A short introduction of Multi-Hop KBQA and the Datasets

Mingchen Li
Department of Computer Science
Georgia State University, Atlanta, GA, USA
lmingchen96@gmail.com

Abstract. Knowledge graph (KG) is known as a useful tool for question answering (QA), since it provides a well-structured graph which is used to obtain the exact information and infer unknown targets. More recently, in order to make KGQA more practicality in real scenarios, many works begin to explore how to progress complex question, that is multi-hop question answering based knowledge graph (multi-KBQA), which requires model and algorithm have inference ability. In this survey, we present the overview of this task, and show the dataset which is used.

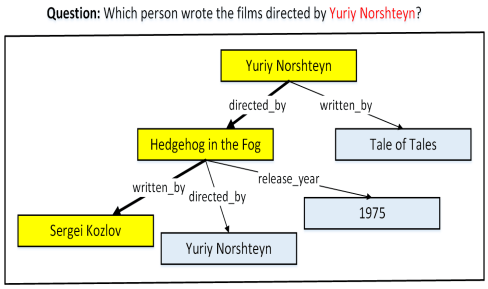

Fig. 1. Inference path for a question, the topic entity is Yuriy Norshteyn, the answer is Sergei Kozlov.

1 Introduction

Knowledge graph is a multi-relational semantic network consisting of many entities (nodes), and relations (edges) between them. It plays a critical role in question answering. The existing knowledge graph includes YAGO [1], DBpedia [2], Wordnet [3], Freebase [4], etc. People can use this structure network get more useful information from natural language question, this progress is called KGQA which already is successfully applied in amounts of voice assistant and search engineer, such as Siri, Microsoft Xiaoice, Amazon Alexa.

The current work on KBQA mainly focused on simple questions which can be answered from a triple (head, relation, tail), the relation links two entities in the knowledge graph (KG). For example, the question "who wrote the screenplay for True Romance?" can be answered by triple (True Romance, written_by, Quentin Tarantino) in KB. However, questions in real scenario can be more complex and require complex relations (multi-hop relations) to answer the questions. Compared with single KBQA, multi-hop KBQA system should have powerful reasoning ability to get the right answer along the edges of KG. One basic reasoning path is shown in Figure 1. In the first, we should find the topic entity (Yuriy Norshteyn) in the question. Second, to choose the relation (directed_by) which most relevent to the question at hop1, using the same method to choose the relation in the next hop to find the right answer in the last. This progress is very intuitive for human, but it's difficult for machine. In this survey, we summary six main challenges for task and explore the resolvents.

Incomplete in knowledge graph. The Knowledge Graphs are often incomplete, this will lead to a poor performance in multi-KBQA. For a question, we assume we know the answer, but this answer is not in knowledge graph, so the model will return False. KG is a semantic network which is built by human. It's a very normal thing which has low coverage.

Interpretable reasoning path. Interpretable inference path is very important. For excample, we expert our math teacher teach give us the solution process, not just an answer.

How to infer, more power inference. We should choose method or algorithm to get the answer in a fast and efficient way.

How to stop, what's the maximum of hops. It is unrealistic to know the what is the maximum number of hops in advance for real world applications.

How to obtain the topic entity, model should progress more than more topic entity. It's the first step in reasoning path. The topic entity and entity mentions are not always same in literal, and this task is related to some downstream tasks, such as entity linking (EL), word sense disambiguation (WSD), name entity recognition (NER), etc. It's also a challeging work.

Dataset A dataset should have multi knowledge types, less noises, plenty of training set and from real scenario.

Table 1. The comparison of different datasets. K types refer to the knowledge type, Numbers refer to the number of questions (all hops, train,dev and test), NL refers to whether the question is natural language, TN refers to the number of topic entity, EM refers to whether topic entity is marked, RP refers to reasoning progress, "–" refers to this data is not open.

Dataset	KG	Numbers	NL	SPARQL	TN	EM	RP
MetaQA	WikiMovies	407,513	No	No	1	Yes	No
WebQSP	FreeBase	5,808	Yes	Yes	1/2	Yes	Yes
PathQuestion	FreeBase	9,731	No	No	1	Yes	Yes
WC2014/WorldCup2014	FIFAKG	18,116	No	No	1	Yes	Yes
QALD-4	DBpedia	250	Yes	Yes	1	No	No
ComplexQuestions (CQ)	FreeBase	2,100	Yes	No	1	No	No
ComplexWebQuestons (CWQ)	FreeBase	34,689	Yes	Yes	1	No	No
LC-QuAD2.0	Wikidata&DBpedia	30,000	Yes	Yes	1/more	Yes	No
CSQA	Wikidata	196,601	No	No	1/more	Yes	No
KQA Pro	Wikidata	117,970	Yes	Yes	–	–	Yes

In this survey, we we mainly give a brief introdution about the relevent dataset.

2 Dataset

The approaches for multi-KBQA have been evaluated on a variety of benchmarks datasets, in this section, we will give a brief introdution about these dataset.

2.1 MetaQA

MetaQA[1] is proposed by [5], it is a large QA dataset which based on the movie knowledge graph MovieQA. This data contains three main components, Vanilla text data, NTM text data and Audio data. Vanilla text data always is applied in this task. In vanilla, MetaQA provides 1-hop, 2-hop, and 3-hop questions. Each sentence just have one topic entity which is annotated by square brackets. MetaQA also provide a KG which has 134741 triples, 29 relations and 43233 entities. The follow is a 2-hops excample: *which person wrote the films directed by [Yuriy Norshteyn]*, the topic entity is *Yuriy Norshteyn*, the answer is *Sergei Kozlov*.

2.2 WebQuestionsSP (WebQSP)

WebQSP [6] is a small multi-KBQA dataset based on Freebase (338580000 triples) with 4737 questions (1-hop and 2-hop). In WebQSP, there are 413 questions have more than one topic entities. Moreover, this data provide the inference chain with relations. In this dataset, entity mention and topic entity all are labeled. Same as Freebase, topic entity and answer is represented by ID which can solve the polysemy, but at the same time, it also increase the difficulty for entity links. You can download this data from Here [2]

2.3 PathQuestion

This dataset is proposed by [7] in 2018, they adopted two subsets of Freebase as knowledge graph to construct PathQuestion (PQ) and PathQuestion-Large (PQL). These datasets both contains two hops and three hops. PQ and PQL are generated by templates and synonyms for relations. The number of this dataset is less 9000. You can download it from Here [3]

2.4 WordCup2014

WC2014 is constructed by [8]. This dataset is constructed by their own knowledge graph (this KG is about football players which participated in FIFA World Cup 2014, we call this dataset FIFAKG-6482 trples). In WC2014, there are two types training set, path queries and conjunctive queries. You can learn it in detail from the appendix Table 4 from [8]. WC2014 is also constructed by template and have one topic entity. You can download it from Here [4].

2.5 QALD-4

QALD-4 is proposed by [9] in 2014, it is built upon DBpedia. The QALD-4 consists three QA dataset, they are Multilingual question answering (MQA), Biomedical question answering and Hybrid question answering. MQA is usually used in multi-KBQA. There are 250 English questions (200 for training, 50 for testing) in it, each question has its SPQRQL structure.

[1] https://github.com/yuyuz/MetaQA

[2] https://www.microsoft.com/en-us/research/publication/the-value-of-semantic-parse-labeling-for-knowledge-base-question-answering-2/

[3] https://github.com/zmtkeke/IRN

[4] https://github.com/zmtkeke/IRN

2.6 ComplexQuestions (CQ)

ComplexQuestions (CQ) is released by [10], this dataset consists of 2100 QA pairs which from three source, WebQuestions, [11] and manually annotated. You can download it from Here [5].

2.7 ComplexWebQuestons (CWQ)

ComplexWebQuestions [12] is a multi-KBQA dataset that contains a large set of complex questions. CWQ has 34,689 QA pairs, each question has a corresponding SPARQL query in the Freebase. Otherwise, this dataset has human-written questions, so its language structure is diverse and fluent. You can download it from Here [6].

2.8 LC-QuAD2.0

LC-QuAD 2.0 [13] is a large Complex QA dataset with 30,000 pairs of question and its corresponding SPARQL query. It has two target knowledge graph, Wikidata and DBpedia. You can download it from Here [7].

2.9 CSQA

CSQA is proposed by [14]. The templates in CSQA are generated from the dialogue between the annotators, and then these templaptes are used to generate more QA pair. In this progress, authors pursue the question can be answered from KG by using logical forms, comparative reasoning. There are 196,601 pairs in CSQA, the average length of user's question is 10 (in words). An excample for multi topic entity : *Does Q95639 have Q2728849 and Q1816795 as a child ?* You can download it from Here [8].

2.10 KQA Pro

KQA Pro [15] is a new complex question answering dataset which show the explicit reasoning progress and provide SPARQL, functional programs. For solve the problem of insufficient data and poor language structure, they use a combination of manual and template. Finally, they get 117,970 instances, 94376 for training, 11797 for validation, 11797 for testing. Unfortunately, this data is not open.

2.11 Summary

Note [9]. We summary 10 multi-KBQA datasets in this survery. By these data we can learn,

(1) Most of data base on the knowledge graph in English, such as Wikidata, Freebase, etc. While data in other language is scarce, such as Chinese. This situation will influence the practical application of multi-KBQA. So a multilingual data is imperative.

(2) In natural language, the topic entity always not just one, so we should label the multi topic entity in the sentence.

(3) A reasoning path is very important, reasoning progress is like the way of human thinking, a better dataset should have the inference path.

(4) Language diversity. There are three kinds of knowledge in KG: Relational (Obama, born in, Hawaii); Attribute (Kobe, height, 198); Factual (Atlanta, population, 47 million). MetaQA and CSQA just consider Relational. Except these knowledge, the dataset should have some QA pairs about arithmetic operation.

3 Basic Concepts

In this section, we will describe some of the basic concepts related to multi-KBQA task.

3.1 Knowledge Graph

knowledge graph is a semantic network where the entities and relations are represented by nodes and edges. Specifically, $\{h,r,t\}$ is represented as a fact in knowledge graph, r refers to the relation between entity h and entity t.

3.2 Knowledge Graph Embedding

Knowledge graph embedding models compute the vector embedding for every node (entity) and ege (relation) in KG. The embedding of node h and t can be represented by e_h^d and e_t^d, where d is the vector dimensionality. Each of method has score function to evaluate the relation between h, r and t.

[5] https://github.com/JunweiBao/MulCQA/tree/ComplexQuestions
[6] https://www.tau-nlp.org/compwebq
[7] http://lc-quad.sda.tech/
[8] https://amritasaha1812.github.io/CSQA/
[9] If you find some wrong with the statistics of data, please contact us

References

[1] Suchanek, F. M., Kasneci, G., and Weikum, G., 2007. "Yago: a core of semantic knowledge". In Proceedings of the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, May 8-12, 2007.

[2] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., and Ives, Z. G., 2007. "Dbpedia: A nucleus for a web of open data". In Semantic Web, International Semantic Web Conference, Asian Semantic Web Conference, Iswc + Aswc, Busan, Korea, November.

[3] Fellbaum, C., and Miller, G., 1998. *WordNet:An Electronic Lexical Database*. MIT Press.

[4] Bollacker, K., Evans, C., Paritosh, P., Sturge, T., and Taylor, J., 2008. "Freebase: A collaboratively created graph database for structuring human knowledge". pp. 1247–1250.

[5] Zhang, Y., Dai, H., Kozareva, Z., Smola, A. J., and Song, L., 2017. "Variational reasoning for question answering with knowledge graph".

[6] Yih, W. T., Richardson, M., Meek, C., Chang, M. W., and Suh, J., 2016. "The value of semantic parse labeling for knowledge base question answering". In Proceedings of the 54th Annual Meeting of the Associa-

tion for Computational Linguistics (Volume 2: Short Papers).

[7] Zhou, M., Huang, M., and Zhu, X., 2018. "An interpretable reasoning network for multi-relation question answering".

[8] Zhang, L., Winn, J., and Tomioka, R., 2016. "Gaussian attention model and its application to knowledge base embedding and question answering".

[9] Unger, C., Forascu, C., Lopez, V., Ngomo, A. C. N., Cabrio, E., Cimiano, P., and Walter, S., 2014. "Question answering over linked data (qald-4)".

[10] Bao, J., Duan, N., Yan, Z., Zhou, M., and Zhao, T., 2016. "Constraint-based question answering with knowledge graph". In COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan, N. Calzolari, Y. Matsumoto, and R. Prasad, eds., ACL, pp. 2503–2514.

[11] Yin, P., Duan, N., Kao, B., Bao, J., and Zhou, M., 2015. "Answering questions with complex semantic constraints on open knowledge bases". pp. 1301–1310.

[12] Talmor, A., and Berant, J., 2018. "The web as a knowledge-base for answering complex questions". In North American Association for Computational Linguistics (NAACL).

[13] Dubey, M., Banerjee, D., Abdelkawi, A., and Lehmann, J., 2019. *LC-QuAD 2.0: A Large Dataset for Complex Question Answering over Wikidata and DBpedia*. 10, pp. 69–78.

[14] Saha, A., Pahuja, V., Khapra, M., Sankaranarayanan, K., and Chandar, S., 2018. "Complex sequential question answering: Towards learning to converse over linked question answer pairs with a knowledge graph".

[15] Shi, J., Cao, S., Pan, L., Xiang, Y., Hou, L., Li, J., Zhang, H., and He, B., 2020. "Kqa pro: A large diagnostic dataset for complex question answering over knowledge base".

[16] Saxena, A., Tripathi, A., and Talukdar, P., 2020. "Improving multi-hop question answering over knowledge graphs using knowledge base embeddings".

[17] Trouillon, T., Welbl, J., Riedel, S., Gaussier, ., and Bouchard, G., 2016. "Complex embeddings for simple link prediction". In International Conference on International Conference on Machine Learning.

[18] Sun, Z., Deng, Z.-H., Nie, J.-Y., and Tang, J., 2019. "Rotate: Knowledge graph embedding by relational rotation in complex space".

[19] Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., and Yakhnenko, O., 2013. "Translating embeddings for modeling multi-relational data".

[20] Balaevi, I., Allen, C., and Hospedales, T., 2019. "Tucker: Tensor factorization for knowledge graph completion".

[21] Xu, W., Zheng, S., He, L., Shao, B., Yin, J., and Liu, T.-Y., 2020. "Seek: Segmented embedding of knowledge graphs".