A Relateness-Based Ranking Method for Knowledge-Based Question Answering

Han Ni[1], Liansheng Lin[2], and Ge Xu[3]

[1] NetDragon Websoft Inc., Fuzhou, China
nihan@nd.com.cn
[2] NetDragon Websoft Inc., Fuzhou, China
linliansheng@nd.com.cn
[3] Minjiang University, Fuzhou, China
xuge@pku.edu.cn

Abstract. In this paper, we report technique details of our approach for the NLPCC 2018 shared task knowledge-based question answering. Our system uses a word-based maximum matching method to find entity candidates. Then, we combine editor distance, character overlap and word2vec cosine similarity to rank SRO triples of each entity candidate. Finally, the object of the top 1 score SRO is selected as the answer of the question . The result of our system achieves 62.94% of answer exact matching on the test set.

Keywords: Question answer, Knowledge base, Entity linking, Relation Ranking

1 Introduction

Automatic open-domain quesion answering has attracted great attention with the development of Natural Language Processing (NLP) and Information Retrieval (IR) techniques. One of the typical tasks named Knowledge-Based Question Answering (KBQA) is defined to retrieve a specific entity from knowledge base as the answer to a given question.

The challenge of retrieval-based KBQA is how to match unstructured natural language questions with structured data in knowledge base. To understand a question, it is necessary to figure out the topic entity and relation chain inside the question. Thus, topic entity linking and relation ranking are the most important modules in our system.

2 Related Work

Knowledge-based question answering is a challenging task in the field of NLP. The mainstream approaches can be divided into three categories: semantic parsing based[1][2][3][4] [5], information extraction based [6][7] [8] and retrieval based [9][10][11].

The semantic parsing based approaches translate natural language questions into a series of semantic representations in logic forms. They query the answer in knowledge base through the corresponding query statement. Yih et al.[12] present a semantic parsing method via staged query graph generation. Convolution neural network is used to calculate the similarities between question and relation chains.

The information extraction based approaches extract topic entities from questions and generate a knowledge base subgraph with the topic entity node as the center. Each node in the subgraph can be used as a candidate answer. By examining the questions and extracted information according to some rules or templates, they obtain the feature vectors of the questions. A classifier is then constructed to filter candidate answers based on input feature vectors. Yao and Van Durme [13] associate question features with answer patterns described by Freebase. They also exploit ClueWeb, mined mappings between knowledge base relations and natural language text, and show that it helps both relation prediction and answer extraction.

The idea of retrieval-based method is similar to that of information extraction based methods. The question and candidate answers are mapped to distributed representation. The distributed representations are trained on labeled data, aiming to optimize the matching function between the question and the correct answer. Zhang et al. [14] combine bi-directional LSTM with an attention mechanism to represent the questions dynamically according to diverse focuses of various candidate answers.

These approaches work well on the English open dataset WebQuestion. However, their performances on a Chinese KBQA dataset have not been presented before.

3 Our Approach

Figure 1 shows the system architecture of our approach. For each question, the system finds the entity candidates firstly. And then entity ranking and relation ranking are conducted seperately to assign each entity candidate and relation a rank score. Finally, in the answer ranking stage, the system finds the top 1 triple according to the entity score and relation score. The object entity of the top 1 triple is the answer of the quesiton.

3.1 Entity Linking

Since the entity in the knowledge base has various name, such as Chinese name, English name, nick name, alias and so on, we build a Entity Map which mapps these names to the original entity.

In order to detect the topic entity in the question, we use a word-based maximum matching method to find entity candidates. First of all, the question is segmented by ltp[4] [15] segmenter. Then, we join the words in the question one

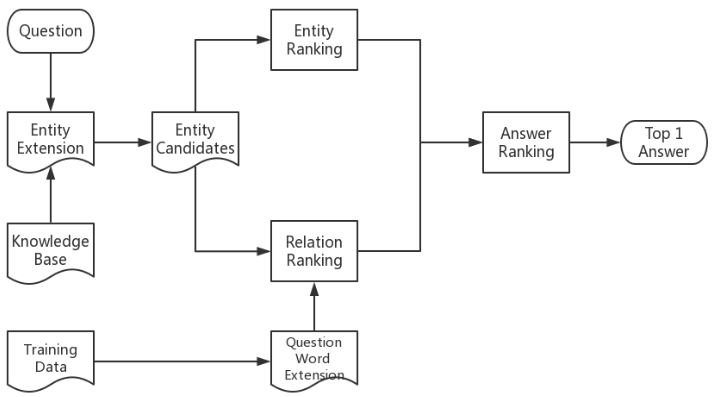

Fig. 1. System architecture

by one and search it in the Entity Map. If it exists in the keys of Entity Map, the corresponding value to the key will be added to a entity candidates list.

Here is an example:

Consider the question, 马丁·泰勒青少年时在哪支球队踢球 . After segmentation, we get a list of words 马丁·泰勒, 青少年, 时, 在, 哪, 支, 球队, 踢球 . Then we filter stop words 时, 在, 支 and question words 哪 , because they are impossible to be part of entity. After that, two sub-list of words are left, which are 马丁·泰勒, 青少年 and 球队, 踢球 . Then, we do word-based maximum matching for each sub-list. For sub-list, 马丁·泰勒, 青少年 , we first join all the words (马丁·泰勒青少年) and search it in the Entity Map. Apparently, it is not a entity. Then, we shorten the string length by one. Now, we search 马丁·泰勒 and 青少年 in the Entity Map separately. They are both existing entity, so they are added to the entity candidates list. For sub-list 球队, 踢球 , we conduct the same operation. In the end, we get entity candidates:

马丁·泰勒
马丁·泰勒(英国足球运动员)
马丁·泰勒(英国足球评论员)
青少年
《青少年》(2007年英国电影)
青少年(按年龄划分的社会群体)
青少年(2011年贾森·雷特曼导演美国电影)
踢球

3.2 Ranking

The knowledge base consists of millions of Subject-Relation-Object (SRO) triples. Each subject entity has dozens of Relation-Object pairs, each relation corre-

[4] http://ltp.ai/

sponding to only one object entity. Therefore, finding the answer to the question is equal to rank the relations of the subject entity and the object entity corresponding to the top one relation is supposed to be the answer.

After entity candidates have been found in the entity linking section, we can now collect all SRO triples of theses entities from the knowledge base. In this section, we combine editor distance score, character overlap score and word2vec cosine similarity score to rank each entity candidate and their relations.

Edit Distance The edit distance is a way of quantifying how dissimilar two strings are to one another by counting the minimum number of operations required to transform one string into the other. The edit distance score we used is a variant of the original edit distance. Suppose that the original edit distance of two strings s_1 and s_2 is notated as $ed(s_1, s_2)$, the edit distance score we use is

$$score_{ed} = 1 - \frac{ed(s_1, s_2)}{max(len(s_1), len(s_2))} \tag{1}$$

Character Overlap The character overlap is the number of overlapped characters in two strings. Greater character overlap suggests that the two strings are more topic related. We notate the character overlap score as $score_{co}$.

$$score_{co} = \frac{|set(s_1)| \cap |set(s_2)|}{|set(s_1)| \cup |set(s_2)|} \tag{2}$$

Word2vec Cosine similarity We train a word2vec model with a 20G chinese news corpus so that we can obtain a vector for each chinese word in the vocabulary. Then the string vector is computed as

$$v(s) = \sum_{w_i \in s} v(w_i) \tag{3}$$

so the word2vec cosine similarity of two strings is computed as the cosine similarity of two string vectors.

$$score_{w2v} = \frac{v(s_1) \cdot v(s_2)}{||v(s_1)|| \cdot ||v(s_2)||} \tag{4}$$

Related score The related score of two string is defined as:

$$score_{related} = score_{ed} + score_{co} + score_{w2v} \tag{5}$$

Entity Ranking We rank entity candidate by how many object entity of the candidate are related with question.

Consider the question 巴西球员托罗是哪天出生的 . Suppose that there are more than one entities 托罗 in the knowledge base and they have different nationalities such as 波多黎各, 墨西哥, 巴西 etc. and different occupations such as 球员, 导演, 演员 etc. When we calculate the related score of each entity candidate and the question, apparently the related score of the entity of which the nationality is 巴西 and the occupation is 球员 will be higher than that of others. In experience, if the related score is greater than a threshold λ, then we think that the object entity is related with the question. So entity score is computed as

$$score_{award} = (1 + score_{ed}) \times (1 + score_{co}) \times (1 + score_{w2v}) \tag{6}$$

$$score_S = 1 \times \prod_{o \in S} score_{award}(O, Q) \tag{7}$$

$$score_{award}(O, Q) = \begin{cases} score_{award}(O, Q) & score_{award}(O, Q) \geq \lambda \\ 1 & score_{award}(O, Q) < \lambda \end{cases} \tag{8}$$

We tune the value of λ from 1.0 to 2.0, gap 0.1, and find that when $\lambda = 1.5$ it achieves the best result on the training data.

Relation Ranking Before calculating the score, we remove the string corresponding to the entity candidate and related object entity for simplifying the computation. For example, after removing the string, the question 巴西球员托罗是哪天出生的 becomes 是哪天出生的 .

In addition, we also do question word extension. In some cases, the relation of entity does not appear in the question. For example, the relation to the question is supposed to be 出生日期 or 生日 (both refers to "birthday" in English), however, neither of them exists in the question. So, we map 哪天出生 to 出生日期 and 生日 , the latter is called the extension of question word.

Finally, we rank relations of each entity candidate by caculating the related score of each relation with the quesiton and the extension of question word.

$$score_R = \alpha * score_{related}(R, Ext) + \beta * score_{related}(R, Q \cup Ext) \tag{9}$$

where, R refers to relation, Ext refers to the extension of question word, and $Q \cup Ext$ refers to the union of question and ext. The α and β are weight factors. We set α to be 0.47 and β to be 0.53 according to the experiment.

Answer Ranking For a SRO triple, we caculate the score as below:

$$score_{SRO} = score_R \times score_S \tag{10}$$

$$score_S = \begin{cases} score_S & score_{award}(O, Q) < \lambda \\ 1 & score_{award}(O, Q) \geq \lambda \end{cases} \tag{11}$$

where O refers to Object of SRO triple, Q refers to the question and λ is set to be 1.5 according to the experiment.

If the $score_{award}(O, Q) \geq \lambda$, it suggests that the object is a known fact and can not be the answer of the question shown as Eq(11).

We rank SRO by multiplying the score of each relation and the score of corresponding entity candidate and get the Object from the top 1 SRO as the answer.

4 Experiments

4.1 Dataset

In this paper, we use the dataset provided by the NLPCC 2018 open domain KBQA shared task. The dataset includes 24,479 single-relation question-answer pairs for training, a Chinese knowledge base with 43M SRO triples, and 7M mapping data from mentions to entities. The test set contains 618 questions.

Since the mapping data is not what our system desires, we rebuild a Entity Map from mentions to entities with no word segmentation.

4.2 Setup

The word embeddings used in our system is pre-trained by gensim[5] . We use the skip-gram model [16] and the dimension is set to be 300.

4.3 Results

The results of our system achieves 62.94% of answer exact matching on the test set, which ranks 3rd place in the final leaderboard.

4.4 Error Analysis

We analyze the causes of the error cases (229 in total). 37.6% of errors are caused by entity linking and 27.9% are caused by relation ranking. 16.6% of errors are attributed to that the desire answer of the quesiton is the subject entity of the SRO triple and we can not use object entity to infer subject entity.

In addition, 5.2% of errors are caused by the confliction of knowledge base. For example, to the question 国脚黄博文在场上司职什么角色？, our answer is 中场, while the official answer is 前卫. However, in the knowledge base, the entity 黄博文 contains both triples 黄博文 ||| 场上位置 ||| 中场 and 黄博文 ||| 位置 ||| 前卫.

For the last 16.6% of errors, in fact, we find the correct answers, but the offical system judge them as incorrect ones. For example, to the question 泰国武里南联队是哪年成立的？, our answer is 1970年, while the official answer is 1970. And in the preprocessing stage, we convert all characters in the knowledge base from full-width to half-width and convert all upper case letter to lower case, which also

[5] https://radimrehurek.com/gensim/

cause the official system to judge our correct answer as wrong one. For example, to the question 牛津联足球俱乐部的主席是谁？, our answer is 达利尔·伊尔斯 (darryl eales), while the official answer is 达利尔·伊尔斯（Darryl Eales）. If these cases caused by wrong judgement and knowledge base confliction are revised, the answer exact matching of our results will be 69.42%.

5 Conclusion

In this paper, we report technique details of our approach for the NLPCC 2018 shared task knowledge-based question answering. Our system uses a word-based maximum matching method to find entity candidates. Then, we combine editor distance, character overlap and word2vec cosine similarity to rank SRO triples of each entity candidate and get the object of the top 1 score SRO as the answer of the question.

We also try to use deep learning in entity linking and question-relation match. However, for entity linking, since the questions of test set are greatly different from that of training set, the model can not generalize from training data to test data. For question-relation match problem, it seems to be quite difficult to match thousands of questions to millions of relations, even by deep learning. And the number of relations in the training set is 4,385, however the number of that in the knowldge base is up to 587,576. It is impractical to train a relation match model from such a small dataset. Even though, after replacing the provided mention2id with the entity extension built by us and revising some errors in the knowledge base, we also achieve good results with statistic and rule-based methods.

References

1. Zettlemoyer, L.S., Collins, M.: Learning to map sentences to logical form: structured classification with probabilistic categorial grammars. In: UAI (2005)
2. Kwiatkowski, T., Zettlemoyer, L., Goldwater, S., Steedman, M.: Inducing probabilistic CCG grammars from logical form with higher-order unification. In: EMNLP (2010)
3. Liang, P., Jordan, M.I., Klein, D.: Learning dependency-based compositional semantics. In: ACL (2011)
4. Berant, J., Chou, A., Frostig, R., Liang, P.: Semantic parsing on freebase from question-answer pairs. In: EMNLP (2013)
5. Berant, J., Liang, P.: Semantic parsing via paraphrasing. In: ACL (2014)
6. Bast, H., Haussmann E.: More accurate question answering on freebase. In: Information and Knowledge Management (2015)
7. Fader, A., Zettlemoyer, L., Etzioni, O.: Open question answering over curated and extracted knowledge bases. In: Knowledge Discovery and Data Mining (2014)
8. Yao, X.: Lean question answering over freebase from scratch. In: ACL (2015)
9. Bordes, A., Chopra, S., Weston, J.: Question answering with subgraph embeddings. In: EMNLP (2014)
10. Bordes, A., Weston, J., Usunier, N.: Open question answering with weakly supervised embedding models. In: ECML (2014)

11. Dong, L., Wei, F., Zhou, M., Xu, K.: Question answering over freebase with multi-column convolutional neural networks. In: ACL (2015)
12. Yih, W., Chang, M.-W., He, X., Gao, J.: Semantic parsing via staged query graph generation: question answering with knowledge base. In: ACL (2015)
13. Yao, X., Van Durme, B.: Information extraction over structured data: question answering with freebase. In: ACL (2014)
14. Zhang, Y., Liu, K., He, S., Ji, G., Liu, Z., Wu, H., Zhao, J.: Question answering over knowledge base with neural attention combining global knowledge information. arXiv:1606. 00979 (2016)
15. Wanxiang Che, Zhenghua Li, Ting Liu. LTP: A Chinese Language Technology Platform. In Proceedings of the Coling 2010:Demonstrations. 2010.08, pp13-16, Beijing, China.
16. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. Comput. Sci. (2013)