Chinese Microblog Entity Linking System Combining Wikipedia and Search Engine Retrieval Results

Zeyu Meng, Dong Yu, and Endong Xun

Inter. R&D center for Chinese Education,
Beijing Language and Culture University, 100083, Beijing, China
mengzeyu_blcu@163.com, {yudong,edxun}@blcu.edu.cn

Abstract. Microblog has provided a convenient and instant platform for information publication and acquisition. Microblog's short, noisy, real-time features make Chinese Microblog entity linking task a new challenge. In this paper, we investigate the linking approach and introduce the implementation of a Chinese Microblog Entity Linking (CMEL) System. In particular, we first build synonym dictionary and process the special identifier. Then we generate candidate set combining Wikipedia and search engine retrieval results. Finally, we adopt improved VSM to get textual similarity for entity disambiguation. The accuracy of CMEL system is 84.35%, which ranks the second place in NLPCC 2014 Evaluation Entity Linking Task.

1 Introduction

In recent years, microblog has provided a convenient and instant platform for information publication and acquisition. Bridging microblog post with knowledge bases (KB) can facilitate kinds of tasks such as news detection, information aggregation and correlation recommendation. Chinese Entity Linking Task is aimed to link mentions in microblog-genre text with the corresponding entities in knowledge base, or return "NIL" if the entity is out of knowledge base. Unlike formal text such as news or papers, microblog post is short (no more than 140 characters), noisy (abbreviated form, typos, network-format words etc.) and real-time (new words occurs, words popularity changes), which make Chinese microblog entity linking task a new challenge.

There are two main issues in entity linking task: mention ambiguity and mention variation. The mention ambiguity refers to polysemy phenomenon of nature language: one mention is potentially related to many different KB entries. The mention variation means that a named entity may be expressed in different ways including nickname, previous name, abbreviation or sometimes misspellings.

In this paper, we adopt a cascade approach to identity links between mentions in microblog and entities in knowledge base. We first construct synonym dictionary to deal with mention variation problem. Then, we process special identifier to identify whether the username after "@" is celebrity or ordinary people. After that we generate candidate set combining Wikipedia and search engine retrieval results. Finally, we adopt improved VSM to get textual similarity for entity disambiguation.

C. Zong et al. (Eds.): NLPCC 2014, CCIS 496, pp. 449–456, 2014.

2 Related Work

One of the classical methods is to extract discriminative features of a mention from its context and an entity from its description, then link a mention to the entity which is most similar with it.[1-3] Zheng et al., Zhang et al. and Zhou et al. adopted learning to rank techniques which utilizing relations among candidate entities.[4-6] Those context similarity based methods heavily rely on features but fail in incorporate heterogeneous entity knowledge.

There are also some inter-dependency based entity linking methods. Those methods assumed that the entities in the same document are related to each other and the referent entity is the entity which is most related to its contextual entities.[7-9] The inter-dependency based methods are usually designed for long and normative documents, but do not suit well for microblog genre text from which few feature or related entities can be extracted due to its short and informal content.

Recently, more and more works have been focusing on short informal text. Guo et al. [10] proposed a context-expansion-based and a graph-based method for microblog entity linking by leveraging extra posts. In NLPCC 2013 evaluation task, Miao et al. [11] introduced a microblog semantic annotation system which includes mention expansion by knowledge repository and entity disambiguation considering lexical matching, popularity probability and textual similarity. Zhu et al. [12] used improved pinyin edit distance, suffix vocabulary matching and entity clustering disambiguation. These methods perform well in microblog genre text but do not fully utilize web sources such as search engine retrieval results and information provided by Weibo platform.

3 The Approach

3.1 Overview of the Framework

The main process of Chinese Microblog Entity Linking system contains 3 modules: preprocessing, candidate generation and entity disambiguation. The preprocessing module includes normalizing and indexing the knowledge base, building the dictionary of synonyms and processing special identifier. Candidate generation module uses voting mechanism to combine the retrieval result of Wikipedia and search engine. At last, entity disambiguation module generates the optimal result adopting improved vector space model. Figure 1 shows the framework of the system.

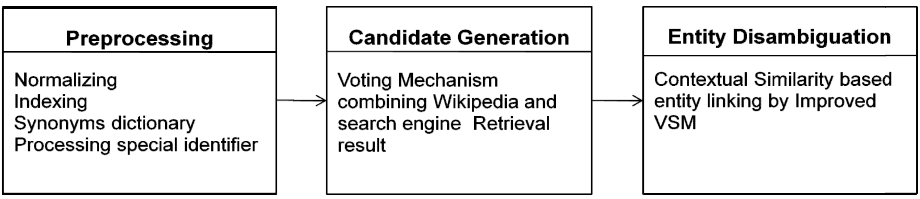

Fig. 1. The framework of Chinese Microblog Entity Linking system

The preprocess module makes preparation for the next two steps. Especially, if there is special identifier in microblog to indicate usernames, we get the user's popularity information from Weibo, thus resolving the problem that common user has the same name with celebrity.

The candidate generation module considers both Wikipedia and search engine retrieval results. In Wikipedia, redirect pages and disambiguation pages could be recognized if a mention is ambiguous. Search engine retrieval result integrates popularity information, and provides error correction function to recall potential candidate.

The entity disambiguation module introduces improved vector space model to avoid word segmentation error which may affect the accuracy of similarity calculation in VSM.

In Microblog Entity Linking task, given a mention, M, the system is a function F(M), which links M to its referent entity in KB, or NIL if outside KB. Figure 2 shows workflow of CMEL system:

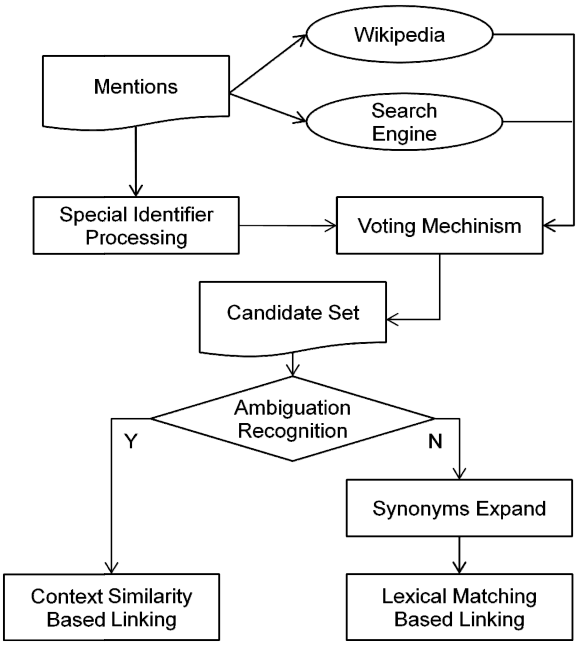

Fig. 2. The workflow of Chinese Microblog Entity Linking system

The implementation details of each step will be described in next 3 section.

3.2 Data Preprocess

The preprocess module makes preparation for next modules, including normalizing and indexing the knowledge base, building the dictionary of synonyms and processing special identifier.

Normalizing and Indexing Knowledge Base

Before we index knowledge base, we have to normalize the knowledge base first. Since the tag name contains both Chinese and English, we use English-Chinese dictionary to normalize all tag names as Chinese. We also normalize punctuations from underline to whitespace and generate a paragraph of description text from original xml format. For word segmentation and stop words filtering job, we adopt SmartChineseAnalyzer carried in Apache Lucene 4.6[1]. Table 1 shows original text in KB with its normalized text.

Table 1. The original knowledge base vs. preprocessed knowledge base

Original knowledge base	Preprocessed knowledge base
<entity enity_id="WKB136" title="自由软件基金会">	[ID]WKB136 [NAME]自由软件基金会
<leader_name>理查德·斯托曼</leader_name>	领导者 姓名 理 查 德 斯 托 曼
<motto>free_software,_free_society</motto>	座右铭 free softwar free societi

Then, we index entities with corresponding normalized description paragraph for KB retrieval later.

Building Synonyms Dictionary

Words in microblog are flexible: nickname, previous name and abbreviation often occurred and the statement style is informal containing large amount of network words. Building a synonyms dictionary is an effective way to deal with mention variation issue. We extract those information whose tag name is “简称”, “本名”, “绰号” etc. from knowledge base first. For instance, we get “皖” as alias of “安徽省” and “李珍基” as previous name of “温流”. Also, redirect pages in Wikipedia give us a good indicator for synonyms. For example, for query “老蒋”, Wikipedia redirect it to the article “蒋中正”. Another resource to build synonyms dictionary is from first paragraph of each article in Wikipedia. We adopt Miao’s methods [16] to extract all those information to build the synonyms dictionary.

Processing Special Identifier

Some mentions in microblog are, or part of, usernames after the special identifier “@”. For example, “@李巍LW1982”, obviously, the mention “李巍” is just a common name, not referring to the football player “李巍”; while in “与中国时尚界最具潜力和最受瞩目的新晋超模 @李丹妮 ”, the mention “李丹妮” refers to the famous model “李丹妮”, indeed. To deal with this kind of mentions, we extract the username’s amount of fans from Weibo platform. The username “李巍LW1982” only has 52 fans while the username “李丹妮” has more than 380 thousand fans. So we set

[1] http://lucene.apache.org/

50 thousand of fans as the threshold to identifying the celebrity's name with ordinary user name.

3.3 Candidate Generation by Voting Mechanism

Since the knowledge base in this task is from Wikipedia, for each mention, we retrieve it in Wikipedia first. However, Wikipedia's searching mechanism is naive without error correction or popularity knowledge: Its redirect page may be incorrect, and sometimes it cannot return any result for misspelling case. For instance, when search the mention "李小露", Wikipedia cannot return any matched result, but actually it is the misspelling form of mention "李小璐" which is supposed to be linked with the famous actress in China. Therefore, we import retrieval result of the largest Chinese search engine, Baidu, into our candidate generation module. Baidu provides error correction function and takes more factors such as popularity reference into account. For the case mentioned above, Baidu can correct the misspelling "李小露" as "李小璐" and return the desired retrieval result.

In this module, we introduce voting mechanism to combine the retrieval results of Wikipedia and Baidu search engine. Through the voting process, each retrieval result will be given a vote score, constituting candidate set of a mention.

Wikipedia Retrieval Result

For a mention, the Wikipedia's returned result will be divided into three types: 1) not existing, 2) only one existing result (may be redirected), 3) disambiguation page. For case 1, we mark Wikipedia retrieval result as NIL with score 0.9; for case 2, we give the result 1.0 score; for case 3, we extract all entities in the disambiguation page and give each of them score of 0.5. Thus, for each mention, M, if it is linked to candidate C, the vote score by Wikipedia is represented as follows,

$$V_{\text{Wiki}}(C|M) = \begin{cases} 0.9, & C = \text{NIL} \\ 1.0, & C = \text{the retrieval result} \\ 0.5, & C \in \text{entities set in disambiguation page} \end{cases} \quad (1)$$

Baidu Search Engine Retrieval Results

We submit mentions with assist query "维基百科" and "中文维基百科" to Baidu API. If top 1 retrieval result is linked to the website "zh.wikipedia.org/wiki/…", we extract the entity name in the title of returned result, regard it as a candidate, C, with score 0.5; while if the top 1 is linked to other website, we regard the mention, M, as an outside KB query, and set score 0.4 to be "NIL". Then, with Baidu search engine, we get 2 votes score expressed as:

$$V_{\text{Baidu}}(C|M) = \begin{cases} 0.5, C = \text{entity name in top 1 result} \\ 0.4, C = \text{NIL} \end{cases} \quad (2)$$

Final Vote Score

Now that given a mention M, we have 1 vote of Wikipedia and 2 votes of Baidu search engine, for each candidate $C_i \in Set_{Wiki} \cup Set_{Baidu}$, the final vote score will be:

$$V_{final}(C_i|M) = \alpha V_{Wiki}(C_i|M) + \beta V_{Baidu1}(C_i|M) + \beta V_{Baidu2}(C_i|M) \quad (3)$$

Testing on the training data set, we assign $\alpha = 1.0, \beta = 0.5$.

To generate final candidate set, if a mention is unambiguous, we choose C_i which has the maximum V_{final} as final candidate, expand it with synonyms dictionary, and search in the knowledge base to get the matched entity. If a mention is ambiguous, we add all the entities in the disambiguation page to candidate set with corresponding vote score, V_{final}.

3.4 Entity Disambiguation Based on Improved VSM

As is mentioned above, after candidate generation module, we get candidate set of the ambiguous mention. Since candidate set contains both the potential named entity and its vote score generated during voting mechanism, we adopt vector space model to get textual similarity between microblog text and content of candidate in knowledge base. The context of mention M and candidate C_i will be expressed as vector ($< t_1, w_1 >, < t_2, w_2 >, \Lambda, < t_i, w_i >, \Lambda, < t_n, w_n >$).

However, in preprocess module, word segmentation error is inevitable. To avoid error in word segmentation affecting the accuracy of similarity in VSM, we take the length of t_i into account and the weight, w_i of each t_i will be

$$w_i = \text{tfidf}(t_i) \cdot \text{length}(t_i) \quad (4)$$

We use cosine distance to measure the similarity. So, in improved VSM, similarity between M and C_i is measured as:

$$Sim(C_i, M) = \frac{C_i \cdot M}{||C_i|| \cdot ||M||} = \frac{\sum_{j=1}^{n}(w_{ij} \cdot w_{Mj})}{\sqrt{\sum_{j=1}^{n} w_{ij}^2} \cdot \sqrt{\sum_{j=1}^{n} w_{Mj}^2}} \quad (5)$$

Finally, the mapping function, F(M), from a mention to its referent entity is expressed as follow:

$$F(M) = \begin{cases} C_i, & \text{if } \text{argmax}(Sim(C_i, M) + V_{final}(C_i|M)) \geq \text{length}(C_i) \\ NIL, & \text{else} \end{cases} \quad (6)$$

4 Experiments

4.1 Experimental Setup

In the experiment, we use the data set published by Natural Language Processing and Chinese Computing Conference (NLPCC) 2014 for Chinese Entity Linking evaluation task. The reference knowledge base used in this task is built from the InfoBoxes

of the Chinese part of Wikipedia dumps in 2013, which contains 400,000 entities with all kinds of tags of properties. The training data set contains 169 microblog post with 250 mentions, and the test data set contains 570 microblog post with 607 mentions. In addition, there exist 2 mentions in training set which are usernames after special identifier "@" and 32 in test data set, showed in Table 2.

Table 2. The statistics of the annotated results

Data set	Microblog posts	Mentions	Special identifier "@"
Training data set	169	250	2
Test data set	570	607	32

4.2 Experimental Results

Table 3 shows the contribution that special identifier "@" makes. By processing special identifier, the system's overall accuracy is improved from 80.07% to 84.35%, which indicates this is an effective method to differentiate celebrity name from common user name.

Table 3. The performance of processing special identifier "@"

System	Accuracy
System without processing "@"	0.8007
System with processing "@"	0.8435

Table 4 reports the performance of our Chinese Microblog Entity Linking (CMEL) system compared with the number 1 system in EL evaluation task. For evaluation, we use the overall micro-averaged accuracy, and further compute the precision, recall, F-1 measures over in-KB entities and NIL entities, respectively.

Table 4. The EL evaluation results

System	Overall	in-KB			NIL		
	Accuracy	Precision	Recall	F1	Precision	Recall	F1
#1	0.8682	0.8078	0.8598	0.8330	0.9202	0.8746	0.8969
CMEL	0.8435	0.8103	0.7765	0.7930	0.8672	0.8950	0.8809

The overall accuracy of our CMEL system is 84.35%, which ranks the second place in Entity Linking evaluation task. From table 4, we can see that, the precision of in-KB result and recall of NIL result are higher than #1's, while the recall of in-KB result and precision of NIL are not that satisfying. It means our system is such strict in candidate generation process that linking some uncertain mention to NIL incorrectly. Generally speaking, combining Wikipedia and search engine can achieve promising result in real world dataset.

5 Conclusion

In this paper we introduce how our Microblog Entity Linking system works to link mentions in microblog posts with named entity in knowledge base. We combine Wikipedia and search engine retrieval results to generate candidate set and adopt improved VSM in entity disambiguation step. In addition, we process the special identifier "@" to differentiate celebrity name from common user name. Experiment result on real world microblog datasets shows this entity linking system is promising and effective. In future work, we will focus on colloquial and casual address of personal name according to its context. About entity disambiguation, finding the strategy to accurately and properly expand microblog-genre text with inter-dependency will be main part of our future work.

References

1. Cucerzan, S.: Large-scale named entity disambiguation based on Wikipedia data. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 708–716 (2007)
2. Mihalcea, R., Csomai, A.: Wikify! Linking Documents to Encyclopedic Knowledge. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management, pp. 233–242 (2007)
3. McNamee, P., Dredze, M., Gerber, A., Garera, N., Finin, T., Mayfield, J., Piatko, C., Rao, D., Yarowsky, D., Dreyer, M.: HLTCOE Approaches to Knowledge Base Population at TAC 2009. In: Proceedings of Text Analysis Conference (TAC) (2009)
4. Zheng, Z., Li, F., Huang, M., Zhu, X.: Learningto Link Entities with Knowledge Base. In: The Proceedings of the Annual Conference of the North American Chapter of the ACL, pp. 483–491 (2010)
5. Zhang, W., Su, J., Tan, C.L., Wang, W.T.: Entity Linking Leveraging Automatically Generated Annotation. In: Proceedings of the 23rd International Conference on Computational Linguistics, pp. 1290–1298 (2010)
6. 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 (2010)
7. Han, X.P., Sun, L.: A Generative Entity-Mention Model for Linking Entities with Knowledge Base. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 945–954 (2011)
8. 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 (2008)
9. Chen, Z., Ji, H.: Collaborative Ranking: A Case Study on Entity Linking. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 771–781 (2011)
10. Guo, Y., Qin, B., Liu, T., Li, S.: Microblog Entity Linking by Leveraging Extra Posts. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 863–868 (2013)
11. Miao, Q., Lu, H., Zhang, S., Meng, Y.: Simple Yet Effective Method for Entity Linking in Microblog-Genre Text. In: Zhou, G., Li, J., Zhao, D., Feng, Y. (eds.) NLPCC 2013. CCIS, vol. 400, pp. 440–447. Springer, Heidelberg (2013)
12. Zhu, M., Jia, Z., Zuo, L., Wu, A., et al.: Research on Entity Linking of Chinese Micro Blog. Journal of Peking University (Natural Science Edition) 01, 73–78 (2014) (in Chinese)