Entity Recognition and Linking in Chinese Search Queries

Jinwei Yuan[1], Yan Yang[1(✉)], Zhen Jia[1], Hongfeng Yin[2], Junfu Huang[1], and Jie Zhu[3]

[1] School of Information Science and Technology,
Southwest Jiaotong University, Chengdu, China
{502966377,990504422}@qq.com, {yyang,zjia}@swjtu.edu.cn

[2] DOCOMO Innovations Incorporation, Palo Alto, USA
hongfeng_yin@yahoo.com

[3] Department of Computer Science, Tibet University, Tibet, China
trocky.jie@gmail.com

Abstract. Aiming at the task of Entity Recognition and Linking in Chinese Search Queries in NLP&CC 2015, this paper proposes the solutions to entity recognition, entity linking and entity disambiguation. Dictionary, online knowledge base and SWJTU Chinese word segmentation are used in entity recognition. Synonyms thesaurus, redirect of Wikipedia and the combination of improved PED (Pinyin Edit Distance) algorithm and LCS (Longest Common Subsequence) are applied in entity linking. The methods of suffix supplement and link value computation based on online encyclopedia are adopted in entity disambiguation. The experiment results indicate that the proposed solutions in this paper are effective for the case of short queries and insufficient contexts.

Keywords: Entity recognition · Entity linking · Entity disambiguation · Suffix supplement · Online encyclopedia

1 Introduction

With the wide application of knowledge graph technology in information retrieval, user modeling, human-computer interaction, question answering and knowledge reasoning, knowledge linking based on structured knowledge base has become a significant task. Entity recognition and linking in Chinese search queries is the task that recognizes all possible entities from queries and links these entities to the target entities in the given knowledge base. It has many difficulties in this task because queries are too short (e.g: in Baidu search engine, queries are limited within 38 words) and often contain a lot of noises (such as spelling mistakes, abbreviations, Internet slangs, nicknames). Several methods are proposed in this paper to solve entity recognition and linking in Chinese search queries. Word dictionary, word popularity, and synonymous thesaurus are constructed based on several online encyclopedias such as Hudongbaike, Baidubaike, and Wikipedia. The solution for entity recognition is mainly based on dictionary, online encyclopedias and SWJTU Chinese word segmentation system [26]. Entity linking is based on synonyms thesaurus, redirection search, and algorithm of combining

© Springer International Publishing Switzerland 2015
J. Li et al. (Eds.): NLPCC 2015, LNAI 9362, pp. 507–519, 2015.
DOI: 10.1007/978-3-319-25207-0_47

improved PED (Pinyin Edit Distance) with LCS (Longest Common Subsequence). Entity disambiguation uses suffix supplement and link value computation methods. The experiment data are acquired from the sample data provided in the task and the queries are extracted from Sogou search logs. Experiment results show that most of the entities in queries can be recognized and can be linked to the entities in the given knowledge base. The average-F1 is 73.3% which ranks first in NLP&CC 2015 evaluation task of entity recognition and linking in Chinese search queries.

2 Related Works

Named entity recognition is the foundation of natural language processing [1] and an important task in many areas [14], including information retrieval, human-machine interaction, machine translation, and question answering. Al-Rfou et al. [2] used online content analysis algorithm to recognize entities. However, this method is only suitable for long texts but not for short texts. Zhao et al. [3] proposed the method of corpus annotation and hierarchical Hidden Markov Model (HMM) to process the recognition of product named entity. Michal et al. [4] introduced a method which uses latent semantic of entities to recognize named entities. Joel et al. [5] recognized named entity by using the content and structure of Wikipedia.

Entity linking refers to linking a given entity to an existing knowledge base [6], [7], [15,16]. Zhu et al. [8] used improved pinyin edit distance and suffix vocabulary matching method to link entities, but this method only suit for Microblog texts. Chen et al. [17] proposed linking with collaborative ranking by using integrated information of entities. But this method is not suitable for short queries because the short texts lack comprehensive information or regularization. Nadeau et al. [25] and Zhang et al. [18] used a supervised machine learning method for entity linking. However, this method is too dependent on semantic information that once extended to other types of corpus, the performance will be affected seriously by noises [19]. Gattani et al. [9] adopted online encyclopedia such as Wikipedia to implement entity linking of social data.

Entity disambiguation can solve the problem that an entity mention refers to multiple real-world things [8], [20], [21]. Yang et al. [10] proposed the method of person name disambiguation based on dependency features of webpage text. Nguyen et al. [11] extracted the contents and characteristics of the corresponding pages from the Wikipedia to carry out the entity disambiguation. Zhu et al. [8] combined entity clustering disambiguation and similar entity disambiguation based on Baidu encyclopedia. Meng et al. [22] adopted improved VSM to get textual similarity for entity disambiguation. Kataria et al. [23] and Sen [24] proposed the methods based on topic models to achieve entity disambiguation, but these methods are not suitable for short texts.

3 Methods

3.1 Problem Analysis and Method Procedure

By analyzing the given sample data and knowledge base in the task, problems are mainly as follow:

1) In some queries, there exist English words and a variety of punctuations.
2) Queries are too short and have no contexts.
3) Entity mentions in queries may be aliases or abbreviations of the target entities in the knowledge base.
4) Queries contain wrongly written characters.
5) Many different entities have the same mentions in the knowledge base. For example, entity mention "步步惊心"(Startling by Each Step) may refer to a book or a TV series.

To solve these problems, this paper proposes the following solutions:

1) Preprocess queries and remove noises in them.
2) Build the word dictionary and use SWJTU Chinese word segmentation system to recognize entities of short phrases in open domains.
3) Construct a synonyms thesaurus to identify aliases or abbreviations of entities.
4) Optimize the knowledge base and remove noises in it.
5) Use LCS algorithm and online encyclopedias to recognize the abbreviations of entities.
6) Combine improved edit distance algorithm [8], [13] with LCS to recognize wrongly written characters.
7) Apply word frequency in Baidu encyclopedias and suffix supplement method to disambiguate entities.

The pipeline of the methods is shown in Figure 1.

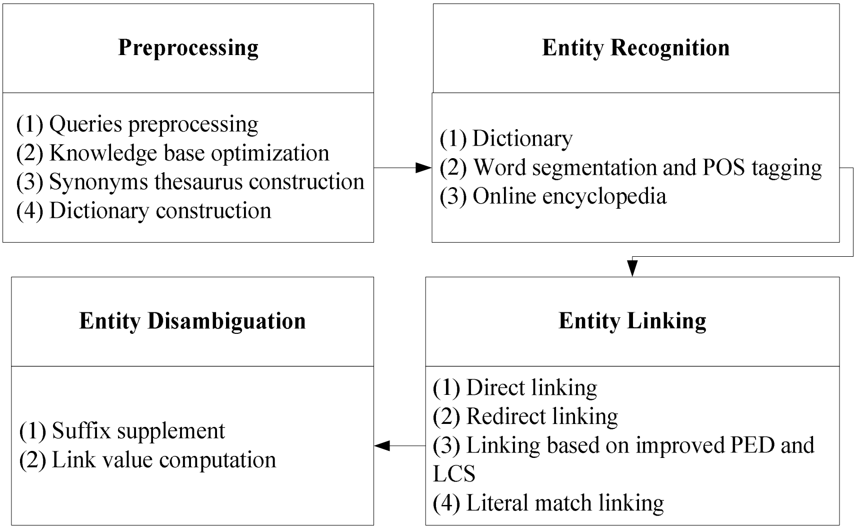
Preprocessing
(1) Queries preprocessing
(2) Knowledge base optimization
(3) Synonyms thesaurus construction
(4) Dictionary construction
Entity Recognition
(1) Dictionary
(2) Word segmentation and POS tagging
(3) Online encyclopedia
Entity Linking
(1) Direct linking
(2) Redirect linking
(3) Linking based on improved PED and LCS
(4) Literal match linking
Entity Disambiguation
(1) Suffix supplement
(2) Link value computation

Fig. 1. Method pipeline

As shown in Figure 1, the methods pipeline includes preprocessing, entity recognition, entity linking and entity disambiguation. Preprocessing consists of query preprocessing, knowledge base optimization, synonyms thesaurus construction and dictionary construction. Entity recognition is mainly based on dictionary, word segmentation and POS tagging, and online encyclopedias. Entity linking has four steps carried out in order to find candidate entities in the knowledge base. Entity disambiguation uses suffix supplement and link value computation method on basis of online encyclopedias.

3.2 Preprocessing

3.2.1 Knowledge Base Preprocessing

It contains many noises in the knowledge base which is provided in the evaluation task, such as disunification in uppercase and lowercase letters in English, mixed use of Chinese and English punctuation, lack of values in knowledge triples. In order to improve accuracy of entity linking, several solutions to remove noises in knowledge base are taken. Firstly, English words are rewritten in capitalization to ensure it's unified. Then, all the punctuations in the knowledge base are turned into unified Chinese punctuation. Finally, the knowledge triples that some values are absent are removed. Examples are as shown below in Table 1.

Table 1. Knowledge base preprocessing

The original knowledge	The knowledge after preprocessing	remark
林纳斯·托瓦兹 employer Linux 基金会 (Linus Torvalds employer Linux foundation)	林纳斯·托瓦兹 EMPLOYER LINUX基金会 (Linus Torvalds EMPLOYER LINUX foundation)	Rewritten English words in capitalization.
南京!南京! (Nanjing! Nanjing!)	南京！南京！ (Nanjing！Nanjing！)	Turn English punctuation into Chinese punctuation.
中华人民共和国abstract null (the People's Republic of China)		Remove this knowledge triple.
《步步惊心》 (《Startling by Each Step》)	步步惊心 (Startling by Each Step)	Delete book title mark punctuations which are in the start or end of knowledge entries

3.2.2 Synonyms Thesaurus Construction

For the length limitation in queries, many entity mentions appear in form of aliases, abbreviations and Internet slangs. In fact, these mentions with variable names refer to one entity. For example, "习大大" (Xi Dada), and "习主席" (Chairman Xi) are different mentions but represent one person. A large number of synonyms are extracted from Hudong encyclopedia, Baidu Encyclopedia, Douban and Wikipedia for constructing a synonyms thesaurus which in total reaches 184,430 synonyms.

3.2.3 Dictionary Construction

The dictionary contains the words are extracted from the knowledge base given in the task, Wikipedia, Baidu encyclopedia, Hudong encyclopedia and Douban. The words collected in the dictionary include nouns, academic terms, idioms, times, aliases, abbreviations, Internet slangs which in total reach about more than 800,000 words.

3.2.4 Query Preprocessing

Before recognizing entities for the given queries, preprocessing need to be done to get the large amount of noise removed. Firstly, filtering punctuations and turning lowercase letters to capital letters in English. Then, using SWJTU Chinese word segmentation system to process queries and remaining the words which are tagged as noun, academic term, abbreviation, idiom and time[1]. Query preprocessing is described in detail by taking one query as example, as shown below in Table 2.

Example Query:

《Linkin Park》歌曲的演唱者麦克·信田 (《Linkin Park》song's singer Mike Shinoda)

3.3 Entity Recognition

3.3.1 Entity Recognition Difficulties

In general, entity recognition is carried out in such cases as entities have full names, but many entities in queries appear in form of aliases or abbreviations because of length limitation. At present, named entity recognition algorithms with better effect are domain related, while the sample data provided in this evaluation task are from open domain, thus it cannot reach good result using traditional machine learning methods(such as CRF) based on labeled training data. To solve this problem, we apply a simple method of combining the existing Chinese word segmentation, dictionary, and online encyclopedia.

3.3.2 Named Entity Recognition

For the given query, word segmentation and POS tagging are performed in the step of query preprocessing. The NLP processing tool is SWJTU Chinese word segmentation system. This system supports the user-defined dictionary, so we add the dictionary

[1] The POS tags are Entity, n, t, j, nnt, nrf etc. Entity is a user-defined POS that refers to the words in the dictionary and possible entities. Other POS descriptions are in the website http://ics.swjtu.edu.cn.

Table 2. Main steps of query preprocessing

Step No.	Preprocessing	Example
1	punctuations filtering and letters unification	LINKIN_PARK歌曲的演唱者麦克·信田 (LINKIN_PARK song’s singer Mike Shinoda)
2	Word segmentation and POS tagging	LINKIN_PARK/Entity 歌曲/n 的/ude1 演唱者/nnt 麦克·信田/nrf(LINKIN_PARK/Entity song/n s/ude1 singer/nnt Mike Shinoda /nrf)
3	POS filtering	LINKIN_PARK/Entity 歌曲/n 演唱者/nnt 麦克·信田/ nrf (LINKIN_PARK/Entity song/n singer/nnt Mike Shinoda /nrf)

into the system and the words in the dictionary are tagged as “Entity”. This system supports two segmentation ways that are coarse-grained and fine-grained segmentations. Coarse-grained segmentation uses longest matching algorithm, but fine-grained segmentation can split the words into smaller units. For example, “西南交通大学” (Southwest Jiaotong University) is cut into one entity in coarse-grained segmentation “西南交通大学/ntu”, but three words in fine-grained segmentation “[西南/ns 交通/n 大学/nis]/ntu”([Southwest/ns Jiaotong/n University/nis]/ntu).

Firstly, the query is segmented in coarse-grained segmentation. If the POS tags of the words are “Entity” (user-defined POS tag), nr (person name), ns (place) or j (abbreviation), they are viewed as entities. While if the POS tags are nt (organization), nz (proper noun), l (idiom), and n (noun), the following steps are applied to determine whether the words are entities.

(1) If the word has redirect in Wikipedia, it can be recognized as an entity directly.

(2) If the word has no redirect and the POS tag is nt or nz, it should be segmented in fine-grained segmentation to smaller unit. If some of the units are in the dictionary, these units are determined named entities.

Here are some detail examples about entities recognition, as shown below in Table 3.

3.4 Entity Linking

The pipeline of entity linking is shown below in Figure 2.

The pipeline includes four main steps which are carried out in order: direct linking, redirect linking, linking based on improved PED and LCS, and literal match linking. The detail process is as follows.

Table 3. Entity recognition examples

Result of word segmentation	Detailed description
赵本山/Entity 否认/v 离婚/n[2] (Benshan Zhao/Entity denied/ divorce/n)	The POS of 赵本山(Benshan Zhao) is **Entity**, so it can be reserved as an entity directly. 离婚(divorce) is tagged as **n**, but redirect in Wikipedia cannot be found, so it be removed.
湖北/ns 石首/ns群体性事件/nz (Hubei/ns Shishou/ns group event/nz)	湖北(Hubei) and 石首(Shishou) are marked as **ns**, so they are reserved as entities directly. 群体性事件(group event) is marked as **nz**, and the redirect is骚乱(riot), so it is reserved as an entity.
巴比伦通天塔/nz (Babylon tower/nz)	巴比伦通天塔(Babylon tower)is tagged as **nz**, but it's redirect cannot be found, then it is segmented using fine-grained segmentation into [巴比伦/**nsf** 通天塔/**nz**]/nz ([Babylon/nsf tower/nz]/nz). The two words "Babylon" and "tower" are in the dictionary, so both of them are recognized as entities.

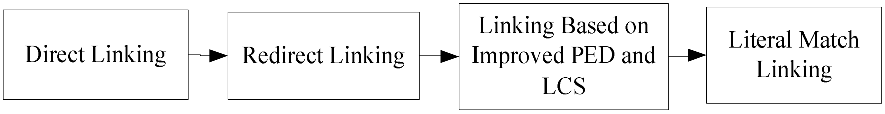

Fig. 2. Entity linking pipeline

1) Direct Linking

Direct linking means searching an entity or its synonyms in the knowledge base directly which is the first step. For improving linking accuracy, when the entity is recognized as person name, especially foreigner name, we first search this name in Hudong encyclopedia to find the name with highest popularity. Namely, the most frequently visited and concerned entry. In Hudong encyclopedia, when users search some entries, the most popular entries will be returned. For example, when "科比" (Kobe) is searched, "科比·布莱恩特"(Kobe Bryant) will be returned. In the general case, "科比" refers to "科比·布莱恩特". Then we link the full name to the knowledge base.

2) Redirect Linking

If the entity cannot be found in the knowledge base after the step above, we search this entity or its synonyms in Wikipedia to get the redirect. If the redirect exists, we use it to search in the knowledge base.

[2] The notation after "/" is POS tag.

3) Linking Based on Improved PED and LCS

Some entities contain wrongly written characters or inconsistence in English to Chinese transliteration so they cannot be linked to the knowledge base. We adopt the method of combining the improved PED(Pinyin Edit Distance) algorithm [8] [13] with LCS to solve this problem.

Firstly, we use PED algorithm to find the candidates set of the entity in the knowledge base, as shown in (1).

$$e=\left\{e_1, e_2, e_3, \ldots, e_i\right\} \tag{1}$$

Where e_i is the i-th candidate entity of e.

In PED algorithm, Spell is a set of consonants and vowels with similar pronunciation as follows.

$$\text{Spell}=\{(l,n),(l,r),(z,zh),(c,ch),(an,ang),(en,eng),(in,ing),(ang,ong), (si,ci)\} \tag{2}$$

Difference between character I and I' is computed according to the consonants and vowels of them, as shown in (3) [8], [13].

$$\text{Difference}(I, I') = \begin{cases} 0.5\ ,\ I \in \text{Spell}_i, I' \in \text{Spell}_i \\ 1, \text{other} \end{cases} \tag{3}$$

Difference between consonants or vowels with similar pronunciation in the same Spell_i is set to 0.5. Difference of other consonants and vowels which are not in Spell set is set to 1.

If the difference between each character of two entity mentions is less than 1, then the entity in knowledge base will be a candidate entity. For example, for the entity "诺维茨基"(Nowitzki) in queries pronouncing nuo wei ci ji in Chinese and the entity in the knowledge base "诺维斯基" pronouncing nuo wei si ji, the differences between each character is less than 1, then "诺维斯基" is put into the candidate set.

If the difference between characters of two entities is no less than 1, the entity cannot be added to the candidate set. For example, the entity in queries "周杰伦"(Jay Chou) pronounces zhou jie lun in Chinese and "周杰峰" in the knowledge base pronounces zhou jie feng, and the difference between "轮"(lun) and "峰"(feng) reaches 1, so "周杰峰" is not the candidate of "周杰伦".

After acquiring the candidate set of the entities, we choose the one with the highest literal similarity as the linking result in the knowledge base. Literal similarity is computed by LCS. The principle idea of LCS is that with the same length, the more the same words are, the higher the similarity will be. The formula is as shown below.

$$\text{Sim}_i = \frac{\text{len}(m, e_i)}{\text{len}(m)} \tag{4}$$

where Sim_i is the similarity between the linking entity m and the candidate entity e_i; len(m) indicates the length of the linking entity m; len(m,e_i) represents the number of same characters between linking entity m and candidate entity $e_{i.}$.

4) Literal Match Linking

Literal match linking is to solve the problem of abbreviations. For example, “中科院” (CAS) in the query is often used to refer to “中国科学院” (Chinese Academy of Sciences). In this paper, when the literal matching degree between two mentions is no less than 0.6, the two entity mentions will be linked.

3.5 Entity Disambiguation

3.5.1 Difficulties in Entity Disambiguation

For the variable meanings of the same word in different contexts, the traditional word disambiguation determines the exact meaning according to the contexts [12]. However, search queries are too short to provide complete contexts of entities, which brings difficulties in disambiguation, mainly shown as follow points:

1) Diversity of Entity Mentions

An entity may have various mentions, such as full name, abbreviation, aliases, etc. For example, the famous American basketball player “科比·布莱恩特” (Kobe Bryant) can be called as “Kobe”, “黑曼巴” (Black Mamba), “小飞侠” (Peter Pan), etc.

2) Ambiguity of Entity in Different Contexts

An entity has different meanings in different contexts. For example, the word “Apple” may refer to a fruit apple or Apple Company.

In the task, we may get several candidate entity links in the knowledge base after the step of entity linking. To disambiguate these candidate entities, we propose the methods of suffix supplement and link value computation based on online encyclopedias.

3.5.2 Suffix Supplement

Some online encyclopedias uses suffix to disambiguate the entities which have the same names. Since the knowledge base is from Wikipedia, it is effective to use the suffix of online encyclopedia to pad and disambiguate entities. First, we search the entities in the Hudong encyclopedia, which can return the most popular entry name with suffix according to visiting frequency of users. The suffix of entry name will be added into the entity. Then, by calculating the similarity between the new entity mentions with suffix and candidate entities in the knowledge base, the candidate entity with the highest similarity will be returned as the final linking result.

3.5.3 Link Value Computation

This method includes three steps:

Step 1, computing word weight of each candidate entities in knowledge base based on Wikipedia.

Step 2, getting popularity weight based on Baidu encyclopedia.

Step 3, computing link value to decide which candidate entity is the linking result.

1) Word Weight Computation

Suppose E is the entity recognized in queries, C is one of the candidate entities of E linking to the knowledge base.

Firstly, we search C in Wikipedia and get the search result webpage of C. Then, we count the word frequency of C sum(C) in the webpage. We else count the word frequency sum(E) in the webpage. If C contains E, that is to say, if E is part of C, we don't count. The word weight W_i of C is calculated using the following formula:

$$W_i = \frac{\mathrm{sum(E)}}{\mathrm{sum(C)}} \tag{5}$$

2) Popularity Weight Acquisition

Second, we search C in Baidu encyclopedia to get the visiting frequency and determine its popularity. A popularity weight value λ is assigned to each C according to its visiting frequency which can be estimated by experience, as shown in Table 4:

Table 4. Weight λ of popularity

Popularity ranking	1	2	3	4	5
Weight λ	1.0	0.9	0.8	0.7	0.6

3) Link Value Computation

According to the word weight and popularity weight of the candidate entity C, we can calculate the link value of C, the calculation method is as shown in formula (6):

$$S_i = \lambda \times W_i \tag{6}$$

where S_i indicates the link value of the candidate linked entities.

Select the candidate linked entity with largest value of S_i as the linking result.

4 Experiments

For the sample data provided in this evaluation task is only 159 queries, we extract 1000 queries from the Sogou search logs, which are similar to the sample data. We manually label the entities and linking results of each query according to the knowledge base. We use these corpus as experiment data to design entity recognition, linking and disambiguating algorithms. The entity recognition result is shown in Table 5. The linking result is shown in Table 6. The entity recognition and linking result for sample data is shown in Table 7.

Table 5. Entity Recognition Result

Marked entity	Precision	Recall	F1
1159	0.844	0.886	0.865

The result in Table 5 shows that using SWJTU Chinese word segmentation and dictionary to recognize entities is effective.

Table 6. Entity Linking Result

Marked entity	Precision	Recall	F1
1159	0.905	0.905	0.905

From Table 6, we noted that the effect is very good using the above entity linking method. That is to say, most of the entities link right.

Table 7. Entity recognition and linking

Precision	Recall	F1
0.742	0.779	0.761

From Table 7, the precision of entity recognition and linking is not particularly high. But it is also noted that the precision of entity recognition has a great effect on the final results.

The testing data is provided by CCF in the evaluation task, among which includes 3849 search queries. The final evaluation result is shown below as Table 8.

Table 8. Evaluation result of entity recognition and linking in Chinese search queries

NO.	Precision	Recall	F1	Average-F1
1	**0.724**	**0.736**	**0.73**	**0.733**
2	0.562	0.695	0.621	0.611
3	0.497	0.704	0.583	0.569

The result of NO.1 represents the performance of our system. Compared with other systems in this evaluation task, our system achieves higher precision, recall, F1, and average-F1 than other systems, which shows the method proposed in this paper is effective.

5 Conclusions and Outlook

In this paper, the method used in the evaluation task of entity recognition and linking in Chinese search queries in NLP&CC2015 is introduced. The result shows that it works fine in entity linking, but the entity recognition and disambiguation need to be improved. In future work, we will use and merge more information, including the results of Internet search engine, to expand the length of Chinese search queries. And then we will conduct semantic analysis of the queries to further improve the effect of named entity recognition and disambiguation.

Acknowledgement. This work is supported by National Natural Science Foundation of China under Grants No.61170111, No. 61202043, No. 61262058, and Fundamental Research Funds for the Central Universities under Grants No. 2682015CX070.

References

1. Lei, J., Tang, B., Lu, X., et al.: A comprehensive study of named entity recognition in Chinese clinical text. Journal of the American Medical Informatics Association **21**(5), 808–814 (2014)
2. Al-Rfou, R., Skiena, S.: Speedread: a fast named entity recognition pipeline (2013). arXiv preprint arXiv: 1301.2857
3. Zhao, J., Liu, F.: Product named entity recognition in Chinese text. Language Resources & Evaluation **42**(2), 197–217 (2008)
4. Konkol, M., Brychcín, T., Konopík, M.: Latent semantics in Named Entity Recognition. Expert Systems with Applications **42**(7), 3470–3479 (2015)
5. Nothman, J., Ringland, N., Radford, W., et al.: Learning multilingual named entity recognition from wikipedia. Artificial Intelligence **194**, 151–175 (2013)
6. Hachey, B., Radford, W., Nothman, J., et al.: Evaluating Entity Linking with wikipedia. Artificial Intelligence **194**(194), 130–150 (2013)
7. Shen, W., Wang, J., Han, J.: Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions. IEEE Transactions on Knowledge and Data Engineering **27**(2), 443–460 (2015)
8. Zhu, M., Jia, Z., Zuo, L., et al.: Research on Entity Linking of Chinese Micro Blog. Acta Scientiarum Naturalium Universitatis Pekinensis **1**, 73–78 (2014). (in Chinese)
9. Gattani, A., Lamba, D.S., Garera, N., et al.: Entity extraction, linking, classification, and tagging for social media: a wikipedia-based approach. Proceedings of the VLDB Endowment **6**(11), 1126–1137 (2013)
10. Yang, X., Li, P., Zhu, Q.: Name Disambiguation Based on Dependency Feature in Web Page Text. Computer Engineering **38**(19), 133–136 (2012). (in Chinese)
11. Nguyen, H.T., Cao, T.H.: Exploring wikipedia and text features for named entity disambiguation. In: Nguyen, N.T., Le, M.T., Świątek, J. (eds.) ACIIDS 2010. LNCS, vol. 5991, pp. 11–20. Springer, Heidelberg (2010)
12. Zhao, J.: A Survey on Named Entity Recognition, Disambiguation and Cross-Lingual Coreference Resolution. Journal of Chinese information Processing **23**(2), 3–13 (2009). (in Chinese)

13. Cao, J., Wu, X., Xia, Y., et al.: Pinyin-indexed method for approximate matching in Chinese. Journal of Tsinghua University (Science and Technology) **49**(S1), 1328–1332 (2009). (in Chinese)
14. 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)
15. Zheng, Z., Li, F., Huang, M., et al.: Learning to link entities with knowledge base. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 483–491 (2010)
16. 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 (2011)
17. Chen, Z., Ji, H.: Collaborative ranking: a case study on entity linking. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 771–781 (2011)
18. Zhang, W., Sim, Y.C., Su, J., Tan, C.L.: Entity linking with effective acronym expansion, instance selection and topic modeling. In: Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, vol. 3, pp. 1909–1914 (2011)
19. Zou, X., Sun, C., Sun, Y., Liu, B., Lin, L.: Linking entities in tweets to wikipedia knowledge base. In: Zong, C., Nie, J.-Y., Zhao, D., Feng, Y. (eds.) NLPCC 2014. CCIS, vol. 496, pp. 368–378. Springer, Heidelberg (2014)
20. Davis, A., Veloso, A., Silva, A.S.D., et al.: Named entity disambiguation in streaming data. In: Proceedings of the Conference on European Chapter of the Association for Computational Linguistics, vol. 1, pp. 815–824 (2012)
21. 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 (2009)
22. Meng, Z., Yu, D., Xun, E.: Chinese microblog entity linking system combining wikipedia and search engine retrieval results. In: Zong, C., Nie, J.-Y., Zhao, D., Feng, Y. (eds.) NLPCC 2014. CCIS, vol. 496, pp. 449–456. Springer, Heidelberg (2014)
23. Kataria, S.S., Kumar, K.S., Rastogi, R.R., et al.: Entity disambiguation with hierarchical topic models. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1037–1045 (2011)
24. Sen, P.: Collective context-aware topic models for entity disambiguation. In: Proceedings of the 21st International Conference on World Wide Web, pp. 729–738 (2012)
25. Nadeau, D., Turney, P.: A supervised learning approach to acronym identification. In: The Eighteenth Canadian Conference on Artificial Intelligence (2005)
26. SWJTU Chinese Word Segmentation System. http://ics.swjtu.edu.cn