Cross-Lingual Entity Query from Large-Scale Knowledge Graphs

Yonghao Su, Chi Zhang, Jinyang Li, Chengyu Wang, Weining Qian(✉), and Aoying Zhou

Institute for Data Science and Engineering, ECNU-PINGAN Innovative Research Center for Big Data, East China Normal University, Shanghai, China
{suyonghao,chizhang,jinyangli,chengyuwang}@ecnu.cn,
{wnqian,ayzhou}@sei.ecnu.edu.cn

Abstract. A knowledge graph is a structured knowledge system which contains a huge amount of entities and relations. It plays an important role in the field of named entity query. DBpedia, YAGO and other English knowledge graphs provide open access to huge amounts of high-quality named entities. However, Chinese knowledge graphs are still in the development stage, and contain fewer entities. The relations between entities are not rich. A natural question is: how to use mature English knowledge graphs to query Chinese named entities, and to obtain rich relation networks. In this paper, we propose a Chinese entity query system based on English knowledge graphs. For entities we build up links between Chinese entities and English knowledge graphs. The basic idea is to build a cross-lingual entity linking model, RSVM, between Chinese and English Wikipedia. RSVM is used to build cross-lingual links between Chinese entities and English knowledge graphs. The experiments show that our approach can achieve a high precision of 82.3 % for the task of finding cross-lingual entities on a test dataset. Our experiments for the sub task of finding missing cross-lingual links show that our approach has a precision of 89.42 % with a recall of 80.47 %.

Keywords: Cross-lingual entity linking · Knowledge graph · Entity disambiguation · Semantic query

1 Introduction

Over the past years, the amount of knowledge grows rapidly, but stored in an unstructured way. Knowledge graphs can describe entities structurally, and we can get attributes about entities. For example, when we query entity "图灵", *Turing* in English, in a Chinese knowledge graph, we may get some information about *Turing*, such as *Turing* was male and came from the UK. But we want to know more about *Turing*, such as which university he graduated from. However, Chinese knowledge graphs contain fewer entities and relations between entities are not rich [14]. We can not fully describe the entity *Turing* in Chinese. English knowledge graphs contain rich entities and relations, which can describe

R. Cai et al. (Eds.): APWeb 2015 Workshops, LNCS 9461, pp. 139–150, 2015.
DOI: 10.1007/978-3-319-28121-6_13

entities comprehensively [3,9,10,18]. But the cross-lingual links between English knowledge graphs and Chinese knowledge graphs are rare. It will lead to a low recall to use English knowledge graph directly. Typically, the cross-lingual links are manually added by authors of articles and are incomplete or erroneous. When the author of an article does not link to an article which expresses to the same concept in an other language version of Wikipedia. This is called a *missing cross-lingual link*.

It is vital to find such entity in English knowledge graphs with the same meaning of the Chinese query entity. It needs to solve two problems: (a) entity disambiguation and (b) cross-lingual entity linking. The challenge (a) has already been addressed by other researchers [4,5,7,11]. A group of highly related works for challenge (b) has been proposed by [1,6,12,15,16]. But these algorithms do not fit our query task. The methods of entity disambiguation are based on the same language version of Wikipedia, but we need solve the entity disambiguation task in two different versions of Wikipedia. Moreover, the cross-lingual algorithms can not deal with the structure challenge in Wikipedia as shown in Fig. 2. The problem of query Chinese entities in English knowledge graphs is non-trivial and challenging, summarized as follows:

Cross-Lingual Entity Disambiguation. We treat cross-lingual entity disambiguation task as two sub tasks: entity disambiguation and re-ranking candidate entities disambiguated in other language version of knowledge graphs. Existing methods for entity disambiguation are for the same language. They can not be used for cross-lingual entity disambiguation directly. We propose a method that use a vector space model to solve entity disambiguation problem, and get a set of candidates. With the help of our cross-lingual entity linking module, we re-rank the candidates to achieve cross-lingual entity disambiguation.

Cross-Lingual Entity Linking. A large amount of new knowledge is frequently added in Chinese knowledge graphs or English knowledge graphs. The structures of these new knowledge graphs are sparse. But exiting methods heavily depend on structure features, thus we must find other unstructured features to describe the relations between cross-lingual entities.

In this paper, we propose a cross-lingual entity query system CLEQS based on Chinese Wikipedia, English Wikipedia and links between them. The novelties of CLEQS as shown below: (a) We can find missing entity links between Chinese knowledge graphs and English knowledge graphs. (b) In entity disambiguation task we obtain a set of candidates instead of only one candidate and (c) we design a method to re-rank the candidates with the help of structure relations in YAGO [9], which contains 1 million entities and 5 million facts, as the English data source. In this way, we get high precision in the cross-lingual query task, especially in the entity disambiguation task. (d) Because the structure features are less important when more and more entities are added into Wikipedia as shown in Fig. 2, we pay more attention to semantic features and design a set of semantic features to improve query accuracy.

More precisely, given a Chinese entity mention and its context, CLEQS uses a vector space model in order to effectively identify a set of candidate entities in Chinese Wikipedia. For each candidate entity in the resulting candidate set, a *ranking SVM* model with a set of structure and semantic features are used to find the cross-lingual links in YAGO, and finally we use structure features in YAGO to re-rank the candidate result sets. We evaluate CLEQS and the two sub tasks of CLEQS on a dataset of 1000 pairs of articles. The result that we obtain show that CLEQS performs very well.

The remainder of the paper is organized as follows. Section 2 outlines some related work. Section 3 formally defines the problem of knowledge linking and some related concepts. Section 4 describes the proposed cross-lingual query approach. Section 5 presents the evaluation results and finally Sect. 6 concludes this work.

2 Related Work

The problem of entity disambiguation has been addressed by many researchers starting from Bagga and Baldwin [4], who use the bag of words and vector cosine similarity to represent the context of the entity mention. Jiang et al. [7] adopt the graph based framework to extend the similarity metric to disambiguate the entity mentions effectively. Researchers have shown a great interest in mapping textual entity mention to its corresponding entity in the knowledge base. Bunescu and Pasca [5] firstly deal with this problem by extracting a set of features derived from Wikipedia for entity detection and disambiguation. They use the bag of words and cosine similarity to measure the relation between the context of the mention and the text of the Wikipedia articles. Shen et al. [11] present a framework named LINDEN, and propose a set of semantic features to disambiguate English entities based on YAGO.

The problem of missing cross-lingual links has attracted increasing attention. A group of highly related work has been proposed based on Wikipedia. ADaFre and Rijke [1] exploit the structure features between Wikipedia to find missing entity links. Wentland et al. [16] extract multilingual contexts for named entities contained in Wikipedia by considering the cross-lingual link structure of Wikipedia. Sorg and Cimiano [12] propose a method, which uses SVM model [6] with structure features, to find missing cross-lingual links between English and German Wikipedia. Wang et al. [15] discover missing cross-lingual links between Chinese Wikipedia and English Wikipedia by using factor graph model.

3 Problem Formulation

In this section, we formally define the cross-lingual query problem. This problem can be decomposed into two sub problems: cross-lingual entity linking and cross-lingual entity disambiguation. Here, we first define the *entity query* and *entity disambiguation* as follows.

Definition 1. Entity query. *Given a knowledge graph K, unstructured text T, entity p in K and entity e in T. If K contains entity p, which can uniquely map to e, we call this process entity query. When the knowledge graph K and unstructured text T are in different languages, we call it cross-lingual entity query.*

Definition 2. Entity disambiguation. *Given a knowledge graph K and named entity sets E= $\{e_1, e_2, ..., e_n\}$ in which elements have the same surface form. If we can find elements in K mapping to E for each element e_i. We call this process entity disambiguation. When the knowledge graph K and named entity sets E are in different languages, we call it cross-lingual entity disambiguation.*

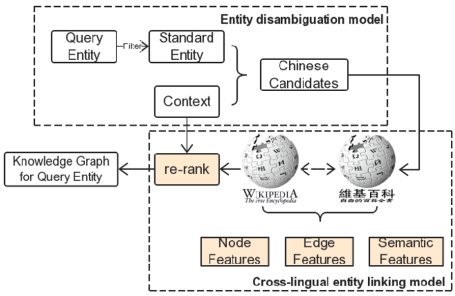

Fig. 1. The framework of the system

We first use existing cross-lingual links in Wikipedia to find out the important factors of knowledge linking, which is the core module in cross-lingual entity query. Here, we download English Wikipedia and Chinese Wikipedia dumps from Wikipedia's website and extract cross-lingual links between them. We extract 450 thousand cross-lingual links (KCL) between Chinese and English Wikipedia. The Chinese version of Wikipedia is considered as a directed graph W_{zh}, The English Wikipedia is considered as W_{en}, where each node n_α represents a Wikipedia article in the language version α of Wikipedia, and has inlinks $in(n_\alpha)$, outlinks $out(n_\alpha)$ and categories $cat(n_\alpha)$.

We first investigate how important structure features are in cross-lingual links prediction. If two articles, n_{zh}, n_{en}, link to two other equivalent articles, we say the two articles have a *common outlink* $col(n_{zh}, n_{en})$. Similarly, if n_{zh}, n_{en} are linked by two other equivalent articles, we say they have a *common inlink* $cil(n_{zh}, n_{en})$. The *common categories* of n_{zh}, n_{en} are called $ccl(n_{zh}, n_{en})$. Because YAGO uses WordNet as its taxonomy instead of Wikipedia category. We only calculate the probabilities of being equivalent conditioned on the number of $col(n_{zh}, n_{en})$ and $cil(n_{zh}, n_{en})$ between n_{zh} in YAGO and n_{en} in English Wikipedia. The number of $col(n_{zh}, n_{en})$ accounts for 85.94 % of total links in YAGO, and the number of $cil(n_{zh}, n_{en})$ is 92.88 %. It is obvious that we can exploit relations between YAGO and Wikipedia to sort the candidate YAGO entities.

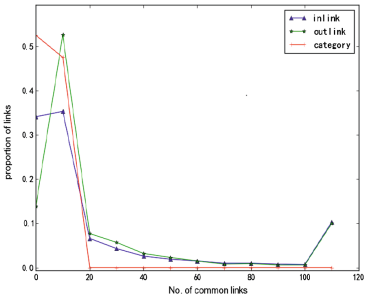

Fig. 2. Distribution of common links between Chinese Wiki. and English Wiki.

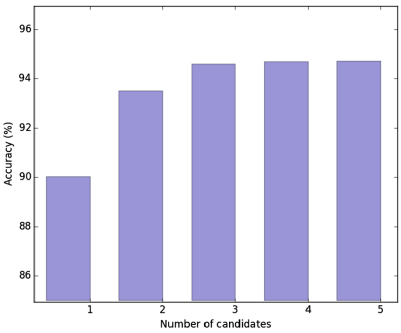

Fig. 3. The cumulative distribution of query entity being found

Due to the fact that more entities are added into Wikipedia, we find that structure features have a smaller effect on cross-lingual entity linking problem. We evaluate the number of $col(n_{zh}, n_{en})$, $cil(n_{zh}, n_{en})$ and $ccl(n_{zh}, n_{en})$ in set KCL. It is obvious that structure features are less important, as shown in Fig. 2.

Figure 1 shows the framework of our system. Where standard entity is the name of the surface form of Chinese query entity m, and context is the bag of words around entity m. We consider two aspects: accurate structure entity links and accurate entity disambiguation. To solve the two problem, we construct entity disambiguation module and cross-lingual entity linking module. The detail description of each module is shown in Sects. 4.1 and 4.2.

In this paper, entity disambiguation is defined as the task to map a textual named entity m, which is already recognized in the unstructured text, to a unique entity e in Wikipedia. We use the bag of words model to represent the context of the entity mention, and obtain the candidate entities in YAGO based on the vector cosine similarity. Cross-lingual entity linking module interacts between Chinese Wikipedia and English Wikipedia. This module finds the English entity describing the same concept of Chinese entity, and finds the missing cross-lingual entity links between Chinese Wikipedia and English Wikipedia. We treat cross-lingual entity linking as a ranking problem, and use Ranking SVM with a set of semantic features and structure features to find cross-lingual entity links.

4 The Proposed Approach

In this section, we describe our proposed method and two modules in detail.

4.1 Entity Disambiguation

Candidate Entity Generation. Given an entity m, the set of candidate entities E_m should have the name of the surface form of m. To solve this problem, we need to build a dictionary that contains vast amount of information

about the surface forms of entities, like abbreviations and nicknames, etc. We use Wikipedia, which contains a set of useful features for the construction of the dictionary we need. We use the following three structures of Wikipedia to build the dictionary about the surface forms of entities:

Entity pages: Each entity page in Wikipedia describes a unique entity and the information focusing on this entity.

Redirect pages: A redirect page maps an alternative name to the page of formal name, such as synonym terms and abbreviations in Wikipedia.

Disambiguation pages: A disambiguation page is created to explain and link to entity pages, which is given the same name in Wikipedia.

Candidate Entity Disambiguation. The named entities express different concepts in different contexts. For example, entity "Michael Jordan" refers to the famous NBA player or the Berkeley professor in different contexts. In this paper, we use vector space model to represent the contexts of the entity mention and use the vector cosine similarity to calculate the similarity between the contexts of query entity and article in Wikipedia. The stop-words in the contexts of query entity lead to a higher vector dimension and make words feature unobvious. Thus, we use TF-IDF [2] to filter out the stop words.

4.2 Cross-Lingual Entity Linking

In this paper, we treat cross-lingual entity linking as a ranking problem, we construct a Ranking SVM with a set of structure features and semantic features between Chinese Wikipedia and English Wikipedia. Ranking SVM is an ranking model using machine learning algorithm, which is proposed by Joachims [8]. The optimization objective of Ranking SVM is to minimize the objective function. Interested readers please refer to [8] for details:

$$Min : V(\overrightarrow{w}, \overrightarrow{\xi}) = \frac{1}{2}\overrightarrow{w}^T \cdot \overrightarrow{w} + C\sum \xi_{i,j,k} \tag{1}$$

Subject to:

$$\begin{aligned} \overrightarrow{w}\Phi(q_1, di) &> \overrightarrow{w}\Phi(q_1, dj) + 1 - \xi_{i,j,1} \\ &... \\ \overrightarrow{w}\Phi(q_n, di) &> \overrightarrow{w}\Phi(q_n, dj) + 1 - \xi_{i,j,n} \end{aligned} \tag{2}$$

$$\forall i \forall j \forall k : \xi_{i,j,k} > 0 \tag{3}$$

where C is a margin size against training error and $\xi_{i,j,k}$ is a parameter of slack variables.

Feature Design. From the observation of Fig. 2, we know the structure features are less important in cross-lingual entity links. Because if one cross-lingual entity link pair describes the same concept, the contents of them are similar with each other. So we design a set of semantic features and some structure features. In the following part, we introduce the definition of structure features and semantic features in detail.

Structure Features. We treat Chinese Wikipedia combining with English Wikipedia as a graph. In the graph, we view articles as nodes, common inlink, common outlink and links to categories as edges.

Common rate feature. (a) We design features for inlinks, outlinks and categories. We use the rate between common links and all links between Chinese Wikipeida entity and English Wikipedia entity formally:

$$\begin{aligned}
&\forall n_\alpha \in W_{zh}, \exists n_\beta \in W_{en}, if\ \exists col(n_\alpha, n_\beta)\ then\ n_\beta \in CI^\beta(n_\alpha)\\
&\forall n_\alpha \in W_{zh}, \exists n_\beta \in W_{en}, if\ \exists cil(n_\alpha, n_\beta)\ then\ n_\beta \in CO^\beta(n_\alpha)\\
&\forall n_\alpha \in W_{zh}, \exists n_\beta \in W_{en}, if\ \exists ccl(n_\alpha, n_\beta)\ then\ n_\beta \in CC^\beta(n_\alpha)
\end{aligned}$$

We define f_{in}, f_{out} and f_{cat} to describe *common rate features*:

$$\begin{aligned}
f_{in} &= \frac{\mid CI^{en}(n_{zh}) \mid}{\mid in_{zh} \mid + \mid in_{en} \mid}\\
f_{out} &= \frac{\mid CO^{en}(n_{zh}) \mid}{\mid out_{zh} \mid + \mid out_{en} \mid}\\
f_{cat} &= \frac{\mid CC^{en}(n_{zh}) \mid}{\mid cat_{zh} \mid + \mid cat_{en} \mid}
\end{aligned} \tag{4}$$

(b) As shown in Fig. 2, every intervals have their own rates for links. We classify links into 10 intervals according to the rate in Fig. 2.

Coherence feature. We have extracted 450 thousand known cross-lingual links between Chinese Wikipedia and English Wikipedia. We calculate the coherence between common links and their links in known cross-lingual links.

$$\begin{aligned}
&\forall n_\alpha \in W_\alpha, \exists col(n_\alpha, n_\beta) \in KCL\ and\ n_\beta \in W_\beta,\ then\ n_\beta \in KCO^\beta(n_\alpha)\\
&\forall n_\alpha \in W_\alpha, \exists cil(n_\alpha, n_\beta) \in KCL\ and\ n_\beta \in W_\beta,\ then\ n_\beta \in KCI^\beta(n_\alpha)\\
&\forall n_\alpha \in W_\alpha, \exists ccl(n_\alpha, n_\beta) \in KCL\ and\ n_\beta \in W_\beta,\ then\ n_\beta \in KCC^\beta(n_\alpha)
\end{aligned}$$

We define fc_{in}, fc_{out} and fc_{cat} to describe *coherence features*:

$$\begin{aligned}
fc_{in} &= \frac{\mid CI^{en}(n_{zh}) \mid}{\mid KCI^{en}(n_{zh}) \mid + \mid KCI^{zh}(n_{en}) \mid}\\
fc_{out} &= \frac{\mid CO^{en}(n_{zh}) \mid}{\mid KCO^{en}(n_{zh}) \mid + \mid KCO^{zh}(n_{en}) \mid}\\
fc_{cat} &= \frac{\mid CC^{en}(n_{zh}) \mid}{\mid KCC^{en}(n_{zh}) \mid + \mid KCC^{zh}(n_{en}) \mid}
\end{aligned} \tag{5}$$

Semantic Features. The articles of Chinese Wikipedia and English Wikipedia are similar with each other in semantics if they describe the same entity. Articles consist of abstract, main body of text and relevant titles of articles in Wikipedia. Every word in one article has its own POS tag, we assume that noun term is the most important part of one article. One problem is that a Chinese noun can be mapped to a set of English noun. To solve this problem, we translate Chinese noun to English noun with the help of Chinese-English Dictionary and WordNet [10]. We find the English translation for the Chinese noun, then we get the Synset the English noun is seated. We use all the words in this Synset as the result. We create three features based on brief introduction, full text and relevant Wikipedia entities.

Abstract similarity feature. We calculate the similarity of the abstract between Chinese Wikipedia article and English Wikipedia article by calculating the count of noun similarity.

Full text similarity feature. We calculate the similarity of full text between Chinese Wikipedia and English Wikipedia.

Entity coherence similarity feature. We caluculate the full text similarity of the entities in the context of Chinese Wikipeida article and English Wikipedia article.

4.3 Cross-Lingual Disambiguation Accuracy Improving

Through the above process, we get some candidate entities in YAGO. Then we re-rank this candidate entities by the relevant entities linked with the candidate entity. We calculate the Semantic Associativity [17] use Eq. (6) between Chinese entity mention and candidate entities in YAGO.

$$SimSA(e_1, e_2) = 1 - \frac{log(max(\mid E_1 \mid, \mid E_2 \mid)) - log(\mid E_1 \cap E_2 \mid)}{log(\mid W \mid) - log(min(\mid E_1 \mid, \mid E_2 \mid))} \tag{6}$$

where e_1 is the English entity which has the same means with Chinese entity. e_2 is the English entity which has the same means with candidate entities in YAGO. E_1 and E_2 are the sets of English Wikipedia entities that link to e_1 and e_2, and W is the set of all entities in English Wikipedia.

5 Expriments

5.1 Datasets

In order to evaluate our approach, we construct two datasets for entity disambiguation task and cross-lingual entity linking task.

EDD. In order to evaluate our approach, we construct a dataset named EDD that contains article pairs from Baidu Baike (a large scale Chinese Wiki knowledge base) and Chinese Wikipedia. We randomly select 1000 article pairs from the dataset to test entity disambiguation module. Each article pair contains a

Baike article and a Wikipedia article, both of which express the same concept. The knowledge graph we adopt in this work is YAGO [9]. The reason why we choose YAGO as the knowledge graph is that YAGO contains more than 1 millon entities and 5 millon facts, and we can use the rich facts to describe one entity comprehensively.

CLD. In cross-lingual entity linking module, we construct a dataset named CLD that contains article pairs form Chinese Wikipedia and English Wikipedia. we selected 1000 English articles with cross-lingual links to Chinese articles from Wikipedia. We generate all possible 1000 × 1000 article pairs from selected articles. 1000 of them in English-Chinese article pairs linked by cross-lingual links are labeled as positive examples and the rest of articles are negative examples. If we choose this training data, the negative pairs are far more than positive. It will cause the problem of overlap. From Fig. 2, we find about 92 % English-Chinese article pairs have common outlinks in KCL. So we restrict the number of inequivalent pairs by common outlinks. After restriction, we control the ratio between positive and negative to be 1:5.

EDD and CLD have the same set of Chinese Wikipedia articles. We can evaluate entity disambiguation task and cross-lingual entity linking task together.

5.2 Performance Evaluation

Evaluation of Entity Disambiguation. The entity disambiguation is a crucial step for CLEQS because it can affect the input of the cross-lingual entity linking module. In particular, we have marked the named entities in unstructured text. The evaluation of this module is to compare the precision. we use EDD to obtain the best number of the candidate entities in Chinese Wikipedia. In Fig. 3 we show the histogram of the accuracy that real entity in these candidates with the grow of the candidate number.

The accuracy in Fig. 3 shows that with the increase of the candidate number, the accuracy of finding article in Chinese Wikipedia which have the same meaning to query entity mention increases slowly. We choose 3 as the number of candidate articles based on the statistics in Fig. 3 and get a precision of 94.6 % based on our entity disambiguation module.

Evaluation of Cross-Lingual Entity Linking. Cross-lingual entity linking module is the core module in our framework. we compare our method with two state-of-the-art cross-lingual linking methods based on CLD. These methods are SVM-S based on the work of Sorg and Cimiano [12] and linkage factor graph model(LFG) based on the work of Wang et al. [15].

- **SVM-S.** This method treat cross-lingual entity linking as a classification problem, and train a SVM with some graph-based and text-based features between Wikipedia articles. They consider the top-k candidates with the respect to a ranking determined on the basis of the distance from the SVM-induced hyperplane.

Table 1. Experiment of cross-lingual entity linking (%).

Model	Precision	recall	F1
SVM-S	68.3	66.9	67.59
LFG	99.1	38.03	54.97
RSVM-G	68.18	69.84	69
RSVM-N	76.3	71.9	74.03
RSVM-Y	89.42	80.47	84.7

Table 2. Contribution analysis of different factors

(a) Semantic factors analysis(%)

Ignored Factor	Precision	recall	F1
Wiki full text	6.35	1.3	3.76
Wiki brief	**6.5**	**1.6**	**3.98**

(b) Graph-based factors analysis(%)

Ignored Factor	Precision	recall	F1
inlink	1.4	0.3	0.83
outlink	11.44	8.49	9.9
category	**20.9**	**15.1**	**17.9**

- **Linkage Factor Graph Model (LFG).** This method presents a factor graph model, and defines some structure features and constraint feature to describe the article in Wikipedia and the relations between articles in two language version of Wikipedia.

Because we have extracted 450 thousand English-Chinese wikipedia article pairs (KCL), so we set up RSVM-Y and RSVM-N. RSVM-Y adds links in KCL into our module, and RSVM-N not. As shown in Fig. 2, the structure information is less and less important for the new articles in Wikipedia. Model like LFG could not deal with these articles. So, we set up RSVM-G to evaluate our module.

Table 1 shows the performance of 3 different methods. According to the result, the LFG method gets really high precision of 99.1 %, but recall is only 38.03 %. Because LFG model ignores the entities with fewer structure features to other entities. RSVM-N outperforms SVM-S 6.44 % in terms of F1. By considering the known cross-lingual links, our method gets a precision of 89.42 %, and a recall of 80.47 %. Therefore, our RSVM model can discover more cross-lingual links, and performs better than SVM-S and LFG.

Overall Performance. We re-rank the 3 candidate YAGO entities by the score of Eq. (6), and get a 82.3 % query precision by CLEQS.

We perform an analysis to evaluate the contribution of different factors. We run RSVM-Y 5 times on evaluation data, and each time we remove one factor. Table 2(a) and Table 2(b) list the result of ignoring different factors. We find that the brief introduction of Wiki articles is more important than full text of Wiki articles. As shown in Table 2(b), outlink and category are more important than inlink in cross-lingual entity linking task. Because the category system changes

less frequently than inlink and outlink, it is more important than inlink and outlink.

6 Conclusion

In this paper, we propose an approach for cross-lingual entity query from Chinese entity in text to the knowledge graph of YAGO. We have published a demo system [13] based on our approach. Our approach is made up of two modules: entity disambiguation module and cross-lingual entity linking module. Our approach uses the result of cross-lingual entity linking module to increase the precision of entity disambiguation module, and get a 82.3 % in query precision. We evaluate the core module and cross-lingual entity linking module, with other approaches. It shows that our approach can achieve higher precision and recall.

Acknowledgement. This work is supported by National Science Foundation of China under grant No. 61170086. The authors would also like to thank Ping An Technology (Shenzhen) Co., Ltd. for the support of this research.

References

1. Adafre, S.F., de Rijke, M.: Finding similar sentences across multiple languages in wikipedia. In: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, ECAL 2006, 3 April - 7 April 2006, Trento, Italy, pp. 62–69 (2006)
2. Albitar, S., Fournier, S., Espinasse, B.: An effective TF/IDF-based text-to-text semantic similarity measure for text classification. In: Benatallah, B., Bestavros, A., Manolopoulos, Y., Vakali, A., Zhang, Y. (eds.) WISE 2014, Part I. LNCS, vol. 8786, pp. 105–114. Springer, Heidelberg (2014)
3. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.G.: DBpedia: a nucleus for a Web of open data. In: Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L.J.B., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) ASWC 2007 and ISWC 2007. LNCS, vol. 4825, pp. 722–735. Springer, Heidelberg (2007)
4. Bagga, A., Baldwin, B.: Entity-based cross-document coreferencing using the vector space model. In: Proceedings of the Conference on 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, COLING-ACL 1998, 10–14 August, 1998, Université de Montréal, Montréal, pp. 79–85. Quebec, Canada (1998)
5. Bunescu, R.C., Pasca, M.: Using encyclopedic knowledge for named entity disambiguation. In: Proceedings on 11th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2006, 3–7 April, 2006, Trento, Italy (2006)
6. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press, New York (2010)

7. Jiang, L., Wang, J., An, N., Wang, S., Zhan, J., Li, L.: GRAPE: a graph-based framework for disambiguating people appearances in web search. In: ICDM 2009, The Ninth IEEE International Conference on Data Mining, Miami, Florida, USA, 6–9 December 2009, pp. 199–208 (2009)
8. Joachims, T.: Optimizing search engines using clickthrough data. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 23–26 July, 2002, Edmonton, Alberta, Canada, pp. 133–142 (2002)
9. Mahdisoltani, F., Biega, J., Suchanek, F.M.: YAGO3: a knowledge base from multilingual wikipedias. In: Seventh Biennial Conference on Innovative Data Systems Research, CIDR 2015, Asilomar, CA, USA, January 4–7, 2015, Online Proceedings (2015)
10. Miller, G.A.: Wordnet: a lexical database for english. Commun. ACM **38**(11), 39–41 (1995)
11. Shen, W., Wang, J., Luo, P., Wang, M.: LINDEN: linking named entities with knowledge base via semantic knowledge. In: Proceedings of the 21st World Wide Web Conference 2012, WWW 2012, Lyon, France, 16–20 April, 2012, pp. 449–458 (2012)
12. Sorg, P., Cimiano, P.: Enriching the crosslingual link structure of wikipedia - a classification-based approach. In: Proceedings of the Aaai Workshop on Wikipedia and Artifical Intelligence (2008)
13. Su, Y., Zhang, C., Cheng, W., Qian, W.: Cleqs: a cross-lingual entity query system based on knowledge graphs. In: NDBC 2015, Chengdu, China (2015)
14. Wang, C., Gao, M., He, X., Zhang, R.: Challenges in chinese knowledge graph construction. In: 31st IEEE International Conference on Data Engineering Workshops, ICDE Workshops 2015, Seoul, South Korea, 13–17 April, 2015, pp. 59–61 (2015)
15. Wang, Z., Li, J., Wang, Z., Tang, J.: Cross-lingual knowledge linking across wiki knowledge bases. In: Proceedings of the 21st World Wide Web Conference 2012, WWW 2012, Lyon, France, 16–20 April, 2012, pp. 459–468 (2012)
16. Wentland, W., Knopp, J., Silberer, C., Hartung, M.: Building a multilingual lexical resource for named entity disambiguation, translation and transliteration. In: Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco (2008)
17. Witten, I.H., Milne, D.N.: An effective, low-cost measure of semantic relatedness obtained from wikipedia links. Proceedings of Aaai (2008)
18. Wu, W., Li, H., Wang, H., Zhu, K.Q.: Probase: a probabilistic taxonomy for text understanding. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2012, Scottsdale, AZ, USA, 20–24 May, 2012, pp. 481–492 (2012)