Research and Applications Analysis of Knowledge Base Question Answering

Jingyi Huang[1,*]

[1] School of Mathematics and Statistics, Wuhan University, Wuhan 430000, China

* Corresponding Author Email: 2014301000011@whu.edu.cn

Abstract. Knowledge Base Question Answering (KBQA) has become one of recent trends in Natural Language Processing (NLP). It helps solve question answering tasks in many fields, such as commerce, medical treatment, etc. This article represents research of KBQA from theory to practice. The concept of knowledge graph in a new way are defined, the steps for building a basic knowledge graph are listed. The category of knowledge base is generalized. This article analyzes the category of knowledge based on systems and introduces the definition and working principle of KBQA. This article also introduces two main approaches used in KBQA, Information Retrieval-based (IR-based) methods and Semantic Parsing-based (SP-based) methods, including summarizing pipeline frameworks of these two approaches and the comparison between them. Two successful applications of KBQA, Meituan and AliMe, including researching on the schema of knowledge graph and pipeline frameworks are discussed in this article. Moreover, this article analyzes the applicable scenes of the two applications and analyzes the main challenges of KBQA.

Keywords: Knowledge Graph, KBQA, IR-based Methods, SP-based Methods, Meituan, AliMe.

1. Introduction

With the rapid development of information technologies, knowledge contained in the internet has grown to an enormous sum, which becomes one of the worthiest parts of the internet. People search for information depending on these huge knowledge bases, however, the traditional searching engines always output answers with low pertinence and large amounts of candidates, specifically, several pages and articles regarding the questions. The emergence of question answering systems can solve this problem. Question answering systems always output a definite answer and it works with a higher efficiency than searching engines. Furthermore, people tend to search for answers for the more complicated questions, and the enterprises tend to manage knowledge data of their own domains to make better use of these data. The birth of knowledge graph caters for these requirements.

A question answering systems based on knowledge graph is called as KBQA. It has achieveda great success on the simple questions and has made great progress on solving complex questions. Some enterprises even have built their own KBQA systems. Research is made on the deep principle of KBQA and some successful practice of it.

2. Knowledge graph

Knowledge graph is a cross field of the applied mathematics, computer graphics, information science and technology, bibliometrics, and citation analysis, which is concerned with representing relations between the entities in the real world. The history of knowledge graph can be traced back to 1960s, when it was firstly known as semantic networks, and the concept of knowledge graph was first introduced by Google in 2012.Since that knowledge graph has been widely applied in many domains, the diverse definitions of knowledge graph have emerged [1]. In this article, the concept of knowledge graph is not different from that of knowledge base, and it is defined as follows:

A knowledge graph/base retrieves information from diverse kinds of data sources and integrates information into an ontology [1] with an architecture consisting of a set of interweaving triples, which describes the relations between entities from the retrieved information.

Each knowledge graph triple is made of three components: entity, property, and property value [2]. The architecture of knowledge graph is shown in Figure 1, the nodes refer to different entities respectively and the edges between them refer to the relation between the two nodes. In Figure 1, Jay Chou, Hannah, and Hathaway are entities referring to human being and Taipei, Taiwan are entities referring to locations. In the red border, the triple is (entity1, relation, entity2), specifically, (Jay Chou, spouse, Hannah), which means the Jay Chou and Hannah are a couple. In the blue border, the triple is (entity, property, value), specifically, (Taipei, the city of, Taiwan). The entity "Taipei" serves as a subject and the entity "Taiwan" serves as a property value used to describe "Taipei".

In Figure 1, the entities are shown in boxes, and the relation is written besides the edge between two boxes. Sub architectures discussed in this article are annotated by blue and red borders.

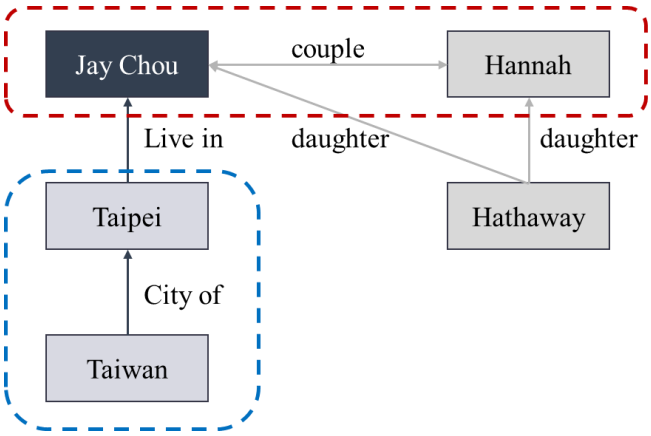

Figure 1. An example of knowledge graph

Constructing a basic knowledge graph consists of following steps:

(1) Data extraction

It is to retrieve data from structured data, semi-structured data, unstructured data by crawling or parsing data from diverse resources such as Wikipedia, Baidu, email, CSV file, etc.

(2) Data normalization

Itis to define the schema design of the knowledge graph depending on retrieved data in step (1) and the application of the final knowledge graph.

(3) Data enrichment

It is to enrich the semi-structured data and unstructured data by Natural Language Processing (NLP) technologies [3] such as Named Entity Recognition (NER), relation extraction and attributive abstraction.

After accomplishing above three steps, knowledge extraction is done.

(4) Knowledge fusion

It is to fuse the extracted knowledge from diverse sources into an ontology by ontology building, data mapping, entity alignment, and entity matching [3].

(5) Knowledge storage

It is to store and manage the knowledge data and solutions for storing and managing depending on the way of interacting with the knowledge graph.

(6) Knowledge update

It is to update data and schema of the knowledge graph for specific applications to advance with the times [2].

Knowledge graph can be categorized as either general-purpose knowledge graph or domain-specific knowledge graph. The former caters for overall internet user; thus, it consists of knowledges from as many fields as possible to chase for higher breadth of knowledge instead of depth. The latter is used to accomplish tasks of specific domain; thus, it mainly consists of professional knowledge from specific domain. In practice, these two kinds of knowledge graphs are often combined to work to take advantage of both.

3. Knowledge base question answering

Knowledge graph can provide Artificial Intelligence (AI) service to solve specific problems and it has become one of the recent trends in accomplishing NLP tasks, such as information retrieval, question answering, chat bot, etc. Such systems working with knowledge graph as its sources are knowledge base systems. A knowledge base system consists of a knowledge base, specifically a knowledge graph, and an inference engine [1].

As for question answering, the history of it could date back to twenty century when Turing firstly introduced Turing test [4]. Recently, technologies regarding question answering have matured considerably after decades of rapid development, and large quantities of question answering systems are proposed. Question answering systems can be categorized, depending on the category of data sources. One relies on unstructured data sources, such as Machine Reading Comprehension (MRC) question answering and Frequently Asked Questions (FAQ) question answering. This kind of question answering systems demands large-scale training data and are often applied to answer generic questions. Another relies on structured data sources, such as Knowledge Base Question Answering (KBQA), also known as Knowledge Graph Question Answering (KGQA). This kind of question answering systems works with higher accuracy and can be applied to answer much more complex questions due to its great performance on reasoning tasks.

KBQA is also a kind of knowledge base system. KBQA provides answers derived from knowledge bases for natural language questions. The architecture of general KBQA is shown in Figure 2. The input natural language question is transformed into the form of tokens; thus, it could be understood by computers. The subject is recognized and is executed against the topic entity that is the counterpart of the subject to figure out the answer derived from the neighborhood of topic entity [5]. The principle of approaches in KBQA will be discussed in Section 4.

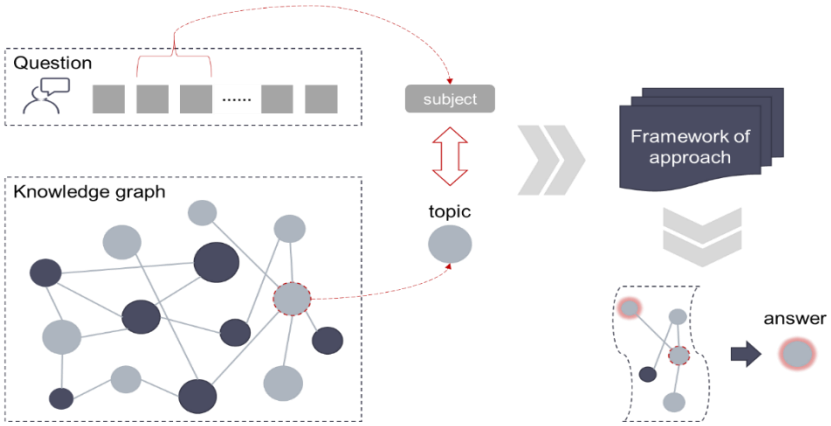

Figure 2. The architecture of KBQA

As for early KBQA, the question is one-hop that there is only one subject and one relation in the question, such as “Who founded Harvard University?”. The entity of the answer is directly connected to the topic entity [6] in the knowledge graph. This kind of KBQA is named as simple KBQA. However, considering the real applications of question answering systems, most questions are multi-hop, so that the question consists of multiple subjects and multiple relations such as “When was the founder of Harvard University born?”. The entity of the answer is often multiple hops or even multiple aggregation away from the topic entity [5].

4. Approaches for KBQA

There are two main approaches for KBQA known as Information Retrieval-based (IR-based) methods and Semantic Parsing-based (SP-based) methods [5]. The former one works as generating a subgraph related to the natural language question from the knowledge graph, retrieving entities of candidate answers from subgraph, and applying ranking generator to get the entity with the highest rank, which is obtained as the predicated answer [5]. The latter one works as transferring the

unstructured natural language question into a query language [7] that can be executed against knowledge base to output the answer [5, 8].

4.1. IR-based methods

The pipeline framework of IR-based methods consists of following modules:

(1) Entity definition

It is to define entities in natural language question and to find the corresponding nodes in knowledge base.

(2) Subgraph construction

In order to decrease the cost of executing the whole knowledge graph, this module is to extract a subgraph with nodes recording all entities in the question and edges recording all relations about the entities.

(3) Question representation

It is to get semantical information of the question by representing it as reasoning instructions through using neural network together with other methods [5].

(4) Graph based reasoning

It is to get the information from neighboring entities and relations [5].

In some recent methods, module (3) and module (4) are repeated for several times.

(5) Answer ranking

It is to rank candidate answers that come from neighboring entities by calculating the matching scores between candidates answers and their questions [7].

4.2. SP-based methods

The pipeline framework of SP-based methods consists of the following modules:

(1) Language understanding

It is to leverage neural networks to obtain the encoded tokens of the input natural language question. The output encoded question should consist of both semantical and syntactic meanings of the question.

(2) Logic forming

It is to apply semantic parsing tools to map the encoded question to an uninstantiated logic form and align the logic form to knowledge base to obtain an instantiated logic form.

(3) Answer executing

It is to execute the logic form against knowledge base to obtain final answer.

4.3. Summary of approaches for KBQA

Due to the recent advance in NLP and deep learning, these two approaches have received the great success in diverse scenarios. However, the difference between the mechanisms makes it necessary to compare the two main approaches when applying them to the specific tasks. The comparison of the two approaches is shown in Table 1.

Table 1. Comparison of IR-based methods and SP-based methods

Dimension	IR-based methods	SP-based methods
Pipeline	extract-reason-rank	parse-execute
Main challenges	Language understanding	Language understanding
	Perturbation of graph context	Logic forms designing
		Large scale of answer executing
Advantages	Optimized easier [5]	Reasoning with high interpretability
Disadvantages	Lacking interpretability	High cost of obtaining an uninstantiated logic form
Applications	Cannot handle complex questions requiring constraint inference [7]	Often applied to domain-specific question answering

Generally, B. et al. [7] pointed out that the results of the SP-based methods are slightly better than most of the IR-based methods. Recently in many scenarios, these two kinds of approaches could be combined to take the advantages of both, thus leading the performance of KBQA to the unparalleled success.

5. Application analysis

According to the state-of-the-art performance of KBQA, many companies have leveraged the KBQA technologies to the commercial and service applications. In this article, three typical applications are discussed.

5.1. KBQA system of Meituan

Meituan is an e-commerce platform that can provide daily life services, and in many business scenarios, question answering systems are needed, specifically, information consultation about commodities on the platform. Thus, the team of Meituan has made a Meituan-specific KBQA system.

As for construction of knowledge graph, Meituan can extract data from its own platform, such as product pages and merchant pages.

As for operating principle of the system, Meituan combined the two mainstream approaches to get a higher interpret ability as well as to decrease the cost of computing. Firstly, it applies the entity recognition and entity linking to obtain the subject entity of the input question. Secondly, it applies dependency analysis to get the relations included in the question and these relations are concerning subject entity. Thirdly, it extracts a subgraph depending on the retrieved entities and relations in previous steps. Finally, it works as module (4) Graph based reasoning and module (5) Answer ranking of IR-based methods to get the highest-ranking answer.

Question answering system of Meituan is a classical application of KBQA. This kind of system can be easily transferred to be applied to other domains on simple questions. The most questions asked in this system come from several structure-fixed options or entities in these questions are easy to be recognized. However, when it comes to scenarios where questions are unstructured and where the platform works with tons of the complicated business data, this kind of system may not ensure the accuracy of answers.

5.2. KBQA system of Alibaba

Alibaba group released an AI shopping assistant named as AliMe in 2015, providing aftermarket service in e-commerce platforms, such as Taobao, Alipay, etc. One function of AliMe is question answering based on a domain-specific knowledge graph.

The knowledge graph of AliMe is based on tons of data from the platforms of Alibaba and the information of merchants on Alibaba. Due to the complexity of questions and knowledge in e-commerce platforms and the difficulty of enumerating every possible scenario, traditional knowledge graph is unable to perform well and accurately the question answering application. Alibaba group has created a new structure of knowledge graph that adds a module of schema to the module of traditional triple representation. Li et al. [9] showed the structure of an evolved triple of this creative knowledge graph and it is shown as Figure 3.

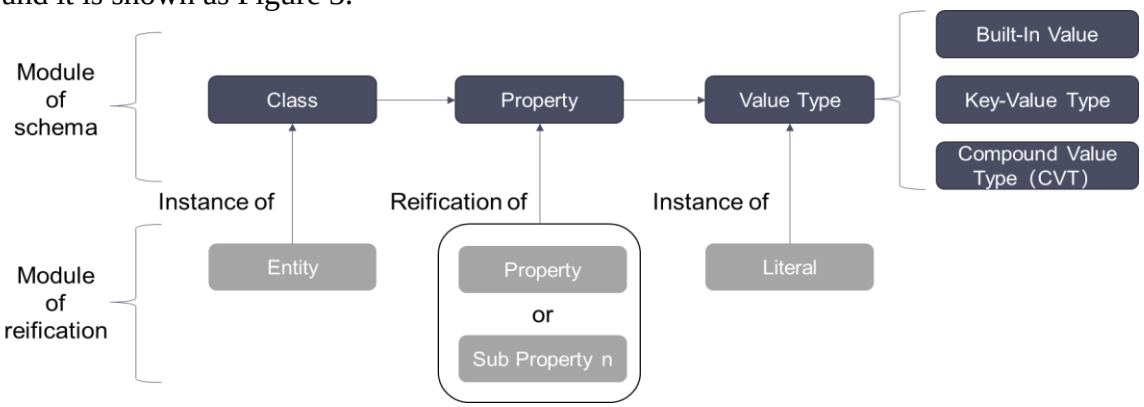

Figure 3. The structured knowledge representation used in AliMe. Here “Property” is shown as nodes, which is shown as edges in traditional structural of knowledge graph. It is only a difference in representation, not in true meaning.

The module of schema consists of "Class", "Property" and "Value Type". "Class" can refer to the category of entities. "Property" can be decomposed into the sub-properties [9]. Entities sharing the same composite property can fit in the same root property and have sub-property respectively, making the final knowledge graph capture both case-level knowledge and category-level knowledge [9]. "Value Type" can refer to the type of property value and is comprised of "Built-In Value", "Key Value Type" and "Compound Value Type(CVT)" in Figure 3. Structured knowledge with these three types of value can capture the constrains of knowledge items and provide answers with the better reading experience [9]. The module of reification consists of the instantiated entity, instantiated property and specific value.

As for operating principle of the system, it works as an evolved IR-based method. Firstly, it recognizes the entity and property in the input question. Secondly, it generates sub-graphs according to entity, property, and the value of property that works as the constrains of knowledge items of the question. Thirdly, in most cases, step two will output the multiple sub-graphs, and a ranking module based on Lambda Rank [10]is needed to rank these sub-graphs. Fourthly, the sub-graph with the highest score is executed to get the answer [9].

This creative structure of knowledge graph does successfully promote accomplishment of basic tasks of question answering. Establishing such a domain-specific knowledge graph costs a lot because lots of resources should be spent on analyzing big data and designing the structure of knowledge representation, specifically, the structure of knowledge graph, which is now manual. An accomplished knowledge graph cannot be easily transferred and be applied in another domain, because the structure of knowledge representation could be totally different due to the different business processes. In conclusion, it is a good deal for big corporations with high business volume like Alibaba to establish such a KBQA system with the high cost and the high accuracy.

5.3. Summary of applications based on KBQA

Both two applications discussed above work on domain-specific knowledge graph, which is common in recent KBQA systems because fixing problems in specific domain is more profitable and practical. The design of the knowledge graph and working pipeline are concerned with the scenarios of questions in specific domain and complexity of source data of the platform.

There remain several challenges in the applications of KBQA. The first one is on how to improve the understanding of complex question including multiple relations [11]. A special structured knowledge graph may work excellently, like AliMe, however, it does cost a lot to achieve such great understanding of questions. Thus, how to make a good balance between the high good performances in complicated scenarios and the affordable costs remains as one of the major challenges in the future. The second one is that many methods could be transferred to the other domain-specific knowledge graphs [12]. The third one is the necessity for reasoning since that even the simplest question needs the multiple steps of reasoning [13]. The fourth one is the lack of training datasets [13] to train these approaches used in KBQA.

6. Conclusions

KBQA has gained a lot of attention these days and has been applied into many fields. This article makes an overview of knowledge graph, specifically, the history and the basic structure of it, as well as steps used to construct a basic knowledge graph. Then, this article reviews the concept of KBQA, which consists of the history, working principle and category of it. To better understand the principle of KBQA, this article also talks about working modules of two main applied approaches. Finally, two successful applications of KBQA, including Meituan and AliMe, are analyzed and discussed. Although KBQA has gained a lot of benefits due the rapid development of NLP, deep learning, and AI, extending KBQA to wider and deeper domains still faces significant challenges. Depending on the analysis of practical applications, this article has concluded several main challenges in KBQA, the main challenges include: complex questions understanding, methods transferring, necessity of

reasoning, and training dataset establishment. These challenges will become the main points of future work.

References

[1] Ehrlinger L. and Wöß W., "Towards a Definition of Knowledge Graphs," p. 4.

[2] Xu Z., Sheng Y., He L., and Wang Y., "Review on Knowledge Graph Techniques," Dianzi Keji Daxue XuebaoJournal Univ. Electron. Sci. Technol. China, Jul. 2016, doi: 10.3969/j.issn.1001-0548.2016.04.012.

[3] Qi G., Gao H., and Wu T., "The Research Advances of Knowledge Graph," Technol. Intell. Eng., vol. 3, no. 1, pp. 4–25, Jan. 2017, doi: 10.3772/j.issn.2095-915x.2017.01.002.

[4] TURING A. M., "COMPUTING MACHINERY AND INTELLIGENCE," Mind, vol. LIX, no. 236, pp. 433–460, Jan. 1950.

[5] Lan Y., He G., Jiang J., Jiang J., Zhao W. X., and Wen J.-R., "Complex Knowledge Base Question Answering: A Survey." arXiv, Apr. 17, 2022. Accessed: Jun. 06, 2022. [Online]. Available: http://arxiv.org/abs/2108.06688

[6] Lan Y., He G., Jiang J., Jiang J., Zhao W. X., and Wen J.-R., "A Survey on Complex Knowledge Base Question Answering: Methods, Challenges and Solutions." arXiv, May 24, 2021. Accessed: Jun. 06, 2022. [Online]. Available: http://arxiv.org/abs/2105.11644.

[7] Fu B., Qiu Y., Tang C., Li Y., Yu H., and Sun J., "A Survey on Complex Question Answering over Knowledge Base: Recent Advances and Challenges." arXiv, Jul. 26, 2020. Accessed: Jun. 24, 2022. [Online]. Available: http://arxiv.org/abs/2007.13069

[8] Wu P., Zhang X., and Feng Z., "A Survey of Question Answering over Knowledge Base," in Knowledge Graph and Semantic Computing: Knowledge Computing and Language Understanding, Singapore, 2019, pp. 86–97.

[9] Li F.-L., Chen W., Huang Q., and Guo Y., "AliMe KBQA: Question Answering over Structured Knowledge for E-Commerce Customer Service," in Knowledge Graph and Semantic Computing: Knowledge Computing and Language Understanding, vol. 1134, X. Zhu, B. Qin, X. Zhu, M. Liu, and L. Qian, Eds. Singapore: Springer Singapore, 2019, pp. 136–148. doi: 10.1007/978-981-15-1956-7_12.

[10] Burges C. J. C., "From RankNet to LambdaRank to LambdaMART: An Overview," p. 19.

[11] Keysers D. et al., "Measuring Compositional Generalization: A Comprehensive Method on Realistic Data." arXiv, Jun. 25, 2020. Accessed: Jul. 02, 2022. [Online]. Available: http://arxiv.org/abs/1912.09713

[12] Yih W., He X., and Meek C., "Semantic Parsing for Single-Relation Question Answering," in Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Baltimore, Maryland, 2014, pp. 643–648. doi: 10.3115/v1/P14-2105.

[13] Kapanipathi P., Abdelaziz I., Ravishankar S., Roukos S., and Yu M., "Question Answering over Knowledge Bases by Leveraging Semantic Parsing and Neuro-Symbolic Reasoning," 2020.