A Joint Model of Entity Linking and Predicate Recognition for Knowledge Base Question Answering

Yang Li[1]✉, Qingliang Miao[1]✉, ChenXin Yin[1], Chao Huo[1], Wenxiang Mao[2], Changjian Hu[1], Feiyu Xu[1]

Building H, No.6, West Shangdi Road, Haidian District Beijing, China
{liyang54,miaoql1,yincx1,huochao2,hucj1,fxu}@lenovo.com
{maowenxaing612}@163.com

Abstract. In the paper, we build a QA system which can automatically find the right answers from Chinese knowledge base. In particular, we first identify all possible topic entities in the knowledge base for a question. Then some predicate scores are utilized to pre-rank all candidate triple paths of topic entities by logistic model. Second, we use a joint training entity linking and predicate recognition model to re-rank candidate triple paths for the question. Finally, the paper selects the answer component from matched triple path based on heuristic rules. Our approach achieved the averaged F1-score of 57.67% on test data which obtained the second place in the contest of CCKS 2018 COQA task.

Keywords: KBQA · Entity Linking · Predicate Recognition · Semantic Matching.

1 Introduction

In the paper, we introduce a system that answers an open domain factoid question in Chinese automatically. Our method recognizes topic entities at first. Then one-hop and two-hop triple paths are selected and pre-ranked for these topic entities. Second, we use a semantic matching model BiMPM [1] to train a joint model for entity linking and predicate recognition to re-rank candidate triple paths. At last, the answer component is selected from matched candidate triple path based on heuristic rules. By pre-processing and analysing the training data, questions only with one-hop or two-hop candidate triple paths account for 90.02%. Thus, the paper concerns with these questions mainly.

2 Related Work

Open domain KBQA is an important task in the field of natural language processing. There are two mainstream approaches: semantic parsing based and retrieval based.

The semantic parsing based method first parses question into a logical form which is a semantic tree explicitly representing the meaning of the question in a compositional manner, and then the logical form is executed based on the knowledge base to get the answer [2]. The logical form in this method is helpful to understand the semantic structure of a question, which also increases the difficulty of this task. Lai [3] uses

word embedding based features to search best subject predicate pair and obtains the first place in NLPCC 2016 KBQA task by rules. Lai [4] proposed a novel method based on deep CNNs to rerank the entity-predicate pairs which generated by shallow features. The approach obtained the first place in the contest of NLPCC 2017 KBQA task. Hu [5] proposed a dynamic query graph matching method to process disambiguation tasks for entities and relationships from a data-driven perspective. In our work, we also use retrieval based approach.

3 The Proposed System

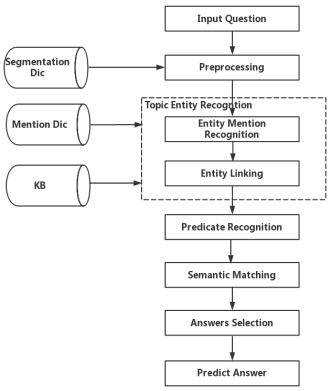

Fig. 1. Architecture of our System.

The architecture of our system is shown in Fig 1. The first step is pre-processing which is word segmentation for input question. Based on the pre-processing results, Topic Entity Recognition module first recognizes topic entity mentions and then link them to knowledge base. In Predicate Recognition module, we pre-rank candidate triple paths with some features. Then BiMPM is utilized to select the matched triple paths. Finally, we select the answer from matched candidate triple path.

3.1 Knowledge Base and Dictionary

In the paper, the used knowledge base (KB) and an entity mention dictionary (Mention Dic) which are provided by CCKS. Besides, we construct a Segmentation Dic for Hanlp to improve the segmentation and entity mention recognition result. Segmentation Dic is composed of all subjects of KB, all entities and its mentions in Mention Dic.

3.2 Topic Entity Recognition

Entity Mention Recognition If the segmentation word of the question in the Segmentation Dic then the word is seemed as entity mention. The recognized entity mentions

have different probabilities of being a topic mention from the perspective of some features. The features used in our system are defined as follows:

F1:The Length of Entity Mention An entity mention with a longer string is more likely to be a topic entity than shorter one.

F2:The TF value of Entity Mention An entity mention with a high Term Frequency (TF) value tends to have a low probability to be a topic entity than lower ones.

F3:The Distance Between the Entity Mention and Interrogative Word Entity mentions in question close to the interrogative word is more likely to be a topic entity.

Entity Linking Entity mentions that recognized by the last step is not the entities in knowledge base so this step is aimed to determine the identity of entity mentions in question. Relations and properties information of an entity are helpful for entity linking so at first we extract two-hop sub-graph of the entity. Based on the selected candidate entity mentions, we use three features below to rank and select the matched topic entity.

F4:Word Overlap Between Question and Triple Paths The more overlap words shared between question and candidate entity's two-hop sub-graph, the bigger probability that the entity mention be a topic entity.

F5:Word Embedding Similarity Between Question and Triple Paths The larger similarity between the question and candidate entity's two-hop sub-graph, the bigger probability that the entity mention be a topic entity.

F6:Char Overlap Between Question and Triple Paths The feature is similar to F4. The only difference is that this feature uses char level instead of word level.

After calculating and normalizing all features a linear weighing method is utilized to rank candidate entities. The score equation is defined as below equation where w_i indicates the weight of feature i.

$$Score_{topicentity} = w_1 * F_1 + w_2 * F_2 + w_3 * F_3 + w_4 * F_4 + w_5 * F_5 + w_6 * F_6$$

3.3 Predicate Recognition

A topic entity can extract about 349.6 candidate triple paths. It's difficult to select the best matched one from such large amount candidate triple paths. Narrowing down candidate triple paths is an important step to improve the final result. In this module, we first extract four features about predicates of triple path. Then logistic regression algorithm is utilized to pre-rank candidate triple paths with below four features and topic entity recognition features. At last, we select top 10 triple paths as candidates for next semantic matching module.

F7:Word Overlap Between Question and Predicates
The more overlap words shared between question and candidate predicates of triple path, the bigger probability that the candidate predicates be truly predicates.

F8:Word Embedding Similarity Between Question and Predicate
The larger similarity between question and candidate predicates, the bigger probability that the candidate predicates be truly predicates.

F9:Char Overlap Between Question and Predicates
This feature is almost same as F7. The only difference is that this feature uses char level instead of word level.

F10:Char Embedding Similarity Between Question and Predicates
This feature is almost same as F8. The only difference is that this feature uses char level instead of word level.

3.4 Semantic Matching

Problem Formalization The goal of this module is to identify the TPi from *n* candidate triple paths $\{TP_1, TP_2, ..., TP_n\}$ that best matches Q. Q is the question of user. TPi is a candidate triple path of Q. In this paper, we use a pairwise scoring function S(TPi,Q) to score and sort all candidate triple paths. In the paper, n is 10.

BiMPM+Fea In this section, we present a innovative solution that incorporate word embedding and all ten features into BiMPM to select the best matched triple path. BiMPM+Fea contains five kernel layers.

(1) Word Representation Layer: The goal of this layer is to represent each word in question and triple path with d-dimensional vector. The word embedding in the paper is pre-trained with Gensim [6] and d is 100.

(2) Context Representation Layer:The purpose of this layer is to incorporate contextual information into the representation of each time step of question and triple path. This paper uses a BiLSTM to encode contextual embeddings for each time step.

(3) Matching Layer: This is the core layer and it is used to obtain the similarity of the question and triple path in time-steps. Moreover, the matching is bi-directional, means that the question and the triple path will match each other and get the matching information from their respective.

(4) Aggregation Layer: This layer is applied to aggregate question and triple path of matching information into fixed-length. The aggregation layer is composed of BiLSTM, and we use the final hidden state to represent the information aggregated.

(5) Feature Aggregation Layer: The layer concatenates the fixed-length tensor of the last layer with our 10 extracted features.

3.5 Answers Selection

Matched triple paths are selected in semantic matching modules. Then we generate the answer based on heuristic rules. The Fig 2 displays examples of our heuristic rules. In the figure, the circle node or rectangle node just represents entity or attribute value and without affecting the rules to select answer. The blue node is the answer.

(1) One-hop Triple Path: In this situation the answer is the component in triple path which does not appear in the question.

(2) Two-hop Triple Path: If both the far right node and the far left node in the triple path do not appear in the question then the middle node is the answer. If either the far right node or the far left node appears in the question then the other one is the answer.

4 Experiments and discussion

We evaluate our approaches by using CCKS data. The data set is published by CCKS 2018 evaluation task which includes a knowledge base, knowledge entity mention file

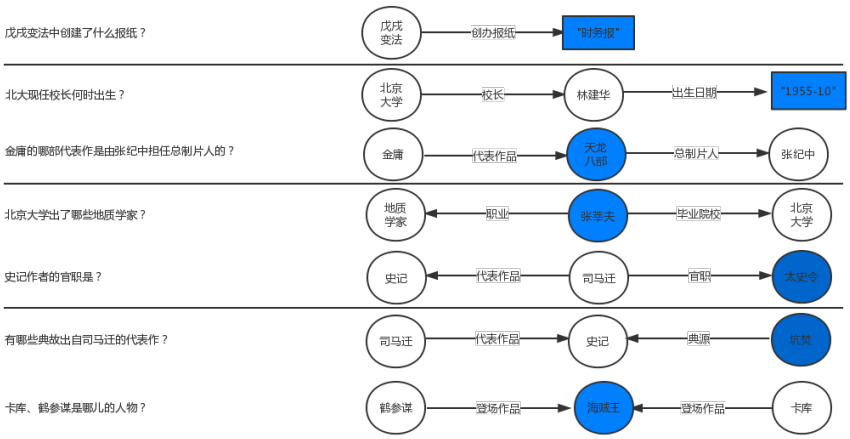

Fig. 2. The different triple path shapes of questions.

and question-answer pairs for training, validation and testing. The knowledge base has 41 million triples. The 2018-Train set, 2018-Val set, 2018-Test set contain 1283,400,400 samples respectively. To obtain negative samples in the training process, for each question, we select top 10 wrong candidate triple paths. To alleviate the impact of unbalanced training data, we oversample positive samples.

4.1 Topic Entity Recognition Result

	2018-Val	**2018-Test**
baseline$_{TE}$	92.58%	90.79%
baseline$_{TE}$+Emb	96.29%	93.28%
baseline$_{TE}$+Emb+Char	98.58%	95.35%

Table 1. Pre@1 of topic entity recognition.

Table 1 shows systems performance for topic entity recognition module. The basic model baseline$_{TE}$ only uses F_1, F_2, F_3, F_4. The second model is baseline$_{TE}$+Emb, which also uses embedding feature F_5. The last one is baseline$_{TE}$+Emb+Char, which also uses embedding feature F_5 and char level feature F_6. From Table 1, it is obvious that embedding feature F_5 and char level feature F_6 all can improve the Pre@1 of topic entity recognition. Hyper-parameters w_i in the model baseline$_{TE}$+Emb+Char is [0.25, 0.37, -0.32, 0.67, 0.71, 0.58].

4.2 BiMPM Re-ranking Result

	2018-Val	**2018-Test**
BiMPM	54.83%	53.85%
BiMPM+Fea	57.15%	56.54%
BiMPM+Fea+CV	58.23%	57.67%

Table 2. F1-score of KBQA.

Table 2 shows our systems performance of experimentation. The basic model BiMPM only use pre-trained word embedding. The second one is BiMPM+Fea, which also uses 10 extracted features based on baseline BiMPM. The last one is BiMPM+Fea+CV, which utilizes 10-fold cross validation based on the BiMPM+Fea. From Table 2, it is obvious that both extracted features and cross validation can improve the F1-score. The result of BiMPM+Fea is about 2.5% higher than BiMPM on the 2018-Test. The reason is that the entity linking scores and predication recognition scores are useful for pre-ranking triple paths.

5 Conclusion

In the paper, we present a joint model of entity linking and predicate recognition for KBQA. The system achieves the F1-score of 57.67% on CCKS 2018 COQA task. For future research, we plan to extend our approach to alleviate unseen predicates issue.

References

1. Wang, Z., Hamza, W., Florian, R. Bilateral Multi-Perspective Matching for Natural Language Sentences. arXiv preprint arXiv:1702.03814, (2017)
2. Bao, J., Duan, N., Zhou, M.:An Information Retrieval-Based Approach to Table-Based Question Answering. In: 6th National CCF Conference on Natural Language Processing and Chinese Computing, pp.601–611. Springer, Cham (2017)
3. Lai, Y., Lin, Y., Chen, J.:Open domain question answering system based on knowledge base. In: 5th National CCF Conference on Natural Language Processing and Chinese Computing, pp.722–733. Springer, Cham (2016)
4. Lai, Y., Jia, Y., Lin, Y.:A Chinese Question Answering System for Single-Relation Factoid Questions. In: 6th National CCF Conference on Natural Language Processing and Chinese Computing, pp.124–135. Springer, Cham (2017)
5. Hu, S., Zou, L., Yu, J. X.: Answering Natural Language Questions by Subgraph Matching over Knowledge Graphs. IEEE Transactions on Knowledge & Data Engineering, 2018: 824-837
6. Řehůřek, R., Sojka, P. Software Framework for Topic Modelling with Large Corpora. In:7th Proceedings of the LREC Workshop on New Challenges for NLP Frameworks, pp.45–50. ELRA, Valletta Malta (2010)