Unsupervised Joint Entity Linking over Question Answering Pair with Global Knowledge

Cao Liu[1,2], Shizhu He[1], Hang Yang[1], Kang Liu[1], and Jun Zhao[1,2]

[1] National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China
[2] University of Chinese Academy of Sciences, Beijing, 100049, China
{cao.liu,shizhu.he,hang.yang,kliu,jzhao}@nlpr.ia.ac.cn

Abstract. We consider the task of entity linking over question answering pair (QA-pair). In conventional approaches of entity linking, all the entities whether in one sentence or not are considered the same. We focus on entity linking over QA-pair, in which question entity and answer entity are no longer fully equivalent and they are with the explicit semantic relation. We propose an unsupervised method which utilizes global knowledge of QA-pair in the knowledge base(KB). Firstly, we collect large-scale Chinese QA-pairs and their corresponding triples in the knowledge base. Then mining global knowledge such as the probability of relation and linking similarity between question entity and answer entity. Finally integrating global knowledge and other basic features as well as constraints by integral linear programming(ILP) with an unsupervised method. The experimental results show that each proposed global knowledge improves performance. Our best F-measure on QA-pairs is **53.7%**, significantly increased **6.5%** comparing with the competitive baseline.

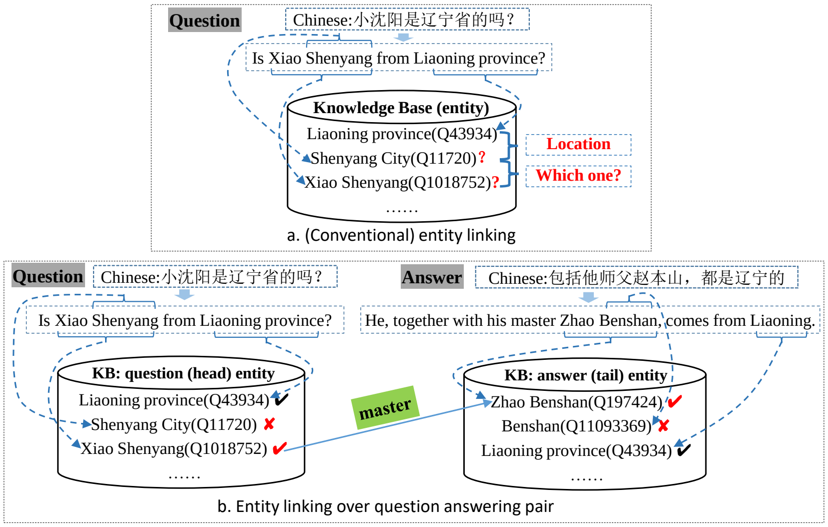
Keywords: joint entity linking, question answering pair, global knowledge, integral linear programming

1 Introduction

Entity Linking(EL) plays an important role in natural language processing, which aims to link text span or name **mention** with **entity** in the knowledge bases [5,7,9,10,2,16]. Entity linking is widely used in Information Extraction(IE), knowledge-based question answering(KB-QA), and some other AI applications. Recently, we witness many large-scale knowledge bases(KBs), such as Freebase[3], DBpedia[1], WikiData[3]. Although they contain lots of structured knowledge in the form of triple(*head entity, relation, tail entity*), there is much missing knowledge in the knowledge bases. On the one hand, entity linking contributes to expanding knowledge bases by extracting unstructured text. On the other hand, entity linking is a key step for developing current knowledge bases to other NLP tasks.

We focus on entity linking over QA-pair, in which the answer is a fluency, correct and coherent response(e.g., answer in Figure 1(b) *He, together with his master Zhao*

[3] https://www.Wikidata.org/wiki/Wikidata:Main_Page


Question
Chinese:小沈阳是辽宁省的吗？
Is Xiao Shenyang from Liaoning province?
Knowledge Base (entity)
Liaoning province(Q43934)
Shenyang City(Q11720)?
Xiao Shenyang(Q1018752)?
……
Location
Which one?
Question
Chinese:小沈阳是辽宁省的吗？
Is Xiao Shenyang from Liaoning province?
Answer
Chinese:包括他师父赵本山，都是辽宁的
He, together with his master Zhao Benshan, comes from Liaoning.
KB: question (head) entity
Liaoning province(Q43934) ✓
Shenyang City(Q11720) ✗
Xiao Shenyang(Q1018752) ✓
……
master
KB: answer (tail) entity
Zhao Benshan(Q197424) ✓
Benshan(Q11093369) ✗
Liaoning province(Q43934) ✓
……

a. (Conventional) entity linking

b. Entity linking over question answering pair

Fig. 1. An illustration of entity linking and entity liking over question answering pair

Benshan, comes from Liaoning.), rather than the solitary entity or phrase. Such answer provides a friendly interaction for human-machine. Furthermore, it provides some explanation of answering process which could be used to answering verification and is better to support downstream tasks such as synthetic speech[12]. These QA-pairs widely appears in community website, such as Quara[4], Wiki.answer[5], Baidu Zhidao[6] and so on. Yin et al proposed generative natural answers in sequence-to-sequence Generative-QA based on Chinese community website including Baidu Zhidao[20]. Entity linking over QA-pair, as a kind of entity linking, not only contributes to the development of entity linking, but also benefits to choose QA-pairs which suit for answering automaticly[20].

Entity linking over QA-pair is different from conventional entity linking. Firstly, entity linking is multi-sentences and multi-entities linking, which inputs at least two sentences including question and answer. So the number of entities is uncertain. The significant difference, entity linking over QA-pair considers the explicit semantic relation in the KB between question entity(entity in the question) and answer entity(entity in the answer), while traditional (collective) entity linking takes the coherent topic or semantic into consideration[9,10], and all entities whether in one sentence or not are considered the same. Therefore, it is lack of constraints on the explicit semantic relation. As for the question *Is Xiao Shenyang from Liaoning province?* shown in Fig 1(a), mention *Xiao Shenyang*, *Shenyang* and *Liaoning province* correspond entity *Xiao Shenyang(Q1018752)*, *Shenyang City (Q11720)* and *Liaoning province(Q43934)* in the KB respectively. Both of *Shenyang City(Q11720)* and *Liaoning province(Q43934)* are locations, and they are close in the topic space. So it is likely to link the wrong entity *Shenyang City(Q11720)* rather than *Xiao Shenyang(Q1018752)* for conventional

[4] https://www.quora.com/

[5] http://www.answers.com/Q/

[6] https://zhidao.baidu.com/

entity linking. As for entity linking over QA-pair, question entity and tail entity are constrained on the explicit semantic relation, e.g., triple *(head entity, relation, tail entity)*. One basic hypothesis is that question entity is one of *head entity* and *tail entity*, and the answer entity is the other. In most cases, question entity and answer entity are *head entity* and *tail entity* respectively[20]. So question entity and answer entity are no longer fully equivalent. As shown in Fig 1(b), there is relation *master* between question entity *Xiao Shenyang(Q1018752)* and answer entity *Zhao Benshan(Q197424)*. If taking such explicit semantic relation into consideration, it is more likely to link the correct entity *Xiao Shenyang(Q1018752)* rather than *Shenyang City(Q11720)*.

In this paper, firstly, we collect 5,546,743 QA-pairs from Baidu Zhidao and get their corresponding triples in Wikidata for all the QA-pairs. Furthermore, we exploit global knowledge of QA-pair in the KB for entity linking. The most significant global knowledge are: 1) The probability of relation between question entity and answer entity. The higher probability means that these entities are more likely to be linked. We train TransE[4] to represent the entity, then using multi-layer perceptrons to calculate the probability of relation between question entity and answer entity. 2) Linking similarity between question entity and answer entity in the KB. We count the same entities which for question entity and answer entity linked to. Finally, Integral linear programming(ILP)[18,11] integrates the above as well as some other basic features and constraints. Specifically, ILP is unsupervised and convenient to increase or decrease features and constraints. The experimental results show that each proposed knowledge improves performance. Our best F-measure on QA-pairs is **53.7%**, significantly increased **6.5%** comparing with the competitive baseline.

2 Task and Data

2.1 Task description

The input, task of entity linking over QA-pairs, is natural question and answer. All mention-entity pairs in the QA-pair should be returned. e.g., as shown in Figure 1(b), **Question:** *Is Xiao Shenyang from Liaoning province?* and **Answer:** *He, together with his master Zhao BenShan, comes from Liaoning*. Mention-entity: *Xiao Shenyang-Xiao Shenyang(Q1018752)* for question, *Liaoning(province)-Liaoning province(Q43934)* for question and answer, and *Zhao BenShan-Zhao BenShan(Q197424)* for answer should be returned. Other mentions such as *Shenyang* and *BenShan* are noise. In fact, their entity *Shenyang City(Q11720)* is nearly to *Liaoning province(Q43934)* in semantic space, and the *Benshan(Q11093369)* is another entity of person.

2.2 Data

To research the task of entity linking over QA-pair, we construct a new database collected from the Internet. The dataset and extracted process as follows:

1. **QA-pairs**: We crawl HTML files from Baidu Zhidao and extract QA-pairs from them. We obtain 5,546,743 QA-pairs(Table 1) after filtering these which either question or answer is longer than 50 in the number of characters. As for the task of entity linking, if question or entity do not contain entity, discarding it.

Baidu Zhidao	extracted knowledge base		KB corresponding to QA-pairs	
#QA-pairs	#triples	#entities	#triples	#entities
5,546,743	80,421,642	22,450,412	3,581,158	1,069,593

Table 1. Dataset of QA-pairs and KB

2. **Candidate mention and entity**: We use the tool FEL[2,16] to get the mentions, entities and their scores($Score_{fel}$). Especially, one mention may correspond more than one entity. Each entity is one to one correspondence on Wikipedia. All of them as candidates.
3. **Knowledge base**: We extract structured triple*(head entity, relation, tail entity)* from Wikidata. In particular, Wikidata is public and convenient to obtain. It is language-independent, which links to hundreds of languages and makes up to the lack of KB in Chinese. We totally get 80,421,642 triples and 22,450,412 entities. Some entities of Wikidata correspond to Wikipedia entity with simplified or complex Chinese. Fortunately, the entity outputted on the Tool *FET* is Wikipedia entity too. So, our entity in QA-pair links to Wikidata by Wikipedia entity. Eventually, these QA-pairs match 3,581,158 triples and 1,069,593 different entities.

After getting QA-pairs, candidate mentions and entities, the key challenge in entity linking is to choose appropriate mentions and their corresponding entities from candidates. Due to the lack of labeled data, it's hard to use supervised or semi-supervised methods. To make use of the question, answer and knowledge bases with unsupervised way, we take advantage of integral linear programming(ILP) to integrate global knowledge between question entity and answer entity on the next section.

3 Methodology

Overall structure of integral linear programming for entity linking over QA-pair illustrates in Figure 2. As for each QA-pair, all candidate mentions and entities are the variables which equal 0 or 1. The objective function contains different features to guarantee that the selected mention or entity are relevant, consistent, correct. Because ILP is unsupervised, we can design different features and constraints to decide which of them are effective. We consider two important features as global knowledge: 1) The probability of relation between question entity and answer entity(noted as $Score_{pro\_rel}$). If there is the semantic relation between question entity and answer entity, such as *(question entity, relation, answer entity)*, these entities are more likely to be linked. 2) Linking similarity between question entity and answer entity($Score_{link\_sim}$). The more same entities which question entity and answer entity link to, the higher possibility that linking to these entities. Beside, there are some basic features: the score of FEL($Score_{fel}$), and the length of mention($Score_{len\_men}$). As for constraints, we consider as follows: Selected mention can not contain or overlap, the maximum number of linked mentions and entities in the question or answer, the number of one mention corresponds an entity at most and so on. Finally, combining all scores of features as optimized objections and constraints to ILP, then obtaining the linked mentions and entities.

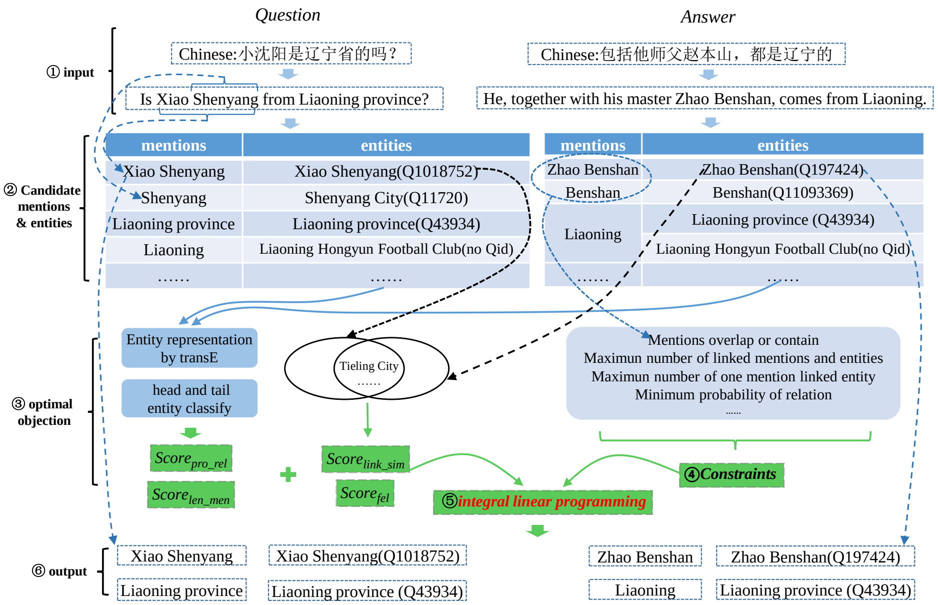

Question
Answer
① input
Chinese:小沈阳是辽宁省的吗？
Is Xiao Shenyang from Liaoning province?
Chinese:包括他师父赵本山，都是辽宁的
He, together with his master Zhao Benshan, comes from Liaoning.
② Candidate mentions & entities
mentions
entities
Xiao Shenyang
Xiao Shenyang(Q1018752)
Shenyang
Shenyang City(Q11720)
Liaoning province
Liaoning province(Q43934)
Liaoning
Liaoning Hongyun Football Club(no Qid)
……
Zhao Benshan
Zhao Benshan(Q197424)
Benshan
Benshan(Q11093369)
Liaoning
Liaoning province (Q43934)
Liaoning Hongyun Football Club(no Qid)
③ optimal objection
Entity representation by transE
head and tail entity classify
Tieling City
Mentions overlap or contain
Maximun number of linked mentions and entities
Maximun number of one mention linked entity
Minimum probability of relation
Score_pro_rel
Score_len_men
+
Score_link_sim
Score_fel
④Constraints
⑤integral linear programming
⑥ output
Xiao Shenyang
Liaoning province
Xiao Shenyang(Q1018752)
Liaoning province (Q43934)
Zhao Benshan
Liaoning
Zhao Benshan(Q197424)
Liaoning province (Q43934)

Fig. 2. Overall structure of integral linear programming for entity linking over QA-pair

3.1 Features

The probability of relation between question entity and answer entity This step aims to calculate the probability of relation between question entity and answer entity. The better probability means that the question entity and answer entity exist more semantic relation in the knowledge base, the two entities are more likely to be linked entities. The main steps are entity representation and classification.

a) Entity representation: Entity representation aims to embed entity into low dimensional space. We use the transE[4]. The basic idea of transE is the relational hypothesis of head entity and tail entity: $h + r \approx t$, where h, r, t denotes head entity, relation and tail entity respectively, such as $Xiao\,Shenyang(Q1018752) + master \approx Zhao\,BenShan(Q197424)$. After getting all the entities to Wikidata for QA-pairs. We train the transE model with the following formulations:

$$L = \sum_{(h,r,t)\in S} \sum_{(h^{'},r,t^{'})\in S^{'}_{h,r,t}} [\gamma + d(h,r,t) - d(h^{'},r,t^{'})]_{+} + \alpha\|\theta\| \tag{1}$$

where $[x]_+$ is $max(0,x)$, $\gamma(\gamma>0)$ denotes margin hyperparameter. $\|\theta\|$is the regular term. S is the positive triples(h,r,t), while $S^{'}$ is negative triple by random replacing h or t, but a negative example only replace one of h and t, as:

$$S^{'} = \{(h^{'},r,t)\} \cup \{(h,r,t^{'})\} \tag{2}$$

The distance between h, r and t notes $d(h,r,t)$, and:

$$d(h,r,t) = \sum_{i\in D}(h_i + r_i - t_i) \tag{3}$$

where D is the dimensionality of entity, we calculate errors of h, r, t directly.

b) Calculating the probability of relation by classifying question entity and answer entity: Entity representation by transE is used to the input of calculating probability of relation. For triple*(head entity, relation, tail entity)*, we view *head entity* combined with *tail entity* as the positive instance, and their expected probability of relation is 1. We random sample negative example by replacing one of head entity and tail entity, and their expected probability is 0.

Here, a *softmax* classifier with two-layer MLP(multi-layer perceptron) is used to calculate the probability of relation. The middle layer used the rectified linear unit (ReLU) as activation function. Finally, we get the score of *head entity* and *tail entity* with relation, noted as $Score_{pro\_rel}$.

We consider question entity and answer entity are head entity and tail entity respectively. Each question entity and answer entity pair, can get the probability of relation. As shown in Figure 2, because of the triple *(Xiao Shenyang(Q1018752), master, Zhao BenShan(Q197424))*, the $Score_{pro\_rel}$ for either question entity *Xiao Shenyang(Q1018752)* or answer entity *Zhao BenShan(Q197424)* is high.

Linking similarity between question entity and answer entity: The probability of relation between question entity and answer entity consider the direct relation. Besides, the same entities which link to both question entity and answer entity are another feature for entity linking. For example(shown in Figure 2), *Xiao Shenyang(Q1018752)* links to *Tieling city(Q75268)*, and *Tieling city(Q75268)* links *Zhao Benshan(Q197424)* too. We count the mutual linked entity for question entity and answer entity. One question entity may correspond more than one answer entities, such as question entity *Xiao Shenyang(Q1018752)* corresponds different answer entities *Zhao BenShan(Q197424)*, *BenShan(Q11093369)* and so on, which each of them is with linking similarity. So do other entities. Extracting the maximum as $Score_{link\_sim}$.

Basic Features: Besides the probability of relation and linking similarity. There are some other features. Firstly, the FEL tool gives each mention-entity pair a confident score when getting the candidate of mention and entity. The score is negative, the more approaching zero means better confident score. Adding a constant and becoming to the positive number. The confidence of FEL notes as $Score_{fel}$. Secondly, the length of mention contributes to entity linking. Intuitively, most entities are linked by mentions which are not too long. While long mentions link to entities usually possess high performance. Basing on the above observation, we add the length of mention as another feature, marked as $Score_{len\_men}$.

3.2 Model: Entity linking over QA-pair by integral linear programming(ILP)

Integral linear programming(ILP) is optimal problem under constraint condition. *'Integral'* means the variable is integral. The variable is usually binary. The binary variable represents selecting the variable or not. The definition of ILP with mathematical formula[15] as follows:

$$\begin{aligned} maximize \quad & c^T x \\ subject\ to \quad & Ax \leq b \\ & x \in \{0,1\} \end{aligned} \tag{4}$$

x is the variable which constraint in 0 or 1. Under the constraint $Ax \leq b$, getting the maximize objection $c^T x$.

As for entity linking over QA-pair by ILP, the above features(Score) can be the optimal objection of ILP. Adding some constraints to constitute the whole ILP. Integrating the different scores for question and answer. The mathematical optimizational objection is:

$$maximize \quad Score_{question} + Score_{answer} \tag{5}$$

where, $Score_{question}$ and $Score_{answer}$ are the total score of question and answering respectively, calculated as follows:

$$Score = c^T [Score_{fel}, Score_{len\_men}, Score_{link\_sim}, Score_{pro\_rel}] \tag{6}$$

c is the weight of features. The constrains of question and answer are:

1. **Mentions overlap or contain**: Selecting overlap or contain mentions is forbidden. For example, the mention *Xiao Shenyang* contains the mention *Shenyang*, so the two mentions are selected one at most, eventually.
2. **Maximun number of linked mentions and entities**: Choosing too many mentions or entities is more likely to bring noisy mentions and entities. It is necessary to set an appropriate threshold for maximum number of mentions and entities. Due to the unsupervised character of ILP, it is easy to change the threshold for different applications.
3. **Maximun number of one mention linked entities**: If mention links more than one entity, the ambiguity still exists. So a mention links one entity at most.
4. **Minimum probability of relation:** If the probabilities of relation for question entity to each answer entity are low, the most possibility is that the candidate question entity is improper. So does the answer entity. For example(shown in Figure 2), the question mention *Shenyang* has a candidate entity *Shenyang Taoxian International Airport*. This entity is low probability of relation to all answer entities. In fact, it is wrong to link it. In our experiment, if the maximum probability of relation is small and less than the threshold, discard it.

Above are the optimal objection and their constraints. They can combine, remove and add randomly. If the entity as well as it's corresponding mention variable equals to 1, these mention-entity pairs are the final outputs.

4 Experiment

4.1 Dataset and Evaluation Metric

We extracted QA-pairs from Baidu Zhidao as the dataset. Due to the unlabeled mentions and entities, we invited the volunteer to label data for evaluation. Different mentions may link to the same entity, such as mention *Liaoning* and *Liaoning Province* are linked to entity *Liaoning province(Q43934)*. To be convenient for evaluation, we just label linked entity on QA-pairs. In fact, if the final entity is correct, the mention is less important. The volunteer labels 200 QA-pairs in total. To evaluate the performance in the question and answer, labeling question entity and answer entity respectively. In special, for testing system on one mention corresponding to one or multi candidate entities(such as: mention *Liaoning* links 2 candidate entities: *Liaoning Province* and *Liaoning Hongyun Football Club*, some mention may correspond only one entity). That one linked mention corresponds to multi-entities notes as **1-m**. And that one linked mention corresponds to only one candidate entity is **1-1**. We distinguish **1-m** and **1-1** in the question and answer by splitting QA-pairs as: 1) **QA:1-1** All linked mentions are one to one for entities in QA-pair. 2) **Q:1-m** Existing **1-m** only in the question. 3) **A:1-m** Existing **1-m** only in the answer. 4) **QA:1-m** Existing **1-m** in both question and answer. Each of them is 50 QA-pairs.

4.2 Evaluation Metric

We utilize standard precision, recall and F-measure to evaluate entity linking performance[7]. Where precision is the proportion for correctly returned entities to all returned entities, recall is the correctly returned entities to all labeling entity, F-measure reconciles precision and recall, they are:

$$precision = \frac{\|List_{return} \bigcap List_{label}\|}{\|List_{return}\|} \tag{7}$$

$$recall = \frac{\|List_{return} \bigcap List_{label}\|}{\|List_{label}\|} \tag{8}$$

$$F_1 = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{9}$$

4.3 Comparison Models

Our candidate mention-entity comes from FEL[2,16]. Mention-entity of FEL as well as confident score is pretty good. ILP with $Score_{fel}$ and constraints(except probability of relation) for candidate mention-entity of FEL is our baseline, noted *FEL* in following Table. *+len_men* uses $Score_{fel}$ and $Score_{len\_men}$ as optimal objection. *+link_sim* optimize $Score_{fel} + Score_{len\_men} + Score_{link\_sim}$. While *pro_rel* continues to add optimal objection $Score_{pro\_rel}$. In particular, each question(or answer) entity corresponds more than one probabilities of relation. That calculating the sum, maximum and average are make sense. If no special explaination, probability of relation is the average. *Questions* and *Answers* stand for evaluating in the question and answer respectively, while *QA-pairs* represent performance on both question and answer. By the way, all the performance is percentage(%). Specially, we compare different methods on QA-pair, single question or answer on the four label datasets.

[7] http://nlp.cs.rpi.edu/kbp/2014/scoring.html

4.4 Overall performance

We evaluate the performance of different methods on the *Questions*, *Answers* and *QA-pairs*. The overall performance on test data is shown in Table 2. The conclusions are:

Methods	*Questions*			*Answers*			*QA-pairs*		
	P	R	F	P	R	F	P	R	F
FEL	46.3	61.7	52.9	33.1	53.3	40.8	39.9	58.0	47.2
+*len_men*	51.8	68.7	59.0	34.8	56.2	43.0	43.5	63.2	51.5
+*link_sim*	51.9	**69.2**	**59.3**	36.5	59.2	45.2	44.4	**64.8**	52.7
+*pro_rel*	**52.5**	65.0	58.0	40.2	**62.1**	**48.8**	**46.4**	63.7	**53.7**

Table 2. Overall performance

1. Each feature improves performance on *QA-pairs*. Taking the length of mention into consideration improves prominently.
2. ***+link_sim*** as well as ***+pro_rel*** contribute to improve performance. Both of them are global knowledge of QA-pair as well as their knowledge in the KB.
3. The entity linking performance on the *Questions* is superior to the *Answers* for the whole data. Intuitively, QA-pairs come from the community website. Asking the question aims at solving the question, The question is usually specific while the answer is uncertain. So entity linking in the *Questions* is easier than entity linking in the *Answers*.
4. The best F-measure on QA-pairs is **53.7%**, improving apparently **6.5%** compared with *FEL* **47.2%**.

4.5 Performance on one mention corresponding to different number of entities

To evaluate performance of **1-m** on the question and answer respectively, we compare our model on **QA:1-1**, **Q:1-m**, **A:1-m** and **QA:1-m**. The detail results are shown in Table 3. We can get:

Methods	Datas	**QA:1-1**			**Q:1-m**			**A:1-m**			**QA:1-m**		
		P	R	F	P	R	F	P	R	F	P	R	F
FEL	*Questions*	55.1	69.0	61.3	50.0	62.2	55.4	44.8	60.6	51.5	44.3	62.3	51.8
	Answers	36.8	53.3	43.5	26.9	44.6	33.6	37.8	54.0	44.4	34.8	**62.0**	44.6
	QA-pairs	46.0	61.8	52.8	38.4	54.6	45.1	41.4	57.5	48.1	39.8	62.2	48.5
+*len_men*	*Questions*	60.7	76.1	67.5	55.4	68.9	61.5	51.6	69.0	59.0	48.5	*68.1*	**56.6**
	Answers	43.7	63.3	51.7	32.3	53.6	40.3	37.4	54.0	44.2	34.8	**62.0**	44.6
	QA-pairs	52.3	70.2	59.9	43.8	62.3	51.4	44.6	61.9	51.9	41.9	65.6	51.2
+*link_sim*	*Questions*	57.3	71.8	63.8	**59.8**	**74.3**	**66.3**	49.0	66.2	56.3	47.4	66.7	55.4
	Answers	39.8	58.3	47.3	32.3	53.6	40.3	**44.6**	**65.1**	**52.9**	32.6	58.0	41.7
	QA-pairs	48.6	65.7	55.8	**46.0**	**65.4**	**54.0**	46.8	65.7	54.7	40.3	**63.0**	49.2
+*pro_rel*	*Questions*	**67.9**	**80.3**	**73.6**	46.6	55.4	50.6	**61.8**	**77.5**	**68.8**	**48.9**	62.3	54.8
	Answers	**47.1**	**66.7**	**55.2**	**37.4**	**60.7**	**46.3**	44.0	63.5	52.0	**39.2**	**62.0**	**48.1**
	QA-pairs	**57.4**	**74.1**	**64.7**	41.9	57.7	48.5	**52.8**	**70.9**	**60.5**	**44.3**	62.2	**51.8**

Table 3. Performance on mention corresponding different number of entities

1. Simple situation(**QA:1-1**) gets better than complex cases (**Q:1-m**, **A:1-m** and **QA:1-m**) for all methods on F-measure. It proves that **1-m** is more challenge than **1-1**.
2. When adding linking similarity, performance on *Questions* improved much for **Q:1-m** while performance on *Answers* is in low level, and performance on *Answers* of **A:1-m** achieved the best performance while performance of *Questions* is low. However, ***+pro_rel*** improves performance on one of *Questions* and *Answers*, and the other maintains good relatively at the same time. It implies that ***+pro_rel*** keeps the balanced performance on the *Questions* and *Answers* when improving one of them.
3. On most of situations, ***+pro_rel*** achieved the best performance. Which proved again that all of our features are effective. Especially, the probability of relation improves performance at last.

4.6 Performance on different forms to the probabilities of relation between question entity and tail entity

The above experiments show that the probability of relation is an important feature. $Score_{pro\_rel}$ can be the sum, maximum and average(noted ***pro_rel_sum***, ***pro_rel_max*** and***pro_rel_ave*** respectively) when question(answer) entity calculates the probability of relation with different answer(question) entities. Table 4 shows the results on different form to calculate the probability of relation. ***pro_rel_ave*** achieved the best performance on the whole situations as well as different evaluation metrics. Intuitively, the sum may bring some noise and the maximum will get good performance. while ***pro_rel_max*** superiors ***pro_rel_sum*** a little and inferiors to ***pro_rel_ave***. One guess is that the maximum is influenced largely by noise. We look forward the performance on the probability of relation between question entity and answer entity. The precisions are 85.6% for positive example, 86.6% for negative example, respectively. Although the performance is pretty good, it still exists noise which make the maximum bad performance.

Methods	*Questions*			*Answers*			*QA-pairs*		
	P	R	F	P	R	F	P	R	F
pro_rel_sum	50.8	62.6	56.1	39.9	61.5	48.4	45.3	62.1	52.4
pro_rel_max	51.7	64.0	57.2	39.9	61.5	48.4	45.8	62.9	53.0
pro_rel_ave	**52.5**	**65.0**	**58.0**	**40.2**	**62.1**	**48.8**	**46.4**	**63.7**	**53.7**

Table 4. Performance on different forms to the probabilities of relation

5 Related Work

Entity linking is a foundational research in natural language processing. Many works researched on entity linking. Mihalcea & Csomai use cosine distance to calculate between mention and entity[6]. Milne et al calculate the mention-to-entity compatibility by using inter-dependency of mention and entity[14]. Zhou et al propose ranking-based and classification-based resolution approaches which disambiguate both entities and word senses[22]. While it is lack of global constraints. Han et al propose Structural Semantic Relatedness and collective entity linking[10,9]. Medelyan et al take the semantic relatedness of candidate entity as well as contextual entities into consideration[13]. These semantic relations of this work are relatively simple. Blanco et al's multilingual entity extraction and linking with fast speed(named as fast entity linking(FEL)) and high performance[2,16]. It divides entity linking into mention detection, candidate entity retrieval, entity disambiguation for mentions with multiple candidate entities and mention clustering for mentions that do not link to any entity. This paper utilizes less feature to realize multilingual, fast and unsupervised entity linking with high performance.

As for entity linking on question answering over knowledge base, [17] using Smart (Structured Multiple Additive Regression Trees) tool[19] for entity linking, which returned all the possible candidate entity for freebase by surface matching and ranking via statistical model. Dai et al realize the importance of entity linking on KB-QA[8]. They explore entity priority or relation priority. The candidate entities are large, while relation is with a small number. Determining firstly relation contributes to entity linking for reducing candidates. Yin et al come up with active entity linker by sequential labeling to search surface pattern in the entity vocabulary lists[21].

In short, these methods consider all entities whether in one sentence or not are the same. However, question entity and answer entity in QA-pair usually represent head entity and tail entity respectively with the explicit semantic relation. So we take the semantic relation of question entity and answer entity into consideration.

6 Conclusion

This paper proposes a novel entity linking over question answer pair. Differring from traditional entity linking which considers the coherent topic or semantic and all the entity are the same. Question entity and answer entity are no longer fully equivalent, and they are constrained with the explicit semantic relation. We collect a large-scale Chinese QA-pairs along with their corresponding triples as knowledge base, and propose unsupervised integral linear programming to get the linked entities of QA-pair. The main steps of our method: 1) Retrieving candidate mentions and entities, 2) Setting optimal objection. The main objections are the probability of relation and linking similarity between question entity and answer entity, which are the global knowledge of QA-pair and could be used to semantic constraints. 3) Adding some constraints of mention and entity. 4) Combining optimal objection and constraints to integer linear programming, and obtaining target mention and entity. The experimental results show that each proposed global knowledge improves performance. Our best F-measure on QA-pairs is **53.7%**, significantly increased **6.5%** comparing with the competitive baseline.

Acknowledgements

This work was supported by the Natural Science Foundation of China (No. 61533018) and the National Basic Research Program of China (No. 2014CB340503). And this research work was also supported by Google through focused research awards program.

References

1. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: Dbpedia: A nucleus for a web of open data. The semantic web pp. 722–735 (2007)

2. Blanco, R., Ottaviano, G., Meij, E.: Fast and space-efficient entity linking in queries. In: Proceedings of the Eight ACM International Conference on Web Search and Data Mining. WSDM 15, ACM, New York, NY, USA (2015)
3. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: a collaboratively created graph database for structuring human knowledge. In: Proceedings of the 2008 ACM SIGMOD international conference on Management of data. pp. 1247–1250. AcM (2008)
4. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. In: Advances in neural information processing systems. pp. 2787–2795 (2013)
5. Bunescu, R.C., Pasca, M.: Using encyclopedic knowledge for named entity disambiguation. In: Eacl. vol. 6, pp. 9–16 (2006)
6. Csomai, A., Mihalcea, R.: Linking documents to encyclopedic knowledge. IEEE Intelligent Systems 23(5) (2008)
7. Cucerzan, S.: Large-scale named entity disambiguation based on wikipedia data (2007)
8. Dai, Z., Li, L., Xu, W.: Cfo: Conditional focused neural question answering with large-scale knowledge bases. arXiv preprint arXiv:1606.01994 (2016)
9. Han, X., Sun, L., Zhao, J.: Collective entity linking in web text: a graph-based method. In: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. pp. 765–774. ACM (2011)
10. Han, X., Zhao, J.: Named entity disambiguation by leveraging wikipedia semantic knowledge. In: Proceedings of the 18th ACM conference on Information and knowledge management. pp. 215–224. ACM (2009)
11. Khachiyan, L.G.: Polynomial algorithms in linear programming. USSR Computational Mathematics and Mathematical Physics 20(1), 53–72 (1980)
12. McTear, M., Callejas, Z., Griol, D.: The conversational interface. Springer (2016)
13. Medelyan, O., Witten, I.H., Milne, D.: Topic indexing with wikipedia. In: Proceedings of the AAAI WikiAI workshop. vol. 1, pp. 19–24 (2008)
14. Milne, D., Witten, I.H.: Learning to link with wikipedia. In: Proceedings of the 17th ACM conference on Information and knowledge management. pp. 509–518. ACM (2008)
15. Papadimitriou, C.H., Steiglitz, K.: Combinatorial optimization: algorithms and complexity. Courier Corporation (1982)
16. Pappu, A., Blanco, R., Mehdad, Y., Stent, A., Thadani, K.: Lightweight multilingual entity extraction and linking. In: Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. WSDM 17, ACM, New York, NY, USA (2017)
17. Xu, K., Reddy, S., Feng, Y., Huang, S., Zhao, D.: Question answering on freebase via relation extraction and textual evidence. arXiv preprint arXiv:1603.00957 (2016)
18. Yahya, M., Berberich, K., Elbassuoni, S., Ramanath, M., Tresp, V., Weikum, G.: Natural language questions for the web of data. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. pp. 379–390. Association for Computational Linguistics (2012)
19. Yang, Y., Chang, M.W.: S-mart: Novel tree-based structured learning algorithms applied to tweet entity linking. arXiv preprint arXiv:1609.08075 (2016)
20. Yin, J., Jiang, X., Lu, Z., Shang, L., Li, H., Li, X.: Neural generative question answering. Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16) Neural (2016)
21. Yin, W., Yu, M., Xiang, B., Zhou, B., Schütze, H.: Simple question answering by attentive convolutional neural network. arXiv preprint arXiv:1606.03391 (2016)
22. Zhou, Y., Nie, L., Rouhani-Kalleh, O., Vasile, F., Gaffney, S.: Resolving surface forms to wikipedia topics. In: Proceedings of the 23rd International Conference on Computational Linguistics. pp. 1335–1343. Association for Computational Linguistics (2010)