Contrastive Semantic Similarity Learning for Multi-Hop Question Answering over Event-Centric Knowledge Graphs

Wei Tang[1,2], Qingchao Kong[1,2,*], Wenji Mao[1,2], Xiaofei Wu[3]

[1]Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China
[3]Beijing Wenge Technology Co., Ltd., Beijing 100080, China
{tangwei2020, qingchao.kong, wenji.mao}@ia.ac.cn, xiaofei.wu@wenge.com

Abstract: Question answering in natural languages provides an intuitive and efficient way to help people access the rich information stored in various kinds of knowledge graphs (KGs). One of the key challenges for question answering over knowledge graphs (KGQA) is to learn a semantic representation of the input question and candidate relation chains over KGs and accurately measure the similarity between them. However, existing methods often failed to capture the semantic similarity for complex question answering, e.g., multi-hop and temporal constrained situations. In addition, existing KGQA related research mostly concentrates on entities while often ignores the events which contain a large portion of the world knowledge. To solve this issue, we propose a Contrastive Semantic Similarity Learning (CSSL) method for multi-hop question answering over event-centric KGs. In this method, for candidate relation chains generation, the retrieval subgraph is first constructed by identifying the topic event or entity in the question. To better accommodate complex questions, we introduce the contrastive learning framework to learn a common semantic space, where the similarity score is finally calculated to select the final answer. The experimental results on the EventQA dataset show that the proposed method achieves superior performances compared to the state-of-the-art baselines.

Keywords: Multi-hop question answering; Semantic similarity; Contrastive learning; Event-centric knowledge graph

1 Introduction

The knowledge graphs (KGs) are defined as a set of nodes and directional edges, where nodes represent entities, events or other concepts in the real world, and directional edges represent the semantic relations between nodes [1]. To access the huge amount of human knowledge stored in KGs, it is usually required to know the schema of KGs and master the Structural Query Languages (SQL), such as SPARQL [1] and Cypher [2] . In contrast, Question Answering in natural language over KGs (KGQA) provides an intuitive and efficient way to help people acquire these information in intelligence analysis and
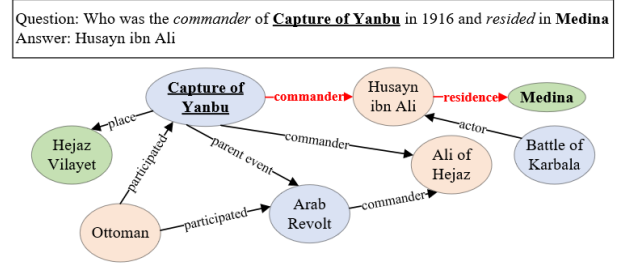

Figure 1 An example question and corresponding retrieval subgraph of the EventKG. The correct relation chain is colored in red.

decision making. One of the key challenges for KGQA is to learn a semantic representation for input questions and relation chains over KGs, which is used to score the similarity between them and select correct answers (see Figure 1 for an illustration). Recently proposed research work usually adopts deep learning-based methods to learn the representation of questions and relations. For example, Bi et al. [2] proposes an unrestricted multi-hop reason network to encode questions and relations. Yan et al. [3] introduces three additional pre-training tasks for BERT, including relation extraction, relation matching and relation reasoning for relation-augmented training and improves relation representation abilities. Zhang et al. [4] utilizes dependence tree, constituency tree and the first token to construct a composited structural attention so as to generate relation features. However, for complex questions such as multi-hop and temporal constrained questions, existing methods often failed to learn a semantic representation for them to conduct similarity matching. In addition, existing KGQA methods usually concentrate on entities and ignore events, which represent the fast-developing world and are also important sources of world knowledge. The main reason is that most KGs are entity-centric, e.g., Wikidata, DBpedia, NELL, YAGO, and OpenKG. To better incorporate event information, some event-centric KGs are recently proposed, such as EventKG [5] and ASER [6].

In this paper, we propose a Contrastive Semantic Similarity Learning (CSSL) method for complex question answering over event-centric KGs. The proposed method

---

first constructs a retrieval subgraph of the KGs using the identified topic event or entity to generate candidate relation chains. To better accommodate complex questions, we adopt the contrastive learning framework to specifically learn a common semantic space for them. Lastly, the final answer is selected with the highest similarity score.

The main contributions of this paper are summarized as follows.

(1) We propose a contrastive semantic similarity learning method for multi-hop question answering over event-centric knowledge graphs, which consists of relation chain generation, semantic similarity learning and similarity score calculation.

(2) To better accommodate complex question answering in multi-hop and temporal constrained situations, we adopt the contrastive learning framework to calculate the similarity between the input questions and candidate relation chains in a common semantic space.

(3) Experimental results on the standard EventQA dataset demonstrate that our proposed method achieves superior performances compared to the state-of-the-art baselines.

The rest of this paper is organized as follows. The related work is introduced in Section 2. The proposed method is described in Section 3. The experiment is presented in Section 4. The conclusion is provided in Section 5.

2 Related Work

There has been numerous research work on question answering over knowledge graphs, which mainly fall into three categories, namely template-based methods, semantic parsing-based methods and information retrieval-based methods.

2.1 Template-based methods

Template-based methods usually construct logic query forms (e.g., SPARQL and Cypher) using predefined question templates [7] and execute them on the KGs to find answers. Wahyudi et al. [8] first matches a question with a template, which is then transformed into a graph, and finally translates the graph into a Cypher query. Athreya et al. [9] regards the question answering task as a classification problem, using recursive neural networks to classify questions and match SPARQL query templates. Reddy and Madhavi [10] integrates templates and convolutional neural network to decompose complex questions. In summary, template-based methods usually have good interpretability, but suffer from poor generalizations because of the limited coverage of pre-defined templates.

2.2 Semantic parsing-based methods

Different from template-based methods, semantic parsing-based methods directly translate questions into logic query forms rather than using pre-defined templates

[11]. Gao et al. [12] combines rules and neural networks to parse the semantic segment sequences and build the semantic query graphs. Liang et al. [13] puts forward a BERT-based semantic query graph extraction model. Sorokin and Gurevych [14] utilizes gated graph neural networks to encode the semantic parsing graph, partly solving the representation problem of complex questions. Semantic parsing-based methods usually require a large number of manually annotated pairs of question and corresponding logical query form, and seldom utilize the content and structure of the KGs.

2.3 Information retrieval-based methods

Information retrieval-based methods usually firstly extract the topic entity and predicates from questions, encode the questions and candidate answers, and finally rank the candidate answers from KGs according to their similarity with the questions. Chen and Li [15] presents a transformer-based deep attentive matching model to identify the relations from questions. Wang et al. [16] proposes a retrieval-and-reranking policy to select the answer from candidates with fine-grained matching. Lan et al. [17] iteratively grows the candidate relation paths which could lead to the answer entities. Yan et al. [3] introduces three auxiliary tasks to augment relation learning, i.e., relation extraction, matching and reasoning. Information retrieval-based methods generally performs better than the above two types of methods, and our proposed method also belongs to this category.

3 Problem Definition

In this paper, the problem is defined as: given an input multi-hop question Q in natural language and the event-centric KGs $\mathcal{K}$ with a collection of triples $< e_1, r, e_2 >$, where $e_1, e_2 \in \mathcal{E}$ are entities or events and $r \in \mathcal{R}$ is the relation, and the goal is to find an entity or event node $a \in \mathcal{E}$ which serves as the answer for the input question Q.

4 Proposed Method

The proposed method CSSL consists of three components (see Figure 2 for an overview), namely, relation chain generation, semantic similarity learning and similarity score calculation. The first component aims to obtain the candidate relation chains relevant to the input question by constructing a retrieval subgraph using the identified topic event or entity. The semantic similarity learning component learns the semantic representations of questions and candidate relation chains through contrastive learning. The third component selects the final answer with the highest similarity score.

4.1 Relation Chain Generation

To construct the retrieval subgraph, the topic event or entity is first identified, which is the focus of the input question. In our proposed method, the sequence labeling technique is adopted to find the topic event and entity. Specifically, the vector representation of input question Q,
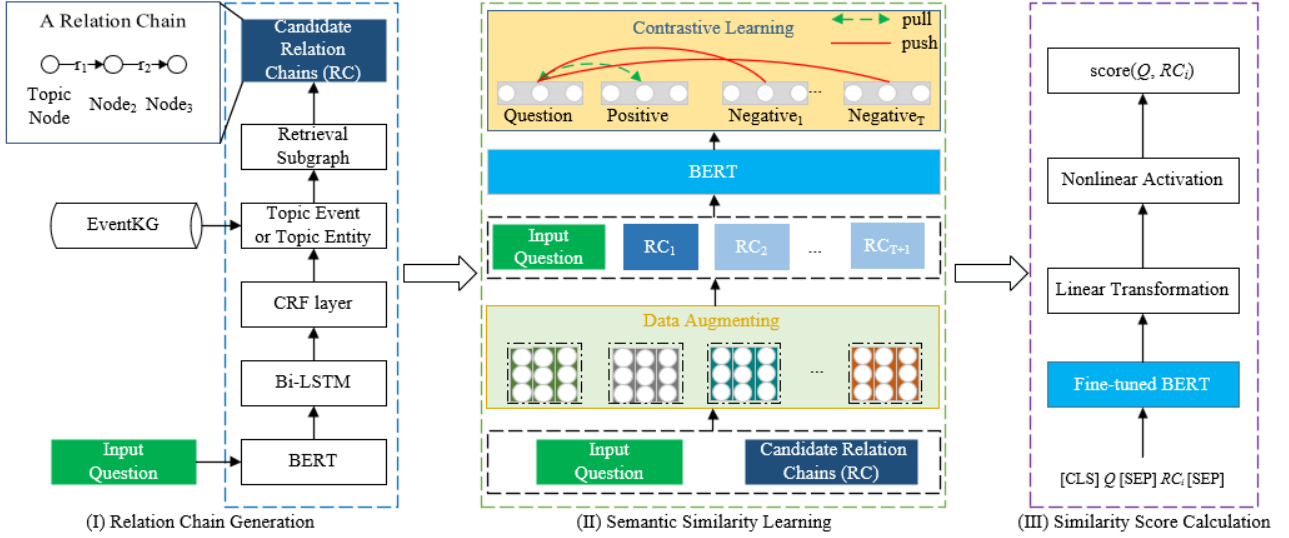
Figure 2 Overview of our proposed CSSL method for multi-hop question answering

denoted as $\{x_t, t = 1, 2, \dots, M\}$ where x_t is a word and M is the maximum length of input questions, is obtained through BERT and Bi-LSTM.

$$\boldsymbol{v} = BERT_{CLS}([CLS]\, x_1, \dots, x_M\, [SEP]) \qquad (1)$$

$$\boldsymbol{h} = Bi\text{-}LSTM(\boldsymbol{v}) \qquad (2)$$

where $\boldsymbol{h} = [h_1, \dots, h_M]$ is the output of the Bi-LSTM layer and h_i is the final hidden representation for word x_i. Then the topic event or entity t_e is obtained through a CRF layer with $\boldsymbol{h}$.

Given by the identified t_e, the retrieval subgraph $\mathcal{K}_{RS} \subseteq \mathcal{K}$ is constructed as follows. The t_e is first linked to $\mathcal{K}$ to locate the topic node, and centering on this topic node, all relations and nodes within the maximum number of hops are retrieved from $\mathcal{K}$ to construct $\mathcal{K}_{RS}$. Finally, all the relation chains centering on the topic node are extracted as the candidate relation chains.

4.2 Contrastive Learning for Semantic Similarity

Intuitively, the correct relation chain denoted as rc_+, which leads to the right answer, should be more semantically similar with the input question Q than the incorrect relation chains denoted as rc_-. To better distinguish the rc_+ and rc_-, the contrastive learning framework is adopted to learn a common semantic space, which aims to make rc_+ closer to Q, and rc_- farther from Q. Specifically, the BERT model is utilized to encode questions and each candidate relation chain (formed as `<E1> Topic Node </E1> r₁ node₂ r₂ <E2> node₃ </E2>`), respectively. The contrastive loss InfoNCE [18] is utilized:

$$\mathcal{L}_{InfoNCE} = -\log \frac{\exp(\boldsymbol{q} \cdot \boldsymbol{rc}_+/\tau)}{\sum_{i=1}^{T+1} \exp(\boldsymbol{q} \cdot \boldsymbol{rc}_i/\tau)} \qquad (3)$$

where $\boldsymbol{q}$ represents the encoded question, T is the number of negative relation chains, $\{\boldsymbol{rc}_i \mid i = 1, \dots T+1\}$ is the set of encoded relation chains, $\boldsymbol{rc}_+$ represents the encoded correct relation chain, τ is the temperature hyper-parameter. Considering that there are varying

number of incorrect relation chains for different questions: if a question has more than T negative relations, the random down-sampling method is applied; if a question has less than T negative relations, the random dropout operation is adopted to generate more samples based on existing negative relation chains [19].

4.3 Similarity Score Calculation

The similarity score between the input question and candidate relation chains is predicted by the BERT model which is already fine-tuned in the semantic similarity learning component. In the training phase, the similarity score is set to 1 for the correct relation chain and 0 otherwise. The loss function can be the cross-entropy loss (CE) or the hinge loss (HL).

$$\boldsymbol{m}_i = BERT_{CLS}([CLS]\, Q\, [SEP]\, rc_i\, [SEP]) \qquad (4)$$

$$t_i = \sigma(\boldsymbol{W} \cdot \boldsymbol{m}_i + \boldsymbol{b}) \qquad (5)$$

$$\mathcal{L}_{CL} = -(y \cdot \log t_i + (1 - y) \cdot \log(1 - t_i)) \qquad (6)$$

$$\mathcal{L}_{HL} = \frac{1}{T} \sum_{i=1}^{T} \max(0, l + s^+ - s^-) \qquad (7)$$

where rc_i is the candidate relation chain, σ is the ReLU activation function, $\boldsymbol{W}$ and $\boldsymbol{b}$ are parameters. For the cross-entropy loss, y is the ground truth label which indicates whether rc_i is the correct relation chain or not. For the hinge loss, l is a margin, s^+ is the similarity score between Q and rc_+, s^- is the similarity score between Q and rc_-.

5 Experiment

5.1 Dataset

We use the EventQA [20] dataset for evaluation, which is the only question answering dataset for event-centric KGs to the best of our knowledge. This dataset includes 1000 two-hop questions, 1005 events, 1655 entities, and 309 predicates. EventQA is created based on EventKG V2.0

[5] which is a multilingual event-centric temporal knowledge graph. EventKG V2.0 has more than 970k contemporary and history events and 2.8 million temporal relations extracted from DBpedia, Wikidata and YAGO knowledge graph as well as several semi-structured sources. The EventQA dataset is further split into three parts, i.e., training, validation and test, which contains 600, 200 and 200 questions, respectively.

5.2 Experimental Settings

The maximum number of hops to construct $\mathcal{K}_{RS}$ is 2 and the maximum number of nodes in the retrieval subgraph is limited to 500. The number of negative relation chains T is 20. The dropout ratio in the random dropout operation is 0.2. The temperature τ of InfoNCE loss is 0.05. The margin l for the hinge loss is 1. The model is optimized by Adam with a learning rate 3e-5 [21]. The metrics used to evaluate the performance in this paper are precision, recall, and F1-score (F1).

5.3 Baseline Methods

Two groups of recently proposed KGQA methods are selected as baseline methods for comparison. The semantic parsing-based methods are not included because of the costly manual annotation process to generate logic query forms for each question.

The first group contains two template-based methods.
- **Athreya et al**. [9] utilizes recursive neural network to classify a question and match a responding template.
- **T-CRNN** [10] adopts the template representation based convolutional recurrent neural network to obtain answers.

The second group includes four information retrieval-based methods, which achieve competitive results.
- **Yan et al**. [3] introduces two external datasets and proposes three auxiliary tasks for relation learning, namely relation extraction, relation matching and relation reasoning to better map the question to reasoning paths in the knowledge graph.
- **Lan et al**. [17] iteratively grows the candidate relation paths based on a topic entity and prunes away less relevant branches.
- **DAM** [15] proposes a transformer-based deep attentive matching model to extract the relations and employs the fine-grained word-level attention to enhance the matching of questions and relations.
- **Wang et al**. [16] applies a retrieval-and-reranking policy to select the answer from candidates with fine-grained matching.

5.4 Experimental Results

Topic event/entity identification result. The identification results of the topic event or entity are reported in Table I. As we can see, the identification results of events are relatively worse than entities, because events in a question may contain much more

information than entities, such as time and location, which is more challenging to extract.

Table I Results of identifying the topic event/entity (%)

	Precision	Recall	F1
Event	76.05	75.52	75.78
Entity	95.07	92.02	93.52

Table II Results of question answering (%)

Method	Precision	Recall	F1
Athreya et al. [9]	40.50	38.77	39.62
T-CRNN [10]	39.23	36.78	37.97
Yan et al. [3]	47.70	45.03	45.36
Lan et al. [17]	48.50	44.09	46.19
DAM [15]	47.50	43.28	45.29
Wang et al. [16]	49.27	<u>47.09</u>	48.16
CSSL+HL	<u>52.90</u>	46.82	<u>49.67</u>
CSSL+CE	**53.50**	**47.73**	**50.45**

Question answering results. Table II shows the question answering results of our proposed methods and different baseline methods, where CSSL+HL and CSSL+CE denote the proposed methods with different loss functions, i.e., hinge loss and cross-entropy loss, for predicting the final similarity score. The best results are in bold and the second-best results are underlined.

In general, Table II shows that information retrieval-based methods perform better than template-based methods, which is also verified in related research [22]. The reason is that the performances of the template-based methods are highly dependent on the coverage of the question templates, while the information retrieval-based methods only need to compare the text similarity between questions and different relation chains (or entities) to find the answer, especially for complex questions. Table II also shows that our proposed methods achieve the best and second-best results in F1 score. Compared to other methods, our methods adopt the contrastive learning framework to learn a common semantic space to better distinguish the correct relation chains and incorrect ones.

5.5 Ablation study

In this section, we evaluate the effectiveness of the semantic similarity learning component. After removing the semantic similarity learning component, the similarity score calculation component is directly used to fine-tune a BERT model to predict the answer using two different loss functions, denoted as CE and HL.

Table III Results of w/ and w/o the semantic similarity learning component (%)

	Precision	Recall	F1
HL	46.68	43.07	44.80
CSSL+HL	<u>52.90</u>	<u>46.82</u>	<u>49.67</u>
CE	47.14	43.68	45.34
CSSL+CE	**53.50**	**47.73**	**50.45**

From Table III we can see that without the semantic similarity learning component, the performances of the proposed methods drop significantly, with a degradation of 4.87% and 5.11% in terms of F1 score for CSSL+HL and CSSL+CE, respectively. This further verifies the importance of learning a common semantic space of questions and relation chains through contrastive learning.

6 Conclusions

This paper proposes a contrastive semantic similarity learning method for complex question answering over event-centric knowledge graphs. The proposed method consists of three components, including the relation chain generation, semantic similarity learning and similarity score calculation. To generate candidate relation chains, the retrieval subgraph is first constructed based on the identified topic event or entity. To better distinguish the semantic similarity between questions and candidate relation chains, the contrastive learning framework is adopted to learn a common semantic space. Experimental results on the EventQA dataset demonstrate the effectiveness of our proposed method.

Acknowledgements

References

[1] D. Fensel, U. Şimşek, K. Angele, E. Huaman, E. Kärle, O. Panasiuk, I. Toma, J. Umbrich, A. Wahler, Introduction: what is a knowledge graph?, in: Knowledge Graphs, Springer, 2020, pp. 1-10.

[2] X. Bi, H. Nie, X. Zhang, X. Zhao, Y. Yuan, G. Wang, Unrestricted multi-hop reasoning network for interpretable question answering over knowledge graph, Knowl Based Syst, 243 (2022).

[3] Y. Yan, R. Li, S. Wang, H. Zhang, D. Zan, F. Zhang, W. Wu, W. Xu, Large-Scale Relation Learning for Question Answering over Knowledge Bases with Pre-trained Language Models, in, Association for Computational Linguistics (ACL), 2021, pp. 3653-3660.

[4] J. Zhang, L. Zhang, B. Hui, L. Tian, Improving complex knowledge base question answering via structural information learning, Knowl Based Syst, 242 (2022).

[5] S. Gottschalk, E. Demidova, EventKG: a multilingual event-centric temporal knowledge graph, in: European Semantic Web Conference, Springer, 2018, pp. 272-287.

[6] H. Zhang, X. Liu, H. Pan, Y. Song, C.W.K. Leung, ASER: A Large-scale Eventuality Knowledge Graph, in, Association for Computing Machinery, Inc, 2020, pp. 201-211.

[7] L. Biermann, S. Walter, P. Cimiano, A guided template-based question answering system over knowledge graphs, in: O. Corby, P. Cimiano (Eds.), CEUR-WS, 2018, pp. 61-64.

[8] Wahyudi, M.L. Khodra, A.S. Prihatmanto, C. Machbub, A Question Answering System Using Graph-Pattern Association Rules (QAGPAR) on YAGO Knowledge Base, in, Institute of Electrical and Electronics Engineers Inc., 2018, pp. 536-541.

[9] R.G. Athreya, S.K. Bansal, A.C.N. Ngomo, R. Usbeck, Template-based Question Answering using Recursive Neural Networks, in, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 195-198.

[10] A.C.O. Reddy, K. Madhavi, Convolutional recurrent neural network with template based representation for complex question answering, Int. J. Electr. Comput. Eng., 10 (2020) 2710-2718.

[11] Y. Lan, G. He, J. Jiang, J. Jiang, W.X. Zhao, J.R. Wen, A Survey on Complex Knowledge Base Question Answering: Methods, Challenges and Solutions, in: Z.H. Zhou (Ed.), International Joint Conferences on Artificial Intelligence, 2021, pp. 4483-4491.

[12] L.J. Gao, W. Zhao, J.F. Zhang, B. Jiang, G2S: Semantic Segment Based Semantic Parsing for Question Answering over Knowledge Graph, Tien Tzu Hsueh Pao, 49 (2021) 1132-1141.

[13] Z. Liang, Z. Peng, X. Yang, F. Zhao, Y. Liu, D.L. McGuinness, BERT-based Semantic Query Graph Extraction for Knowledge Graph Question Answering, in: O. Seneviratne, C. Pesquita, J. Sequeda, L. Etcheverry (Eds.), CEUR-WS, 2021.

[14] D. Sorokin, I. Gurevych, Modeling semantics with gated graph neural networks for knowledge base question answering, in: E.M. Bender, L. Derczynski, P. Isabelle (Eds.), Association for Computational Linguistics (ACL), 2018, pp. 3306-3317.

[15] Y. Chen, H. Li, DAM: Transformer-based relation detection for Question Answering over Knowledge Base, Knowl Based Syst, 201-202 (2020).

[16] Z. Wang, P. Ng, R. Nallapati, B. Xiang, Retrieval, re-ranking and multi-task learning for knowledge-base question answering, in, Association for Computational Linguistics (ACL), 2021, pp. 347-357.

[17] Y. Lan, S. Wang, J. Jiang, Multi-hop knowledge base question answering with an iterative sequence matching model, in: J. Wang, K. Shim, X. Wu (Eds.), Institute of Electrical and Electronics Engineers Inc., 2019, pp. 359-368.

[18] A.v.d. Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, arXiv preprint arXiv:1807.03748, (2018).

[19] T. Gao, X. Yao, D. Chen, SimCSE: Simple Contrastive Learning of Sentence Embeddings, in, Association for Computational Linguistics (ACL), 2021, pp. 6894-6910.

[20] T. Souza Costa, S. Gottschalk, E. Demidova, Event-QA: A Dataset for Event-Centric Question Answering over Knowledge Graphs, in, Association for Computing Machinery, 2020, pp. 3157-3164.

[21] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980, (2014).

[22] L. Jiang, R. Usbeck, Knowledge Graph Question Answering Datasets and Their Generalizability: Are They Enough for Future Research?, in, Association for Computing Machinery, Inc, 2022, pp. 3209-3218.