An Effective Framework for Question Answering over Freebase via Reconstructing Natural Sequences

Bin Yue, Min Gui, Jiahui Guo
CCCE&CS
Nankai University, China
{yuebin,nk_guimin,guojiahui}@mail.nankai.edu.cn

Zhenglu Yang*, Jin-Mao Wei
CCCE&CS
Nankai University, China
{yangzl,weijm}@nankai.edu.cn

Shaodi You
Data61-CSIRO Australian National University, Australia
Shaodi.You@data61.csiro.au

ABSTRACT

Question answering over knowledge bases has rapidly developed with the continuous expansion of resources. However, how to match the natural language question to the structured answer entities in the knowledge bases remains a major challenge. In this paper, we propose an effective framework that bridges the gap between the given question and the answer entities, by reconstructing the intermediate natural sequences on the basis of the entities and relations in knowledge bases. The intuitive idea is that these intermediate sequences may encode rich semantic information that can identify the candidate answer entities. Experimental evaluation is conducted on a benchmark dataset *WebQuestions*. Results demonstrate the effectiveness of our proposed framework, i.e., it outperforms state-of-the-art models by up to 6.8% in terms of F1 score.

Keywords

question answering; knowledge base; CNN

1. INTRODUCTION

Knowledge base (KB) based question answering (KB-QA) has attracted considerable attention due to the ubiquity of the World Wide Web and the rapid development of the artificial intelligence (AI) technology. Large scale structured KBs, such as DBpedia, Freebase, and YAGO, provide abundant resources and rich general human knowledge which can be used to respond to users' queries across open domains. However, how to bridge the gap between natural language questions and structured entity data in KBs remains to be a huge challenge. The previous work on KB-QA can be broadly classified into two main categories, namely, semantic parsing based and information retrieval based. These methods commonly transform or interpret a question into a database query depending heavily on the structured information and handcrafted features.

*Corresponding author

WWW'17 Companion, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3054216

We propose a novel framework that establishes the semantic similarity between the question and the structured answer entities in Freebase, by reconstructing natural language sequences on the basis of the entities and relations. We introduce effective deep learning models (i.e., CNN) to calculate the similarities between the intermediate sequences and the question to obtain the final answer. As illustrated in Fig. 1, given a question *when did avatar release in uk*, our framework generates a natural sequence, i.e., *Avatar release date 2009-12-17 United Kingdom* from Freebase. Such sequence not only includes nearly all useful information linked to the topic entity *Avatar* in the question, but also extends the concise answer *2009-12-17* to a more complete sentence that is semantically similar to the question.

2. OUR PROPOSED MODEL

We present a detailed overview of our framework with an concrete example (*When did Avatar release in uk*, *2009-12-17*). Our framework is implemented to find the correct answer in four steps (as shown in Fig. 1).

2.1 Entity Linking and Relation Selection

First, the results obtained from the entity linking tool S-MART [5] and Freebase Search API are combined to link the topic entities in the question (i.e.,*Avatar*) to their corresponding named entities (i.e., m.0bth54) in Freebase. All the entities and relations linked to the named entities in their 2-hops neighbors are collected to construct the subset of Freebase. Subsequently, the candidate answer sets are formed by the relation selection step, which selects the top-N relations in 1-hop neighbors of each named entity, based on the ranked similarity score between the question and the 1-hop relations learned by a CNN model [4]. Through this process, as shown in the bottom of Fig. 1, the relation *film.film.release_date_s* is retained and *film.film.directed_by* is filtered out.

2.2 Natural Sequence Reconstruction

We argue that a 1-hop relation describes an event of the named entity, whereas 2-hop relations and its linked nodes represent different aspects of the event, i.e., time, place or other details. Intuitively, all relevant data are integrated to reconstruct the natural sequences with regard to the manner of human language. The rich semantic information encoded by these sequences facilitates the distinction among candidate answers. Some previous work, i.e., [5], utilizes the structured information of Freebase to obtain answers from human defined aspects such as relation path, answer type,

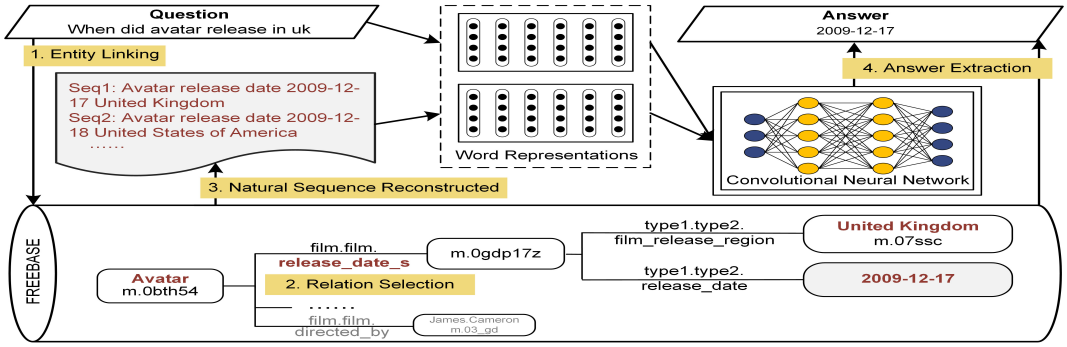

Figure 1: Overview of the proposed framework for the given question *when did Avatar release in UK*; the framework includes four steps to obtain the correct answer *2009-12-17*

and so forth. It retrieves *2009-12-17* and *2009-12-18* as candidate answers without distinction in Fig.1. In contrast, our model accurately obtains *2009-12-17* as the final answer because of the disparity between sequences, i.e., *United Kingdom* in *Seq 1* and *United States of America* in *Seq 2*, which is learned by the CNN model [4]. Overall, our reconstructed sequences bridge the gap between the natural language question and structured answer entities in Freebase.

2.3 Answer Extraction

The answer extraction in our framework is intended to restore the sequence to the corresponding answer entity in Freebase. For example, both *United Kingdom* and *2009-12-17* are named entities in *Seq 1*; thus, the similarity between the question and the 2-hop relation (i.e., type1.type2.release_date) is calculated to select *2009-12-17* as the final answer.

3. EXPERIMENTAL EVALUATION

We evaluate our framework on the *WebQuestions* dataset [1] with regard to the metric of F1 score. As illustrated in Table 1, our model outperforms the state-of-the-art ones by up to 6.8% on F1 score. Previous methods, including both semantic parsing based [1, 2] and information retrieval based [3, 5, 6], analyzes the question based on the structured information in Freebase with additional manually-defined features. By contrast, our model reconstructs the natural sequences from the combination of entities and relations in Freebase, to bridge the gap between the natural language question and the structured answer entities in the Freebase. Moreover, our framework does not require expensive processing (i.e., parsing) or manual features.

We define six strategies to generate sequences, and evaluate on the validation set by 5-fold cross-validation to obtain the optimal method. Intuitively, the sequence combined with strategy *e1+r1+e2+e3* is the most natural one (i.e., *avatar release date 2009-12-17 united kingdom*). As illustrated in Fig. 2, strategy *e1+r1+e2+e3* performs best with the F1 score being 53.9%.

4. CONCLUSIONS

We have proposed a framework for QA over Freebase which bridges the gap between a question and the corresponding structured answer entities, by reconstructing the intermediate sequences. The effectiveness of our model has been experimentally demonstrated on *WebQuestions*.

Table 1: F1 score comparison of our model and state-of-the-art models

Models	Average F1
Berant et al. (2013) [1]	31.4
Yao and Van Durme (2014) [6]	33.0
Borders et al. (2014) [2]	39.2
Dong et al. (2015) [3]	40.8
Xu et al.(2016) (structured+joint) [5]	47.1
This Work	**53.9**

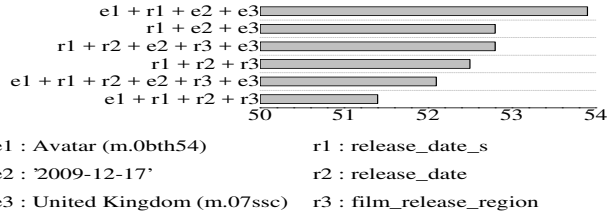

Figure 2: F1 score comparison of different natural sequence reconstructed strategies

Acknowledgments: This research is supported by the National Natural Science Foundation of China: U1636116, 11431006, 61070089, Research Fund for International Young Scientists: 61650110510, National Undergraduate Innovative Experiment Project: 201510055065, Ministry of Education of Humanities and Social Science: 16YJC790123, Natural Science Foundation of Tianjin: 14JCYBJC15700, 15JCYBJC46600.

5. REFERENCES

[1] J. Berant, A. Chou, R. Frostig, and P. Liang. Semantic parsing on freebase from question-answer pairs. In *EMNLP*, 2013.

[2] A. Bordes, S. Chopra, and J. Weston. Question answering with subgraph embeddings. In *EMNLP*, 2014.

[3] L. Dong, F. Wei, M. Zhou, and K. Xu. Question answering over freebase with multi-column convolutional neural networks. In *ACL*, 2015.

[4] A. Severyn and A. Moschitti. Learning to rank short text pairs with convolutional deep neural networks. In *SIGIR*, 2015.

[5] K. Xu, S. Reddy, Y. Feng, S. Huang, and D. Zhao. Question answering on freebase via relation extraction and textual evidence. In *ACL*, 2016.

[6] X. Yao and B. V. Durme. Information extraction over structured data: Question answering with freebase. In *ACL*, 2014.