Complex Knowledge Base Question Answering: A Survey

Yunshi Lan, Gaole He [ID], Jinhao Jiang, Jing Jiang,
Wayne Xin Zhao [ID], *Member, IEEE*, and Ji-Rong Wen [ID], *Senior Member, IEEE*

Abstract—Knowledge base question answering (KBQA) aims to answer a question over a knowledge base (KB). Early studies mainly focused on answering simple questions over KBs and achieved great success. However, their performances on complex questions are still far from satisfactory. Therefore, in recent years, researchers propose a large number of novel methods, which looked into the challenges of answering complex questions. In this survey, we review recent advances in KBQA with the focus on solving complex questions, which usually contain multiple subjects, express compound relations, or involve numerical operations. In detail, we begin with introducing the complex KBQA task and relevant background. Then, we present two mainstream categories of methods for complex KBQA, namely semantic parsing-based (SP-based) methods and information retrieval-based (IR-based) methods. Specifically, we illustrate their procedures with flow designs and discuss their difference and similarity. Next, we summarize the challenges that these two categories of methods encounter when answering complex questions, and explicate advanced solutions as well as techniques used in existing work. After that, we discuss the potential impact of pre-trained language models (PLMs) on complex KBQA. To help readers catch up with SOTA methods, we also provide a comprehensive evaluation and resource about complex KBQA task. Finally, we conclude and discuss several promising directions related to complex KBQA for future research.

Index Terms—Knowledge base question answering, knowledge base, question answering, natural language processing, survey

✦

1 INTRODUCTION

KNOWLEDGE base (KB) is a structured database that contains a collection of facts (alias triples) in the form (*subject*, *relation*, *object*). Large-scale KBs, such as Freebase [1], DBPedia [2], Wikidata [3], and YAGO [4], have been constructed to serve many downstream tasks. Among them, knowledge base question answering (KBQA) is a task that aims to answer natural language questions with KBs acting as its knowledge source. Nowadays, KBQA has attracted intensive attention from researchers as it plays an important

role in many intelligent applications such as Apple Siri, Microsoft Cortana and so on [5].

Early work on KBQA focused on answering a simple question, where only a single fact is involved. For example, *"Who was the nominee of The Jeff Probst Show?"* is a simple question which includes the subject *"The Jeff Probst Show"*, the relation *"nominee"* and queries about the object entity *"Jeff Probst"* of fact *"(The Jeff Probst Show, nominee, Jeff Probst)"* in KBs. It is not trivial to retrieve the correct entity from large-scale KBs, which consist of millions or even billions of facts. Therefore, researchers have spent much effort in proposing different models to answer simple questions over KBs [6], [7], [8], [9], [10].

Recently, researchers started paying more attention to answering ***complex questions*** over KBs, i.e., the complex KBQA task [11], [12]. Complex questions usually contain multiple subjects, express compound relations, or include numerical operations. Take the question in Fig. 1 as an example. This example question starts with the subject *"The Jeff Probst Show"*. Instead of querying a single fact, the question requires the composition of two relations, namely, *"nominee"* and *"spouse"*. This query is also associated with an entity type constraint *"(Jeff Probst, is a, TV producer)"*. The final answer should be further aggregated by selecting the possible candidates with the earliest marriage date. Generally, complex questions are questions involving ***multi-hop reasoning***, ***constrained relations*** or ***numerical operations***.

Tracing back to the solutions for simple KBQA task, a number of studies from two mainstream approaches have been proposed. We show the overall architecture of simple KBQA systems in Fig. 2. These two approaches first recognize the subject in a question and link it to an entity in the KB (referred to as the ***topic entity***). Then they derive the answers

• *Yunshi Lan is with the School of Data Science and Engineering, East China Normal University, Shanghai 200050, China. E-mail: yslan@dase.ecnu.edu.cn.*
• *Gaole He is with the Web Information Systems group of the Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS/EWI), Delft University of Technology, 2628 CD Delft, The Netherlands. E-mail: g.he@tudelft.nl.*
• *Jinhao Jiang, Wayne Xin Zhao, and Ji-Rong Wen are with the Gaoling School of Artificial Intelligence, Renmin University of China, Beijing 100872, China, and Beijing Key Laboratory of Big Data Management and Analysis Methods, Beijing 100872, China. E-mail: {jiangjinhao, jrwen}@ruc.edu.cn, batmanfly@gmail.com.*
• *Jing Jiang is with the School of Computing and Information System, Singapore Management University, Singapore 188065. E-mail: jingjiang@smu.edu.sg.*
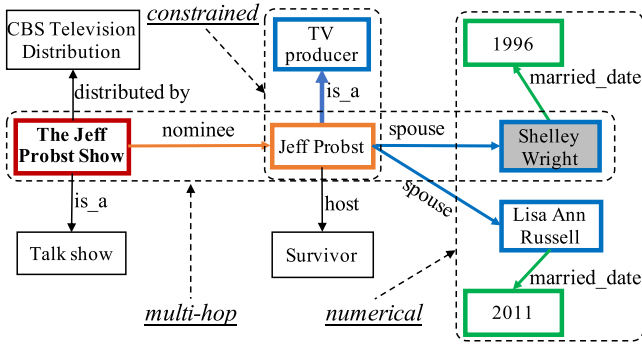
Fig. 1. We present the related KB subgraph for the question "*Who is the first wife of TV producer that was nominated for The Jeff Probst Show?*". The topic entity and the answer entity are shown in the bold font and shaded box respectively. "multi-hop", "constrained", and "numerical" are highlighted in black dotted box. Different colors indicate different reasoning hops heading to the answer.

within the neighborhood of the topic entity by either executing a parsed logic form or reasoning in a question-specific graph extracted from the KB. The two categories of methods are commonly known as *semantic parsing-based* (SP-based) methods and *information retrieval-based* (IR-based) methods in prior work [6], [7], [10], [13]. They design different working mechanisms to solve the KBQA task. The former approach represents a question by a symbolic logic form, and then executes it against the KB to obtain the final answers. The latter approach constructs a question-specific graph delivering comprehensive information related to the question, and generates the final answers based on the extracted graph.

However, when applying the two mainstream approaches to the complex KBQA task, complex questions bring in challenges on different parts of the approaches:

- Parsers used in existing SP-based methods are difficult to cover diverse complex queries (e.g., multi-hop reasoning, constrained relations). Similarly, previous IR-based methods may fail to answer a complex query, as the answer is generated without traceable reasoning.
- More relations and subjects in complex questions indicate a larger search space of potential logic forms for parsing, which will dramatically increase the computational cost. Meanwhile, more relations and subjects prevent IR-based methods from retrieving all relevant facts for reasoning, which makes the common incomplete KB issue become severer.
- When questions become complicated from both semantic and syntactic aspects, models are required

to have strong capabilities of natural language understanding and generalization. Comparing the question "*Who is the first wife of TV producer that was nominated for The Jeff Probst Show?*" with another question "*Who is the wife of the first TV producer that was nominated for The Jeff Probst Show?*", the models are supposed to understand that the ordinal number "*first*" is used to constrain "*wife*" or the phrase "*TV producer*".

- Generally, only question-answer pairs are provided. This indicates SP-based methods and IR-based methods have to be trained without the annotation of correct logic forms and reasoning paths. Such weak supervision signals bring difficulties to both approaches due to the lack of guidance in intermediate reasoning process.

Regarding the related surveys, we observe Wu et al. [14] and Chakraborty et al. [15] reviewed the existing work on simple KBQA. Gu et al. [16] provided a semantic parsing perspective to KBQA task. Furthermore, Fu et al. [17] investigated the current advances in complex KBQA. They provided a general view of advanced methods only from the perspective of techniques and more focused on application scenarios in e-commerce domain. Different from these surveys, our work tries to identify the challenges encountered in previous studies, and extensively discuss existing solutions in a comprehensive and well-organized manner. It is worth noting that this survey is an extended version of the short survey [18]. As a comparison, this survey has several main differences: (1) We add more recently-published papers and refine the description of challenges as well as solutions with a fine-grained taxonomy. (2) We provide deep discussions of the two mainstream categories including a comprehensive comparison of their core modules and a unified paradigm of neural symbolic reasoning. (3) We add a new section to discuss the role of cutting-edge pre-trained language techniques for complex KBQA and give a more thorough outlook on several promising research directions.

The remainder of this survey is organized as follows. We first introduce preliminary knowledge in Section 2. Next, we describe the two mainstream categories of methods for complex KBQA in Section 3. Following the categorization, we figure out typical challenges and solutions for SP-based and IR-based methods in Sections 4 and 5, respectively. We highlight the impact of pre-trained language models on complex KBQA in Section 6. We summarize datasets and relevant resources in Section 7. Finally, we discuss recent trends and conclude in Sections 8 and 9.
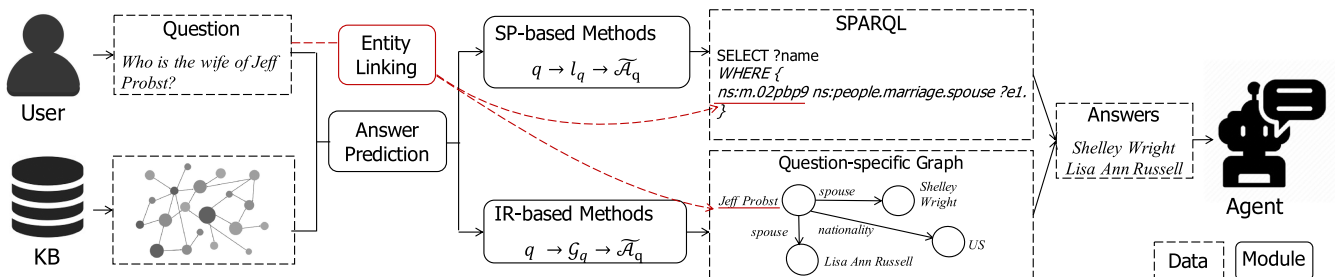

Fig. 2. Architecture of KBQA systems. The entity linking procedure is shown in red color.

2 PRELIMINARY

In this section, we first briefly introduce KBs and the task formulation of KBQA, then we talk about the traditional approaches for KBQA systems.

2.1 Knowledge Base

As mentioned earlier, KB is usually in the format of triples. They are designed to support modeling relationships between entities. Take Freebase [19] as an example for KB. Each entity in Freebase has a unique ID (referred to as **mid**), one or more types, and uses properties from these types in order to provide facts [3]. For example, the Freebase entity for person Jeff Probst has the mid "$m.02pbp9$"[1] and the type "$people.marriage$" that allows the entity to have a fact with "$people.marriage.spouse$" as the property and "$m.0j6d0bg$" (psychotherapist Shelley Wright) as the value. Freebase incorporates compound value types (CVTs) to represent n-ary ($n > 2$) relational facts [1] like "*Jeff Probst was married to Shelley Wright in 1996*", where three entities, namely "*Jeff Probst*", "*Shelley Wright*", and "*1996*", are involved in a single statement. Different from entities that can be aligned with real world objects or concepts, CVTs are artificially created for such n-ary facts.

In practice, large-scale open KBs (e.g., Freebase and DBPedia) are published under Resource Description Framework (RDF) to support structured query language [3], [20]. To facilitate access to large-scale KBs, the query language SPARQL is frequently used to retrieve and manipulate data stored in KBs [3]. In Fig. 2, we have shown an executable SPARQL to obtain the spouses of entity "*Jeff Probst*".

Different KBs are designed with different purposes and have varying properties under different schema designs. For example, Freebase is created mainly by community members and harvested from many resources including Wikipedia. YAGO [21] takes Wikipedia and WordNet [22] as the knowledge resources and covers taxonomy of more general concepts. WikiData [3] is a multilingual KB which integrates multiple resources of KBs with high coverage and quality. A more comprehensive comparison between open KBs is available at [23].

2.2 Task Formulation

Formally, we denote a KB as $\mathcal{G} = \{\langle e, r, e' \rangle | e, e' \in \mathcal{E}, r \in \mathcal{R}\}$, where $\langle e, r, e' \rangle$ denotes that relation r exists between subject e and object e', $\mathcal{E}$ and $\mathcal{R}$ denote the entity set and relation set, respectively.

Given the available KB $\mathcal{G}$, this task aims to answer natural language questions $q = \{w_1, w_2, ..., w_m\}$ in the format of a sequence of tokens (typically organized with a unique vocabulary $\mathcal{V}$) and we denote the predicted answers as $\tilde{\mathcal{A}}_q$. Specially, existing studies assume the correct answers $\mathcal{A}_q$ can be derived from the entity set $\mathcal{E}$ of the KB or a natural language sequence (i.e., the surface name of entities). Unlike answers of simple KBQA task, which are entities directly connected to the topic entity, the answers of the complex KBQA task are entities multiple hops away from the topic entities or even the aggregation of them. Generally, a KBQA system is trained using a dataset $\mathcal{D} = \{(q, \mathcal{A}_q)\}$.

2.3 Traditional Approaches

General KBQA systems for simple questions have a pipeline framework as displayed in Fig. 2. The preliminary step is to identify the topic entity e_q of the question q, which aims at linking a question to its related entities in the KBs. In this step, named entity recognition, disambiguation and linking are performed. It is usually done using some off-the-shelf entity linking tools, such as S-MART [24], DBpedia Spotlight [25], and AIDA [26]. Subsequently, an answer prediction module is leveraged to predict the answer $\tilde{\mathcal{A}}_q$ taking q as the input.

For simple KBQA task, the predicted answers are usually located within the neighborhood of the topic entities. Different features, as well as methods, are proposed to rank these candidate entities. Early attempts on solving simple KBQA task employed existing semantic parsing tools to parse a simple natural language question into an uninstantiated logic form, and then adapted it to KB schema by aligning the lexicons. This step results in an executable logic form l_q for q. In detail, the existing semantic parsing tools usually follow Combinatory Categorical Grammars (CCGs) [27], [28], [29] to build domain-independent logic forms. Then different methods [27], [28], [29], [30], [31] are proposed to perform schema matching and lexicon extension, which results in logic forms grounded with KB schema. For simple KBQA task, this logic form is usually a single triple starting from the topic entity and connecting to the answer entities. As early methods heavily rely on rule-based mapping, which is hard to generalize to large-scale datasets [32], [33], [34]. Follow-up work proposed some scoring functions to automatically learn the lexicon coverage between the logic forms and the questions [35], [36]. With the development of deep learning, several advanced neural networks such as Convolutional Neural Network [37], Hierarchical Residual BiLSTM [8], Match-Aggregation Module [38], and Neural Module Network [39] are utilized to measure the semantic similarities. This line of work is known as semantic parsing-based methods.

Information retrieval-based methods were also developed over the decades. They retrieve a question-specific graph $\mathcal{G}_q$ from the entire KB. Generally, entities one hop away from the topic entity and their connected relations form the subgraph for solving a simple question. The question and candidate answers in the subgraph can be represented as low-dimensional dense vectors. Different ranking functions are proposed to rank these candidate answers and top-ranked entities are considered as the predicted answers [6], [9], [40], [41]. Afterwards, Memory Network [42] is employed to generate the final answer entities [43], [44]. More recent work [7], [45], [46] employs attention mechanism or multi-column modules to this framework to boost the ranking accuracy. In Fig. 2, we have displayed different pipelines and intermediate outputs of the two methods.

3 TWO MAINSTREAM APPROACHES

Complex KBQA systems follow the same overall architecture as shown in Fig. 2, where the entity linking is first performed. Subsequently, as introduced in Section 1, SP-based and IR-based methods are two mainstream approaches to

---

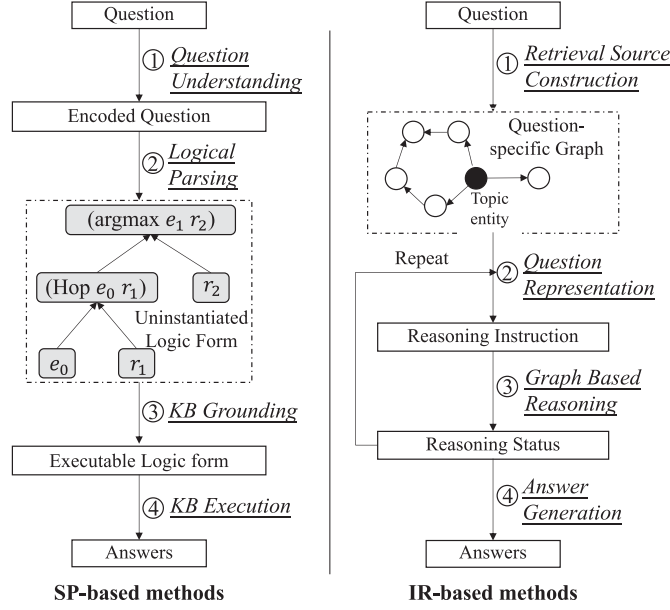1. PREFIX: $<http://rdf.freebase.com/ns/>$

Fig. 3. Illustration of two mainstream approaches for complex KBQA.

answering complex questions. SP-based methods parse a question into a logic form and execute it against KBs for finding the answers. IR-based methods retrieve a question-specific graph and apply some ranking algorithms to select entities from top positions or directly generate answers with a text decoder. To summarize, the two approaches follow either a *parse-then-execute* paradigm or a *retrieve-and-generate* paradigm. To show the difference between the two paradigms, we illustrate their question answering procedures with detailed modules in Fig. 3.

3.1 Semantic Parsing-Based Methods

As shown in Fig. 3, we summarize the procedure of SP-based methods into the following four modules:

1) They understand a question via a <u>question under-standing</u> module, which is to conduct the semantic and syntactic analysis and obtain an encoded question for the subsequent parsing step. We denote this module as follows:

$$\tilde{q} = Question\_Understanding(q),$$

where $\tilde{q}$ is the encoded question that captures semantic and syntactic information of the natural language question. It can be distributed representation, structural representation, or their combination. Intuitively, neural networks (e.g., LSTM [47], GRU [48], and PLMs) are employed to act as the question understanding module and obtain hidden states to represent the question. Meanwhile, some syntactic parsing is performed to extract structural properties of the question.

2) A <u>logical parsing</u> module is utilized to transfer the encoded question into an uninstantiated logic form:

$$\bar{l}_q = Logical\_Parsing(\tilde{q}),$$

where $\bar{l}_q$ is the uninstantiated logic form without the detailed entities and relations filled in. The grammar

and constituents of logic forms can be different with specific designs of a system. Here, $\bar{l}_q$ can be obtained by either generating a sequence of tokens or ranking a set of candidates. In practice, Seq2seq models or feature-based ranking models are employed to generate $\bar{l}_q$ based on the encoded question.

3) To execute against KBs, the logic form is further instantiated and validated by conducting some semantic alignments to structured KBs via <u>KB grounding</u>. Note that, in some work [35], [49], the logical parsing and KB grounding are simultaneously performed, where logic forms are validated in KBs while partially parsed:

$$l_q = KB\_Grounding(\bar{l}_q, \mathcal{G}).$$

After this step, $\bar{l}_q$ is instantiated with the entities and relations in $\mathcal{G}$ so that we obtain an executable logic form l_q. It is worth noting that l_q always contains e_q, which are detected via entity linking module. Its format is not restricted to the SPARQL query but is always transferable to SPARQL.

4) Eventually, the parsed logic form is executed against KBs to generate predicted answers via a <u>KB execution</u> module:

$$\tilde{\mathcal{A}}_q = KB\_Execution(l_q),$$

where $\tilde{\mathcal{A}}_q$ is the predicted answers for the given question q. This module is usually implemented via an existing executor.

During training, the logic form l_q is treated as the intermediate output. The methods are trained using the KBQA datasets in the format of $\mathcal{D} = \{(q, \mathcal{A}_q)\}$, where the objective is set to generate a logic form matching the semantics of the question and resulting in correct answers.

3.2 Information Retrieval-Based Methods

Similarly, we summarize the procedure of IR-based methods into four modules as illustrated in Fig. 3:

1) Starting from the topic entity e_q, the system first extracts a question-specific graph from KBs. Ideally, this graph includes all question-related entities and relations as nodes and edges respectively. Without explicitly generating an executable logic form, IR-based methods perform reasoning over the graph. We represent a <u>retrieval source construction</u> module taking as input of both the question and KB as:

$$\mathcal{G}_q = Retrieval\_Source\_Construction(q, \mathcal{G}),$$

where $\mathcal{G}_q$ is the question-specific graph extracted from $\mathcal{G}$. As the size of subgraph grows exponentially with the distance to topic entities, some filtering tricks (e.g., personalized Pagerank) are adopted to keep the graph size in a computation-affordable scale [50], [51].

2) Next, the system encodes input questions via a <u>question representation</u> module. This module analyzes the semantics of the question and outputs reasoning instructions, which are usually represented as vectors. Typically, question q is encoded into hidden

vectors q with neural networks (e.g., LSTM, GRU, and PLMs) and then combined with other methods (e.g., attention mechanism) to generate a vector as instruction:

$$\boldsymbol{i}^{(k)} = Question\_Representation(\boldsymbol{i}^{(k-1)}, q, \mathcal{G}_q)$$

Here, $\{\boldsymbol{i}^{(k)}, k = 1, ..., n\}$ is the instruction vector of k-th reasoning that encodes the semantic and syntactic information of the natural language question. Both multi-step reasoning and one-step matching are applicable, which results in varying reasoning steps n.

3) A _graph based reasoning_ module conducts semantic matching via vector-based computation to propagate and aggregate the information along the neighboring entities within the graph. The reasoning status $\{s^{(k)}, k = 1, ..., n\}$, which has diverse definitions in different methods (e.g., distributions of predicted entities and representations of relations), is updated based on the reasoning instruction:

$$s^{(k)} = Graph\_Based\_Reasoning(s^{(k-1)}, \boldsymbol{i}^{(k)}, \mathcal{G}_q),$$

where $s^{(k)}$ is the reasoning status which is considered as the status of k-th reasoning step on the graph. Recently, several studies [44], [46] repeat steps (2) and (3) multiple times to perform the reasoning.

4) An _answer generation_ module is utilized to generate answers according to the reasoning status at the end of reasoning. There are mainly two types of such generators: (1) entity ranking generator which ranks the entities to obtain top-ranked entities as predicted answers, (2) text generator which generates free text answers with vocabulary $\mathcal{V}$. This module can be formalized as:

$$\tilde{\mathcal{A}}_q = Answer\_Generation(s^{(n)}, \mathcal{G}_q, \mathcal{V}),$$

where $s^{(n)}$ denotes the reasoning status at the last step.

In the entity ranking paradigm, the entities contained in $\mathcal{G}_q$ are candidates for answer prediction $\tilde{\mathcal{A}}_q$. In many cases, $\tilde{\mathcal{A}}_q$ is obtained through selecting the entities with a score larger than the pre-defined threshold, where the score is derived from $s^{(n)}$. While in text generation paradigm, the answers are generated from vocabulary $\mathcal{V}$ as a sequence of tokens.

During training, the objective of entity ranking generator is usually to rank the correct entities higher than others in $\mathcal{G}_q$. In comparison, the text generator is usually trained to generate gold answers (name of correct entities).

3.3 A Comparison of Core Modules

Comparing the procedures of SP-based and IR-based methods, we note that these two methods have different designs of core modules and working mechanisms, but they also share similarities in multiple aspects.

Difference: SP-based methods rely heavily on the logical parsing module, which produces an expressive logic form for each question. In practice, many commercial KBQA systems developed upon SP-based methods require expertise to provide feedback to the generated logic forms so that the system can be further improved [52]. However, considering the expensive cost and expertise required for obtaining the annotated logic form, SP-based methods are usually trained under a weak supervision setting in research. Compared with IR-based methods, SP-based methods have the advantage of showing interpretability with explicit evidence about reasoning and defending perturbation of the question. However, the logical parsing module is bound by the design of the logic form and the capability of the parsing techniques, which is the key of performance improvement.

As a comparison, IR-based methods first employ a retrieval module to obtain the question-specific graph, and then conduct complex reasoning on the graph structure with graph based reasoning module. The answers are eventually predicted via an answer generation module. The performance of the IR-based methods partially depends on the recall of the retrieval module as the subsequent reasoning takes the retrieved graph as input. Meanwhile, the graph based reasoning and answer generation module play key roles in making accurate predictions. Instead of generating logic forms, IR-based methods directly generate entities or free text as predictions. So they naturally fit into the end-to-end training paradigm and could be optimized easier compared with SP-based methods. Nevertheless, the blackbox style of the reasoning module makes the reasoning process less interpretable, which decreases the robustness and hinders users from interacting with the system.

Similarity: SP-based and IR-based methods both contain parameter-free modules, which are KB grounding, KB execution modules, and retrieval source construction module, respectively. While they are generally not learned from KBQA datasets, their performance has a great impact on the final KBQA performance. Both categories of methods make use of detected topic entities. SP-based methods leverage them to instantiate the logic form in the KB grounding module, while IR-based methods utilize them to narrow down the reasoning scale in the retrieval source module. Besides, both SP-based and IR-based methods emphasize the importance of natural language understanding with a question understanding (representation) module. The output of such modules substantially influences the subsequent parsing or reasoning process.

3.4 A Unified Paradigm - Neural Symbolic Reasoning

In recent years, neural symbolic reasoning (NSR) has become a hot topic in machine learning. _It describes a type of hybrid systems that apply the high efficiency of connectionism (the neural system) and the generalization of symbolism (the symbolic system) to integrate learning and reasoning effectively_ [70], [71], [72]. Relevant techniques are widely applied to intelligent applications like question answering [73] and semantic parsing [74]. As for complex KBQA, NSR techniques are helpful in addressing some challenges for SP-based and IR-based methods. Furthermore, NSR may be a potential paradigm to unify both SP-based and IR-based methods.

The symbolic system and neural system play different roles in KBQA tasks. The symbolic system usually takes the
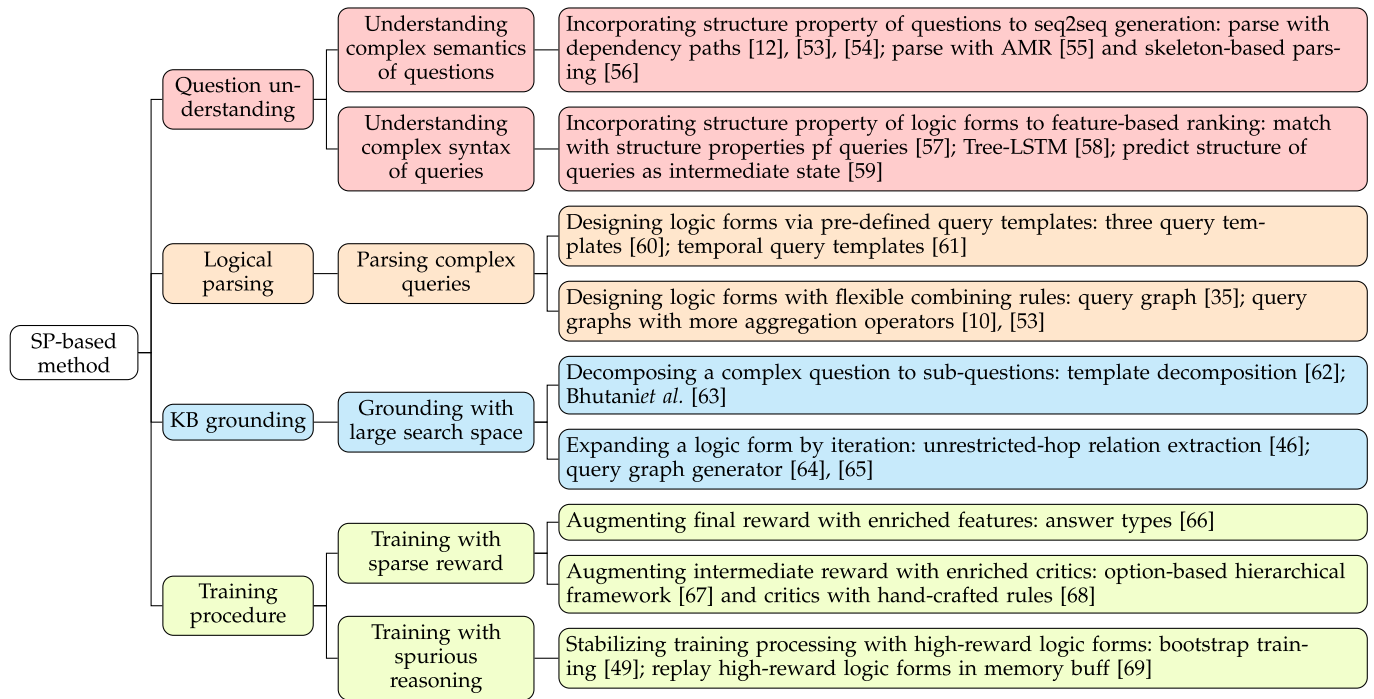
Fig. 4. The main content of SP-based methods. The hierarchical structure is arranged with: SP-based method → module → challenge → solution.

KBs and grammar rules as inputs, searches for the solution space for question answering, and performs reasoning for the results. In comparison, the neural system takes the natural language questions and elements in KB as inputs, learns a neural network model for a specific task, and serves the reasoning in latent space. In this way, NSR-based methods could reason with powerful neural networks in latent space, meanwhile providing explicit inference evidence to explain the results as well as the reasoning process. For SP-based methods, the logical parsing and KB grounding modules act as the symbolic system to interact with KBs and search for potential logical forms. The question understanding module usually acts as the neural system to learn the semantic matching between the given question and the potential logical forms [12], [34], [49]. For IR-based methods, the retrieval source construction and graph based reasoning modules link to the symbolic system, while the neural system usually consists of question representation and graph based reasoning modules [43], [50], [51].

The benefits of applying neural symbolic reasoning to complex KBQA come from the following aspects: *1) Symbolic system facilitates discrete inferences on the structured data.* Symbolic system helps narrow down the search space for complex questions, increase the interpretability of reasoning process [75], and compositional generalization capability of systems [13]. For example, Bao et al. [75] and Lan et al. [65] developed symbolic systems to cope with a large range of questions and integrate diverse modalities. *2) Neural system facilitates modeling heterogeneous and imperfect data.* Neural system deals with the diversity of natural language expressions of complex questions [50], manipulates with heterogeneous data [76] (e.g., complex questions, entities, relations, and even generated templates) and even infers relations that are missing from the incomplete KB [77]. It has been proven to be effective in solving the above issues in both

SP-based and IR-based methods [34], [50]. *3) Neural symbolic reasoning (NSR) can take advantage of both neural systems and symbolic systems.* For SP-based methods, Liang et al. [49] designed the subset of λ-calculus as the search space of logical forms in the symbolic system to fit the nature of complex queries in complex KBQA, which further constrains the generation of programs through seq2seq neural models. Similarly, in IR-based methods, Sun et al. [78] developed PullNet, a neural symbolic machine that conducts (symbolic) reasoning graph expanding along with the graph-based (neural) reasoning process, and achieved promising performance.

As we can see, complex KBQA systems show a trend to connect with and benefit from neural symbolic reasoning. The two mainstream approaches can be unified with the neural symbolic reasoning paradigm, which differs in the detailed designs of the symbolic system and neural system. More discussion can be found in recent study [72].

4 SEMANTIC PARSING-BASED METHODS

In this part, we discuss the challenges and solutions for semantic parsing-based methods. The taxonomy of challenges and solutions can be visualized with Fig. 4.

4.1 Overview

As introduced in Section 3, SP-based methods follow a parse-then-execute procedure via a series of modules, namely question understanding, logical parsing, KB grounding, and KB execution. These modules encounter different challenges for complex KBQA. First, question understanding becomes more difficult when the questions are complicated in both semantic and syntactic aspects. Second, logical parsing has to cover diverse query types of complex questions. Moreover, a complex question involving more relations and subjects will
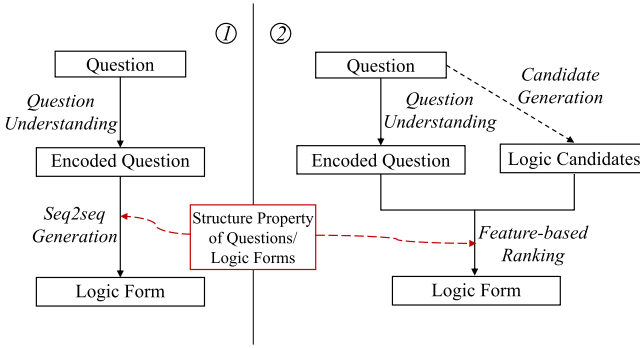
Fig. 5. Illustration of two lines of research which leverage structure properties for better understanding of complex question.

dramatically increase the possible search space for parsing. Third, the manual annotation of logic forms is expensive and labor-intensive, and it is challenging to train the SP-based methods with weak supervision signals (i.e., question-answer pairs).

In the following parts, we will introduce how prior studies deal with these challenges and summarize advanced techniques proposed by them.

4.2 Understanding Complex Semantics and Syntax

As the first step of SP-based methods, question understanding module converts unstructured text into encoded question, which benefits the downstream parsing. Compared with simple questions, complex questions are featured with compositional semantics and more complex query types, which increase the difficulty in linguistic analysis.

4.2.1 Understanding Complex Semantics of Questions

The complex semantics of questions indicates a complex dependency pattern of sentences, which expresses the relation between constituents. Knowing the core part of the sentence structure could be beneficial for question understanding. Incorporating structure property of questions is an intuitive strategy to achieve this goal.

Incorporating Structure Property of Questions to seq2seq Generation. Many existing methods rely on syntactic parsing, such as dependencies [12], [53], [54] and Abstract Meaning Representation (AMR) [55], to provide better alignment between question constituents and logic form elements (e.g., entity, relation, entity types, and attributes). This line of research is illustrated at the left side of Fig. 5. In order to represent long-range dependencies between the answer and the topic entity in question, Luo et al. [12] extracted the dependency path between them. By encoding the directional dependency path, they concatenated both syntactic features and local semantic features together to form global question representations. Similarly, Abujabal et al. [53] leveraged dependency parse to cope with compositional utterances and only focused on important tokens contained by parsed dependency path when creating query templates. Instead of directly creating logic forms upon the dependency paths, Abujabal et al. [54] leveraged dependency parse to analyze the composition of the utterances and aligned it with the logic form. Kapanipathi et al. [55] introduced AMR to help understand questions. The benefits are two-fold: (1) AMR is effective in disambiguating natural

language utterances. (2) AMR parsing module is highly abstract and helps to understand the questions in a KB-independent way. However, the accuracy of producing syntactic parsing is still not satisfying on complex questions, especially for those with long-distance dependency.

In order to alleviate the inaccurate syntactic parsing of complex questions, Sun et al. [56] leveraged the skeleton-based parsing to obtain the trunk of a complex question, which is a simple question with several branches (i.e., head word of original text-spans) to be expanded. For example, the trunk for question *"What movie that Miley Cyrus acted in had a director named Tom Vaughan?"* is *"What movie had a director?"*, and attributive clauses in question will be regarded as the branches of the trunk. Under such a skeleton structure, only simple questions are to be parsed further, which is more likely to obtain accurate parsing results.

4.2.2 Understanding Complex Syntax of Queries

It is important to understand questions by analyzing their complex semantics. It is also crucial to analyze the syntax of queries and ensure that the generated logic forms could meet the complex syntax of queries. While the above methods generate logic forms with Seq2seq framework, another line of work (shown at the right side of Fig. 5) focuses on leveraging structural properties (e.g., tree structure or graph structure of logic forms) for ranking candidate parsing.

Incorporating Structure Property of Logic Forms to Feature-Based Ranking. Maheshwari et al. [79] proposed a novel ranking model which exploits the structure of query graphs and uses attention weights to explicitly compare predicates with natural language questions. Specifically, they proposed a fine-grained slot matching mechanism to conduct hopwise semantic matching between the question and each predicate in the core reasoning chain. Instead of capturing semantic correlations between a question and a simple relation chain, Zhu et al. [57] focused on structure properties of query and conducted KBQA with query-question matching. They employed a structure-aware encoder to model entity or relation context in a query, promoting the matching between queries and questions. Similarly, Zafar et al. [58] incorporated two Tree-LSTMs [80] to model dependency parse trees of questions and tree structure of candidate queries respectively, and leveraged structural similarity between them for comprehensive ranking.

Traditional methods adopted a state-transition strategy to generate candidate query graphs. As this strategy ignores the structure of queries, a considerable number of invalid queries will be generated as candidates. To filter these queries out, Chen et al. [59] proposed to predict the query structure of the question and leverage the structure to restrict the generation of the candidate queries. Specifically, they designed a series of operations to generate placeholders for types, numerical operators, predicates, and entities. After that, they can ground such uninstantiated logic forms with KBs and generate executable logic forms.

4.3 Parsing Complex Queries

To generate an executable logic form, traditional methods first utilized the existing parsers to convert a question into CCG derivation which is then mapped to a SPARQL via
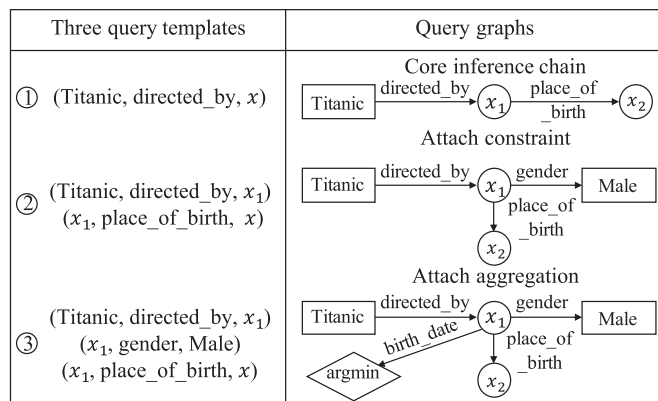
Three query templates	Query graphs
① (Titanic, directed_by, x)	Core inference chain Titanic →directed_by x_1 →place_of_birth x_2
② (Titanic, directed_by, x_1) (x_1, place_of_birth, x)	Attach constraint Titanic →directed_by x_1 →gender Male x_1 →place_of_birth x_2
③ (Titanic, directed_by, x_1) (x_1, gender, Male) (x_1, place_of_birth, x)	Attach aggregation Titanic →directed_by x_1 →gender Male x_1 →birth_date argmin x_1 →place_of_birth x_2

Fig. 6. The illustration of possible parsing targets for a complex question "*What's the birth place of the youngest male director of Titanic?*". x denotes the satisfied entity to be queried and x_1 denotes the intermediate entity included in the query.

aligning predicates and arguments to relations and entities in the KBs [27]. Such methods are sub-optimal for complex questions due to the ontology mismatching problem [28]. Thus it is necessary to leverage the structure of KBs for accurate parsing, where parsing is performed along with the grounding of the KB.

Designing Logic Forms via Pre-Defined Query Templates. To satisfy the compositionality of the complex questions, researchers have developed diverse expressive logic forms as parsing targets. Recalling the topic entities recognized in the preliminary step, Bast et al. [60] started from the topic entities and designed three query templates as the parsing targets. We list these three query templates in Fig. 6. The first two templates return entities which are 1-hop and 2-hop away from the topic entities "*Titanic*". The third template returns entities that are two hops away from the topic entities and constrained by another entity. A follow-up study concentrated on designing templates to answer temporal questions [61]. Although such template-based methods can successfully parse several types of complex questions, it suffers from the limited coverage issue.

Designing Expressive Logic Forms With Flexible Combining Rules. To design more expressive logic forms, Yih et al. [35] proposed query graph as the expressive parsing target. A query graph is a logic form in graph structure which closely matches the KB schemas and is an alternative to an executable SPARQL. It consists of entities, variables, and functions, which correspond to grounded entities mentioned in questions, variables to query and aggregation operations, respectively. As illustrated in Fig. 6, a set of core inference chains [35] starting from the topic entity are first identified. Constraint entities and aggregation operators are further attached to the path chains to make them adapt to more complex questions. Unlike the pre-defined templates, query graphs are not limited to the hop and constraint numbers. They have shown strong capabilities to express complex questions while they are still incapable of dealing with long-tail complex question types. Based on more observations towards the long-tail data samples, follow-up work tried to improve the formulation of query graphs by involving syntactic annotation to enhance the structural complexity of the query graph [53], applying more aggregation operators such as merging, coreference resolution [11] to fit

complex questions. Compared with query templates, logic forms with flexible combining rules could fit into a large variety of complex queries. A more expressive logic form indicates a more robust KBQA system which can handle questions with greater diversity.

4.4 Grounding With Large Search Space

To obtain executable logic forms, KB grounding module instantiates possible logic forms with a KB. As one entity in the KB can be linked to hundreds or even thousands of relations, it is unaffordable to explore and ground all the possible logic forms for a complex question considering both computational resource and time complexity.

Decomposing a Complex Question to Sub-Questions. Instead of enumerating logic forms with a single pass, researchers try to propose methods to generate the complex queries with multiple steps. Zheng et al. [62] proposed to first decompose a complex question into multiple simple questions, where each simple question was parsed into a simple logic form. The final answers are obtained with either the conjunction or composition of the partial logic forms. This *decompose-execute-join* strategy can effectively narrow down the search space. A similar approach was studied by Bhutani et al. [63]. As decomposing questions costs manual efforts, they reduced human annotation and identify the composition plan through an augmented pointer network [81]. The final answers are obtained via conjunction or composition of the answers of decomposed questions.

Expanding a Logic Form by Iteration. Unlike decomposing a complex question to sub-questions, a number of studies adopted the *expand-and-rank* strategies to reduce the search space by expanding the logic forms in an iterative manner. Specifically, they collected all the query graphs that are 1-hop neighborhood of the topic entities as the candidate logic forms at the first iteration. These candidates are ranked based on their semantic similarities with the question. Top-ranked candidates are kept to do further expansion while low-ranked candidates are filtered out. At the following iterations, each top-ranked query graph in the beam is extended, which results in a new set of candidate query graphs that are more complicated. This procedure will repeat until the best query graph is obtained. Chen et al. [46] first utilized the hopwise greedy search to expand the most-likely query graphs. Lan et al. [64] proposed an incremental sequence matching module to iteratively parse the questions without revisiting the generated query graphs at each searching step. Above expansion is conducted in a linear manner, which is only effective in generating multi-hop relations. Lan et al. [65] defined three expansion actions for each iteration, which are extending, connecting, and aggregating to correspond to multi-hop reasoning, constrained relations, and numerical operations, respectively. Examples in Fig. 7 show the different principles of these two strategies.

4.5 Training Under Weak Supervision Signals

To cope with the issue of unlabeled reasoning paths, reinforcement learning (RL) based optimization has been used to maximize the expected reward [49], [67]. However, the
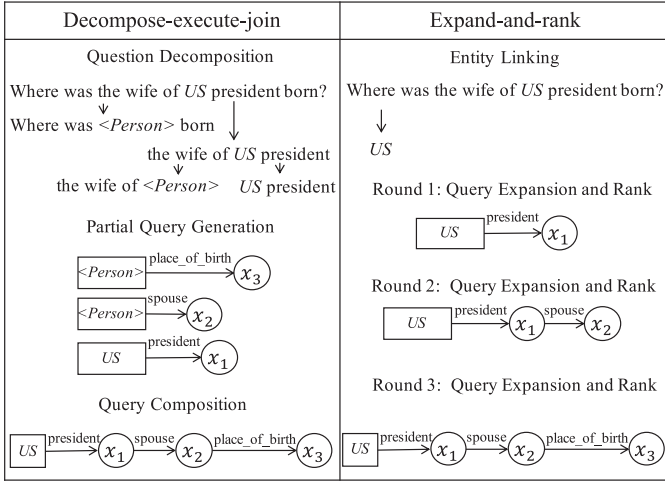
Fig. 7. Illustration of two grounding strategies for a complex question "*Where was the wife of US president born?*".

insufficient training data makes it a challenge to train under weak supervision.

4.5.1 Training With Sparse Reward

Training via RL indicates that SP-based methods can only receive feedback after the execution of the complete parsed logic form. This leads to a long exploration stage with severe sparse positive rewards. To tackle this issue, methods are proposed to augment the final reward or intermediate reward.

Augmenting Final Reward With Enriched Features. Some research work adopted reward shaping strategy for parsing evaluation. Specifically, researchers augment the reward of a logic form by involving more information of answers as the enriched features of the final prediction. Saha et al. [66] rewarded the model with additional feedback when the predicted answers have the same type as ground truth. In this way, even if the predicted answers are not exactly the ground truth, they could also encourage the model to search for the right answer type. This helps to avoid the sparse positive rewards during the exploration stage.

Augmenting Intermediate Reward With Enriched Critics. Besides rewards derived from the final prediction, intermediate rewards during the semantic parsing process may also help address this challenge. Recently, Qiu et al. [67] formulated query graph generation as a hierarchical decision problem, and proposed an option-based hierarchical framework to provide intermediate rewards for low-level agents. Through options over the decision process, the high-level agent sets goals for the low-level agent at intermediate steps. To evaluate whether the intermediate states of the low-level agent meet the goal of the high-level agent, they measured the semantic similarity between the given question and the generated triple. To provide the policy with effective intermediate feedback, Qiu et al. [67] augmented the critic of query graphs with hand-crafted rules.

4.5.2 Dealing With Spurious Reasoning

At the early stage of training, it is difficult to find a logic form with positive rewards. Moreover, random exploration at the early stage easily leads to spurious reasoning,

where logic forms result in correct answers but are semantically incorrect. Therefore, the early supervision of high-quality logic forms could be conducted to speed up the training and prevent models from misguiding of spurious reasoning.

Stabilizing Training Processing With High-Reward Logic Forms. To accelerate and stabilize the training process, Liang et al. [49] proposed to maintain pseudo-gold programs found by an iterative maximum-likelihood training process to bootstrap training. The training process contains two steps: (1) leveraging beam search mechanism to find pseudo-gold programs, and (2) optimizing the model under the supervision of the best program found in history. Hua et al. [69] followed a similar idea to evaluate the generated logic form by comparing it with the high-reward logic forms stored in the memory buffer. To make a trade-off between exploitation and exploration, they proposed the proximity reward and the novelty reward to encourage remembering the past high-reward logic forms and generating new logic forms to alleviate spurious reasoning respectively. Combining such bonus with terminal reward, models can obtain dense feedback along the learning phrase.

5 INFORMATION RETRIEVAL-BASED METHODS

In this section, we summarize the main challenges brought by complex questions for different modules of IR-based methods. The taxonomy of challenges and solutions can be visualized with Fig. 8.

5.1 Overview

The overall procedure typically consists of the modules of retrieval source construction, question representation, graph based reasoning, and answer generation. These modules will encounter different challenges for complex KBQA. First, the retrieval source module extracts a question-specific graph from KBs, which includes both relevant facts and a wide range of noisy facts. Due to unneglectable incompleteness of source KBs [94], the correct reasoning paths may be absent from the extracted graph. The two issues are more likely to occur in the case of complex questions. Second, question representation module understands the question and generates instructions to guide the reasoning process. This step becomes challenging when the question is complicated. After that, reasoning on the graph is conducted through semantic matching. When dealing with complex questions, such methods rank answers through semantic similarity without traceable reasoning in the graph, which hinders reasoning analysis and failure diagnosis.

The following parts illustrate how prior work deals with these challenges and the utilized advanced techniques.

5.2 Reasoning Under Imperfect KB

In general, IR-based methods find answers by conducting reasoning on a graph structure. This graph structure is a question-specific graph extracted from a KB in most cases. However, such question-specific graphs are never perfect, due to incompleteness of KBs and the noisy graph context brought by heuristic graph generation strategy.
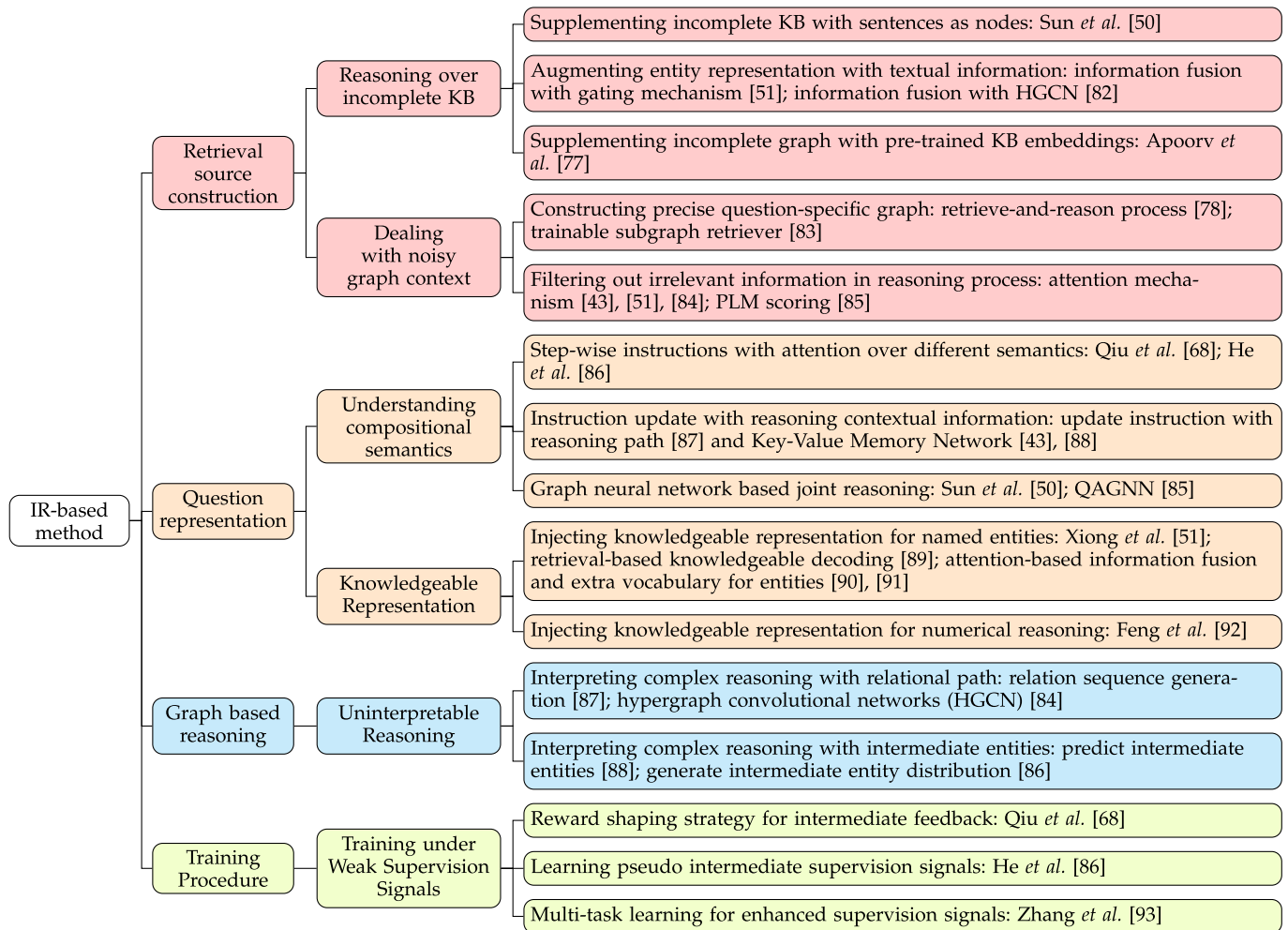
Retrieval source construction
- Reasoning over incomplete KB
  - Supplementing incomplete KB with sentences as nodes: Sun *et al.* [50]
  - Augmenting entity representation with textual information: information fusion with gating mechanism [51]; information fusion with HGCN [82]
  - Supplementing incomplete graph with pre-trained KB embeddings: Apoorv *et al.* [77]
- Dealing with noisy graph context
  - Constructing precise question-specific graph: retrieve-and-reason process [78]; trainable subgraph retriever [83]
  - Filtering out irrelevant information in reasoning process: attention mechanism [43], [51], [84]; PLM scoring [85]

Question representation
- Understanding compositional semantics
  - Step-wise instructions with attention over different semantics: Qiu *et al.* [68]; He *et al.* [86]
  - Instruction update with reasoning contextual information: update instruction with reasoning path [87] and Key-Value Memory Network [43], [88]
  - Graph neural network based joint reasoning: Sun *et al.* [50]; QAGNN [85]
- Knowledgeable Representation
  - Injecting knowledgeable representation for named entities: Xiong *et al.* [51]; retrieval-based knowledgeable decoding [89]; attention-based information fusion and extra vocabulary for entities [90], [91]
  - Injecting knowledgeable representation for numerical reasoning: Feng *et al.* [92]

Graph based reasoning
- Uninterpretable Reasoning
  - Interpreting complex reasoning with relational path: relation sequence generation [87]; hypergraph convolutional networks (HGCN) [84]
  - Interpreting complex reasoning with intermediate entities: predict intermediate entities [88]; generate intermediate entity distribution [86]

Training Procedure
- Training under Weak Supervision Signals
  - Reward shaping strategy for intermediate feedback: Qiu *et al.* [68]
  - Learning pseudo intermediate supervision signals: He *et al.* [86]
  - Multi-task learning for enhanced supervision signals: Zhang *et al.* [93]

IR-based method

Fig. 8. The main content of IR-based methods. The hierarchical structure is arranged with: IR-based method → module → challenge → solution.

5.2.1 Reasoning Over Incomplete KB

It is vital for the question-specific graph to obtain a high recall of correct reasoning paths. Since simple questions only require 1-hop reasoning on the neighborhood of the topic entity in the KB, IR-based methods are less likely to suffer from the inherent incompleteness of KBs [94] when solving simple questions. By contrast, the correct reasoning paths for complex questions are of high probability to be absent from the question-specific graph and it turns out to be a severe issue. To tackle with this challenge, researchers utilize auxiliary information to supplement the knowledge source. We divide the different supplementary methods into three categories and show the core differences in Fig. 9.

Supplementing Incomplete KB With Sentences as Nodes. Intuitively, a large amount of question-relevant text corpus retrieved from Wikipedia can provide a wide range of unstructured knowledge as supplementary evidence. Based on this observation, Sun et al. [50] proposed to complement the graph with extra question-relevant sentences as nodes and reason on the augmented heterogeneous graph (i.e., the left side of Fig. 9). According to the entities mentioned in sentences, they linked them to corresponding entities on the graph and viewed them as nodes.

Augmenting Entity Representation With Textual Information. Instead of directly complementing sentences to the question-

specific graph as nodes, Xiong et al. [51] and Han et al. [82] proposed to fuse extra textual information into the entity representation as the second way (shown in the middle of Fig. 9). Xiong et al. [51] designed a novel conditional gating mechanism to obtain knowledge-aware information of sentences under the guidance of text-linked entity representations extracted with a subgraph reader. Such knowledge-aware information of sentences is further aggregated to enhance the entity representations to complement incomplete KB. Similarly, Han et al. [82] fused textual information of sentences into entity representations. In their settings, every sentence is regarded as a hyperedge connecting all of its involved entities, and a document can be viewed as a hypergraph. Based on hypergraph convolutional networks (HGCN) [95], they encoded the sentences in the document and fused sentence representations into sentence-linked entity representations.

Supplementing Incomplete Graph With Pre-Trained KB Embeddings. In knowledge base completion (KBC) task, knowledge base embeddings have been adopted to alleviate the sparsity of KB by performing missing link prediction. Inspired by that, Apoorv et al. [77] utilized pre-trained knowledge base embeddings to address the incomplete KB issue as shown at the right side of Fig. 9. Specifically, they pre-trained KB embeddings (i.e., entity and relation embeddings) with ComplEX [96] approach and predicted the
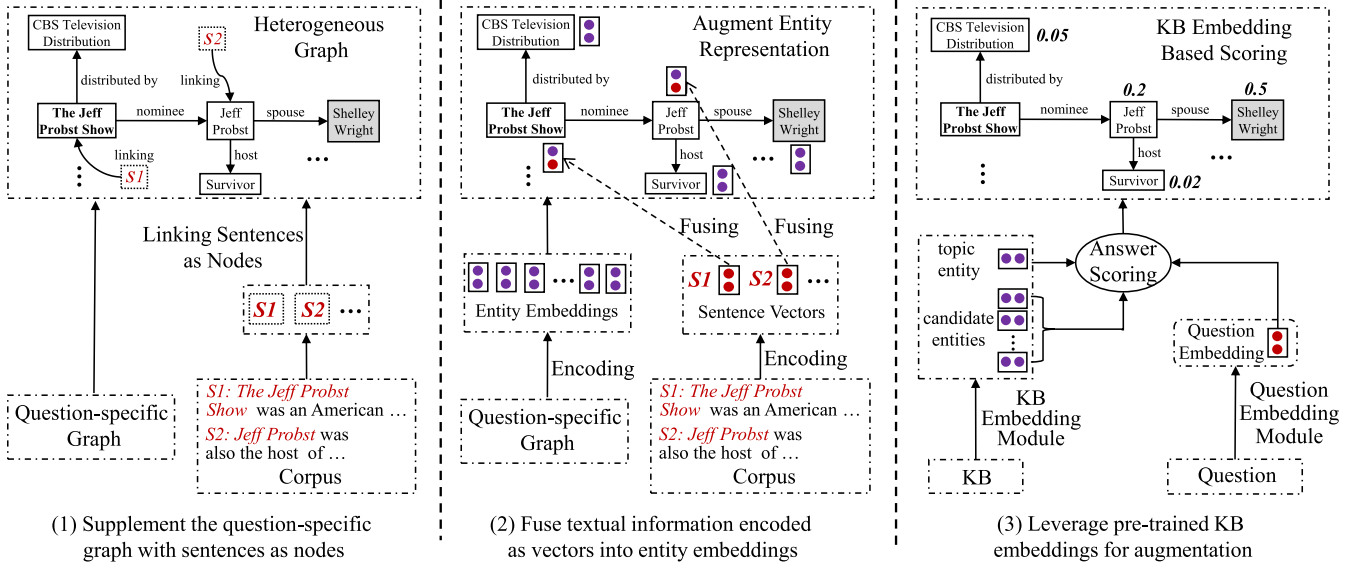
Fig. 9. Illustration of three categories of methods to supplement the incomplete KB. All subfigures are drawn in a bottom-up style, where the input is placed at the bottom and supplemented graph is placed on the top. The topic entity and the answer entity are shown in the bold font and shaded box respectively.

answer via a triple scoring function taking the triples in the format of (topic entity, question, answer entity) as inputs. To make questions fit into original ComplEX scoring function, they map Roberta [97] embeddings of question into the complex space of same dimension. By leveraging the pre-trained knowledge from global KBs, they implicitly complemented the incomplete question-specific graph.

5.2.2 Dealing With Noisy Graph Context

Since question-specific graphs are always constructed with heuristics [50], it may introduce redundant and even question-irrelevant noisy graph context (both entities and sentence nodes). Compared with simple questions which require only 1-hop reasoning, the question-specific graphs constructed for complex questions are more likely to involve noisy graph context. Reasoning over such noisy graphs poses a great challenge for complex questions, meanwhile it also reduces the efficiency of model training.

Constructing Precise Question-Specific Graph. An intuitive idea is to construct a relatively small and precise graph for downstream reasoning. To achieve this goal, Sun et al. [78] proposed to build the heterogeneous graph with an iterative retrieve-and-reason process under the supervision of shortest paths between the topic entities and answer entities. In recent work, Zhang et al. [83] proposed a trainable subgraph retriever (SR) which retrieves relevant relational paths for subsequent reasoning. And their experimental results proved such precise graphs can bring substantial performance gains for IR-based methods.

Filtering Out Irrelevant Information in Reasoning Process. Besides constructing small and precise graphs for subsequent reasoning, some research work proposed to filter irrelevant information out along the reasoning process. Attention mechanisms, which are effective in eliminating irrelevant features, have been adopted by existing IR-based methods [43], [51], [84] to reserve relevant information during the reasoning process. Similarly, Yasunaga et al. [85] adopted pre-trained language model scoring of each node conditioned on question

answering context as relevance scores to guide subsequent reasoning process.

5.3 Understanding Complex Semantics

Understanding complex questions is the prerequisite for subsequent reasoning. However, complex questions contain compositional semantics and require specific knowledge (e.g., named entities, ordinal reasoning) to answer. Due to such intrinsic properties of complex questions, methods designed for simple question understanding may be not fit for complex questions.

5.3.1 Understanding Compositional Semantics

IR-based methods usually generate initial question representation q by directly encoding questions as low-dimensional vectors through neural networks (e.g., LSTM and GRU). Static reasoning instruction (e.g., final hidden states of q) obtained through above approach can not effectively represent the compositional semantics of complex questions, which poses challenges to guide the reasoning over the question-specific graph. In order to comprehensively understand questions, some studies dynamically update the reasoning instruction during the reasoning process.

Step-wise Instructions With Attention Over Different Semantics. To make the reasoning models aware of the reasoning step, Qiu et al. [68] proposed to learn a step-aware representation through transforming initial question representation q with a single-layer perceptron. After obtaining step-aware question representation, attention mechanism is further incorporated to select useful information to generate instruction vectors. Similarly, He et al. [86] proposed to focus on different parts of the question with dynamic attention mechanism. Based on both step-aware question representation and previous reasoning instruction $i^{(k-1)}$, they generated attention distribution over tokens of the question and updated the instruction vector.

Instruction Update With Reasoning Contextual Information. Besides explicitly recording the analyzed part of the question

via attention, some other work proposed to update the instruction with information retrieved along the reasoning process. A typical example is generating explicit reasoning paths and updating instruction with generated paths. Zhou et al. [87] designed a model that takes the current reasoning instruction $i^{(k)}$ as the input, and then predicts the intermediate relation $r^{(k)}$ from all relations in KB. After obtaining the predicted relation, the model updated the instruction vector as: $i^{(k+1)} = i^{(k)} - r^{(k)}$, where the subtraction is meant to omit the analyzed information from the question. Thus, the updated reasoning instruction can hold unanalyzed parts of the question in the subsequent reasoning process.

Instead of generating explicit reasoning paths, Xu et al. [88] and Miller et al. [43] employed key-value memory network to achieve similar dynamic instruction update. Specifically, they first included all KB facts that contain one of the topic entities as subject into the memory. Then, they indexed the keys and values in the key-value memory, where keys are (subject, relation) pairs and values are corresponding object entities. A key addressing process is conducted to find the most suitable key and the corresponding value for the instruction. With the addressed key and value, they concatenated their representations with the previous step reasoning instruction $i^{(k)}$ and performed a linear transformation to obtain the updated reasoning instruction $i^{(k+1)}$ to guide the next hop reasoning. In this way, the reasoning instruction will be updated over the memory.

Graph neural Network Based Joint Reasoning. Besides instruction update, another line of research addresses such compositional semantics with graph neural network (GNN) based reasoning. Sun et al. [50] proposed a GNN-based model Graft-Net to reason complex questions over heterogeneous information sources. Through iterative GNN reasoning steps, the entity representations and reasoning instruction get updated in turn. The reasoning instruction conveys the knowledge of the topic entity which is dynamically updated over the reasoning process. Despite iterative update of reasoning instruction and graph neural network, Yasunaga et al. proposed [85] QAGNN model which reasoned complex questions with single graph neural network based joint reasoning. They constructed the question-specific graph with an extra question-answering context node which connects with all other nodes in the graph. All nodes are uniformly encoded with pre-trained language models (PLMs) as initial representation, and get updated along with graph neural network reasoning.

5.3.2 Knowledgeable Representation

Apart from compositional semantics, complex questions may also contain knowledge-intensive tokens or phrases (e.g., named entities, ordinal constraint), which hinders natural language understanding for text-based semantic understanding. Besides question text, external knowledge is taken as input to help understand these complex questions.

Injecting Knowledgeable Representation for Named Entities. In the natural language questions, the topic entities are always named entities which are not informative enough for understanding. To cope with such named entities, some existing work proposed to inject more informative representations obtained from knowledge bases. As a typical example, Xiong et al. [51] proposed to reformulate query representation in latent space with knowledge representation learned from the graph context of topic entities. Through an ablation study, they verified the effectiveness of injecting such knowledgeable representation into question representation. Similar ideas were also adopted in knowledge-enhanced language model pre-training [98], [99].

While natural answers can be generated from popular seq2seq text generation framework, it is still hard to directly generate the named entities from token vocabulary. To address this gap, He et al. [89] first proposed a copying and retrieving mechanism to generate the natural answers from extra vocabulary for question tokens and entities in the question-specific graph. Similarly, Yin et al. [90] and Fu et al. [91] fed relational facts into structured memory slots, which served as extra vocabulary to generate named entities, and generate knowledgeable representation with attention-based information fusion.

Injecting Knowledgeable Representation for Numerical Reasoning. While multiple solutions are proposed to conduct multi-hop reasoning, little attention is paid to solving complex questions with numerical operations. To empower IR-based methods with numerical reasoning capability, Feng et al. [92] proposed to encode numerical properties (i.e., the magnitude and ordinal properties of numbers) into entity representations. First, they manually defined a list of ordinal determiners (e.g., first, largest) to detect ordinal constrained questions. For these detected questions, they enrich their question-specific graphs with extra numerical attribute triplets. Encoding these numerical attribute triplets with pretrained number encoding modules, extra number embeddings can be used as model-agnostic plugins to conduct numerical reasoning for IR-based methods.

5.4 Uninterpretable Reasoning

Since the complex questions usually query multiple facts in sequence, the system is supposed to accurately predict answers over the graph based on a traceable reasoning process. While neural networks are powerful, blackbox style of the reasoning module makes the reasoning process less interpretable and hard to incorporate user interaction for further improvement. To derive a more interpretable reasoning process, the reasoning is performed with a multistep intermediate prediction. Along the reasoning process, the KBQA model generates a series of reasoning status $\{s^{(k)}, k = 1, ..., n\}$. While the final status is leveraged to generate the answer prediction, the intermediate status may help generate intermediate predictions (i.e., matched relations or entities) for better interpretability. More importantly, intermediate predictions make detecting spurious or error reasoning easier with user interaction.

Interpreting Complex Reasoning With Relational Path. Existing studies adopted different designs of reasoning status and reasoning modules to interpret the reasoning process. Specifically, Zhou et al. [87] formulated the multi-hop reasoning process as relation sequence generation and represented reasoning status using a vector. For each step, instruction vector and status vector are matched with relation candidates to generate a probability distribution over all relations in KB. And weighted relation representation is then leveraged to update the status. By repeating this

process, the model can achieve an interpretable reasoning process. Inspired by above work, Han et al. [84] proposed an interpretable model based on hypergraph convolutional networks (HGCN) to predict relation paths for explanation. They constructed a dense hypergraph by pinpointing a group of entities connected via same relation, which simulated human's hopwise relational reasoning. To train these two models, gold relation paths are leveraged. However, gold relation path annotations are unavailable in most cases, which makes their methods inapplicable to general datasets.

Interpreting Complex Reasoning With Intermediate Entities. Apart from relation paths, some research work predicted question-relevant entities at intermediate steps to explain multi-hop reasoning process. Xu et al. [88] elaborately adopted key-value memory network to achieve a traceable reasoning process. In their work, status $s^{(k)}$ is defined as the weighted sum of value representation, the weight of which is derived from key-instruction matching. To predict intermediate entities, their model followed traditional IR-based methods to score candidates given query $s^{(k)} + s^{(k-1)}$. As spurious long paths may connect topic entities with answer entities in KB, during training, they proposed to supervise intermediate entity prediction with the final answers. Such objective encourages the model to generate shortest reasoning path. Besides explicitly generating intermediate entities, He et al. [86] proposed to generate intermediate entity distribution to indicate the reasoning process. Their experimental results also showed that such intermediate supervision signals can effectively reduce spurious reasoning.

5.5 Training Under Weak Supervision Signals

Similar to the SP-based methods, it is difficult for IR-based methods to reason the correct answers without any annotations at intermediate steps, since the model cannot receive any feedback until the end of reasoning. It is found that this case may lead to spurious reasoning [86]. Due to the lack of intermediate state supervision signals, the reward obtained from spurious reasoning may mislead the model.

Reward Shaping Strategy for Intermediate Feedback. To train model under weak supervision signals, Qiu et al. [68] formulated multi-hop reasoning process over KBs as a process of expanding the reasoning path on the graph. Based on the encoded decision history, the policy network leveraged attention mechanism to focus on the unique impact of different parts of a given question over triple selection. To alleviate the delayed and sparse reward problem caused by weak supervision signals, they adopted reward shaping strategy to evaluate reasoning paths and provide intermediate rewards. Specifically, they utilized semantic similarity between the question and the relation path to evaluate reasoning status at intermediate steps.

Learning Pseudo Intermediate Supervision Signals. Besides evaluating the reasoning status at intermediate steps, a more intuitive idea is to infer pseudo intermediate status and augment model training with such inferred signals. Inspired by bidirectional search algorithm on graph, He et al. [86] proposed to learn and augment intermediate supervision signals with bidirectional reasoning process. Taking entity distribution as suitable supervision signals at

intermediate steps, they proposed to learn and leverage such signals under teacher-student framework.

Multi-task Learning for Enhanced Supervision Signals. While most of existing work focused on enhancing the supervision signals at intermediate steps, few work paid attention to the entity linking step. Most of existing work utilized off-the-shelf tools to locate the topic entity in question, causing error propagation. In order to accurately locate the topic entity without annotations, Zhang et al. [93] proposed to train entity linking module through a variational learning algorithm which jointly models topic entity recognition and subsequent reasoning over KBs. They also applied the REINFORCE algorithm with variance reduction technique to make the system end-to-end trainable.

6 PLM APPLICATIONS ON COMPLEX KBQA

Unsupervised pre-training language models on large text corpora then fine-tuning pre-trained language models (PLMs) on downstream tasks has become a popular paradigm for natural language processing [100]. Furthermore, due to the powerful performance obtained from broad data at scale and capability to serve a wide range of downstream tasks, PLMs are recognized as "foundation models" [101] for many tasks, including complex KBQA task. Therefore, some recent SP-based and IR-based methods have widely incorporated PLMs in their pipelines.

For SP-based methods, PLMs are always used to simultaneously optimize the trainable modules (i.e., question understanding, logical parsing, KB grounding), which facilitate the generation of executable programs (e.g., SPARQL) in a seq2-seq framework. With such a unified paradigm, transferable knowledge across tasks can be leveraged to mitigate the data sparsity issue in low-resource scenarios. For IR-based methods, PLMs help with precise source construction and further enhance the unified reasoning ability. On one hand, PLMs provide powerful representation ability to retrieve semantically relevant information from KBs. On the other hand, PLMs can help unify the representation of questions and KBs, which contributes to the reasoning capability.

6.1 PLM for SP-Based Methods

Equipped with powerful PLMs, the logic form generation modules benefit from their strong generation and understanding capabilities obtained via unsupervised pre-training. Under a unified seq2seq generation framework, PLMs provide transferable knowledge to help effective model training with limited data.

PLM for Enhancing the Logic Form Generation. To get the executable programs (e.g., SPARQL), traditional SP-based methods parse the question to a logic form and instantiate it via KB grounding. This process can be well formalized under *knowledge-enhanced text generation* [102] framework (i.e., from user requests to executable programs). Therefore, some work [76], [103] leveraged the PLMs, which are typically neural encoder-decoder models, to directly generate the executable programs given the question and other related KB information. To get the related KB information input, Das et al. [103] retrieved similar cases from a cases-memory where each case is a pair of question and its gold executable program. Ye et al. [76] directly retrieved the top-

k relevant logic forms from the candidates enumerated by searching over the KB based on predefined rules. Substantial improvement in model performance has proved the effectiveness of such usages of PLMs.

PLM for Low-Resource Training. The robust and transferable natural language understanding capability obtained from PLMs empowers KBQA methods to conquer the unaffordable need for training data in low-resource scenarios. In a recent study, Shi et al. [104] fine-tuned a pre-trained sequence-to-sequence model on KQA Pro dataset [104] to generate SPARQLs and programs. While no external knowledge was incorporated to enhance the generation, the BART-based generator reached near-human performance and showed robustness to sparse training data.

Besides, similar to complex KBQA, there are a series of structure parsing tasks (e.g., Text2SQL, tabular question answering, semantic parsing over database) can be formed as the *knowledge-enhanced text generation* framework. Motivated by this, Xie et al. [105] proposed structured knowledge grounding (SKG) to unify a series of structure parsing tasks and achieved (near) state-of-the-art performance with PLM model T5 [106] on 21 benchmarks. With this PLM-based general-purpose approach, the challenges of precise semantic parsing in low-resource tasks can be solved by knowledge sharing and cross-task generalization.

6.2 PLM for IR-Based Methods

With the strong representation ability of PLMs, we can augment the retrieval of the question-specific graphs and relieve the incompleteness of KBs during retrieval source construction. Besides, PLMs provide a unified way to model unstructured text and structured KB information in a unified semantic space, which in turn improves question-specific graph reasoning.

PLM for Augmenting the Source Construction. To cover the answers as complete as possible, traditional heuristic-based methods like personalized pagerank [107] would recall a large and noisy question-specific graph [83], which hinders subsequent reasoning. Therefore, Zhang et al. [83] trained a path retriever based on PLMs to retrieve the hop-wise question-related relations. At each step, the retriever ranks top-k relations based on the question and the selected relations of the previous step. This method successfully filters noisy graph context out and keeps a high recall of expected reasoning paths for answers.

Besides building up a precise source construction module, PLMs also provide the potential to mitigate the incompleteness of KB. PLMs have shown their capabilities to answer "fill-in-the-blank" cloze statements [108], [109], [110], which indicates that they may learn relational knowledge from unsupervised pre-training. Such key findings indicate that PLMs have great potential in serving as knowledge sources for question answering, which may play a complementary role for source construction with incomplete KB.

PLM for Precise and Unified Reasoning. Attracted by the powerful pre-trained language models, some researchers made adaptions to complex reasoning over graph structure for further involvement of PLM. While traditional reasoning over KB relies on the embeddings learned for entities and relations, such embeddings may fail to identify relevant

parts of question answering context. To filter out noisy graph context in the retrieved subgraph, Yasunaga et al. [85] adopted PLM similarity scores to identify relevant knowledge given the question. For further joint reasoning of question answering context (i.e., question-answer sequence) and knowledge graph, the node representations in the retrieved subgraph were initialized with PLM encoding of the concatenated sequence of question, answer, and node surface name. With the augmentation of PLMs, the GNN model gets substantial performance improvement [85].

7 EVALUATION AND RESOURCE

In this section, we first introduce the evaluation protocol of KBQA systems. And then, we summarize some popular benchmarks for KBQA. At last, for tracking the research progress conveniently, we make a leaderboard for these benchmark datasets, which contains the evaluation results and resource links of the corresponding publications. We also attach a companion page[2] for a comprehensive collection of the relevant publications, open-source codes, resources, and tools for KBQA.

7.1 Evaluation Protocol

In order to comprehensively evaluate KBQA systems, effective measurements from multiple aspects should be taken into consideration. Considering the goals to achieve, we categorize the measurement into three aspects: reliability, robustness, and system-user interaction [52].

Reliability: For each question, there is an answer set (one or multiple elements) as the ground truth. The KBQA system usually predicts entities with the top confidence score to form the answer set. If an answer predicted by the KBQA system exists in the answer set, it is a correct prediction. In previous studies [35], [49], [53], there are some classical evaluation metrics such as Precision, Recall, F_1, and Hits@1. For a question q, its Precision indicates the ratio of the correct predictions over all the predicted answers. It is formally defined as:

$$\text{Precision} = \frac{|\mathcal{A}_q \cap \tilde{\mathcal{A}}_q|}{|\tilde{\mathcal{A}}_q|},$$

where $\tilde{\mathcal{A}}_q$ is the predicted answers, and $\mathcal{A}_q$ is the ground truth. Recall is the ratio of the correct predictions over all the ground truth. It is computed as:

$$\text{Recall} = \frac{|\mathcal{A}_q \cap \tilde{\mathcal{A}}_q|}{|\mathcal{A}_q|}.$$

Ideally, we expect that the KBQA system has higher Precision and Recall simultaneously. Thus F_1 score is most commonly used to give a comprehensive evaluation:

$$F_1 = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}.$$

Some other methods [43], [50], [51], [86] use Hits@1 to assess the fraction that the correct prediction rank higher than other entities. It is computed as:

2. https://github.com/RUCAIBox/Awesome-KBQA

TABLE 1
Several KBQA Benchmark Datasets Involving Complex Questions

Datasets	KB	Size	LF	CO	NL	NU	SP-based			IR-based		
							Top-1	Top-2	Top-3	Top-1	Top-2	Top-3
WebQuestions [31]	Freebase	5,810	No	Yes	No	Yes	62.9$^\heartsuit$ [62]	54.8$^\heartsuit$ [67]	54.6$^\heartsuit$ [88]	48.6$^\triangle$ [88]	-	-
ComplexQuestions [75]	Freebase	2,100	No	Yes	No	Yes	71.0$^\heartsuit$ [62]	54.3$^\heartsuit$ [11]	45.0$^\heartsuit$ [67]	-	-	-
WebQuestionsSP [34]	Freebase	4,737	Yes	Yes	Yes	Yes	76.5$^\heartsuit$ [111]	75.0$^\heartsuit$ [112]	74.0$^\heartsuit$ [65]	74.3$^\triangle$ [86]	71.4$^\triangle$ [113]	69.5$^\triangle$ [83]
ComplexWebQuestions[81]	Freebase	34,689	Yes	Yes	Yes	Yes	70.4$^\triangle$ [103]	44.1$^\triangle$ [114]	39.4$^\triangle$ [115]	53.9$^\triangle$ [86]	45.9$^\triangle$ [78]	-
QALD series [116]	DBpedia	-		Yes	Yes	Yes	Yes	-	-	-	-	-	-
LC-QuAD [117]	DBpedia	5,000	Yes	Yes	Yes	Yes	75.0$^\heartsuit$ [58]	74.8$^\heartsuit$ [59]	71.8$^\heartsuit$ [118]	33.0$^\heartsuit$ [119]	-	-
LC-QuAD 2.0 [120]	DBpedia & Wikidata	30,000	Yes	Yes	Yes	Yes	59.3$^\heartsuit$ [121]	52.6$^\heartsuit$ [65]	44.9$^\heartsuit$ [59]	-	-	-
MetaQA Vanilla [93]	WikiMovies	400k	No	No	No	No	99.6$^\triangle$ [64]	-	-	100.0$^\triangle$ [113]	99.3$^\triangle$ [122]	98.9$^\triangle$ [86]
CFQ [123]	Freebase	239,357	Yes	Yes	No	No	67.3$^\triangle$ [124]	18.9$^\triangle$ [123]	-	-	-	-
GrailQA [13]	Freebase	64,331	Yes	Yes	Yes	Yes	74.4$^\heartsuit$ [76]	65.3$^\heartsuit$ [125]	58.0$^\heartsuit$ [13]	-	-	-
KQA Pro [104]	Wikidata	117,970	Yes	Yes	Yes	Yes	89.7$^\heartsuit$ [104]	-	-	-	-	-

"LF" denotes whether the dataset provides Logic Forms like SPARQL, "CO" denotes whether the dataset contains questions with COnstraints, "NL" represents whether the dataset incorporates crowd workers to rewrite questions in Natural Language and "NU" denotes whether the dataset contains the questions which require NUmerical operations. Typically, SP-based methods adopt F_1 score as evaluation metric, while IR-based methods adopt $Hits@1$ (accuracy) as evaluation metric. The symbol of $\triangle$ and $\heartsuit$ indicates evaluation metric of $Hits@1$ (accuracy) and F_1 score respectively.

$$\text{Hits@1} = \mathbb{I}(\tilde{a}_q \in \mathcal{A}_q),$$

where $\tilde{a}_q$ is the top 1 prediction in $\tilde{\mathcal{A}}_q$.

Robustness: Practical KBQA models are supposed to generalize to out-of-distribution questions at test time [13]. However, current KBQA datasets are mostly generated based on templates and lack of diversity [52]. And, the scale of training datasets is limited by the expensive labeling cost. Furthermore, the training data for KBQA system may hardly cover all possible user queries due to broad coverage and combinatorial explosion of queries. To promote the robustness of KBQA models, Gu et al. [13] proposed three levels of generalization (i.e., *i.i.d.*, *compositional*, and *zero-shot*) and released a large-scale KBQA dataset GrailQA to support further research. At a basic level, KBQA models are assumed to be trained and tested with questions drawn from the same distribution, which is what most existing studies focus on. In addition to that, robust KBQA models can generalize to novel compositions of seen schema items (e.g., relations and entity types). To achieve better generalization and serve users, robust KBQA models are supposed to handle questions whose schema items or domains are not covered in the training stage.

System-User Interaction: While most of the current studies pay much attention to offline evaluation, the interaction between users and KBQA systems is neglected. On one hand, in the search scenarios, a user-friendly interface and acceptable response time should be taken into consideration. To evaluate this, the feedback of users should be collected and the efficiency of the system should be judged. On the other hand, users' search intents may be easily misunderstood by systems if only a single round service is provided. Therefore, it is important to evaluate the interaction capability of a KBQA system. For example, to check whether they could ask clarification questions to disambiguate users' queries and whether they could respond to the error reported from the users [54], [126]. So far, there is a lack of quantitative measurement of system-user interaction capability of the system, but human evaluation can be regarded as an efficient and comprehensive way.

7.2 Datasets And Leaderboard

Datasets. Over the decades, much effort has been devoted to constructing datasets for complex KBQA. We list the representative complex KBQA datasets for multiple popular KBs (e.g., Freebase, DBpedia, Wikidata, and WikiMovies) in Table 1. In order to serve realistic applications, these datasets typically contain questions which require multiple KB facts to reason. Moreover, they might include numerical operations (e.g., counting and ranking operations for comparative and superlative questions, respectively) and constraints (e.g., entity and temporal keywords), which further increase the difficulty in reasoning the answers from KBs.

Overall, these datasets are constructed with the following steps: Given a topic entity in a KB as a question subject, simple questions are first created with diverse templates. Based on simple questions and neighborhood of topic entities in a KB, complex questions are further generated with predefined composition templates, and another work [104] also generates executable logic forms with templates. Meanwhile, answers are extracted with corresponding rules. In some cases, crowd workers are hired to paraphrase the canonical questions, and refine the generated logic forms, making the question expressions more diverse and fluent.

Leaderboard. In order to show the latest research progress in these KBQA benchmark datasets, we offer a leaderboard including the top-3 KBQA systems with respect to IR-based methods and SP-based methods. To give a fair comparison, the results are selected following three principles: 1) If one dataset has an official leaderboard, we only refer to the public results listed on the leaderboard. 2) otherwise, we select the top-3 results from the published papers accepted formally by conferences or journals before March 2022. 3) besides, we keep the experimental setup of each dataset consistent with the other datasets. Exceptionally, we do not report the results on the QALD series for easy display because they have multiple different versions, and we only report the 3-hop split of MetaQA Vanilla because it is more challenging than 1-hop and 2-hop splits. For LC-QuAD 2.0, we select the results reported in Zou et al. [121]. We leave the slots blank if there is no sufficient result according to the above principles. For more comprehensive evaluation of

KBQA methods on all benchmarks, please refer to our companion page.

Analysis and Discussions. Based on Table 1, we have following observations: (1) Both SP-based and IR-based methods are developed to handle complex KBQA challenges and there is no absolute agreement on which category is better. (2) While SP-based methods cover most benchmarks, IR-based methods focus on benchmarks which are mainly composed of multi-hop questions. The reason why SP-based methods are more commonly used in answering complex questions may be that SP-based methods generate flexible and expressive logic forms which are capable of covering all types of questions (e.g., boolean, comparative). (3) We also observe for each category, the methods achieving outstanding performance are usually equipped with advanced techniques. The SP-based methods on the leaderboard leverage powerful question encoders (e.g., PLMs) to help understand the questions and expressive logic forms to help parse complex queries. For IR-based methods, most SOTA methods adopt the step-wise dynamic instruction in question representation module and conduct multi-step reasoning with relational path modeling or GNN-based reasoning.

8 RECENT TRENDS

In this section, we discuss several promising future directions for complex KBQA task:

Evolutionary KBQA Systems. Existing KBQA systems are usually trained offline with specific datasets and then deployed online to handle user queries. However, most existing KBQA systems neglect to learn from failure cases or unseen question templates received after deployment. At the same time, most existing KBQA systems fail to catch up with the rapid growth of world knowledge and answer new questions. Therefore, a practical KBQA system is imperative to get performance improvement over time after online deployment. Online user interaction may provide deployed KBQA systems an opportunity to get further improvement. Based on this motivation, some work leverages user interaction to rectify answers generated by the KBQA system and further improve itself. With user feedback, Abujabal et al. [54] presented a continuous learning framework to learn new templates that capture previously unseen syntactic structures. Besides increasing the model's template bank, user feedback can also be leveraged to clarify ambiguous questions (e.g., ambiguous phrases or ambiguous entities) [127]. Above methods provide an initial exploration to construct evolutionary KBQA systems with user feedback. Such approaches are effective and practical (i.e., acceptable user cognitive burden and running cost), which may serve industrial needs. Due to the wide applications of KBQA systems, more work and designs of user interaction with KBQA systems are in urgent need.

Robust KBQA Systems. Existing studies on KBQA have conducted with the ideal hypothesis, where training data is sufficient and its distribution is identical with test set. However, this may not be desirable in practice due to data insufficiency and potential data distributional biases. To train robust KBQA systems in low-resource scenarios, meta-learning techniques [128] and knowledge transfer from high-resource scenarios [129] have been explored. We also highlighted the potential impact of PLMs in low-resource training and cross-task generalization (see Sec 6). As manual annotations for KBQA systems are expensive and labor-intensive, there is a need for more studies about training robust KBQA systems in low-resource scenarios. Meanwhile, although existing methods usually hold the i.i.d. assumption, they may easily fail to deal with out-of-distribution (OOD) issue [130], [131], [132] on KBQA. With a systematic evaluation of GrailQA [13] dataset, Gu et al. [13] pointed out that existing baseline methods are vulnerable to compositional challenges. To promote a higher level of robustness, researchers may gain more insights by addressing the three levels of generalization (i.e., *i.i.d.*, *compositional*, and *zero-shot*) proposed by Gu et al. [13]. There is few work investigating robustness on complex KBQA task. It is still an open question of building robust KBQA systems with stronger generalization capability.

Conversational KBQA Systems. Recent decades have seen the rapid development of AI-driven applications (e.g., search engines and personal assistants) which are supposed to answer factoid questions. As users typically ask follow-up questions to explore a topic, deployed models are supposed to handle KBQA task in a conversational manner. In initial explorations of conversational KBQA, several pieces of work [133], [134], [135], [136] focused on ambiguity and difficulties brought by coreference and ellipsis phenomena. To track the focus of conversational KBQA, Lan et al. [114] proposed to model the flow of the focus via an entity transition graph. For a comprehensive understanding of conversation context, Plepi et al. [135], [137] leveraged Transformer [138] architecture taking as input of the previous turn of conversation history. While these studies addressed some challenges for conversational KBQA, it is still far from achieving human-level performance. More critical challenges should be identified and solved in the following research. Up to now, conversational KBQA is quite a new and challenging task, it may play an important role in future search engines and intelligent personal assistants.

Neural Symbolic KBQA Systems. While some recent work[49], [50] has proposed to equip KBQA systems with neural symbolic reasoning (NSR) techniques, the promising potential of such powerful paradigm has not been explored thoroughly. For example, while neural networks have been proved to be effective in conduct multi-hop reasoning on KBs[50], [85], [86], such neural modules can not explicitly consider logical operations (e.g., numerical, boolean). To mitigate such drawbacks while keeping power of neural networks in reasoning, we can introduce a symbolic module coupled with existing neural reasoning modules [139]. Several practices in neural programming[140], [141] have demonstrated this can be effective in empowering blackbox neural networks with mathematical and logical reasoning capabilities. In general, researchers appreciate the interpretability of SP-based methods (i.e., generating logic forms according to grammar rules) and the powerfulness of IR-based methods (i.e., precise reasoning on subgraph with neural networks). As discussed in Section 3.4, both SP-based and IR-based methods can be unified paradigm — neural symbolic reasoning. Thus, NSR provides a potential way to unify the two categories of methods and gather their advantages, which deserves further research.

More General Knowledge Bases. Due to KB incompleteness, researchers incorporated extra information (such as text [142],

images [143], and human interactions [144]) to complement the knowledge bases, which would further address the information need of complex KBQA task. As text corpus is rich in semantics and easy to collect, researchers are fascinated by the idea of extracting knowledge from text corpus and answering questions with extracted knowledge. Researchers have explored various forms of knowledge obtained from text corpus, such as traditional relational triplet [145], virtual knowledge base (VKB) [146] which is stored as key-value memory, and PLMs as implicit knowledge base [108]. With these elaborate designs, more flexible and complementary knowledge can be obtained to solve complex KBQA tasks. Recently, a neglectable trend is to unify similar tasks with general architecture and achieve cross-task knowledge transfer [105]. In the future, more related tasks may be explored with a general definition of KBs, such as synthetic, multilingual, and multi-modal KBs.

9 CONCLUSION

This survey attempted to provide an overview of typical challenges and corresponding solutions on complex KBQA. Particularly, task-related preliminary knowledge and traditional methods were first introduced. Then, we summarized the widely employed semantic parsing-based methods and information retrieval-based methods. We specified the challenges for these two categories of methods based on their working mechanism, and explicated the proposed solutions. Along with the taxonomy, we provide technical summaries to shed light on the applied advanced techniques for these two categories. Most of existing complex KBQA methods are generally summarized into these two categories. Please be aware that there are some other methods like [81], which focus on question decomposition instead of KB based reasoning or logic form generation. In the last section, we investigated several research trends related to complex KBQA task and emphasized many challenges are still open and under-explored. We believe that complex KBQA will continue to be an active and promising research area with wide applications, such as natural language understanding, compositional generalization, multi-hop reasoning. We hope this survey will give a comprehensive picture of cutting-edge methods for complex KBQA and encourage further contributions in this field.

REFERENCES

[1] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: A collaboratively created graph database for structuring human knowledge," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2008, pp. 1247–1250.
[2] J. Lehmann et al., "DBpedia–A large-scale, multilingual knowledge base extracted from wikipedia," *Semantic Web*, vol. 6, pp. 167–195, 2015.
[3] T. P. Tanon, D. Vrandecic, S. Schaffert, T. Steiner, and L. Pintscher, "From freebase to wikidata: The great migration," in *Proc. Int. World Wide Web Conf.*, 2016, pp. 1419–1428.
[4] F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago: A core of semantic knowledge," in *Proc. Int. World Wide Web Conf.*, 2007, pp. 697–706.
[5] H. Zhou, T. Young, M. Huang, H. Zhao, J. Xu, and X. Zhu, "Commonsense knowledge aware conversation generation with graph attention," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 4623–4629.
[6] A. Bordes, N. Usunier, S. Chopra, and J. Weston, "Large-scale simple question answering with memory networks," 2015, *arXiv:1506.02075*.
[7] L. Dong, F. Wei, M. Zhou, and K. Xu, "Question answering over freebase with multi-column convolutional neural networks," in *Proc. Assoc. Comput. Linguistics*, 2015, pp. 260–269.
[8] M. Yu, W. Yin, K. S. Hasan, C. N. dos Santos, B. Xiang, and B. Zhou, "Improved neural relation detection for knowledge base question answering," in *Proc. Assoc. Comput. Linguistics*, 2017, pp. 571–581.
[9] M. Petrochuk and L. Zettlemoyer, "Simplequestions nearly solved: A new upperbound and baseline approach," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2018, pp. 554–558.
[10] S. Hu, L. Zou, J. X. Yu, H. Wang, and D. Zhao, "Answering natural language questions by subgraph matching over knowledge graphs," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 5, pp. 824–837, May 2018.
[11] S. Hu, L. Zou, and X. Zhang, "A state-transition framework to answer complex questions over knowledge base," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2018, pp. 2098–2108.
[12] K. Luo, F. Lin, X. Luo, and K. Q. Zhu, "Knowledge base question answering via encoding of complex query graphs," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2018, pp. 2185–2194.
[13] Y. Gu et al., "Beyond I.I.D.: Three levels of generalization for question answering on knowledge bases," in *Proc. Int. World Wide Web Conf.*, 2021, pp. 3477–3488.
[14] P. Wu, X. Zhang, and Z. Feng, "A survey of question answering over knowledge base," in *Proc. China Conf. Knowl. Graph Semantic Comput.*, 2019, pp. 86–97.
[15] N. Chakraborty, D. Lukovnikov, G. Maheshwari, P. Trivedi, J. Lehmann, and A. Fischer, "Introduction to neural network based approaches for question answering over knowledge graphs," 2019, *arXiv:1907.09361*.
[16] Y. Gu, V. Pahuja, G. Cheng, and Y. Su, "Knowledge base question answering: A semantic parsing perspective," in *Proc. Conf. Automated Knowl. Base Construction*, 2022, pp. 4–8.
[17] B. Fu, Y. Qiu, C. Tang, Y. Li, H. Yu, and J. Sun, "A survey on complex question answering over knowledge base: Recent advances and challenges," 2020, *arXiv:2007.13069*.
[18] Y. Lan, G. He, J. Jiang, J. Jiang, W. X. Zhao, and J.-R. Wen, "A survey on complex knowledge base question answering: Methods, challenges and solutions," in *Proc. Int. Joint Conf. Artif. Intell.*, 2021.
[19] Google, "Freebase data dumps," 2016, pp. 1–7.
[20] R. Cyganiak, D. Wood, and M. Lanthaler, "Rdf 1.1 concepts and abstract syntax," in *Proc. Int. World Wide Web Conf.*, 2014, pp. 18–31.
[21] T. Rebele, F. M. Suchanek, J. Hoffart, J. Biega, E. Kuzey, and G. Weikum, "YAGO: A multilingual knowledge base from wikipedia, wordnet, and geonames," in *Proc. Int. Semantic Web Conf.*, 2016, pp. 177–185.
[22] G. A. Miller, "WordNet: A lexical database for english," *Commun. ACM*, vol. 38, pp. 39–41, 1995.
[23] M. Färber, B. Ell, C. Menne, and A. Rettinger, "A comparative survey of dbpedia, freebase, opencyc, wikidata, and yago," *Semantic Web*, vol. 1, pp. 1–5, 2015.
[24] Y. Yang and M. Chang, "S-MART: Novel tree-based structured learning algorithms applied to tweet entity linking," in *Proc. Assoc. Comput. Linguistics*, 2015, pp. 504–513.
[25] J. Daiber, M. Jakob, C. Hokamp, and P. N. Mendes, "Improving efficiency and accuracy in multilingual entity extraction," in *Proc. Int. Conf. Semantic Syst.*, 2013, pp. 121–124.
[26] M. A. Yosef, J. Hoffart, I. Bordino, M. Spaniol, and G. Weikum, "AIDA: An online tool for accurate disambiguation of named entities in text and tables," *Proc. VLDB Endowment*, vol. 4, pp. 1450–1453, 2011.
[27] Q. Cai and A. Yates, "Large-scale semantic parsing via schema matching and lexicon extension," in *Proc. Assoc. Comput. Linguistics*, 2013, pp. 423–433.
[28] T. Kwiatkowski, E. Choi, Y. Artzi, and L. Zettlemoyer, "Scaling semantic parsers with on-the-fly ontology matching," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2013, pp. 1545–1556.

[29] S. Reddy, M. Lapata, and M. Steedman, "Large-scale semantic parsing without question-answer pairs," *Trans. Assoc. Comput. Linguistics*, vol. 2, pp. 377–392, 2014.

[30] C. Unger, L. Bühmann, J. Lehmann, A. N. Ngomo, D. Gerber, and P. Cimiano, "Template-based question answering over RDF data," in *Proc. Int. World Wide Web Conf.*, 2012, pp. 639–648.

[31] J. Berant, A. Chou, R. Frostig, and P. Liang, "Semantic parsing on freebase from question-answer pairs," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2013, pp. 1533–1544.

[32] W. Yih, X. He, and C. Meek, "Semantic parsing for single-relation question answering," in *Proc. Assoc. Comput. Linguistics*, 2014, pp. 643–648.

[33] X. Yao and B. V. Durme, "Information extraction over structured data: Question answering with Freebase," in *Proc. Assoc. Comput. Linguistics*, 2014, pp. 956–966.

[34] W. Yih, M. Richardson, C. Meek, M. Chang, and J. Suh, "The value of semantic parse labeling for knowledge base question answering," in *Proc. Assoc. Comput. Linguistics*, 2016, pp. 201–206.

[35] W. Yih, M. Chang, X. He, and J. Gao, "Semantic parsing via staged query graph generation: Question answering with knowledge base," in *Proc. Assoc. Comput. Linguistics*, 2015, pp. 1321–1331.

[36] S. Reddy et al., "Transforming dependency structures to logical forms for semantic parsing," *Trans. Assoc. Comput. Linguistics*, vol. 4, pp. 127–140, 2016.

[37] W. Yin, M. Yu, B. Xiang, B. Zhou, and H. Schütze, "Simple question answering by attentive convolutional neural network," in *Proc. Int. Conf. Comput. Linguistics*, 2016, pp. 1746–1756.

[38] Y. Lan, S. Wang, and J. Jiang, "Knowledge base question answering with a matching-aggregation model and question-specific contextual relations," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 10, pp. 1629–1638, Oct. 2019.

[39] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Learning to compose neural networks for question answering," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics - Hum. Lang. Technol.*, 2016, pp. 1545–1554.

[40] A. Bordes, J. Weston, and N. Usunier, "Open question answering with weakly supervised embedding models," in *Machine Learning and Knowledge Discovery in Databases*. Berlin, Germany: Springer, 2014.

[41] A. Bordes, S. Chopra, and J. Weston, "Question answering with subgraph embeddings," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2014, pp. 615–620.

[42] S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus, "End-to-end memory networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 2440–2448.

[43] A. H. Miller, A. Fisch, J. Dodge, A. Karimi, A. Bordes, and J. Weston, "Key-value memory networks for directly reading documents," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2016, pp. 1400–1409.

[44] S. Jain, "Question answering over knowledge base using factual memory networks," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics - Hum. Lang. Technol.*, 2016, pp. 109–115.

[45] Y. Hao et al., "An end-to-end model for question answering over knowledge base with cross-attention combining global knowledge," in *Proc. Assoc. Comput. Linguistics*, 2017, pp. 221–231.

[46] Z.-Y. Chen, C.-H. Chang, Y.-P. Chen, J. Nayak, and L.-W. Ku, "UHop: An unrestricted-hop relation extraction framework for knowledge-based question answering," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics - Hum. Lang. Technol.*, 2019, pp. 345–356.

[47] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, 1997, pp. 37–45.

[48] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder–decoder approaches," in *Proc. 8th Workshop Syntax, Semantics Struct. Statist. Transl.*, 2014, pp. 103–111.

[49] C. Liang, J. Berant, Q. Le, K. D. Forbus, and N. Lao, "Neural symbolic machines: Learning semantic parsers on Freebase with weak supervision," in *Proc. Assoc. Comput. Linguistics*, 2017, pp. 23–33.

[50] H. Sun, B. Dhingra, M. Zaheer, K. Mazaitis, R. Salakhutdinov, and W. W. Cohen, "Open domain question answering using early fusion of knowledge bases and text," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2018, pp. 4231–4242.

[51] W. Xiong, M. Yu, S. Chang, X. Guo, and W. Y. Wang, "Improving question answering over incomplete KBs with knowledge-aware reader," in *Proc. Assoc. Comput. Linguistics*, 2019, pp. 4258–4264.

[52] D. Diefenbach, V. López, K. D. Singh, and P. Maret, "Core techniques of question answering systems over knowledge bases: A survey," *Knowl. Inf. Syst.*, vol. 55, pp. 529–569, 2018.

[53] A. Abujabal, M. Yahya, M. Riedewald, and G. Weikum, "Automated template generation for question answering over knowledge graphs," in *Proc. Int. World Wide Web Conf.*, 2017, pp. 1191–1200.

[54] A. Abujabal, R. S. Roy, M. Yahya, and G. Weikum, "Never-ending learning for open-domain question answering over knowledge bases," in *Proc. Int. World Wide Web Conf.*, 2018, pp. 1053–1062.

[55] P. Kapanipathi et al., "Leveraging abstract meaning representation for knowledge base question answering," in *Proc. Findings Assoc. Comput. Linguistics*, 2021, pp. 3884–3894.

[56] Y. Sun, L. Zhang, G. Cheng, and Y. Qu, "SPARQA: Skeleton-based semantic parsing for complex questions over knowledge bases," in *Proc. Conf. Assoc. Advance. Artif. Intell.*, 2020, pp. 8952–8959.

[57] S. Zhu, X. Cheng, and S. Su, "Knowledge-based question answering by tree-to-sequence learning," *Neurocomputing*, vol. 372, pp. 64–72, 2020.

[58] H. Zafar, G. Napolitano, and J. Lehmann, "Formal query generation for question answering over knowledge bases," in *Proc. Eur. Semantic Web Conf.*, 2018, pp. 714–728.

[59] Y. Chen, H. Li, Y. Hua, and G. Qi, "Formal query building with query structure prediction for complex question answering over knowledge base," in *Proc. Int. Joint Conf. Artif. Intell.*, 2020, pp. 3751–3758.

[60] H. Bast and E. Haussmann, "More accurate question answering on freebase," in *Proc. Conf. Inf. Knowl. Manage.*, 2015, pp. 1431–1440.

[61] Z. Jia, A. Abujabal, R. Saha Roy, J. Strötgen, and G. Weikum, "TEQUILA: Temporal question answering over knowledge bases," in *Proc. Conf. Inf. Knowl. Manage.*, 2018, pp. 1807–1810.

[62] W. Zheng, J. X. Yu, L. Zou, and H. Cheng, "Question answering over knowledge graphs: Question understanding via template decomposition," 2018, pp. 1373–1386.

[63] N. Bhutani, X. Zheng, and H. V. Jagadish, "Learning to answer complex questions over knowledge bases with query composition," in *Proc. Conf. Inf. Knowl. Manage.*, 2019, pp. 739–748.

[64] Y. Lan, S. Wang, and J. Jiang, "Multi-hop knowledge base question answering with an iterative sequence matching model," in *Proc. Int. Conf. Des. Mater.*, 2019, pp. 359–368.

[65] Y. Lan and J. Jiang, "Query graph generation for answering multi-hop complex questions from knowledge bases," in *Proc. Assoc. Comput. Linguistics*, 2020, pp. 969–974.

[66] A. Saha, G. A. Ansari, A. Laddha, K. Sankaranarayanan, and S. Chakrabarti, "Complex program induction for querying knowledge bases in the absence of gold programs," *Trans. Assoc. Comput. Linguistics*, vol. 7, pp. 185–200, 2019.

[67] Y. Qiu et al., "Hierarchical query graph generation for complex question answering over knowledge graph," in *Proc. Conf. Inf. Knowl. Manage.*, 2020, pp. 1285–1294.

[68] Y. Qiu, Y. Wang, X. Jin, and K. Zhang, "Stepwise reasoning for multi-relation question answering over knowledge graph with weak supervision," in *Proc. Int. Conf. Web Search Data Mining*, 2020, pp. 474–482.

[69] Y. Hua, Y. Li, G. Qi, W. Wu, J. Zhang, and D. Qi, "Less is more: Data-efficient complex question answering over knowledge bases," *J. Web Semant.*, vol. 65, 2020, Art. no. 100612.

[70] L. Galárraga, C. Teflioudi, K. Hose, and F. M. Suchanek, "Fast rule mining in ontological knowledge bases with AMIE++," *VLDB Journal.*, vol. 24, no. 6, pp. 707–730, 2015.

[71] L. C. Lamb, A. D. Garcez, M. Gori, M. O. Prates, P. H. Avelar, and M. Y. Vardi, "Graph neural networks meet neural-symbolic computing: A survey and perspective," in *Proc. Int. Joint Conf. Artif. Intell.*, 2020, pp. 4877–4884.

[72] J. Zhang, B. Chen, L. Zhang, X. Ke, and H. Ding, "Neural, symbolic and neural-symbolic reasoning on knowledge graphs," *AI Open*, vol. 2, pp. 14–35, 2021.

[73] L. A. Galárraga, C. Teflioudi, K. Hose, and F. Suchanek, "AMIE: Association rule mining under incomplete evidence in ontological knowledge bases," in *Proc. 22nd Int. Conf. World Wide Web*, 2013, pp. 413–422.

[74] Y. Sun et al., "Neural semantic parsing in low-resource settings with back-translation and meta-learning," in *Proc. Conf. Assoc. Advance. Artif. Intell.*, 2020, pp. 8960–8967.

[75] J. Bao, N. Duan, Z. Yan, M. Zhou, and T. Zhao, "Constraint-based question answering with knowledge graph," in *Proc. Int. Conf. Comput. Linguistics*, 2016, pp. 2503–2514.

[76] X. Ye, S. Yavuz, K. Hashimoto, Y. Zhou, and C. Xiong, "RNG-KBQA: Generation augmented iterative ranking for knowledge base question answering," in *Proc. Assoc. Comput. Linguistics*, 2022, pp. 6032–6043.

[77] A. Saxena, A. Tripathi, and P. P. Talukdar, "Improving multi-hop question answering over knowledge graphs using knowledge base embeddings," in *Proc. Assoc. Comput. Linguistics*, 2020, pp. 4498–4507.

[78] H. Sun, T. Bedrax-Weiss, and W. W. Cohen, "PullNet: Open domain question answering with iterative retrieval on knowledge bases and text," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2019, pp. 2380–2390.

[79] G. Maheshwari, P. Trivedi, D. Lukovnikov, N. Chakraborty, A. Fischer, and J. Lehmann, "Learning to rank query graphs for complex question answering over knowledge graphs," in *Proc. Int. Semantic Web Conf.*, 2019, pp. 487–504.

[80] K. S. Tai, R. Socher, and C. D. Manning, "Improved semantic representations from tree-structured long short-term memory networks," in *Proc. Assoc. Comput. Linguistics*, 2015, pp. 1556–1566.

[81] A. Talmor and J. Berant, "The web as a knowledge-base for answering complex questions," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics - Hum. Lang. Technol.*, 2018, pp. 641–651.

[82] J. Han, B. Cheng, and X. Wang, "Open domain question answering based on text enhanced knowledge graph with hyperedge infusion," in *Proc. Conf. Finding Empir. Methods Natural Lang. Process.*, 2020, pp. 1475–1481.

[83] J. Zhang et al., "Subgraph retrieval enhanced model for multi-hop knowledge base question answering," 2022, *arXiv:2202.13296*.

[84] J. Han, B. Cheng, and X. Wang, "Two-phase hypergraph based reasoning with dynamic relations for multi-hop KBQA," in *Proc. Int. Joint Conf. Artif. Intell.*, 2020, pp. 3615–3621.

[85] M. Yasunaga, H. Ren, A. Bosselut, P. Liang, and J. Leskovec, "QA-GNN: Reasoning with language models and knowledge graphs for question answering," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics - Hum. Lang. Technol.*, 2021, pp. 535–546.

[86] G. He, Y. Lan, J. Jiang, W. X. Zhao, and J. Wen, "Improving multi-hop knowledge base question answering by learning intermediate supervision signals," in *Proc. Int. Conf. Web Search Data Mining*, 2021, pp. 553–561.

[87] M. Zhou, M. Huang, and X. Zhu, "An interpretable reasoning network for multi-relation question answering," in *Proc. Int. Conf. Comput. Linguistics*, 2018, pp. 2010–2022.

[88] K. Xu, Y. Lai, Y. Feng, and Z. Wang, "Enhancing key-value memory neural networks for knowledge based question answering," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics - Hum. Lang. Technol.*, 2019, pp. 2937–2947.

[89] S. He, C. Liu, K. Liu, and J. Zhao, "Generating natural answers by incorporating copying and retrieving mechanisms in sequence-to-sequence learning," in *Proc. Assoc. Comput. Linguistics*, 2017, pp. 199–208.

[90] J. Yin, X. Jiang, Z. Lu, L. Shang, H. Li, and X. Li, "Neural generative question answering," in *Proc. Int. Joint Conf. Artif. Intell.*, 2016, pp. 2972–2978.

[91] Y. Fu and Y. Feng, "Natural answer generation with heterogeneous memory," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics - Hum. Lang. Technol.*, 2018, pp. 185–195.

[92] Y. Feng et al., "A pretraining numerical reasoning model for ordinal constrained question answering on knowledge base," in *Proc. Findings Conf. Empir. Methods Natural Lang. Process.*, 2021, pp. 1852–1861.

[93] Y. Zhang, H. Dai, Z. Kozareva, A. J. Smola, and L. Song, "Variational reasoning for question answering with knowledge graph," in *Proc. Conf. Assoc. Advance. Artif. Intell.*, 2018, pp. 6069–6076.

[94] B. Min, R. Grishman, L. Wan, C. Wang, and D. Gondek, "Distant supervision for relation extraction with an incomplete knowledge base," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics - Hum. Lang. Technol.*, 2013, pp. 777–782.

[95] Y. Feng, H. You, Z. Zhang, R. Ji, and Y. Gao, "Hypergraph neural networks," in *Proc. Conf. Assoc. Advance. Artif. Intell.*, 2019, pp. 3558–3565.

[96] T. Trouillon, J. Welbl, S. Riedel, É. Gaussier, and G. Bouchard, "Complex embeddings for simple link prediction," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 2071–2080.

[97] Y. Liu et al., "Roberta: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[98] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu, "ERNIE: Enhanced language representation with informative entities," in *Proc. Assoc. Comput. Linguistics*, 2019, pp. 1441–1451.

[99] M. E. Peters et al., "Knowledge enhanced contextual word representations," in *Proc. Conf. Empir. Methods Natural Lang. Process. Joint Conf. Natural Lang. Process.*, 2019, pp. 43–54.

[100] R. Bommasani et al., "On the opportunities and risks of foundation models," 2021, *arXiv:2108.07258*.

[101] X. Han, "Pre-trained models: Past, present and future," 2021, *arXiv:2106.07139*.

[102] W. Yu et al., "A survey of knowledge-enhanced text generation," 2020, *arXiv:2010.04389*.

[103] R. Das et al., "Case-based reasoning for natural language queries over knowledge bases," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2021, pp. 9594–9611.

[104] J. Shi et al., "Kqa pro: A large diagnostic dataset for complex question answering over knowledge base," 2020, *arXiv:2007.03875*.

[105] T. Xie et al., "Unifiedskg: Unifying and multi-tasking structured knowledge grounding with text-to-text language models," 2022, *arXiv:2201.05966*.

[106] C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, pp. 1–67, 2020.

[107] R. Andersen, F. R. K. Chung, and K. J. Lang, "Local graph partitioning using pagerank vectors," in *Proc. Annu. IEEE Symp. Found. Comput. Sci.*, 2006, pp. 475–486.

[108] F. Petroni et al., "Language models as knowledge bases?," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2019, pp. 2463–2473.

[109] Z. Bouraoui, J. Camacho-Collados, and S. Schockaert, "Inducing relational knowledge from BERT," in *Proc. Conf. Assoc. Advance. Artif. Intell.*, 2020, pp. 7456–7463.

[110] Z. Jiang, F. F. Xu, J. Araki, and G. Neubig, "How can we know what language models know," *Trans. Assoc. Comput. Linguistics*, vol. 8, pp. 423–438, 2020.

[111] S. Cao, J. Shi, Z. Yao, L. Hou, J. Li, and J. Xiao, "Program transfer and ontology awareness for semantic parsing in KBQA," 2021, *arXiv:2110.05743*.

[112] J. Zhang, L. Zhang, B. Hui, and L. Tian, "Improving complex knowledge base question answering via structural information learning," *Knowl.-Based Syst.*, vol. 242, 2022, Art. no. 108252.

[113] J. Shi, S. Cao, L. Hou, J. Li, and H. Zhang, "TransferNet: An effective and transparent framework for multi-hop question answering over relation graph," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2021, pp. 4149–4158.

[114] Y. Lan and J. Jiang, "Modeling transitions of focal entities for conversational knowledge base question answering," in *Proc. Assoc. Comput. Linguistics*, 2021, pp. 3288–3297.

[115] Y. Lan, S. Wang, and J. Jiang, "Knowledge base question answering with topic units," in *Proc. Int. Joint Conf. Artif. Intell.*, 2019, pp. 5046–5052.

[116] Orange, "Qald data challenge," 2011. [Online]. Available: http://qald.aksw.org/

[117] P. Trivedi, G. Maheshwari, M. Dubey, and J. Lehmann, "LC-QuAD: A corpus for complex question answering over knowledge graphs," in *Proc. Int. Semantic Web Conf.*, 2017, pp. 210–218.

[118] W. Zheng and M. Zhang, "Automated query graph generation for querying knowledge graphs," in *Proc. Conf. Inf. Knowl. Manage.*, 2021, pp. 2698–2707.

[119] S. Vakulenko, J. D. F. Garcia, A. Polleres, M. de Rijke, and M. Cochez, "Message passing for complex question answering over knowledge graphs," in *Proc. Conf. Inf. Knowl. Manage.*, 2019, pp. 1431–1440.

[120] M. Dubey, D. Banerjee, A. Abdelkawi, and J. Lehmann, "LC-QuAD 2.0: A large dataset for complex question answering over wikidata and DBpedia," in *Proc. Int. Semantic Web Conf.*, 2019, pp. 69–78.

[121] J. Zou et al., "A chinese multi-type complex questions answering dataset over wikidata," 2021, *arXiv:2111.06086*.

[122] J. Cai, Z. Zhang, F. Wu, and J. Wang, "Deep cognitive reasoning network for multi-hop question answering over knowledge graphs," in *Proc. Findings Assoc. Comput. Linguistics*, 2021, pp. 219–229.

[123] D. Keysers et al., "Measuring compositional generalization: A comprehensive method on realistic data," in *Proc. Int. Conf. Learn. Representations*, 2020, pp. 5–8.

[124] Y. Guo, Z. Lin, J. Lou, and D. Zhang, "Hierarchical poset decoding for compositional generalization in language," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 6913–6924.

[125] S. Chen, Q. Liu, Z. Yu, C.-Y. Lin, J.-G. Lou, and F. Jiang, "ReTraCk: A flexible and efficient framework for knowledge base question answering," in *Proc. Assoc. Comput. Linguistics*, 2021, pp. 325–336.

[126] Z. Yao, Y. Tang, W. Yih, H. Sun, and Y. Su, "An imitation game for learning semantic parsers from user interaction," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2020, pp. 6883–6902.

[127] W. Zheng, H. Cheng, J. X. Yu, L. Zou, and K. Zhao, "Interactive natural language question answering over knowledge graphs," *Inf. Sci.*, vol. 481, pp. 141–159, 2019.

[128] Y. Hua, Y.-F. Li, G. Haffari, G. Qi, and T. Wu, "Few-shot complex knowledge base question answering via meta reinforcement learning," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2020, pp. 5827–5837.

[129] Y. Zhou, X. Geng, T. Shen, W. Zhang, and D. Jiang, "Improving zero-shot cross-lingual transfer for multilingual question answering over knowledge graph," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics - Hum. Lang. Technol.*, 2021, pp. 5822–5834.

[130] M. Gardner et al., "Evaluating nlp models via contrast sets," 2020, *arXiv:2004.02709*.

[131] D. Kaushik, E. Hovy, and Z. Lipton, "Learning the difference that makes a difference with counterfactually-augmented data," in *Proc. Int. Conf. Learn. Representations*, 2020, pp. 1–3.

[132] I. Oren, J. Herzig, N. Gupta, M. Gardner, and J. Berant, "Improving compositional generalization in semantic parsing," in *Proc. Conf. Finding Empir. Methods Natural Lang. Process.*, 2020, pp. 2482–2495.

[133] D. Guo, D. Tang, N. Duan, M. Zhou, and J. Yin, "Dialog-to-action: Conversational question answering over a large-scale knowledge base," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 2946–2955.

[134] P. Christmann, R. S. Roy, A. Abujabal, J. Singh, and G. Weikum, "Look before you hop: Conversational question answering over knowledge graphs using judicious context expansion," in *Proc. Conf. Inf. Knowl. Manage.*, 2019, pp. 729–738.

[135] J. Plepi, E. Kacupaj, K. Singh, H. Thakkar, and J. Lehmann, "Context transformer with stacked pointer networks for conversational question answering over knowledge graphs," in *Proc. Eur. Semantic Web Conf.*, 2021, pp. 356–371.

[136] M. Kaiser, R. S. Roy, and G. Weikum, "Reinforcement learning from reformulations in conversational question answering over knowledge graphs," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2021, pp. 459–469.

[137] E. Kacupaj, J. Plepi, K. Singh, H. Thakkar, J. Lehmann, and M. Maleshkova, "Conversational question answering over knowledge graphs with transformer and graph attention networks," in *Proc. Eur. Conf. Field Comput. Linguistics*, 2021, pp. 850–862.

[138] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[139] L. C. Lamb, A. S. d'Avila Garcez, M. Gori, M. O. R. Prates, P. H. C. Avelar, and M. Y. Vardi, "Graph neural networks meet neural-symbolic computing: A survey and perspective," in *Proc. Int. Joint Conf. Artif. Intell.*, 2020, pp. 4877–4884.

[140] F. Arabshahi, S. Singh, and A. Anandkumar, "Combining symbolic expressions and black-box function evaluations in neural programs," in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 3–7.

[141] G. Lample and F. Charton, "Deep learning for symbolic mathematics," in *Proc. Int. Conf. Learn. Representations*, 2020, pp. 2–7.

[142] R. Xie, Z. Liu, J. Jia, H. Luan, and M. Sun, "Representation learning of knowledge graphs with entity descriptions," in *Proc. Conf. Assoc. Advance. Artif. Intell.*, 2016, pp. 2659–2665.

[143] R. Xie, Z. Liu, H. Luan, and M. Sun, "Image-embodied knowledge representation learning," in *Proc. Int. Joint Conf. Artif. Intell.*, 2017, pp. 3140–3146.

[144] G. He, J. Li, W. X. Zhao, P. Liu, and J. Wen, "Mining implicit entity preference from user-item interaction data for knowledge graph completion via adversarial learning," in *Proc. Int. World Wide Web Conf.*, 2020, pp. 740–751.

[145] X. Lu, S. Pramanik, R. S. Roy, A. Abujabal, Y. Wang, and G. Weikum, "Answering complex questions by joining multi-document evidence with quasi knowledge graphs," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Informat. Retrieval*, 2019, pp. 105–114.

[146] B. Dhingra, M. Zaheer, V. Balachandran, G. Neubig, R. Salakhutdinov, and W. W. Cohen, "Differentiable reasoning over a virtual knowledge base," in *Proc. Int. Conf. Learn. Representations*, 2020, pp. 1–6.
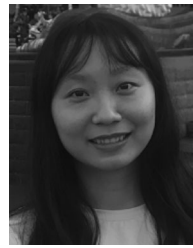
Yunshi Lan received the PhD degree from the School of Computing and Information Systems, Singapore Management University in 2020. She is currently an associate professor with the School of Data Science and Engineering of East China Normal University. Her research interests lie in natural language processing with a focus on knowledge base question answering.


Gaole He received the master's degree from the School of Information, Renmin University of China in 2021. He is currently working toward the PhD degree with the Delft University of Technology. His research focuses on graph mining, knowledge representation and knowledge reasoning.

Jinhao Jiang received the bachelor's degree from the school of information and software engineering, University of Electronic Science and Technology of China in 2021. He is currently working toward the PhD degree with the Renmin University of China. His research focuses on natural language processing, knowledge base and knowledge reasoning.

Jing Jiang received the PhD degree in computer science from the University of Illinois at Urbana-Champaign in 2008. Currently, she is a professor with the School of Computing and Information Systems, the Singapore Management University. Her research interests include natural language processing, text mining and machine learning.
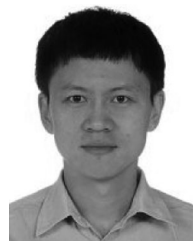
Wayne Xin Zhao (Member, IEEE) received the PhD degree from Peking University in 2014. He is currently a tenured associate professor with the Gaoling School of Artificial Intelligence, Renmin University of China. His research interests are web text mining and natural language processing. He has published a number of papers in international conferences and journals such as ACL, SIGIR, SIGKDD, WWW, *ACM Transactions on Information Systems*, and *IEEE Transactions on Knowledge and Data Engineering*.

Ji-Rong Wen (Senior Member, IEEE) received the PhD degree from the Chinese Academy of Science in 1999. He is a professor with the Renmin University of China. He was a senior researcher and research manager with Microsoft Research from 2000 to 2014. His main research interests include web data management, information retrieval (especially web IR), and data mining.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/csdl.