COKG-QA: Multi-hop Question Answering over COVID-19 Knowledge Graphs (Postprint)

Authors: Huifang, Du, Zhongwen, Le, Haofen, Wang, Yunwen, Chen, Jing, Yu, Haofen Wang

Date: 2022-11-28T00:00:00+00:00

Abstract

COVID-19 evolves rapidly and an enormous number of people worldwide desire instant access to COVID-19 information such as the overview, clinic knowledge, vaccine, prevention measures, and COVID-19 mutation. Question answering (QA) has become the mainstream interaction way for users to consume the ever-growing information by posing natural language questions. Therefore, it is urgent and necessary to develop a QA system to offer consulting services all the time to relieve the stress of health services. In particular, people increasingly pay more attention to complex multi-hop questions rather than simple ones during the lasting pandemic, but the existing COVID-19 QA systems fail to meet their complex information needs. In this paper, we introduce a novel multi-hop QA system called COKG-QA, which reasons over multiple relations over large-scale COVID-19 Knowledge Graphs to return answers given a question. In the field of question answering over knowledge graph, current methods usually represent entities and schemas based on some knowledge embedding models and represent questions using pre-trained models. While it is convenient to represent different knowledge (i.e., entities and questions) based on specified embeddings, an issue raises that these separate representations come from heterogeneous vector spaces. We align question embeddings with knowledge embeddings in a common semantic space by a simple but effective embedding projection mechanism. Furthermore, we propose combining entity embeddings with their corresponding schema embeddings which served as important prior knowledge, to help search for the correct answer entity of specified types. In addition, we derive a large multi-hop Chinese COVID-19 dataset (called COKG-DATA for remembering) for COKG-QA based on the linked knowledge graph OpenKG-COVID19 launched by OpenKG , including comprehensive and representative information about COVID-19. COKG-QA achieves quite competitive performance in the 1-hop and 2-hop data while obtaining the best result with significant improvements in the 3-hop. And it is more efficient to be used in the QA system for

users. Moreover, the user study shows that the system not only provides accurate and interpretable answers but also is easy to use and comes with smart tips and suggestions.

Full Text

Preamble

DATA PAPER

COKG-QA: Multi-hop Question Answering over COVID-19 Knowledge Graphs

Huifang Du[1], Zhongwen Le[2], Haofen Wang[1]†, Yunwen Chen[3], Jing Yu[3]

[1]College of Design and Innovation, Tongji University, Shanghai 200092, China
[2]School of Computer Science, Fudan University, Shanghai 200433, China
[3]DataGrand Inc., Shanghai 201203, China

Keywords: COVID-19; Question answering; Knowledge graph; Knowledge embedding; Pre-trained model; Multi-hop KGQA

Citation: Du, H.F., et al.: COKG-QA: Multi-hop question answering over COVID-19 knowledge graphs. Data Intelligence 4(3), 471-492 (2022). doi: 10.1162/dint_a_{00154}

Data Citation: Du, H.F., et al.: COKG-QA: Multi-hop question answering over COVID-19 knowledge graphs. Data Intelligence 4(3), 2022. doi: https://doi.org/10.57760/sciencedb.02062

Received: Jan. 4, 2022; **Revised:** Apr. 10, 2022; **Accepted:** May 1, 2022

ABSTRACT

COVID-19 evolves rapidly, and an enormous number of people worldwide desire instant access to COVID-19 information such as overviews, clinical knowledge, vaccines, prevention measures, and COVID-19 mutations. Question answering (QA) has become the mainstream interaction method for users to consume ever-growing information by posing natural language questions. Therefore, it is urgent and necessary to develop a QA system to offer consulting services around the clock to relieve the stress on health services. In particular, people increasingly pay more attention to complex multi-hop questions rather than simple ones during this lasting pandemic, but existing COVID-19 QA systems fail to meet these complex information needs. In this paper, we introduce a novel multi-hop QA system called COKG-QA, which reasons over multiple relations in large-scale COVID-19 Knowledge Graphs to return answers given a question. In the field of question answering over knowledge graphs, current methods usually represent entities and schemas based on knowledge embedding models and represent questions using pre-trained models. While it is convenient to represent different knowledge (i.e., entities and questions) using specified embeddings, an issue arises that these separate representations come from heterogeneous vector spaces. We align question embeddings with knowledge embeddings in a common semantic space through a simple but effective embedding projection mechanism. Furthermore, we propose combining entity embeddings with their corresponding schema embeddings, which serve as important prior knowledge, to help search for the correct answer entity of specified types. In addition, we derive a large multi-hop Chinese COVID-19 dataset (called COKG-DATA for reference) for COKG-QA based on the linked knowledge graph OpenKG-COVID19 launched by OpenKG, including comprehensive and representative information about COVID-19. COKG-QA achieves quite competitive performance on 1-hop and 2-hop data while obtaining the best result with significant improvements on 3-hop questions. Moreover, it is more efficient for use in QA systems for users. The user study shows that the system not only provides accurate and interpretable answers but is also easy to use and comes with smart tips and suggestions.

1. INTRODUCTION

The serious situation of COVID-19 is ongoing. By January 16, 2022, more than 5.54 million people had died from the pandemic, raising increasing anxiety about health problems among individuals. The pandemic has severely affected people' s lives, and people dramatically demand accurate, efficient, and instant access to epidemic information. However, the vast amount of COVID-19 information on various websites is not well organized and not specialized for the general public. Question Answering systems based on COVID-19 knowledge have become a convenient interaction method popular among more and more people. There are two existing paradigms for COVID-19 QA: Information Retrieval Question Answering (IRQA) and Question Answering over Knowledge Graph (KGQA). IRQA systems for COVID-19 are based on textual question-answer pairs [1, 2, 3, 4], retrieving answers by computing similarity between the asked question and questions/answers in the dataset. IRQA systems can naturally answer simple questions that people frequently ask.

In contrast, KGQA methods over COVID-19 datasets [5, 6, 7, 8] answer complex questions covering multiple relations over structural KGs. Besides, KGQA techniques can reason for new knowledge in QA tasks. On the other hand, the pandemic has been spreading for a long time, and people now have some basic understanding of COVID-19. As a result, people are no longer satisfied with asking simple questions like "what are the clinical symptoms of patients with COVID-19?" They are more inclined to express complex multi-hop questions, such as the 2-hop question "What are the related diseases having similar symptoms to COVID-19?" and the 3-hop question "how to check the related diseases having similar symptoms to COVID-19?" Therefore, we choose to use multi-hop KGQA techniques to build a COVID-19 QA system.

However, existing KGQA techniques and current COVID-19 KGQA datasets have several limitations. Existing methods [9, 10] often represent knowledge graphs and questions using separate models, raising issues because heterogeneous embeddings from different spaces must be fitted to a common space. Additionally, schemas that define useful, high-level structures of KGs have been neglected in current multi-hop KGQA tasks [11]. Schema information as important prior knowledge can help search for correct entities of specified types. Moreover, public COVID-19 KGs [12, 13, 14] suffer from knowledge sparsity, especially regarding knowledge people would like to ask about daily, which further affects the quality of downstream QA tasks.

In this paper, we improve KGQA performance by proposing COKG-QA: multi-hop Question Answering over COVID-19 Knowledge Graphs. COKG-QA proposes several improvements to address these constraints. The architecture of our system is illustrated in Figure 1 [Figure 1: see original paper], and the main contributions of our paper are as follows:

1) We introduce COKG-QA to demonstrate the importance of the embedding projection mechanism and schema information in multi-hop KGQA tasks. More precisely, embeddings of entities, schemas, and questions from different spaces are transferred into one common space through a projection method to align important features. Furthermore, entity embeddings are incorporated with their type embeddings to predict answers of specified types.

2) There rarely exist comprehensive KGQA datasets managed for COVID-19, especially lacking multi-hop questions. Benefiting from OpenKG-COVID19 [15], we derive a large multi-hop Chinese COVID-19 KGQA dataset, COKG-DATA. It consists of abundant knowledge, providing an important foundation for building a superior question answering system.

3) Experiments in the paper prove that COKG-QA is of high quality and also robust enough to generalize to new knowledge. To facilitate people' s demand for COVID-19 consulting services, we develop a user-friendly interactive application based on COKG-QA. The application not only provides accurate and interpretable answers but is also easy to use and has functions for smart tips and recommendations.

2. RELATED WORK

At the moment of the epidemic, researchers have released relevant datasets and built question answering systems based on natural language processing techniques to help people conveniently obtain information about COVID-19. We introduce these significant efforts and KGQA techniques in this section.

COVID-19 datasets and QA systems: Some useful KGs have been launched to advance COVID-19 research during the ongoing pandemic. However, the published COVID-19 KGs have limited data size and are more academically medical, making them not applicable for users' daily consulting needs. For example, the coronavirus Knowledge Graph [13] has 27 relations and limited entity types.

Publications, case statistics, and molecular data are structured [12] to explore biomedical knowledge, such as specific genes and proteins. KG-COVID-19 [14] also focuses on SARS-CoV-2 and COVID-19 related heterogeneous biomedical data to construct KGs. Based on these public KGs, some KGQA systems have been developed. A template matching method [7] using the Naive Bayes algorithm over a KG is adopted to establish a COVID-19 QA system. A QA system like [16] employs a rule-based classifier for recognizing users' intentions and also adopts templates to parse natural questions from users. To make the framework not limited to predefined rules, some work like [17] introduces a relatively general framework based on the knowledge embedding method TransE [18]. Although these QA systems are developed for COVID-19, they fail to provide optimal performance for users' diverse questions.

KGQA: There are many state-of-the-art KGQA methods, and we briefly review these three types [19]: (1) logic-based methods; (2) path-based methods; (3) embedding-based methods. Logic-based methods are widely discussed due to their advantages of high accuracy and strong interpretability. GQE (Graph Query Embedding) [20], Query2Box [21], and BetaE [22] represent the query as a directed acyclic computational graph to generate logic form query embedding. Path-based methods take the topic entity in the question to search along multiple triples of the KG to find the answer entity or relation. To alleviate the issue that the search space of the Path-Ranking Algorithm [23] is large, DeepPath [24] allows the path attributes to be controllable. A teacher-student network is adopted in NSM [25] to learn intermediate supervision signals. Some other works like [26, 27] regard KG reasoning as a sequential path decision process. Embedding-based methods [11, 28] measure the similarity between question embeddings and candidate answer embeddings to get the right answer. For example, the state-of-the-art method EmbedKGQA represents questions by a pre-trained model and represents knowledge graph embeddings by ComplEx [29], selecting answers through the score function of ComplEx. The Relational Graph Convolutional Networks (R-GCN) method [30] aggregates embeddings of specific multiple relations in KG to predict answers. Research that incorporates KGs with text corpus based on embedding methods [9, 10, 31] also attracts much attention.

3. TASK DEFINITION

To alleviate people' s anxiety about health problems caused by the COVID-19 pandemic, we are determined to develop an effective KGQA system focusing on complex multi-hop questions. In addition, the functions of smart tips and recommendations make the QA system consumer-friendly. Considering that questions tend to be asked daily, data derived from OpenKG-COVID19 will be curated elaborately and finally formed into the multi-hop KGQA dataset, COKG-DATA. Moreover, we propose COKG-QA extending the state-of-the-art EmbedKGQA model with simple and efficient modifications so that it can achieve superior and practical performance for the QA system. In the following sections, we first describe the extension in COKG-QA and the details of COKG-DATA. Based on these two modules, the performance of the KGQA system will be demonstrated finally.

4. COKG-QA

As mentioned in the related work, EmbedKGQA [11] is a good work considering multi-hop reasoning. We extend it in several aspects to achieve better performance in terms of accuracy and coverage in the context of COVID-19 question answering. We first give a brief introduction of EmbedKGQA and then describe our improvements in detail in the following subsections.

4.1 Preliminary

An instance triple in a KG can be represented as h, r, t , where h represents the head entity and t represents the tail entity linked by relation r. Given a set of entities E and relations R, a Knowledge Graph G is a set of triples K such that K E × R × E. The KGQA task searches for an answer entity for a natural language question q including multi-hop relations over a KG. Inspired by EmbedKGQA, we also employ KG Embedding Module, Question Embedding Module, and Answer Selection Module in our method. In this paper, we extend EmbedKGQA over COKG-DATA by adding Embedding Projection and Schema-Aware Module. In addition, we also add a Topic-Entity-Aware Filter at inference to predict answer entities only related to the topic entity in the question. The architecture can be seen in Figure 2 [Figure 2: see original paper]. Details are described as follows.

4.2 Embedding Projection

We regard embeddings generated by different models as heterogeneous. Like triples at the instance level, s_h, r, s_t is a triple at the schema level, where s_h represents the head type and s_t stands for the tail type linked by relation r. Schema embeddings of s_h, s_t E' are also trained by the ComplEx [29] method to enhance searching for answers, but the schema model and instance model are trained separately. Moreover, question embedding is produced by the pre-trained model RoBERTa [32], which leverages quite another technical paradigm. Therefore, these three embeddings are heterogeneous. Even though it helps to maintain their characteristics of schema, instance, and question through separate models, it is hard to model representations in the final KGQA model. Fully Connected (FC) linear layers like "firewalls" can maintain and project important features in transfer learning [33], especially when the source domain and target domain are quite different. Therefore, it is reasonable to project these embeddings before being transferred into one common space. We respectively define question embedding, entity embedding, and schema embedding.

4.3 Schema-Aware Module

Existing KGQA methods only focus on instance facts in the KG, which ignores the well-constructed prior knowledge in the schema. The schema contains valuable structure information of a KG, which defines concepts and properties of these concepts. Entities in KG are linked to their corresponding concepts by entity types [34, 35]. We add a Schema-Aware Module by combining entity embedding with corresponding entity type embedding, which will be helpful to filter answer entities of specified types. This is sufficient for the model to understand which type the topic entity is and which type the answer entity will be.

Specifically, the topic entity representation in the question and the tail entity representation as the answer are constructed by adding the corresponding entity type embedding. Question representation embedded using RoBERTa cannot encode relation embedding at the schema level because there is no relation type label for questions in real applications. However, we concatenate entity type with the given question to imply that the question is relevant to a certain entity type, as shown in the input in Figure 2.

4.4 Topic-Entity-Aware Filter

Because the COKG-DATA we collect is very large, it is necessary to add a filter to get entities related to the topic entity, including 1-hop, 2-hop, and 3-hop entities at inference, like EmbedKGQA, to predict more relevant answer entities. We first create a map between topic entities and their multi-hop entities within 3 hops, and then we predict answers among these multi-hop entities based on the best-trained model.

5. COKG-DATA CONSTRUCTION

Existing COVID-19 QA systems fail to perform complex reasoning with a non-KG dataset. We organize COKG-DATA based on seven sub-KGs (i.e., encyclopedia, prevention, goods, medical, epidemiology, health, character) of OpenKG-COVID19 launched by OpenKG, which people are more prone to ask about daily. COKG-DATA is a new challenging question-answer benchmark that contains single-hop questions and multi-hop questions concerning diseases, symptoms, drugs, etc. The overview of the selected graphs by COKG-DATA is depicted in Appendix A.1. With the large and diverse COKG-DATA, multi-hop KGQA is an appealing and useful task to satisfy people’ s complex query needs during the pandemic. We spend considerable time cleaning data based on OpenKG-COVID19 and collecting multi-hop questions. Details are shown in A.2.

5.1 Human Check

To ensure that the questions in COKG-DATA are natural and meaningful, we recruited four volunteers whose research fields are all Knowledge Graph and

Question Answering to check the quality of the dataset. We obtained random samples in proportion to the number of questions sorted by each relation defined in the cleaned OpenKG-COVID19. These four volunteers were asked to rate the sampled questions with three choices: 1 for Weird; 2 for Natural; 3 for Meaningful. We optimized COKG-DATA four times by removing or modifying the weird question-answer pairs through the scoring process. The sampled number for the last round is 4,000, and the average score by volunteers is 2.8, demonstrating the high quality of COKG-DATA. The final statistics for each hop of questions in COKG-DATA are shown in Table 1 . COKG-DATA will keep up with OpenKG-COVID19 to update for more sufficient knowledge for users.

6. EXPERIMENTS

In this section, we first present the experimental setup, the COKG-QA results on COKG-DATA, and then analyze answer errors.

6.1 Experimental Settings

We follow the same split proportion (i.e., 3:1:1) of train/validation/test for all datasets of 1-hop, 2-hop, and 3-hop questions. The number of each hop of questions is summarized in Table 1. We choose batch sizes of 90, 64, 32 and corresponding learning rates of 5e-5, 2e-5, 1e-6 for training the model across 2 NVIDIA RTX2080ti GPUs. Additionally, we set the patience number as 10, meaning that training will stop when the accuracy score has not improved for ten consecutive checks, and the maximum epoch limit is 100. ComplEx embedding was obtained based on OpenKE, and the dimension of ComplEx embedding and question embedding in COKG-QA are both 400. Weight decay, as a popular and necessary regularization technique, was set as $1e{-}1$.

6.2 Baselines

We compare our model with two state-of-the-art models, including EmbedKGQA [11] and TransferNet [28]. Since EmbedKGQA reasons about answers through link prediction, which can alleviate the KG incompleteness problem and avoid the problem of uneven data distribution, we take extensions over it in our implementations. TransferNet is an effective method and competitive enough as a baseline, achieving best performance on public multi-hop datasets such as MetaQA [41], WebQSP [42], and CompWebQ [43].

EmbedKGQA [11] regards the multi-hop KGQA task as link prediction and searches for answer entities based on question embedding and knowledge embeddings, which mitigates the problem of KG incompleteness and can predict answers among unlimited neighbors.

TransferNet [28] proposes a unified architecture for label and text data. In this framework, TransferNet calculates the relations corresponding to different positions of the question under an attention mechanism at each step and further gets the answer entity.

6.3 Main Results

In Table 2 , we compare EmbedKGQA and TransferNet with COKG-QA on our COKG-DATA datasets. COKG-QA performs better than EmbedKGQA in all hop data, while TransferNet outperforms COKG-QA in 1-hop and 2-hop questions. However, TransferNet obtains the lowest accuracy in 3-hop questions. TransferNet attends to different parts of the question to search for the corresponding relation at each step, which makes it sensitive to both the quality and quantity of each-hop relations in the graph. Therefore, we assume that the small amount of 3-hop data in COKG-DATA causes the poor performance for TransferNet. However, both EmbedKGQA and COKG-QA regard the multi-hop KGQA task as link prediction, which takes a multi-hop relation as a single relation in the KG Embedding Module. For example, each relation of "complication||commonly used medicine||usage and dosage" , "medication||medication ingredient" , and "precaution" is equally seen as a single relation to put in one triple. So COKG-QA avoids the problem of data imbalance, which is very common in the real world and poses challenges to neural models. Moreover, TransferNet has high computational complexity and large memory storage problems because it computes the probability of an entity being activated as the answer entity multiple times, which also affects inference speed.

6.4 Ablation Studies

Table 4 shows ablation studies of the effects of adding the Schema-Aware Module, adding Embedding Projection, and Topic-Entity-Aware Filter. We demonstrate the importance of each improvement by leveraging the same training set, validation set, test set, and hyperparameters. We briefly analyze the effect of each component in this section.

6.4.1 Effect of Embedding Projection Since all entity embeddings are frozen during COKG-QA training, as EmbedKGQA does, the features of entity embeddings are quite different from question embeddings. Besides, entity embeddings and type embeddings are also learned from different trained models. Therefore, it is necessary to build a projection to transform these important features from different vector spaces into a common vector space. The comparison results of projection (in the COKG-QAep row) and without projection can be seen in Table 3 . Although Embedding Projection does not provide as much improvement as the Schema-Aware Module, the 2.61% absolute improvement in 1-hop questions and slightly better performance in other questions demonstrates that Embedding Projection advances the capability of COKG-QA compared to EmbedKGQA.

6.4.2 Effect of Schema-Aware Module We concatenate entity embedding with the corresponding entity type embedding to build a contextual KG embedding for COKG-QA. Furthermore, an ablation test was performed to evaluate the effect of only the Schema-Aware Module. The results listed in Table 3 marked by COKG-QAsam show that the Schema-Aware Module leads to better performance with an average increase of 1.82%, which indicates the effectiveness of enriching entity embedding by adding schema information.

6.4.3 Effect of Topic-Entity-Aware Filter To select an answer entity within the range of the 3-hop neighborhoods of the topic entity, the filter could competitively deliver better inference results with more than a 10% increase, which further ensures the provision of a robust QA system on COKG-DATA.

6.5 Answer Analysis

To fully analyze the experimental results, we collect all wrong answers from the test set to try to find useful patterns. Through observation, we find that wrong samples containing digital numbers (in digit or word form) account for 33.92%. Additionally, about 11.94% of entities in the selected sub-graphs include numbers, which is not a negligible data size. Numerical reasoning or discrete reasoning is a more challenging task [36] with only question-answer pair supervision. Therefore, we experimented with two types of data: numerical question-answer pairs (inserted with numbers) and non-numerical question-answer pairs. We also tested their corresponding 2-hop and 3-hop questions. Table 4 shows the results for different types of datasets using our model.

Numerical Data Analysis. COKG-QA over only numerical data reaches 49.19% hits@1 in the 1-hop data with a 30.87% absolute decrease compared to the model with all data (without Topic-Entity-Aware Filter condition). This highlights the fact that it is harder to model text with numbers. Besides, entity and relation distributions in the numerical dataset are also observed and show that uneven distributions may be another key factor for the worse performance. The right histogram in Figure 3 [Figure 3: see original paper] shows entities and relations distributions of Numerical Data.

Non-numerical Data Analysis. As expected, non-numerical data with large samples is still hard to optimize because non-numerical data accounts for the majority and the distribution of the non-numerical dataset is similar to all data. However, without numerical problems, the experimental results of non-numerical data are better than those of all data. The left histogram in Figure 3 presents the visualization of COKG-DATA distribution according to the first 30 multi-hop relations sorted by entity number. We can see that both numerical and non-numerical data have long-tail data problems, for which data augmentation to compensate [37] or enhancing the recall of long-tail entities [38] are directions that can be considered.

7. COKG-QA PERFORMANCE

The superior performance of COKG-QA illustrated by the extensive experiments above promises an effective QA system. Therefore, we devise an interactive Web QA application based on COKG-QA for people. A friendly QA system design can improve user experience [39, 40]. We discuss the design considerations of the QA application in this section.

7.1 Interpretability of Answers

Unlike most KGQA systems that give direct answers, our system explains the intermediate context for multi-hop questions to make the answers interpretable. An answer is inferred based on the best-trained model by computing the ComplEx score. However, the answer based on the EmbedKGQA model is not understandable. For example, the answer to the 2-hop question "What are the types of drugs recommended for pediatric intracranial tumors" is "Chemical drugs, prescription drugs, and medical insurance drug for work-related injury," which would pose users a question like "what are the respective recommended drugs corresponding to the drug types mentioned in the answer above?" In other words, people not only want to achieve the final answer but also want to figure out what the intermediate results are. Therefore, we offer an interpretable answer: "The recommended drug for pediatric intracranial tumors, glycerol fructose injection, is a chemical drug; the recommended drug for pediatric intracranial tumors, piracetam glucose injection, is a medical insurance work injury drug···" The process for the interpretable response is as follows: (1) When the QA system gets a multi-hop question, the topic entity will be recognized first. (2) Subsequently, the indirect tail answer is obtained by ranking scores based on the question and the recognized head. (3) To get an interpretable final answer, we need to search out the intermediate relations and get intermediate entities. Questions and corresponding multi-hop relations having the same head and answer labeled in the dataset are filtered out. Furthermore, we select the interpretable answer corresponding to the question in the dataset that has the same multi-hop relations or is most similar to the user' s question as the final response.

7.2 Sources of Answers

We provide the sources of answers with corresponding URLs to help users trace the context, which also increases the credibility of the system. The answer sources in our system give evidence by offering graph names in selected subgraphs. Multiple graph names are shown if the user' s question covers multiple linked graphs. An example can be seen in Figure 4 [Figure 4: see original paper].

7.3 User Feedback

We design thumbs-up and thumbs-down buttons to encourage users to provide feedback, which will be used to improve the COKG-QA model. When users give positive feedback, the system will randomly generate a thank-you sentence.

When users thumb down, a bubble will pop up and display three options: Incorrect answer, Incomplete answer, and Customized opinions. The custom options provide space for users to flexibly provide suggestions that further benefit the improvement of the QA system' s effectiveness.

7.4 Ease of Use

Many medical terms are uncommon or difficult for users to remember, such as disease names and treatments. The automatic input prompt function is significant and practical for improving the usability of the system. Our system supports autocompletion in many scenarios. For example, users can use a single word, pinyin, first letters of multiple words, or even fuzzy search. Tips in the input box can expand the focus of users' queries to help complete questions that users want to ask, as shown in Figure 5 [Figure 5: see original paper]. Besides, our system can also recommend questions relevant to the topic entity, which allows users to explore more about the original question.

8. CONCLUSIONS

In this paper, we introduce a multi-hop KGQA method named COKG-QA to develop a QA system for COVID-19 consulting services and meet people' s tailored medical information needs. Multi-hop KGQA techniques have attracted increasing attention from researchers due to their ability to handle complex multi-hop questions and reasoning. We extend the state-of-the-art method EmbedKGQA by adding Embedding Projection and Schema-Aware Module. EmbedKGQA represents knowledge graph embeddings based on ComplEx and represents questions using RoBERTa. Although it is reasonable and convenient to represent different specified embeddings, these representations come from heterogeneous vector spaces, which influences optimal performance. We adapt the important features of questions and knowledge embeddings from different spaces into a common semantic one by adopting an embedding projection mechanism. Moreover, current KGQA methods ignore schema implications for entity representation. COKG-QA learns entity embeddings by summing their corresponding type information to help search for the right answer entity of specified types. To ensure superior performance, we also add a Topic-Entity-Aware Filter to select the answer from the topic entity' s neighbor entities within the 3-hop relation range. Furthermore, we publish a large multi-hop Chinese COVID-19 KGQA dataset COKG-DATA under the open license CC BY SA to provide a comprehensive knowledge foundation for COKG-QA. Extensive experimental results show that COKG-QA is robust as a QA engine and can further generalize to new fields. Based on COKG-QA, we also develop a user-friendly interactive application. The application can generate interpretable answers and is easy to use with functions for smart tips and recommendations.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available in the ScienceDB repository: https://doi.org/10.57760/sciencedb.02062. To reuse the data, please cite the data as: Du, H.F., et al.: COKG-QA: Multi-hop question answering over COVID-19 knowledge graphs. Data Intelligence 4(3), 2022. doi: https://doi.org/10.57760/sciencedb.02062.

ACKNOWLEDGEMENTS

This work was supported by the Fundamental Research Funds for the Central Universities with grant Nos. 22120220069 and the National Nature Science Foundation of China with Grant No. 62176185, and was supported in part by the Shanghai Artificial Intelligence Innovation and Development Fund grant 2020-RGZN-02026.

AUTHOR CONTRIBUTIONS

Du H.F. (duhuifang@tongji.edu.cn), Wang H.F. (carter.whfcarter@gmail.com) designed the model architecture. Le Z.W. (20210240064@fudan.edu.cn) participated in the discussion about the task definition of the project and led the implementation of the collection of COKG-DATA and the training of COKG-QA with Du H.F. Du H.F. also developed the Web application. Chen Y.W. (chenyunwen@datagrand.com) and Yu J. (yujing@datagrand.com) added the contrast experiments and made result analysis. All the authors have made valuable contributions to writing and revising the manuscript.

REFERENCES

[1] Zhang, Y., et al.: WULAI-QA: Web understanding and learning with AI towards document-based question answering against COVID-19. In: Proceedings of the 14th ACM International Conference on Web Search and Data Mining, pp. 898–901 (2021)

[2] Su, D., et al.: CAiRE-COVID: A question answering and query-focused multi-document summarization system for covid-19 scholarly information management. InProceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP (2020)

[3] Moller, T., et al.: COVID-QA: A question answering dataset for COVID-19. InProceedings of the 1st Workshop on NLP for COVID-19 at ACL (2020)

[4] Lee, J., et al.: Answering questions on COVID-19 in real-time. InProceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP (2020)

[5] Ding, K., et al.: Research on question answering system for COVID-19 based on knowledge graph. In 40th Chinese Control Conference, pp. 4659–4664 (2021)

[6] Michel, F., et al.: Covid-on-the-Web: Knowledge graph and services to advance COVID-19 research. International Semantic Web Conference, pp. 294–310 (2020)

[7] Sun, Y., et al.: The COVID-19 question answering system based on knowledge graph. In IEEE/ACIS 20th International Fall Conference on Computer and Information Science, pp. 215–220 (2021)

[8] He, L., et al.: Optimizing automatic question answering system based on disease knowledge graph. Data Analysis and Knowledge Discovery 5(5), 115–26 (2021)

[9] Sun, H., et al.: Open domain question answering using early fusion of knowledge bases and text. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (2018)

[10] Sun, H., Bedrax-Weiss, T., Cohen, W.: PullNet: Open domain question answering with iterative retrieval on knowledge bases and text. InProceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 2380–2390 (2019)

[11] Saxena, A., Tripathi, A., Talukdar, P.: Improving multi-hop question answering over knowledge graphs using knowledge base embeddings. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4498–4507 (2020)

[12] Domingo-Fernandez, D., et al.: COVID-19 Knowledge Graph: a computable, multi-modal, cause-and-effect knowledge model of COVID-19 pathophysiology. Bioinformatics 37(9), 1332–4 (2021)

[13] Zhang, P., et al.: Toward a coronavirus knowledge graph. Genes 12(7), 998 (2021)

[14] Reese, J.T., et al.: KG-COVID-19: A framework to produce customized knowledge graphs for COVID-19 response.

[15] Wang, H., et al.: Construction of A Linked Dataset of COVID-19 Knowledge Graphs: Development and Applications. JMIR Medical Informatics 26(04), 37215 (forthcoming/in press) (2022)

[16] Ding, K., et al.: Research on question answering system for COVID-19 based on knowledge graph. In 40th Chinese Control Conference, pp. 4659–4664 (2021)

[17] Pei, Z., et al.: A general framework for Chinese domain knowledge graph question answering based on TransE. InJournal of Physics: Conference Series 1693(1), 012136 (2020)

[18] Bordes, A., et al.: Translating embeddings for modeling multi-relational data. Advances in neural information processing systems 26 (2013)

[19] Du, H., et al.: Progress, challenges and research trends of reasoning in multi-hop knowledge graph based question answering. Big Data Research 7(3), 2021026 (2021)

[20] Hamilton, W.L., et al.: Embedding logical queries on knowledge graphs. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, pp. 2030–2041 (2018)

[21] Ren, H., Hu, W., Leskovec, J.: Query2box: Reasoning over knowledge graphs in vector space using box embeddings. InInternational Conference on Learning Representations (2018)

[22] Ren, H., Leskovec, J.: Beta embeddings for multi-hop logical reasoning in knowledge graphs. Advances in Neural Information Processing Systems, p. 33 (2020)

[23] Gardner, M., Talukdar, P., Kisiel, B., Mitchell, T.: Improving learning and inference in a large knowledge-base using latent syntactic cues. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 833–838 (2013)

[24] Xiong, W., Hoang, T., Wang, W.Y.: DeepPath: A reinforcement learning method for knowledge graph reasoning. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 564–573 (2017)

[25] He, G., et al.: Improving multi-hop knowledge base question answering by learning intermediate supervision signals. In Proceedings of the 14th ACM International Conference on Web Search and Data Mining, pp. 553–561 (2021)

[26] Meilicke, C., Chekol, M.W., Ruffinelli, D., Stuckenschmidt.: Anytime bottom-up rule learning for knowledge graph completion. International Joint Conference on Artificial Intelligence, pp. 3137–3143 (2019)

[27] Lin, X.V., Xiong, C., Socher, R., Stuckenschmidt.: Multi-hop knowledge graph reasoning with reward shaping. United States Patent Application 16(051), 309 (2019)

[28] Shi, J., et al.: TransferNet: An Effective and Transparent Framework for Multi-hop Question Answering over Relation Graph. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 4149–4158 (2021)

[29] Trouillon, T., et al.: Complex embeddings for simple link prediction. Proceedings of the 33rd International Conference on International Conference on Machine Learning 48, 2071–2080 (2016)

[30] Dong, L., Wei, F., Zhou, M., Xu, K.: Question answering over Freebase with multi-column convolutional neural networks. InProceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Volume 1: Long Papers, pp. 260–269 (2015)

[31] Bordes, A., Weston, J., Usunier, N.: Open question answering with weakly supervised embedding models. InJoint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 165–180 (2014)

[32] Liu, Y., et al.: RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019)

[33] Zhang, C.L., et al.: In defense of fully connected layers in visual representation transfer. InPacific Rim Conference on Multimedia, pp. 807–817 (2017)

[34] Wang, P., Zhou, J., Liu, Y., Zhou, X.: TransET: Knowledge graph embedding with entity types. Electronics 10(12), 1407 (2021)

[35] Moon, C., Jones, P., Samatova, N.F.: Learning entity type embeddings for knowledge graph completion. InProceedings of the 2017 ACM on Conference on Information and Knowledge Management, pp. 2215–2218 (2017)

[36] Saxton, D., Grefenstette, E., Hill, F., Kohli, P.: Analysing mathematical reasoning abilities of neural models. In International Conference on Learning Representations (2018)

[37] Wang, Z., et al.: Tackling long-tailed relations and uncommon entities in knowledge graph completion. InProceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 250–260 (2019)

[38] Yamada, I., et al.: LUKE: Deep contextualized entity representations with entity-aware self-attention. InProceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pp. 6442–6454 (2020)

[39] Vtyurina, A., Savenkov, D., Agichtein, E., Clarke, C.L.: Exploring conversational search with humans, assistants, and wizards. InProceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems, pp. 2187–2193 (2017)

[40] Podgorny, I., Khaburzaniya, Y., Geisler, J.: Conversational agents and community question answering. In Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (2019)

[41] Zhang, Y., et al.: Variational reasoning for question answering with knowledge graph. In Thirty-Second AAAI Conference on Artificial Intelligence (2018)

[42] Yih, W.T., et al.: The value of semantic parse labeling for knowledge base question answering. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 201–206 (2016)

[43] Talmor, A., Berant, J.: The Web as a Knowledge-Base for Answering Complex Questions. InProceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 641–651 (2018)

APPENDICES

A. Details of COKG-DATA

A.1 Overview of COKG-DATA We elaborately select seven sub-graphs that contain topics people are more concerned about during the COVID-19 epidemic. The specific graphs selected by COKG-DATA are demonstrated as follows.

The **encyclopedia KG** gives us a general understanding of SARS-CoV-2 and COVID-19, along with relevant viruses and diseases information.

The **prevention KG** provides prevention guidance published by the government for individuals and organizations in different places.

The **goods KG** is expanded around materials supply status during the epidemic, covering daily protective equipment, medical devices, and drugs.

The **medical KG** and the **health KG** are complementary to exploit COVID-19 related knowledge about various diseases, drugs, symptoms, examination methods, and hospitals.

The **epidemiology KG** employs general techniques of epidemiology to study the distribution of diseases and influencing factors, exploring the causes of disease and clarifying the laws of epidemics for controlling and eradicating diseases effectively.

The **character KG** records concepts such as characters, battles, achievements during the pandemic, articles, resumes of heroes, etc.

A.2 Data Curation **Data cleaning.** To ensure the quality of the QA dataset, we cleaned some bad cases in OpenKG-COVID19 and removed triples that are not practical for QA: (1) some triples contain empty strings, punctuation entities, or useless numbers; (2) some triples are weird for composing natural questions, e.g., Doctors of Xinhua hospital, work in, Xinhua hospital ; (3) the head entity is the same as the tail entity in some triples, such as triples with "alias" relation. We filter out and remove these bad triples described above. In addition, relation patterns including symmetry and inversion exist in OpenKG-COVID19. We extend triples for these relation patterns of OpenKG-COVID19. After data cleaning and relation extension, the knowledge graph dataset contains 112,246 entities, 209 relations, and 787,056 triples.

Multi-hop Questions Collection. We leverage fact triples in the selected subgraphs of OpenKG-COVID19 as single-hop data. Further, we manually design 47 relations for 2-hop questions and 23 relations for 3-hop questions, in which the combined relations must be reasonable and natural. Specifically, the range of the front relation must be the same as the domain of the back relation in a 2-hop relation. For example, the range of the "selected drug" relation is "drug," which must be consistent with the domain of "usage and dosage" in the 2-hop relation "Selected drug || Usage and dosage." The same rule applies to the 3-hop relations collection process. Similar to the multi-hop dataset MetaQA [41], we employ neural translation models in the Helsinki-NLP Opus-MT project to introduce more diverse and natural statements with the same meaning. The opus-mt-zh-en model is leveraged to translate sentences from Chinese to English, and then opus-mt-en-zh is used to translate back to Chinese. Furthermore, to create a large-scale unified knowledge base from the top level, entity alignment and relation alignment have been completed to eliminate inconsistency problems.

AUTHOR BIOGRAPHY

Huifang Du is a Ph.D student in the College of Design and Innovation, Tongji University. Her research areas are knowledge graph and question answering. ORCID: 0000-0002-5241-7620

Zhongwen Le is a graduate student in the School of Computer Science, Fudan University. His research areas are knowledge representation and question answering. ORCID: 0000-0001-9364-1218

Haofen Wang has long served as Chief Technology Officer in a first-line artificial intelligence company and has rich experience in AI R&D management. He is one of the founders of OpenKG, the world' s largest Chinese open knowledge graph alliance. He is responsible for participating in a number of provincial and ministerial AI-related projects and has published more than 100 high-level papers in the AI field, which have been cited more than 2,300 times and the H-Index has reached 23. He built the world' s first interactive virtual idol— "Amber · Xuyan" ; the intelligent customer service robot he built has served more than 1 billion users. Currently, he is the Deputy Director of the Terminology Committee of China Computer Society, the Deputy Secretary General of the Language and Knowledge Computing Committee of Chinese Information Society, and distinguished research fellow in College of Design Innovation, Tongji University. ORCID: 0000-0003-3018-3824

Yunwen Chen received the Ph.D. degree from Fudan University, Shanghai, China. He is the founder and the CEO of DataGrand Inc., Shanghai, a leading AI company in China. He had been the Chief Data Officer of Shanda, Inc., Burlington, IA, USA, the Senior Director of Tencent, Inc., Shenzhen, China, and a Researcher of Baidu, Inc., Beijing, China. He has 32 patents and several academic publications. His current research interests include data mining, natural language processing, search and recommend systems, and knowledge graphs. Dr. Chen was a recipient of the Distinguished Graduate Student in 2008. He is a Senior Member of the CCF and a member of the ACM. ORCID: 0000-0003-4513-9439

Jing Yu received the M.S. degree from Tongji University, Shanghai, China. He is currently the co-founder and the department manager of DataGrand Inc. in China. His research interests include search and recommender systems, knowledge graph, natural language processing, and machine learning. He has tens of

patents and several academic publications in artificial intelligence. He is a Senior Member of the CCF and a member of the ACM. ORCID: 0000-0003-3655-2811

Note: Figure translations are in progress. See original paper for figures.