Collective Entity Linking Method in Chinese Text Based on Topic Consistency

Yi Chen, Qingbo Wu, Yusong Tan, Wei Wang

School of Computer, National University of Defense Technology, Changsha, China 0731
chenyi15a@nudt.edu.cn, wuqingbo@ubuntukylin.com, tanyusong@kylinos.cn, wangwei15a@nudt.edu.cn

Abstract: Entity Linking refers to the task of linking entity mentions in the given text with their referent entities in a knowledge base, which is a key technology of knowledge base expansion. However, the performance of traditional Chinese entity linking methods are affected by the incomplete Chinese knowledge base. Also they rarely use the semantic relevance between entities. Therefore, we propose a Chinese collective entity linking method based on the consistency of the topic, which considers both the content similarity and topic relevance of the co-occurrence entities, and propose a method for calculating the topic consistency of entities. This method implements batch links for multiple ambiguous entities that appear in the same text, and reduces the reliance on the local knowledge base by using the combination of the local knowledge base and the external knowledge base. Experimental results show that our method performs well over the traditional methods. And it is potentially effective.

1. Introduction

Nowadays, we are in such an era of information explosion. How to effectively obtain valuable information from massive network data is a hot research problem in the field of Internet information extraction technology. The main problems are the diversity and ambiguity of natural language. The diversity of natural language refers to that the same meaning can be expressed in many different forms. This is due to the elasticity of expression of natural language. E.g., people call Kobe Bryant for Peter Pan, Ke God and so on. As for the ambiguity of the natural language, it means that the same word, phrase, or sentence have different meanings in different contexts. Because of the existence of ambiguity and ambiguity, it is difficult for the users to retrieve the relevant information of the target entity quickly and accurately when searching the information on the network [1].

Entity linking provides a solution for the both problems above. It mainly focuses on the representation of the entity in the text, and links the entity representation of the text in the natural language with the entries in the knowledge base [2]. If the page is pre-processed in the way of entity linking, the user can quickly and accurately find the relevant information of the corresponding entity. The application area of entity linking is relatively broad, and is applied to the knowledge base expansion, open field online knowledge question answering system, information filtering and information retrieval, machine translation and online advertising, etc [3] [4]. The existing research contributions of entity linking are mainly in English. In contrast, the development of Chinese entity linking technology is somewhat lagging behind. The main reasons are as follow. Firstly, the construction of Chinese open source knowledge base is in the initial stage, which restricts the development of Chinese entity linking work. Secondly, entity extraction technology is constrained by the word segmentation technology. Thirdly, as the Chinese grammar is more flexible, and the semantics is more abundant, which make Chinese entity disambiguation more difficult than English [5]. The current mainstream entity linking method is the context similarity between an entity reference term and its corresponding candidate entity, and the candidate entity with the highest similarity is the link target. This approach does not take into account the semantic relevance of the entities in the text, which leads to the waste of information and the decrease of the accuracy of entity linking.

In this paper, a Chinese collective entity linking method based on topic consistency is proposed. In our method we model each entity mention in the given text and its candidate entity as distinct nodes in a graph, and model topic consistency by links between nodes. We call this graph topic relation graph. For the entity mentions to be linked, we extract the topic of the contexts of the mentions. Finally, the candidate entities we choose are the most relevant to the topic of the given text. So we calculate the topic consistency between the context of the mention and the Baidu encyclopedia page of the candidate entity to determine the weight of the edge of mention to candidate entity firstly. And then we calculate the topic relevance between the candidate entities corresponding to different mentions. Finally, we transform the problem of entity linking into the topic consistency order problem of the candidate entity using the integration algorithm, and choose a set of candidate entities with the highest degree of topic consistency and content similarity as the target of entity linking.

The main contributions of this paper are summarized as follows. The topic relation graph we proposed takes into account both the content similarity and the subject relevance, and uses the collective algorithm of the batch link, to improve the accuracy and efficiency of the link. In the process of constructing the topic relation graph, for the construction of Chinese knowledge base incomplete problem, we combining the local knowledge with external knowledge to reduce the dependence on local knowledge base effectively. The experiment results show that our method achieve dramatic accuracy and recall rate. We conclude that the proposed entity linking method is potentially effective.

The rest parts are organized as follows: We introduce the related works in section 2. In section 3, we describe our linking method in detail, and provide the experimental results and evaluation in section 4. Section 5 is a conclusion.

2. Related Work

As an important basic research value for the expansion of knowledge base, entity linking technology has received extensive attention in academia in recent years. Early entity linking work is focused on a single entity. With the development of a series of collective entity linking methods, this method has become a research hotspot.

The idea of early entity linking research is to select the candidate entity with the largest similarity as the target object, by calculating the context similarity between the text of the entity mention and its corresponding candidate entity. The typical work is a computational model based on context similarity proposed by Bunescu, which has attracted much attention in the similarity computation method [6]. Nguyen adds the contextual characteristics and the page structure of the candidate entity in the Wikipedia page to calculate the similarity, thus effectively improving the accuracy [7]. However the accuracy of such algorithms is susceptible to lack of contextual information [8]. Zeng proposes the use of external knowledge to extend the eigenvector to improve the distinction between candidate entities [9]. Zhang uses the wiki concept similarity as a measure [10]. The main disadvantage of the similarity based method is to ignore the semantic correlation between entities, which is often helpful for distinguishing ambiguous entities. In addition, some scholars use the method of machine learning. For example, Zuo proposes a voting model [11]. The drawback of the machine learning methods is that the performance is subject to the quality and scope of the training corpus. Therefore, in order to overcome the shortcomings of these two kinds of methods, Kulkarni et al. have proposed the idea of collective entity linking [12].

The collective entity linking method solves the link problem of all entity mentions in the text at one time, avoiding the single-threaded processing mode in which the entities to be scanned and disambiguated. The basic idea is to build a graph model using the relationship between entities [13]. Han regards Wikipedia as a local knowledge base, treating entity references and candidate entities as vertices in the graph, establishing entity referent graph [14]. And then he uses the random walk algorithm to sort the candidate entities in the graph. On the basis of the above work, Ayman combines the context similarity and the popularity of the entity into the PageRank algorithm by modifying the initial probability value of the vertex [15]. However, the common problem of these methods is that they rely entirely on Wikipedia pages as the source of knowledge, and performs poorly on non-well-known entities. To solve this problem, Andrea utilizes the BabelNet semantic network to achieve the entity disambiguation by extracting the dense subgraph [16]. But the Babelfy algorithm relies too much on the local knowledge base.

3. Approach

In this section, we proposed our method which consists of three parts: candidate generation, relation graph construction, and candidate entities disambiguation.

3.1 Candidate Generation

The main function of the candidate entity generation module is to identify the entity mention for the given input text, and search the knowledge base to get the set of candidate entities corresponding to each entity mention.

In the first step, for the input text D, we use the toolkit THULAC issued by the Tsinghua University natural language processing and social humanities computing laboratory for word segmentation, and realize the entity recognition according to the output of the parting result. The principle of extraction is to extract only the named entities whose names are marked as people names, place names or institution names, and filter out the general term. Finally, we get the set of all the entity mention $M = \{m_1, m_2, \cdots, m_n\}$ in the input text D. The second step is to generate candidate entity sets. We search the knowledge base to find all the candidate entities that have the same name as the entity mention m_i in the knowledge base. All candidate entities corresponding to this entity mention are used as the initial candidate entity set $C_i = \{c_{i1}, c_{i2}, \cdots, c_{ik}\}$ for this entity. Where c_{ik} represents the kth candidate entity of the entity mention m_i. For the entity mention m_i that does not find the candidate entity in the knowledge base, we define that the candidate entity set C_i of m_i is null. Thus the initial candidate entity set of the entity mention set M is $C' = \{C_1, C_2, \cdots, C_n\}$.

We use Fudan CN-DBpedia Chinese knowledge map as the basis. This knowledge map now has 9,455,262 entities, and offers a full suite of APIs for free. As the

Chinese open source knowledge base construction in the initial stage, we use Baidu Encyclopedia as a supplement to make full use of Encyclopedia website knowledge update quickly to improve the accuracy. Baidu Encyclopedia is now the world's largest Chinese encyclopedia. As of March 2017, more than 1409,000 entries are collected by 124493000 multiple edits, and the content is updated very frequently. The ambiguous entities with the same name in CN-DBpedia have a suffix tag. For example, when we search for "Li Na" in CN-DBpedia, there are 27 objects with the same name. The suffix labels are Chinese women's tennis star, pop singer, Chinese stage director, and Guangxi Art Institute professor etc. It can be seen that the initial set size after the initial search may be too large. If the set of candidate entities is made too large, the calculation is more complicated, so the initial set must be filtered.

We use two steps to filter the initial set. Firstly, traversing each candidate entity set in the knowledge base to obtain the attributes of the candidate entity. Filtering out no attribute and uncommon entities to reduce the size of the initial set preliminarily. Then we calculate the cosine similarity between the input text D of the entity mention and the Baidu Encyclopedia page of the corresponding candidate entity as:

$$\text{Cos\_}\,Sim = \frac{m \cdot c}{|m||c|}$$

Where *m* represents the eigenvector of the text of the entity mention, an *c* represents the eigenvector of the encyclopedia page of the candidate entity. If the candidate entity does not have the corresponding encyclopedia page, the cosine similarity is defined as 0. The threshold τ =0.3 is defined by cross experiment. The candidate entity whose cosine similarity is less than the threshold is filtered to further narrow the candidate entity set size. So that we get the final candidate entity set $C = \{C_1, C_2, \cdots, C_n\}$.

3.2 Topic Relation Graph

In this section, we discuss the method of constructing an entity topic relation graph for the input text D to achieve the goal of a collective linking. The topic relation graph is an undirected graph *G = (V, E)* with weight where *V* is the node set of the graph and E is the set of edges between nodes. The final result of the entity linking is based on the topology of the graph, and the quality of the graph affects the performance of the algorithm to a great extent. So the construction of entity relation graph is very important. We divide the graph into two parts: the outer circle graph and the inner circle graph. The part of all the entities mention is called the outer circle graph, while the part of all the candidate entities is called the inner circle graph. An example of a topic relation graph is shown in Fig.1.

Construct of vertex set: *Firstly, we construct the vertex collection. The* input *text D has a set of mentions* $M = \{m_1, m_2, \cdots, m_n\}$ *,while the set of candidates for the mentions are* $C = \{C_1, C_2, \cdots, C_n\}$ *. So the vertex set V of the entity relation graph is defined as the union of the entity mention set* M *and the candidate entity set* C *. For the sake of convenience, we define the subset* $V_1 = \{m_i \mid \forall m_i \in M\}$ *as the vertex in the outer circle graph, and the subset* $V_2 = \{c_{ij} \mid \forall c_{ij} \in C_i, \forall C_i \in C\}$ *of V is the vertex of the inner circle graph.*

Construct of edges set: We construct the graph *G* as an undirected graph, which is divided into two kinds of edges. The first kind of edge is the edge of the entity mention and its corresponding candidate entity, that is, the connection between outer circle graph and the inner circle graph. The second is the edge of the interconnection between the candidate entities, namely the edge of inner circle graph. And the specific structure is as follows.

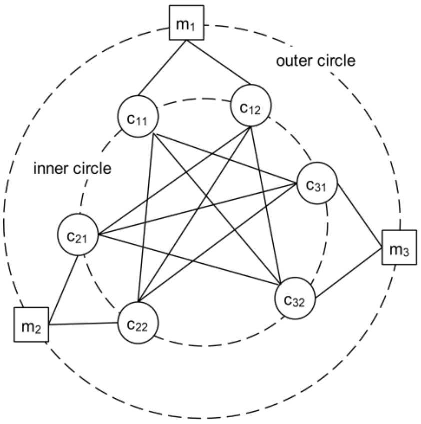

Figure 1. An example of topic relation graph.

The edge of the inner circle graph is called *Local Topic*. That is, the edge between an entity mention to be linked and one of its candidate entity. The value of local topic $Local(c_{ij})$ is calculated as: All the entities extracted from the input text D are formed as a vector $\vec{M} = \{m_1, m_2, \cdots, m_n\}$. For a candidate entity c_{ij} of one of the entity mention m_i , we grab the anchor text on the Baidu Encyclopedia page to form a vector $\vec{c}_{ij} = \{anchor_1, anchor_2, \cdots\}$. Then we use edit distance to calculate the number of the same items in the two vectors, and normalize the calculated values for all candidate entities of the same entity mention. In particular, for an entity that has only one candidate entity, the weight of the edge between them is 1, that is, the value of its local topic is 1. The value of local topic represents the relevance of the subject semantics of a candidate entity to its corresponding entity mention.

The edge of the inner circle graph is called *Global Topic*. If the vertexes in the set V_2 corresponding contents of the Baidu Encyclopedia page exist similarity, then there would be an edge to connect them in the graph. But in fact for the sake of calculation, we will be fully connected to the vertex in the construction phase of the inner circle graph. We use the weight 0 of the edge to determine whether the two vertices of the edge have a semantic association, and use the weight to determine the size of the semantic correlation between the two vertices. Particularly, there is no semantic correlation between candidate entities corresponding to the same entity mention. So we define that the weight of the edges between them are 0 to indicate that there is no association. The value of global topic is calculated by taking the candidate entity set of any two entity mentions m_s and m_t as examples. The candidate entity set of m_s is $C_s = \{c_{s1}, c_{s2}, \cdots, c_{sp}\}$, while the candidate entity set of m_t is $C_t = \{c_{t1}, c_{t2}, \cdots, c_{tq}\}$. Firstly, we initialize the topic relation graph. The vertices in C_s and C_t of the inner circle graph are fully connected, and the weight on the edges are set to 0. Secondly, we crawl the Baidu Encyclopedia pages of the vertex c_{s1} in the set C_s and the vertex c_{t1} in the set C_t separately. In order to determine the topic content of the page, all words in each page are weighted using the TF-IDF schema. Finding the keywords of the Baidu page where the two candidate entities are located, and making the collection of keywords which can represent the topic of each page. And then we generate their own word frequency vector, using the bag of words model to calculate the correlation of the two word frequency vector as the semantic relevance between c_{s1} and c_{t1}.If they are not related, the value is recorded as 0. Calculating the semantic similarity between c_{s1} and $c_{t1}, c_{t2}, \cdots, c_{tq}$ in turn to get q values, and then we normalize these q values to get the results as the global topic of c_{s1} and $c_{t1}, c_{t2}, \cdots, c_{tq}$, using $Global(c_{s1}, c_{tj})$ to express, where $j \in [1, 2, \cdots, q]$. The rest of the global topic values are calculated similarly. So the global topic of the vertexes between the set C_s and C_t are represented as $Global(c_{si}, c_{tj})$, where $i \in [1, 2, \cdots, p]$ and $j \in [1, 2, \cdots, q]$.

At this point, the vertices, edges and weights of the topic relation graph are constructed completely.

3.3 Collective Entity Linking

In this section, we propose a collective linking algorithm. The basic thought of the collective entity linking algorithm is to implement the one-time batch link for the multiple entities that co-occurrence in the given input text, using their semantic links and content dependency auxiliary disambiguation. The idea is to transform the entity linking problem into the rank of the topic consistency of candidate entities. According to the similarity obtained by section 3.2 and the collective entity linking algorithm introduced in the next step, we choose a set of candidate entities with the highest topic consistency of the entity mentions to be linked as the final link object. The process of collective entity linking algorithm consists of three steps.

First of all, for the entity relation graphs constructed in section 3.2, we perform the scoring calculation in the vertex set V_2 of the inner circle graph where $V_2 = \{c_{ij} \mid \forall c_{ij} \in C_i, \forall C_i \in C\}$. As for each $C_i \in C$, each time in the set C_i select and select only one vertex c_{ij}, where $i \in [1, n]$ and $j \in [1, length(C_i)]$. The inner circle graph score of c_{ij} is the score for the star chart with c_{ij} as the head node and the other selected nodes as component. Fig.2 is an example of star diagram. And the formula is as:

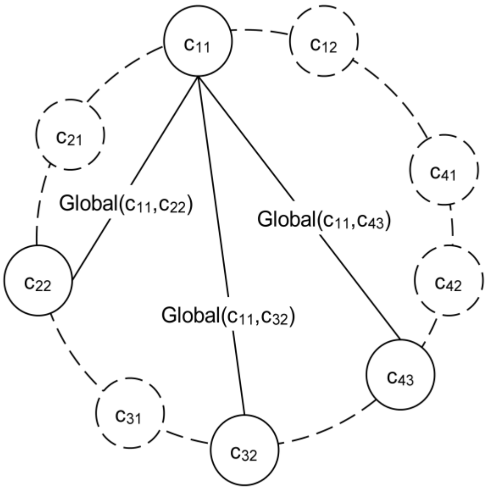

Figure 2. An example of star diagram in topic relation graph.

$$Trail(c_{ij}) = \max\{\sum_{k=1,k\neq i}^{n} Global(c_{ij}, C^{1}_{length(C_k)} C_k)\} .(2)$$

where $C^{1}_{length(C_k)} C_k$ represents a single vertex selected from the set C_k (except the set C_i) which is the only one vertex selected from set C_k at one time. And $Global(c_{ij}, C^{1}_{length(C_k)} C_k)$ represents the global topic of c_{ij} to these selected n-1 vertices, respectively. The n-1 global topic values are then added together, and the result is called a score of c_{ij}. We choose to record the maximum score of c_{ij} as the score $Trail(c_{ij})$ simultaneously.

Secondly, using the local topic between the entity mention and its candidate entity calculated in the section 3.2, and the trail value obtained in the previous step, we derive the semantic consistency score between the entity mention and the candidate entity by the following formula.

$$Consistency(m_i, c_{ij}) = Local(c_{ij}) + Trail(c_{ij}) \,(3)$$

$Consistency(m_i, c_{ij})$ represents the whole topic consistency between entity mention m_i and its corresponding candidate entity c_{ij}, which is composed of the local topic score and the trail score. The value of $Local(c_{ij})$ indicates the relevance of the topic semantics between the encyclopedia page of the candidate entity

c_{ij} and the text where the entity mention m_i appears. $Trail(c_{ij})$ represents the context correlation score between candidate entity c_{ij} and other entities in the current text.

Finally, after calculating the topic consistency score between the entity mention and all of its candidate entities, the candidate entities are ranked according to the score from high to low. We choose the highest ranked candidate entity as the linking result of the entity mention. The linking result is as:

$$Link(m_i, result_i) = \underset{c_{ij} \in C_i}{\arg\max}(Consistency(m_i, c_{ij})) \; (4)$$

where right side indicates that for a given entity mention m_i, the candidate entity c_{ij} with the highest score of the semantic consistency of m_i is selected from its candidate entity set C_i, and $Link(m_i, result_i)$ indicates that the pending entity mention m_i is linked to the candidate entity $result_i$ in the knowledge base finally. $result_i$ is the return value on the right side of the equation.

4. Experiments

In this section, we introduce the data set and baseline algorithm firstly, and then test the performance of our method and compare it with baseline algorithms.

4.1 Experiment Settings

In the experiment, we use CN-DBpedia as the local knowledge base. CN-DBpedia is a large-scale general field structured encyclopedia developed and maintained by Fudan University, which predecessor is Fudan GDM Chinese knowledge map. This map now has up to 9,455,262 entities, and offers a full suite of APIs for free. In order to make full use of Encyclopedia site knowledge update quickly to improve the accuracy, we use Baidu Encyclopedia as a supplement, which has been collected more than 14111 million entries.

We have selected two sets of public available corpus for performance testing of our collective entity linking algorithm based on content consistency.

The first group was compiled by NLP&CC 2014 Entity Linking contest. This test corpus includes 570 microblogging documents. Since the knowledge base that we use is not the NLP&CC Entity Linking contest, it is necessary to manually label the 570 micro-blog documents. And a total of 607 entities to be linked are marked.

The second group of data is a random crawl to 150 news texts from Phoenix News website information and Sina News, including military news, sports news, entertainment news, social news and other categories. Meanwhile we also manually extract the entity, and then label the extracted entities based on the local knowledge base. A total of 587 entities are marked.

4.2 Baselines

Champion_2014: This algorithm is an entity linking algorithm based on context similarity calculation, which integrates Sina microblogging user information based on the return of Baidu Encyclopedia.

Babelfy: This is an entity linking system based on graph which uses the BabelNet semantic network to disambiguate, and supports multi-lingual (including English, Chinese, Russian, etc.) entity disambiguation tasks. For the pending text, Babelfy first constructs its semantic relation graph based on semantic signature, and then discards the ambiguous entity by extracting the dense subgraph.

4.3 Evaluation Method

In order to evaluate the experimental results, we take the algorithm accuracy and performance metrics that include precision, recall and F_1-value as the criterions.

Firstly, we manually link the entity mentions to the corresponding entries in the knowledge base. Set M_E indicates the entity mention set, each entry of which can find the corresponding candidate entity in the knowledge base. For entities that are not linked in the knowledge base, they are marked as NIL manually. We call these entities as set M_N. Then we test the corpus using our proposed algorithm. Set R_E indicates the kind of entity mention set, each entry of which can find the corresponding candidate entity in the knowledge base by using proposed algorithm. The set R_N represents those entities that are not linked in the knowledge base by the outputs of the algorithm, that is, the set of entities labeled NIL. The accuracy rate indicates that how many results of our algorithm are the same as the results of manual labeling. The formula as:

$$Accuracy = \frac{|M_E \cap R_E| + |M_N \cap R_N|}{|M_E \cup M_N|} \times 100\% \; (5)$$

Simply, the meaning of precision is the exact proportion of the linked entities, while the meaning of the recall is the proportion of all the exact entities being linked. In general, the increase in precision means a reduction in recall, so the F_1-value is used as a compromise. The formula is as follows.

$$Precision = \frac{|M_E \cap R_E|}{R_E} \times 100\% \; .(6)$$

$$Recall = \frac{|M_E \cap R_E|}{M_E} \times 100\% \, . \; (7)$$

$$F_1 = \frac{2 * Precision * Recall}{Precision + Recall} \times 100\% \; .(8)$$

4.4 Evaluation Results

We test the performance of the algorithm that we proposed on the two corpus sheets. In contrast, we also test the algorithms Champion_2014 and Babelfy on the same data set. Table 1 shows the precision, recall and F_1-value of the algorithm Champion_2014, Babelfy and our method in NLP&CC 2014 corpus. Table 2 test results for the algorithm Champion_2014, Babelfy and our methods on the news texts which randomly crawl from the Phoenix information news website and the Sina News Network.

As seen from table 1, although our algorithm performs much higher than Babelfy in the accuracy, precision, recall and F_1-value, but lower than Champion_2014 algorithm that has the best performance on NLP&CC 2014 entity linking contest. This is because the micro-blog short text is too short and has its own features, and the statements of users contain a variety of noise factors. However, our algorithm does not have the special treatment for micro-blog short texts as Champion_2014 algorithm does, which has a certain impact on our algorithm. Inspired by this, our next step is to consider how to reduce the impact of noise on our algorithm.

Table 2 shows that the accuracy, precision, recall, and F_1-value of our proposed algorithm in news texts are higher than Champion_2014 and Babelfy algorithms. The news texts have relatively less noise, which is conducive to our entity recognition. So we have better performance in the news texts than in the micro-blog texts. Although the Champion_2014 algorithm has a good performance in the micro-blog texts, but the performance of the algorithm model is highly dependent on the size and type of the samples, so the performance of the Champion_2014 algorithm in the news texts is lower than our method. Meanwhile, the performance of our algorithm in micro-blog texts is considerable in table 1, which indicates that our algorithm has good generalization performance. When compared with the representative work of the collective entity linking method Babelfy, our performance is significantly improved. This is because the performance of the Babelfy algorithm is highly dependent on the knowledge content of the BabelNet semantic network. For an entity that is not in the knowledge base, the Babelfy only assigns an abstract entity with no specific information. And our source of knowledge that combines the Fudan CN-DBpedia with the world's largest Chinese Encyclopedia Baidu encyclopedia, covering almost all areas, to provide effective support for the accuracy of entity linking.

In view of the Chinese word segmentation challenge, we use the THULAC word segmentation kit which is strong, and trained by the world's largest artificial word segmentation and POS tagging Chinese corpus (About 58 million words). Fig.3 shows the effectiveness of the word segmentation tool. We compare the performance of THULAC with the domestic representative of the word segmentation software: LTP-3.2.0, ICTCLAS and jieba. As seen from the Fig.3, the accuracy and F_1-value of the THULAC are higher than those of other three representative work in the test data of the news text corpus. It indicates that the use of THULAC toolkit for good word segmentation also improved our algorithm performance.

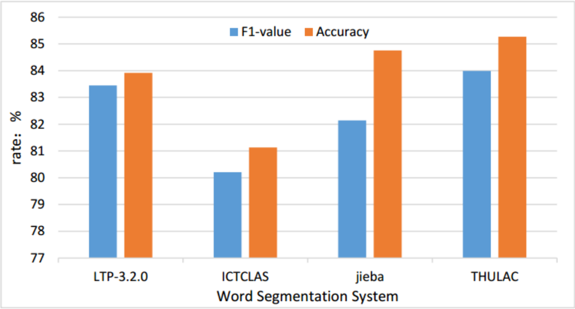

Figure 3. Effectiveness of the word segmentation tool.

Table1. Effectiveness on NLP&CC 2014 Corpus

System	Accuracy	Precision	Recall	F1-value
Champion_2014	86.82	80.78	85.98	83.30
Babelfy	71.82	70.36	75.71	72.93
Our method	82.37	79.88	82.69	81.26

Table2. Effectiveness on News Corpus

System	Accuracy	Precision	Recall	F1-value
Champion_2014	82.45	79.28	81.86	80.55
Babelfy	74.35	71.89	72.92	72.39
Our method	85.69	82.83	85.19	83.99

5. Conclusion and Future Work

In this paper, we propose a Chinese collective entity linking algorithm based on topic consistency and propose a method for calculating the topic consistency of the entity. For the entity mention to be linked, we extract the topic of its context. Finally, the candidate entity we choose is the most relevant to the topic of the context of the entity mention to be linked.

This method uses the combination of local knowledge base and external knowledge base to make up for the problem of imperfect Chinese local knowledge base and improve the accuracy. In addition, two layers of the selection method are used to control the size of candidate entities, and the computational complexity is reduced. Moreover, the computation of the correlation between the entity mention and the candidate entity takes full account of the topic relevance and the content relevance between the co-occurrence entities, which improves the precision and recall. Finally, using the semantic correlation between entities, we adopt the graph-based collective entity linking method to batch link the entity mention into the knowledge base. Our proposed method is refined, easy to accomplish and is effective from the experimental results. But the processing capacity for micro-blog and other social networking texts has yet to be improved. So in future work we intend to improve our algorithm for short texts and noise processing capabilities. In addition, in view of the inherent limitations of Chinese text processing, the next step is to explore more effective approach to further improve the accuracy, precision and recall.

Acknowledgment

This work was supported by National Core Electronic Devices, High-end Generic Chips and Basic Software (CHBS) Project "xx Mobile Terminal Operating System" under Grant No. 2016ZX01040101.

References

[1] M. Sozio and A. Gionis, "The community-search problem and how to plan a successful cocktail party," ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, Dc, Usa,July,2010, pp. 939-948.

[2] W. Shen, J. Wang, and J. Han, "Entity linking with a knowledge base: Issues, techniques, and solutions," IEEE Transactions on Knowledge and Data Engineering, 2015, 27(2) , pp. 443-460.

[3] P. Fafalios, P. Papadakos, and Y. Tzitzikas, "Enriching textual search results at query time using entity mining, linked data and link analysis," International Journal of Semantic Computing, 2014, 8(04) , pp. 515-544.

[4] S. Guo, M. Chang, and E. Kiciman, "To Link or Not to Link? A Study on End-to-End Tweet Entity Linking," HLT-NAACL. 2013, pp. 1020-1030.

[5] W. Li, D. Qian, and Q. Lu, "Detecting, categorizing and clustering entity mentions in Chinese text," Proc. ACM SIGIR conference on Research and development in information retrieval, 2007, pp. 647-654.

[6] R. Bunescu and M. Pasca, "Using Encyclopedic Knowledge for Named entity Disambiguation," Eacl 2006, Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, Trento, Italy. DBLP, 2006, pp. 9–16.

[7] H. Nguyen and T. Cao, "Exploring wikipedia and text features for named entity disambiguation," Intelligent Information and Database Systems, 2010, pp. 11-20.

[8] S. Gottipati and J. Jiang, "Linking entities to a knowledge base with query expansion," Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (ACL), 2011, pp. 804-813.

[9] Y. Zeng, D. Wang and T. Zhang, "Linking entities in short texts based on a Chinese semantic knowledge base," Natural Language Processing and Chinese Computing. Springer Berlin Heidelberg, 2013, pp. 266-276.

[10] T. Zhang, K. Liu and J. Zhao, "A graph-based similarity measure between Wikipedia concepts and its application in entity linking system," Journal of Chinese Information Processing, 2015, pp. 58-67.

[11] Z. Zuo, G. Kasneci and T. Gruetze, "BEL: Bagging for Entity Linking," International Conference on Computational Linguistics. 2014, pp. 2075-2086.

[12] S. Kulkarni, A. Singh and G. Ramakrishnan, "Collective annotation of Wikipedia entities in web text," ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, June 28 - July. DBLP, 2009, pp. 457-466.

[13] Z. Guo and D. Barbosa, "Robust Entity Linking via Random Walks," ACM International Conference on Conference on Information and Knowledgemanagement, CIKM. ACM, 2014, pp. 499-508.

[14] X. Han, L. Sun and J. Zhao, "Collective entity linking in web text: a graph-based method," International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2011, pp. 765-774.

[15] A. Alhelbawy and R. Gaizauskas, "Graph Ranking for Collective Named Entity Disambiguation," Meeting of the Association for Computational Linguistics. 2014, pp. 75-80.

[16] A. Moro, A. Raganato, R. Navigli, "Entity linking meets word sense disambiguation: a unified approach," Transactions of the Association for Computational Linguistics, 2014, pp. 231-244.