Unsupervised Resource-Free Entity Discovery and Linking in Natural Language Questions

Shu Guo[1,2], Jiangxia Cao[3], Quan Wang[1,2]*, Lihong Wang[4], and Bin Wang[1,2]

[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]School of Cyber Security, University of Chinese Academy of Sciences
[3]School of Computer Scince & Technology, Heilongjiang University
[4]National Computer Network Emergency Response Technical Team
Coordination Center of China

Abstract. We present the solution to the CCKS 2017 question entity discovery and linking (QEDL) task. This task is to discover and link entity mentions in natural language questions with their referent entities in a knowledge base (KB). For entity discovery, we devise recognition patterns based on word segmentation and POS tagging. For entity linking, we leverage contextual similarity refined by rich side information contained in the KB. Our solution is fully unsupervised and resource-free, requiring neither labeled data nor auxiliary resources. Experimental results show that our solution is simple yet effective, achieving an F1-score of 44.3% which ranks the third in the QEDL task.

Keywords: Entity discovery, entity linking, natural language questions

1 Introduction

The CCKS 2017 question entity discovery and linking (QEDL) task[1] is to recognize entity mentions from natural language questions, and link them with their referent entities in a given knowledge base (KB), i.e., CN-DBpedia [1]. For example, given a question “吴晓敏演过什么电视剧?/What TV shows did Xiaomin Wu play in?”, we should recognize the mention “吴晓敏/Xiaomin Wu” and link it to its referent entity “吴晓敏(演员)/Xiaomin Wu (actress)” in the KB. Such linking results are extremely useful for answering these questions [2].

Entity discovery and linking has long been regarded as a challenging task in natural language processing (NLP) [3, 4]. The specific scenario of CCKS 2017 QEDL further poses new challenges to this traditional NLP task.

- Entities are no longer restricted to the three classical types of person, location, and organization, but instead could be more generic, e.g., “手指/finger” and “发型/hairstyle”. Most of the currently available well-performing systems (usually trained from massive labeled data) can only recognize entities of the three classical types, and hence fail to work here.

* Corresponding author: Quan Wang (wangquan@iie.ac.cn)

[1] http://www.ccks2017.com/?page_id=51

- Questions are too short, containing 13 Chinese characters on average, which cannot provide sufficient contextual information for entity linking.
- Only a small number of training instances are provided, i.e., 1,400 questions with 1,980 entities manually annotated, some of which might even be mislabeled. For instance, in the question "像我这脸型适合剪什么发型?/`What hairstyle fits my facial shape?`", "发型/`hairstyle`" is labeled as an entity while "脸型/`facial shape`" is not, although both of them have redirects in CN-DBpedia. This limited (and potentially inconsistent) supervision makes it difficult to train supervised models for both entity discovery and linking.

To address these challenges, we devise a fully unsupervised method for QEDL. In our approach, entity discovery is conducted based solely on the results of word segmentation and POS tagging. Entity linking is performed by measuring contextual similarity between entity mentions and their referent in the KB. As questions are short with insufficient contexts, we further leverage side information, e.g., titles, types, and primary tags of entities, to refine contextual similarity. Our approach is fully unsupervised and resource-free, requiring neither labeled training data nor auxiliary resources like hand-crafted dictionaries or thesauri. Our approach is simple yet effective, achieving an F1-score of 44.3% which ranks the third in the CCKS 2017 QEDL task.

2 Related Work

Entity discovery is closely related to named entity recognition (NER) which recognizes entities of specific types (person, location, and organization). Studies on NER roughly fall into three categories: 1) rule-based methods which use hand-crafted rules and dictionaries to design recognition patterns; 2) machine learning-based methods which pose NER as a sequence classification problem, solved by hidden Markov models or conditional random fields; 3) hybrid methods which combine rule-based and machine learning-based approaches. For more details about NER methods, please refer to [3].

Entity linking is to link textual mentions with their referent entities in a given knowledge base. Existing approaches can be roughly categorized into two groups: 1) supervised methods which rely on massive annotated data to learn how to rank candidate entities for each textual mention; 2) unsupervised methods which do not require any annotated data to train the ranking model. See [4] for a thorough review of entity linking techniques.

Given that only a small number of annotated data is provided in the QEDL task, we employ a rule-based method for entity discovery, and an unsupervised method for entity linking.

3 Our Approach

Fig. 1 provides a simple illustration of our approach. Given a question, entity discovery is first conducted by using recognition patterns devised on the basis of

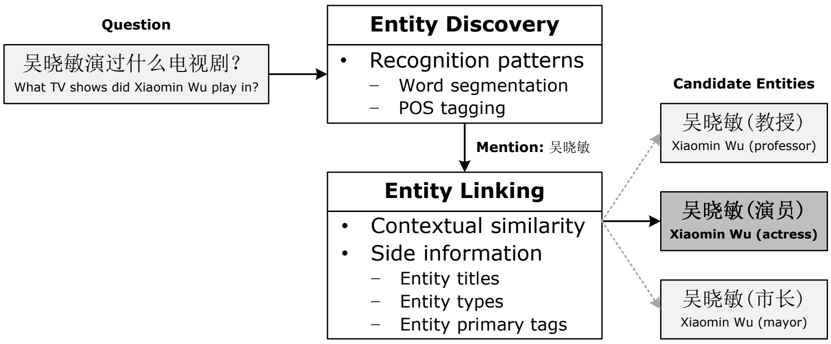

Fig. 1. Simple illustration of our approach.

word segmentation and POS tagging. Entity linking is then performed by measuring contextual similarity between textual mentions and their referent entities, refined by rich side information in the KB. Our approach is fully unsupervised and resource-free, requiring neither labeled training data nor auxiliary resources.

3.1 Entity Discovery

Given an input natural language question, we employ the SWJTU Chinese word segmentation system to perform word segmentation and POS tagging. This system supports two segmentation manners, i.e., coarse-grained and fine-grained segmentations. The former uses a longest matching algorithm, and the latter can split the words into smaller units. For example, "新浪微博/`Sina Microblog`" is segmented into a single word "新浪微博/nt" in coarse-grained segmentation, but two separate words "[新浪/ntc 微博/n]/nt" in fine-grained segmentation. Here, "nt", "ntc", and "n" are POS tags.[2] After word segmentation and POS tagging, we detect entity mentions as follows (see Table 1 for concrete examples).

Rule 1 Words with coarse-grained POS tags of nr (person name), ns (place), and nt (organization) are recognized as entity mentions, e.g., "霍建华", "宁波", and "普陀山".

Rule 2 Words with coarse-grained POS tags of nz (proper noun) and n (noun) are recognized as entity mentions if they have redirects in the KB, e.g., "湘潭火车站". Otherwise, any fine-grained units therein that have redirects are determined as entity mentions, e.g., "中国" in the coarse-grained segmentation "[中国/ns 领土/n]/nz".

Rule 3 Words with coarse-grained POS tags of nz (proper noun) can further be concatenated with their antecedent or succedent words. If the combined words have redirects in CN-DBpedia, they are also identified as entity mentions, e.g., "qq木马病毒" and "三星note2".

[2] They stand for organization, company name, and noun, respectively. A full description of POS tags is available at `http://ics.swjtu.edu.cn`.

Table 1. Examples of recognition rules and corresponding discovered entity mentions.

	Word segmentation & POS tagging results	Entity mentions
Rule 1	霍建华/nr 演/v 过/uguo 哪些/ry 电视剧/n	霍建华
	从/p 宁波/ns 到/v 普陀山/ns 怎么走/nz 最/d 方便/a	宁波, 普陀山
Rule 2	[湘潭/ns 火车站/n]/nz 什么/ry 时候/n 通车/vi	湘潭火车站
	求/v 钓鱼岛/ns 属/v [中国/ns 领土/n]/nz 的/ude 资料/n	中国, 钓鱼岛
Rule 3	qq/x [木马/n 病毒/n]/nz 怎么/ryv 编写/v	qq木马病毒
	三星/nz note2/x 电池/n 怎么样/ryv	三星note2

3.2 Entity Linking

Entity linking consists of three modules: candidate entity selection, candidate entity ranking, and NIL (unlinkable entities) detection, detailed as follows.

Candidate entity selection. For each recognized mention, we query it directly in the CN-DBpedia search engine[3] and retrieve a list of relevant entities. These entities are taken as candidates for that mention.

Candidate entity ranking. We rank candidates for each mention by measuring their contextual similarity. Specifically, given a mention m and a candidate entity e, we construct two feature vectors $\mathbf{m}$ and $\mathbf{e}$ for them. The former is composed of context words of the mention in the question, and the latter context words of the entity in its abstract. Here only words that are tagged as noun and verb are considered. The contextual similarity between m and e can be calculated as, e.g., the dot product of $\mathbf{m}$ and $\mathbf{e}$, i.e., $s(m,e) = \langle \mathbf{m}, \mathbf{e} \rangle$. However, since both the question and the entity abstract are short, we might not get enough contextual information in $\mathbf{m}$ and $\mathbf{e}$. So we propose to further use side information in CN-DBpedia, and calculate a refined contextual similarity $\tilde{s}(m,e) = w \times s(m,e)$. Candidate with the largest $\tilde{s}(m,e)$ score will be selected as the true referent. Three types of side information are considered to calculate the refining factor w, including entity title, entity type, and primary tag.

Entity title is the title of an entity page in CN-DBpedia. The intuition here is that referent entities are usually those that have similar titles with their mentions (string matching). For example, given the mention "格林豪泰", the entity "格林豪泰" is more likely to be the true referent than "林豪泰". So we define the refining factor as string similarity between the mention m and entity title e, i.e.,

$$w_1 = 1 - \frac{\text{edit}(m,e)}{\max(|m|,|e|)},$$

where $|\cdot|$ is the length of a string, and $\text{edit}(\cdot,\cdot)$ the edit distance.

Entity type is the category to which an entity belongs, denoted as $type(e)$. Usually, a mention can only be linked to entities of certain types, e.g., `China`/ns should be linked to countries. So we specify a type set $\mathcal{T}$ for each POS tag, e.g.,

[3] `http://knowledgeworks.cn:30001/?p=**`
`http://knowledgeworks.cn:20313/cndbpedia/api/entity?mention=**`

$\mathcal{T} = \{\texttt{Place}, \texttt{Country}, \texttt{City}\}$ for ns (space), and define the refining factor as

$$w_2 = \begin{cases} 1, & type(e) \in \mathcal{T}, \\ \alpha_1, & type(e) = \emptyset, \\ \alpha_2, & type(e) \notin \mathcal{T} \text{ and } type(e) \neq \emptyset, \end{cases}$$

where $0 \leq \alpha_2 < \alpha_1 < 1$. Here we specify type sets for only three POS tags, i.e., nr (person name), ns (space), and nt (organization).

Primary tags indicate the most popular entities of the given mentions. Candidates with primary tags are more likely to be the true referent. So we define the refining factor according to the presence or absence of primary tags, i.e.,

$$w_3 = \begin{cases} 1, & \text{if } e \text{ has a primary tag}, \\ \beta, & \text{otherwise}, \end{cases}$$

where β is a parameter in the range of $[0, 1)$. These three weights can further be aggregated together, giving a combined refining factor. For example, aggregating all the three weights gives a combined refining factor of $w = w_1 \times w_2 \times w_3$.

NIL detection. Not all mentions have referent entities in the KB. To detect such unlinkable mentions, we use a simple heuristic: mentions with no candidates after performing the candidate entity selection module are predicted to be NIL. To yield more accurate NIL, we do this only for mentions discovered by Rule 1.

4 Experiments

Datasets and evaluation metrics. The training set consists of 1,400 questions with 1,917 mentions manually linked with their referent entities, and 63 mentions labeled as NIL. The test set consists of 749 unlabeled questions. As our approach is fully unsupervised, we use the training data as a development set only for parameter tuning. Submissions are finally evaluated on the test set. Three metrics Precision, Recall, and F1-score are used for the QEDL task.

Implementation details. We use Rule 1, Rule 2, and Rule 3 to detect entity mentions (Section 3.1). And for entity linking (Section 3.2), we test different settings. In the calculation of contextual similarity, we use different term weighting schemes including Boolean, TF, and TF-IDF [5] to compute feature vectors, and we explore two similarity measures, i.e., cosine similarity (Cos) and dot product (Dot). In the calculation of refining factor, we apply each of the three types of side information (Title, Type, and Tag) alone, and get a refining factor of w_1, w_2, and w_3, respectively. We also test all possible combinations, e.g., Title+Type, and get a refining factor of, e.g., $w_1 \times w_2$. Due to the space limitation, we only report the combination with the highest entity linking F1-score on the training set, i.e., Title+Tag which gives a refining factor of $w_1 \times w_3$. All parameters in our approach are determined by maximizing entity linking F1-score on the training set. The optimal configurations are: $\alpha_1 = 0.8$, $\alpha_2 = 0.75$, and $\beta = 0.3$.

Results. Results of entity discovery and linking are shown in Table 2. For discovery, we can see that the recognition rules are simple yet effective in recognizing

Table 2. Results of entity discovery and entity linking.

		Train			Test		
		Precision	Recall	F1	Precision	Recall	F1
Entity Discovery		0.515	0.715	0.599	0.530	0.744	0.619
Entity Linking	Boolean (Cos)	0.193	0.291	0.232	0.246	0.348	0.288
	Boolean (Dot)	0.273	0.376	0.316	0.295	0.415	0.345
	TF (Cos)	0.225	0.310	0.261	0.296	0.419	0.347
	TF (Dot)	0.305	0.424	0.355	0.337	0.477	0.395
	TF-IDF (Cos)	0.283	0.389	0.328	0.298	0.421	0.349
	TF-IDF (Dot)	0.245	0.337	0.283	0.248	0.351	0.291
	TF (Dot)+Title	0.315	0.437	0.366	0.345	0.487	0.404
	TF (Dot)+Type	0.309	0.428	0.359	0.349	0.493	0.408
	TF (Dot)+Tag	0.315	0.437	0.366	0.348	0.491	0.408
	TF (Dot)+Title+Tag	0.328	0.456	0.382	0.378	0.534	0.443

most entity mentions in short questions. However, the rules we created may still miss some difficult cases such as "红/a 米/n note2/x", which will be studied in our future work. For linking, we can see that: 1) the TF term weighting scheme combined with the dot product similarity measure performs the best in calculating contextual similarity; 2) incorporating each of the three types of side information alone can further improve contextual matching; 3) among all possible combinations of side information, Title+Type performs the best, achieving an F1-score of 44.3% on the test set;[4] 4) the performance on the test set is better than that on the training set, which might indicate higher annotation quality of the test data.

5 Conclusion

This paper introduces our solution to the CCKS 2017 QEDL task. We first devise recognition patterns based on word segmentation and POS tagging to discover mentions. Then, we utilize contextual similarity refined by rich side information for entity linking. Our solution is simple yet effective for short questions, achieving an F1-score of 44.3% which ranks the third in the QEDL task.

References

1. B. Xu, Y. Xu, J. Liang, C. Xie, B. Liang, W. Cui, Y. Xiao: CN-DBpedia: A Never-Ending Chinese Knowledge Extraction. In: Proceedings of IEA/AIE, pp. 428-438 (2017)
2. C. Welty, J. W. Murdock, A. Kalyanpur, J. Fan: A comparison of hard filters and soft evidence for answer typing in watson. In: Proceedings of ISWC, pp. 243-256 (2015)
3. A. Mansouri, L. S. Affendy, A. Mamat: Named Entity Recognition Approaches. IJCSNS, vol. 8, no. 2, pp. 339-344, 2008.
4. W. Shen, J. Wang, J. Han: Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions. IEEE T KNOWL DATA EN, 27(2), pp. 443-460 (2015)
5. G. Salton and C. Buckley. Term Weighting Approaches in Automatic Text Retrieval. INF PROCESS MANAGE, 24(5), pp. 513–523 (1988)

[4] During the test phase, we refine the outputs using labeled data in the training set.