

A Joint Framework for Entity Discovery and Linking in Chinese Questions

Ziqi Lin^{1,2}, Wancheng Ni¹, Haidong Zhang¹, Yu Liu¹, Yiping Yang¹

1. Integrated Information System Research Center, CASIA, Beijing, China

2. University of Chinese Academy of Science, Beijing, China

e-mail: {linziqu2013, wancheng.ni, haidong.zhang, yu.liu, yiping.yang}@ia.ac.cn

Abstract—Entity discovery and linking can help to understand the questions semantics and infer answers in question answering systems. In this paper, we study the characteristics of Chinese questions and propose a joint framework that leverages the mutual dependency between entity discovery and linking to enhance their performances. It jointly connects a joint parsing method based on concept knowledge tree for entity discovery, with candidate entity generation of entity linking. And we also investigate conditional random fields to detect entity mentions and filter them with candidate entity generation. Experiments show that our proposed method outperforms state-of-the-art methods on a real dataset.

Keywords-entity discovery; entity linking; joint method; question representation; concept knowledge tree;

I. INTRODUCTION

Entity discovery and linking (EDL) tasks refer to discovering entity mentions in text, and further linking them to an existing knowledge base. In question answering systems, these two tasks can find possible entities in questions and benefit answers inference. EDL contributes to solving the semantic gap between natural language questions and knowledge bases, which have become a key component for question answering systems.

Entity discovery methods can be divided into three categories: rule-based methods [1] by extracting entities' patterns, sequence labeling methods [2] (e.g., Conditional Random Fields (CRFs), Semi-Markov Conditional Random Fields (Semi-CRFs)), and neural networks [3] incorporating with sequence labeling models. Due to its deep architecture and complicated parameters, the neural network does not perform well on the small-scale dataset. Entity linking consists of candidate entities generation and entity disambiguation. The former [4] searches all possible entities of extracted mentions in the knowledge base, while the latter [5] leverage the context to rank them and disambiguate the mismatched entities.

The joint model for entity discovery and linking tasks can enhance the performance with each other because of their mutual dependencies [6]. For example, Luo et al. [7] proposed a joint model for Chinese short text, which built a ranking model to make a joint prediction from the over-generation of candidate mention-entity pairs. Luo et al. [6] utilized mutual dependency between named entity recognition (NER) and disambiguation to extend Semi-CRFs for jointly predicting consistent results.

Chinese questions usually have few words and contain Chinese and English mixed name variations, which lack available context and morphology variations. And the annotated corpus is insufficient for Chinese QEDL. To further improve the joint framework and apply EDL in Chinese questions, we integrate candidate entity generation

of entity linking into entity discovery to build a joint framework for Chinese QEDL. In entity discovery, we jointly connect the candidate entity generation with the joint parsing method based on question representations and incorporate its extracted mentions into CRF. We also leverage candidate entities generation to further filter the extracted mentions by CRF and employ supervised machine learning methods to rank them. To validate our joint method, we perform experiments on a real dataset, and the results show that the joint framework can enhance the performances of these two tasks.

II. OVERVIEW OF OUR FRAMEWORK

In this paper, we build a joint framework that combines entity discovery (including a joint parsing method and CRF) and entity linking in Chinese QEDL, as shown in Fig. 1. This framework firstly preprocesses Chinese questions (including word segmentation, part-of-speech tagging, and dependency syntactic parsing), and extracts features. It utilizes a joint parsing method to extract mentions from Chinese questions. We integrate the extracted mentions into a CRF as features, and filter generated mentions with candidate entities.

In entity discovery, the joint parsing method builds the rules of generating candidate mentions from the question representations to extract mentions by the feedback of the knowledge base and online encyclopedia. And we add the extracted mentions from the joint parsing method into the CRF model for learning more patterns of mentions. The details are shown in Section III and IV.

Moreover, we construct a mention filter model and a mention merging algorithm to improve the confidence of mentions. The mention filter model uses gradient boosting decision tree (GBDT), support vector machine (SVM) and Decision Tree (DT) to build a voting model, which combines mentions and candidate entities features. And the mention filter model would filter mentions, if and only if all models think the mention should be filtered. To make use of the advantages of different methods, we design a mention merging algorithm. The main idea of mention merging algorithm includes: choose the longer mentions; merge neighbor mentions into a longer mention; filter overlapping mentions.

In our work, entity linking consists of candidate entities generation and a linker model. The former module searches related entities of extracted mentions from a knowledge base and an online encyclopedia, while the latter one considers the mentions' context and employs GBDT to disambiguate the mismatched entities.

III. JOINT PARSING METHOD FOR ENTITY DISCOVERY

A. Parsing Algorithm

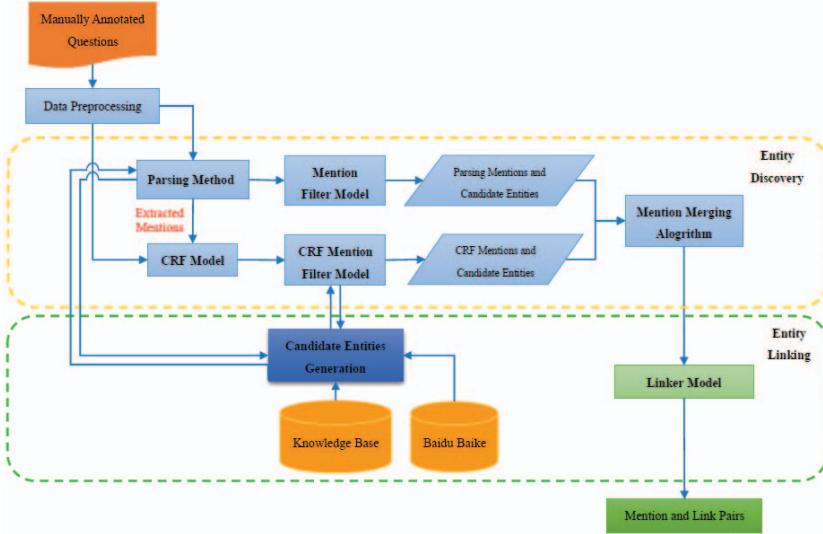


Figure 1. The joint framework for Chinese QEDL.

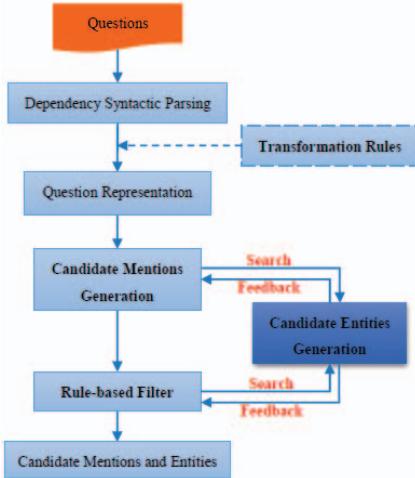


Figure 2. The parsing algorithm for entity discovery.

The joint parsing method is a rule-based method, which combines question representations and candidate entities generation to detect possible mentions. The parsing algorithm is depicted in Fig. 2.

The major process of parsing algorithm is as follows:

1) The generation of question representations:

We build Chinese question representations based on three basic semantic representations of concept knowledge tree. Through the dependency syntactic parsing of questions, we design some transformational rules to obtain question representations. Due to space limitations, we omit the description of the transformational rules.

2) The generation of candidate mentions:

The generation of candidate mentions is the core of this algorithm, which combines question representations and the feedback of candidate entities generation to get candidate mentions. We construct the domain-independent generating rules for candidate mentions from question representations. We use these rules to obtain possible mentions. Then we use the component of candidate entities

generation to search candidate entities for these mentions. Through the feedback of candidate entities generation, we select these mentions that have candidate entities as detected mentions. The next section details the rules of generating candidate mentions.

3) The filtering of candidate mentions:

We design a rule-based filter method to combine the features of mentions and candidate entities for filtering common words or concepts. Through the feedback of candidate entities generation, we try to merge adjacent mentions for filtering separated mentions.

B. Question Representations and Concept Knowledge Tree

Concept knowledge tree [8, 9] is a knowledge representation model, which can express concepts, knowledge and semantic relations. To describe language concepts, CKT defines three basic formal semantic representations, including semantic constraint, semantic logic, and semantic state.

The semantic constraint uses constraint concept and core concept to express concepts modified and attributes constraint relationships for the compound concepts, as shown in (1):

$$\begin{aligned} \text{Semantic Constraint} = & \\ & \langle \text{Constraint Concept} : \text{Core Concept} \rangle \end{aligned} \quad (1)$$

The semantic logic defines five logic relations to describes the logic combination relation of concepts. The logical relations include “entailment”, “list”, “and”, “or” and “not”; “entailment” describes the progressive relationship and causality of concepts; “list” indicates a simple arrangement of several concepts. The semantic logic consists of logical relations and concepts list (2):

$$\text{Semantic Logic} = (\text{Logic Type}, \text{Concepts List}) \quad (2)$$

The semantic state uses predicate concepts, subject concepts, object concepts and state concepts to denote events, as shown in (3):

$$\begin{aligned} \text{Semantic State} = & \langle \text{Subject Concepts} \rangle \{ \text{State Concepts} \} \\ & \text{Predicate Concepts} < \text{Object Concepts} \rangle \end{aligned} \quad (3)$$

Moreover, we extract 15 kinds of primary Chinese phrases based on Chinese grammar books and build their semantic representation models. Based on the primary Chinese phrases semantic representations, we decompose questions into words and express as the three basic formal representations. Thereinto, words are represented as single semantic nodes; these semantic nodes can form phrases, then these semantic nodes and phrases are nested with each other to form questions.

TABLE I. THE RULES OF GENERATING CANDIDATE MENTIONS

Representation	Candidate Mention
Semantic Constraint	<u>word1</u> <bind node:core node> <u>word2</u>
	<u>word1</u> <bind node:core node>
	<bind node:core node> <u>word2</u>
	< bind node:core node >
	<<bind_node1:<core_node1>:<core_node2>
	<bind_node1:<bind_node2:<core_node2>>
	<<bind_node1:<core_node1>:<bind_node2:<core_node2>>
	<node1:char1 char2...>
	<bind node:core_node>
	< bind_node:core_node >
Semantic Logic	<u>word1</u> (logic node, member nodes) <u>word2</u>
	<u>word1</u> (logic node, member nodes)
	(logic node, member nodes) <u>word2</u>
	member node
Semantic State	<u>word1</u> [<sub_node>verb_node<obj_node>{state_node}] <u>word2</u>
	<u>word1</u> [<sub_node>verb_node<obj_node>{state_node}]
	[<sub_node>verb_node<obj_node>{state_node}] <u>word2</u>
	[<sub_node>verb_node<obj_node>{state_node}]
Semantic Node	node (if node.pos in ['nh', 'ni', 'ns', 'j'])
	node (node!=root_node or root_pos in ['n','ws']) and (node isn't interrogative pronouns) and (node.pos not in ['a', 'd', 't', 'nd', 'p', 'q']) and (node.length > 1)

C. The domain-independent generating rules

The question representations describe the relations between elements of questions, and we leverage these relations to build the domain-independent generating rules. These rules are shown in Table I, where the underline and bold fonts denote the elements of generating candidate mentions. The domain-independent generating rules give higher priority to generating longer mentions. If the longer mentions are detected, the algorithm will not detect the shorter mentions.

For example, “胡歌的发型怎么做？” (means “how to do the hairstyle of Hu Ge?”) is expressed as: [<<<胡:歌>:发型>><怎么:做>] ([<<<Hu:Ge>:hairstyle>><how:do>]). The first three rules of semantic state detect whether the partial representation of question merges the left word or the right can construct a mention. For the subject part of

question representation, “胡歌” (Hu Ge) and “发型” (hairstyle) are detected mentions by using the tenth and eleventh rules of semantic constraint. Owing to containing the interrogative pronoun (“怎么” (how)), the algorithm doesn't detect the predicate part of question representation. The golden result of dataset only contains “胡歌” (Hu Ge).

IV. INTEGRATED EXTENSIONAL MENTIONS CRF MODEL

CRF [2] is a probabilistic framework to segment and label sequence data, which is widely used in NER. Considering the characteristics of Chinese questions, we design some features, as shown in Table II. The extracted mentions by joint parsing method are incorporated into CRF as one of the features because they contain the understanding of language cognition. The extracted mentions make CRF model learn more patterns of mentions and alleviate the sparsity of annotated dataset.

We describe the details of features as follows: “SHI” is the word “是” (“is”); “DE” is the word “的” (means modified or constraint); “HED” denotes the root node of dependency syntactic parsing. We use unigram and bigram features for the word feature; unigram, bigram, and tri-gram features use to the part of speech feature; bigram and tri-gram features use to the dependency relation; other features build unigram, bigram, and tri-gram features.

TABLE II. THE FEATURES OF CRF MODEL

No.	Description
1	Word
2	Part of speech
3	Dependency relations
4	The position of interrogative pronouns
5	The position of the word “SHI”
6	The position of the word “DE”
7	The distance of “HED” and word
8	The index of word in the question
9	The index of dependency parent node
10	The union of feature 8 and 9
11	The extracted mentions of parsing method

V. EXPERIMENTS

A. Experimental Setup

To validate our method, we compare it with two baselines (i.e., CRF [2] and BI-LSTM CRF [3]) on the CCKS QEDL dataset¹ which is a competition dataset. The statistics of QEDL dataset is described in Table III. Moreover, we use Baidu Baike² as the online encyclopedia and CN-DBpedia [10] as the knowledge base. We use Language Technology Platform [11] to make data preprocessing, including word segmentation, part-of-speech tagging, and dependency syntactic parsing. And we use precision, recall, and F1 to evaluate the results of entity discovery and linking.

In experiments, BI-LSTM CRF uses 200 dimensions word embeddings and the features in Table II. We set hidden layer size to 256, the dropout rate to 0.5, and use a

¹ http://www.ccks2017.com/?page_id=51

² <https://baike.baidu.com/>

learning rate of 0.002 to train this model. We use word2vec with the training objective of the Skip-gram model and negative sampling [12] to train 200 dimensions word embeddings. According to the analysis of features, CRF model uses 1-7 and 10-11 features.

TABLE III. STATISTICS OF THE QEDL DATASET.

Dataset	Questions	Words	Entities	NIL Entities ^a
Train	1400	28035	1898	66
Test	749	15743	888	44
Total	2149	43778	2786	110

a. NIL entities are the non-existent entities that don't exist in the knowledge base.

B. Results and Analysis

Table IV gives the results of entity discovery, and we can observe that our method achieves about 3.29% performance improvement over the BI-LSTM CRF and 2.86% over the CRF method. To validate the effectiveness of the joint parsing method, we perform these three methods under the condition of with and without the generated features by the joint parsing method. We can find these methods with parsing features has 8.53%~9.98% increase than without parsing features. And the mention filter model improves the CRF performance by 1.76% and the joint parsing method 4.62%. The merging filtered results of CRF and the joint parsing method obtains the best performance, and the F1-measure reaches to 62.27%.

TABLE IV. THE RESULTS OF ENTITY DISCOVERY

Method	Precision (%)	Recall (%)	F1 (%)
BI-LSTM CRF (without parsing feature)	51.17	47.0	49.0
CRF (without parsing feature)	50.58	51.18	50.88
BI-LSTM CRF	56.68	61.48	58.98
CRF	57.96	60.94	59.41
Filtered CRF	63.13	59.33	61.17
Joint Parsing	37.87	69.21	48.95
Filtered Joint Parsing	46.97	62.34	53.57
Filtered CRF + Filtered Joint Parsing	60.78	63.84	62.27

We input the extracted mentions by different entity discovery methods into entity linking of our methods, and compare the linking results, as shown in Table V. We observe that the mention filter model also improves the performance of entity linking. Experimental results show that the mutual information of the two tasks can enhance the overall performance in Chinese QEDL.

TABLE V. THE RESULTS OF ENTITY LINKING

Method	Precision (%)	Recall (%)	F1 (%)
Joint Parsing	25.66	46.89	33.17
Filtered Joint Parsing	32.42	43.03	36.98
Filtered CRF + Filtered Joint Parsing	40.76	42.81	41.76

VI. CONCLUSIONS

In this paper, we propose a joint framework for Chinese question entity discovery and linking in the small-scale dataset, which incorporates the information of entity discovery and linking to enhance the overall performance

in the entity discovery stage. We propose a joint parsing method based on question representations, which combines the domain-independent generating rules and the feedback of candidate entities generation to detect possible mentions. Moreover, we integrate the joint parsing method into CRF model to learn more mentions' patterns. We evaluate the performance on the CCKS QEDL dataset. Experimental results show the effectiveness of the joint framework.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their valuable comments. This work was supported in part by the National Natural Science Foundation of China under Grant 61379099.

REFERENCES

- [1] G. Petasis, F. Vichot, F. Wolinski, G. Paliouras, V. Karkaletsis, and C.D. Spyropoulos, "Using machine learning to maintain rule-based named-entity recognition and classification systems," in Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, ACL Press, 2001, pp. 426-433, doi:10.3115/1073012.1073067.
- [2] L. John D., A. McCallum, and F.C.N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in Proceedings of the Eighteenth International Conference on Machine Learning, Morgan Kaufmann Publishers Inc. Press, 2001, pp. 282-289.
- [3] Z. Huang, W. Xu, and K. Yu, "Bidirectional lstm-crf models for sequence tagging," in arXiv:1508.01991v1, 2015.
- [4] W. Zhang, Y.C. Sim, J. Su, and C.L. Tan, "Entity linking with effective acronym expansion, instance selection and topic modeling," in Proceedings of the Twenty-Second international joint conference on Artificial Intelligence - Volume Volume Three, AAAI Press Press, 2011, pp. 1909-1914.
- [5] T. Zhang, K. Liu, and J. Zhao, "A graph-based similarity measure between wikipedia concepts and its application in entity linking system," Journal of Chinese Information Processing, vol.29, 2015, pp. 58-67.
- [6] G. Luo, X. Huang, C.Y. Lin, and Z. Nie, "Joint named entity recognition and disambiguation," in Conference on Empirical Methods in Natural Language Processing, ACL Press, 2015, pp. 879-888.
- [7] X. Luo, P. Hu, X. Huang, and T. He, "A joint model for entity recognition and linking in chinese short text," International Journal of Advanced Intelligence, vol.8, 2015, pp. 64-71.
- [8] Y. Gao, "A concept-based knowledge representation system," Microelectronics & Computer, vol.21, 2004, pp. 71-74, doi:10.3969/j.issn.1000-7180.2004.09.019.
- [9] M. Zhao, "Research on semantic understanding of the meaning of chinese dictionary (med)," Phd. Thesis, Institute of Automation, Chinese Academy of Sciences, Beijing, China, 2015.
- [10] B. Xu, Y. Xu, J. Liang, C. Xie, B. Liang, W. Cui, et al., "Cndbpedia: A never-ending chinese knowledge extraction system," in Advances in artificial intelligence: From theory to practice: 30th international conference on industrial engineering and other applications of applied intelligent systems, iea/aie 2017, arras, france, june 27-30, 2017, proceedings, part ii, S. Benferhat, K. Tabia, and M. Ali, Eds, Springer International Publishing: Cham, 2017, pp. 428-438.
- [11] W. Che, Z. Li, and T. Liu, "Ltp: A chinese language technology platform," in Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations, ACL Press, 2010, pp. 13-16.
- [12] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in Proceedings of the 26th International Conference on Neural Information Processing Systems, Curran Associates, Inc. Press, 2013, pp. 3111-3119.