. In order to capture contexts of an entity or a relation in a query during the encoding phase, a tree-based bi-directional LSTM is used. A tree-based LSTM runs from all leaves to the root while the other one runs reversely. During decoding, a generating mode and a referring mode are mixed to capture different levels of correlations between queries and questions.

The key idea behind [15] is to leverage graphs to represent questions. This paper conceptualizes semantic parsing as a graph matching problem. Questions are parsed using combinatory categorial grammar and then ungrounded semantic graphs are created. Next, the created ungrounded semantic graphs are mapped to the KG subgraphs through mapping edge labels to KG relations, type nodes to KG entity types, and entity nodes to KG entities. The most pertinent semantic graph is selected among mapped candidates and finally is converted to an executable query to fetch the answer.

3 Problem statement and contribution

This research delves into examining the problem of question answering over knowledge graphs and the following research challenges need to be dealt with:
Challenge 1: complex semantic information in multi-hop questions.
Challenge 2: sparsity and incompleteness in KG.

Challenge 3: high time complexity to detect answers in large-scaled KGs.

Basically, complex semantic information leads to poor performance in analyzing of multi-hop questions. Moreover, KGs are often incomplete and sparse with the sparsity resulting in low recall for multi-hop questions. Extracting an optimal question subgraph which contains the answer and is small enough, results in reducing time complexity.

Different from existing methods, the current research intends to propose a solution to improve the performance of multi-hop QA over KGs. To propose the solution, the following questions should be addressed.

Question 1: which methods can be used to build the optimal subgraph?

Question 2: how to consider the semantic information in multi-hop questions?

Question 3: which techniques should be applied to encode questions and entities of KGs?

Question 4: how to leverage graphs to represent questions and graph alignment to map question graphs to KGs?

Question 5: how to select the answer(s) among the answer candidates?

Therefore, the underlying hypothesis of this research is as follows:

Hypothesis: neural networks, graph alignment, graph embedding and natural language processing can improve the performance of question answering over knowledge graphs in terms of time complexity, recall and precision especially in multi-hop questions.

The main contributions can be summarized into (i) reducing search space to find answers in KGs through building optimal subgraphs, (ii) using graph embedding techniques in order to address incompleteness and sparsity in KGs and (iii) employing graph alignment and neural networks to answer multi-hop questions.

4 Methodology

The current study intends to present a new guided approach to QA over KGs that improves answering natural language questions posed over the KG, either simple questions or multi-hop questions. This approach includes three main steps: identification of optimal-subgraphs, creation of candidates, answer selection.

4.1 Identification of optimal subgraphs

In this step, the entity of interest in the natural language question, namely topic entity, is recognized and then linked to the KG. The linked part of the KG as the optimal subgraph is more likely to contain the answer. Based on the category which the proposed approach belongs to, a variety of techniques can be employed in this step including: part-of-speech tagging, entity recognition, graph embedding, heuristic algorithms and neural networks.

4.2 Creation of candidates

The candidate answers are generated based on the identified optimal subgraph. To achieve this goal, graph alignment, beam search, neural networks, computing semantic similarity and other technologies should be examined.

4.3 Answer selection

According to the generated answers, the more accurate response is selected. Some techniques such as neural networks, reinforcement learning, comparing embeddings of answers to the natural question can be leveraged to select the more pertinent answer.

5 Evaluation Plan

In order to evaluate how the proposed approach outperforms the available state-of-the-art approaches, the benchmark datasets including WebQuestionsSP [21] and MetaQA [22] would be applied. The purpose of this evaluation is to substantiate the claim to be able to improve answering multi-hop questions over knowledge graphs. The evaluation would be carried out using metrics such as recall, precision and F1-score. For the practical evaluation, the proposed approach can be applied in the WordLiftNG project and KI-Net project.

6 Conclusion

QA over KGs has emerged as a significant research area over the last few years. KG QA aims to automatically answer natural language questions via well-structured relation information between entities stored in a KG. This study leverages various techniques in various fields including KGs, graph alignment, natural language processing and neural networks to provide a new guided approach which consists of three main steps: identification of optimal subgraphs, creation of candidates and answer selection. Experiments are to be performed to show that the proposed approach is competitive compared to the state-of-the-art.

Acknowledgements

This research has been supported by the project WordLiftNG within the Eureka, Eurostars Programme (grant agreement number 877857 with the Austrian Research Promotion Agency (FFG)) and the project KI-NET within the Interreg Österreich-Bayern 2014-2020 programme (grant agreement number AB 292). I would like to express my gratitude to Assoc.-Prof. Dr. Anna Fensel for her support and insightful comments.

References

1. Stroh, E., Mathur, P.: Question Answering Using Deep Learning, https://cs224d.stanford.edu/reports/StrohMathur.pdf, last accessed 1.6.21 (2016).
2. Lehmann, J., Isele, R., M. Jakob, Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., Kleef, P.V., Auer, S. and Bizer, C.: DBpedia–A Large-Scale, Multilingual Knowledge Base Extracted From Wikipedia, Semantic Web, 6, pp. 167–195 (2015).

3. Bollacker, K., Evans, C., Paritosh, P., Sturgem T., Taylor, J.: Freebase: a collaboratively created graph database for structuring human knowledge, In: the 2008 ACMSIGMOD International Conference on Management of Data, pp. 1247–1250 (2008).
4. Ait-Mlouk, A., Jiang, L.: KBot: A Knowledge Graph Based ChatBot for Natural Language Understanding Over Linked Data, in IEEE Access, vol. 8, pp. 149220-149230 (2020).
5. Huang, X., Zhang, J., Li, D., Li, P.: Knowledge Graph Embedding Based Question Answering, In: the Twelfth ACM International Conference on Web Search and Data Mining, pp. 105–113 (2019).
6. WordLift New Generation, https://wordlift.io/ng/, last accessed 1.6.21 (2020).
7. KI-NET: AI-based optimization in production, https://scch.at/en/das-projects-details/ki-net, /, last accessed 1.6.21.
8. Fensel, A., Akbar, Z., Kärle, E., Blank, C., Pixner, P., Gruber, A.: Knowledge Graphs for Online Marketing and Sales of Touristic Services, Information, 11(5), 253 (2020).
9. Qiu, Y., Wang, Y., Jin, X., Zhang, K.: Stepwise Reasoning for Multi-Relation Question Answering over Knowledge Graph with Weak Supervision, In: the 13th International Conference on Web Search and Data Mining, pp. 474-482 (2020).
10. Fu, B., Qiu, Y., Tang, C., Li, Y., Yu, H., Sun, J.: A Survey on Complex Question Answering over Knowledge Base: Recent Advances and Challenges, ArXiv, vol. abs/2007.13069 (2020).
11. Abujabal, A., Yahya, M., Riedewald, M., Weikum, G.: Automated Template Generation for Question Answering over Knowledge Graphs, In: the 26th International Conference on World Wide Web, pp. 1191–1200 (2017).
12. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data, In: the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing, pp. 1003-1011 (2009).
13. Sun, H., Bedrax-Weiss, T., Cohen, W.: PullNet: Open do-main question answering with iterative retrieval on knowledgebases and text, In: the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 2380–2390 (2019).
14. Zhu, S., Cheng, X., Su, S.: Knowledge-based question answering by tree-to-sequence learning, Neurocomputing, vol. 372, pp. 64-72 (2020).
15. Reddy, S., Lapata, M., Steedman, M.: Large-scale Semantic Parsing without Question-Answer Pairs, Transactions of the Association for Computational Linguistics, pp. 377-392 (2014).
16. Usbeck, R., Ngomo, A-CN., Haarmann, B., Krithara, A., Röderm M., Napolitano, G.: 7th open challenge on question answering over linked data (qald-7), Semantic Web Challenges, pp. 59–69 (2017).
17. Trivedi, P., Maheshwari, G., Dubey, M., Lehmann, J.: Lc-quad: A corpus for complex question answering over knowledge graphs, The Semantic Web-ISWC, pp. 210–218 (2017).
18. Zhang, Y., Dai, H., Kozareva, Z., Smola, A. J., Song, L.: Variational reasoning for question answering with knowledge graph, Thirty-Second AAAI Conference on Artificial Intelligence (2018).
19. Yih, W., Chang, M.W., He, X., Gao, J.: Semantic Parsing via Staged Query Graph Generation: Question Answering with Knowledge Base, ACL (2015).
20. Talmor, A., Berant, J.: The Web as a Knowledge-Base for Answering Complex Questions, In NAACL-HLT, pp. 641–651 (2018).
21. Yih, W., Richardson, M., Meek, C., Chang, M., Suh, J.: The value of semantic parse labeling for knowledge base question answering, In: the 54th Annual Meeting of the Association for Computational Linguistics, vol. 2, pp 201–206 (2016).

22. Zhang, Y., Dai, H., Kozareva, Z., Smola, A., Song, L.: Variational reasoning for question answering with knowledge graph, In: Association for the advancement of artificial Intelligence AAAI (2018).