Simple Yet Effective Method for Entity Linking in Microblog-Genre Text

Qingliang Miao, Huayu Lu, Shu Zhang, and Yao Meng

Fujitsu Research & Development Center CO., LTD,
No. 56 Dong Si Huan Zhong Rd, Chaoyang District, Beijing, China
{qingliang.miao,zhangshu,mengyao}@cn.fujitsu.com,
lvhuayu@gmail.com

Abstract. Semantic analysis microblog data is a challenging, emerging research area. Unlike news text, microblogs pose several new challenges, due to their short, noisy, contextualized and real-time nature. In this paper, we investigate how to link entities in microblog posts with knowledge base and adopt a cascade linking approach. In particular, we first use a mention expansion model to identify all possible entities in the knowledge base for a mention based on a variety of sources. Then we link the mentions with the corresponding entities in the knowledge base by collectively considering lexical matching, popularity probability and textual similarity.

1 Introduction

With the emergence of knowledge base population projects like DBPedia [9] and YAGO [14], more and more large-scale knowledge bases are available. These knowledge bases include rich semantic knowledge about entities, their properties and relationships. Ideally, automatically linking web data with knowledge bases can facilitate many applications such as entity retrieval, advertising and product recommendation. On the other hand, creating links between web data and knowledge bases could enrich the knowledge base as well.

A key technology to implement the above vision is entity linking, which aims to link the mentions in a document with corresponding entities in the knowledge base. Given a mention *m*, a document *d* and a knowledge base *KB* including a set of entities $\{e_1, e_2 \ldots e_n\}$, an entity linking system is a function $f: m \rightarrow e_i$ which links mention *m* with corresponding entity e_i in *KB* [15]. Figure 1 illustrates the entity linking task. The linking process includes two steps, first identify all the entity candidates that may link with mention "Apple", and then identify which entity should be linked with the mention. The entity linking task, however, can be no-trivial due to the mention ambiguity and variation issues [15].

Recently, microblogs have become an important web data due to its real-time nature. In this paper, we analyze the challenges of entity linking in microblog-genre text and adopt a cascade approach to create links between mentions and entities in knowledge base.

G. Zhou et al. (Eds.): NLPCC 2013, CCIS 400, pp. 440–447, 2013.
© Springer-Verlag Berlin Heidelberg 2013

The rest of the paper is structured as follows. In the following section we review the existing literature. We introduce the proposed approach in section 3. We conduct comparative experiments and present the experiment results in section 4. At last, we conclude the paper with a summary of our work and give our future working directions.

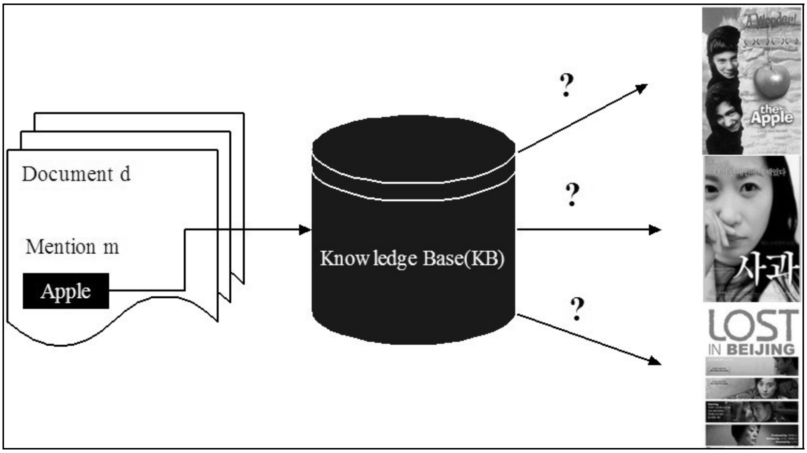

Fig. 1. An example of Entity Linking

2 Related Work

Generally speaking, entity linking is a kind of semantic annotation [6], which is characterized as the dynamic creation of interrelationships between entities in knowledge base and mentions in unstructured or semi-structured documents [5].

In particular, most existing semantic annotation approaches annotate documents with links to Wikipedia or DBPedia. For example, Mihalcea and Csomai [7] first propose Wikify system to use Wikipedia to annotate text. Milne and Witten [8] implement a similar system called Wikipedia Miner, which adopts supervised disambiguation approach using Wikipedia hyperlinks as training data. Han and Sun [15] propose a generative probabilistic model, called entity mention model, which can leverage entity popularity knowledge, name knowledge and context knowledge for the entity linking task. In practical applications, several entity linking systems have been developed [8] [9] [10] [11] [12]. DBpedia Spotlight [10] is a system for automatically annotating text documents with DBpedia URIs. TagMe [9] system adopts a collective disambiguation approach, which computes agreement score of all possible bindings, and uses heuristics to select best target. The disambiguation model of Illinois Wikifier [11] is based on weighted sum of features such as textual similarity and link structure. AIDA [12] is a robust system based on collective disambiguation exploiting the prominence of entities, context similarity between the mention and its candidates, and the coherence among candidate entities for all mentions.

Recently, more and more works have been focusing on entity linking in short informal texts (e.g. tweets) [1][3]. Stephen Guo et al. [2] propose a structural SVM algorithm for entity linking that jointly optimizes mention detection and entity disambiguation. Cassidy et al. [4] mainly test the effects of two tweet context expansion methods, based on tweet authorship and topic-based clustering.

3 The Approach

3.1 Preprocess

In the preprocessing step, we first group each microblog post p_i according to their topic $TP(p_i)$, and then we index the microblog posts and the textual contents describing the entities. The index is implemented by Lucene index API. Due to the creative language usage in microblog posts, the mentions are informal. For example, some mentions are in traditional Chinese characters, such as “鄭州”, some mentions are mixed with Pinyin and Chinese characters, such as “fudan大學”. Consequently, we have to normalize these informal mentions first. We also normalize the punctuations such as common, French quotes, whitespaces and correct the misspelling mentions if any.

3.2 Knowledge Repository

After preprocessing step, we build a knowledge repository of entities that contains vast amount of name variations of entities such as acronyms, confusable names, spelling variations, nick names etc. We use Wikipedia, BaiduBaike and the Web to build the knowledge repository. In particular, we utilize the following resources to build the knowledge repository.

Redirect Pages

Redirections in Wikipedia and other encyclopedias like BaiduBaike are good indicators for synonyms. For example, Wikipedia page “湖人” redirect to “洛杉矶湖人”. In this paper, we use redirect pages to identify alternative names, synonyms, abbreviations, scientific or common terms and alternative spellings etc.

Bold Phrases

The bold phrases in the first paragraph usually summarize name variants of the entity [13], e.g. full names, nick names, alias names etc. For example, in Wikipedia page about “IBM”, we could obtain variants such as “国际商业机器股份有限公司”, “International Business Machines Corporation” and “万国商用机器公司”.

Disambiguation Pages

Disambiguation pages are used for ambiguous entities, which consist of links to Wikipedia pages defining the different meanings of the same mention. They are useful in homonym resolution and help in extracting abbreviations etc. For example, disambiguation page “詹姆斯” contains more than 40 persons such as “詹姆斯·加菲尔德”, the twenty-president of the United States and “勒布朗·詹姆斯”, American professional basketball player.

Anchor-Entity Association

Anchor links could also be used to trace to which entity the mention links. In this work, we use anchor texts from inter WikiPedia links. In addition, we quantify the strengths of associations between entity and mention pairs using basic statistics. The score of strength is computed as the number of times that mention m links to entity e divided by the total number of anchors with mention m.

Besides the above four sources, we also extract alias from the attributes parts $A(e_i)$ of the given knowledge base. For example, we can obtain alias “京”, “Peking” and “Municipality of Beijing” for mention “北京” from the attributes parts of the knowledge base.

3.3 Candidate Generation

In this module, we use heuristics to expand the mentions and obtain all possible variants of the mention from the knowledge repository. Besides the variants expanded by the knowledge repository, we also use contextual information and the Web to expand the mentions. The contextual content of mention usually contains rich information about its entity candidate, especially for abbreviation name mentions. For example, given the following microblog post “北京时间3月12日，2013亚冠联赛小组赛第二轮，广州恒大足球俱乐部客场挑战全北现代，广州恒大首发已经公布”. We can identify the entity of all above abbreviations using simple rules, for example, “广州恒大” refers to “广州恒大足球俱乐部”.

Even though we expand the mention with knowledge repository and contextual content, we still cannot exhaustively detect all the entity candidates of mention. Therefore, we try to exploit the whole web information for detecting the candidates through web search. Given a mention, we submit it with string "维基百科" or "百度百科" to the Google API and retrieve only the web pages within these encyclopedias. For example, given the mention "詹皇", we submit the queries like "詹皇 百度百科" and retrieve the search result "勒布朗·詹姆斯".

3.4 Microblog Post Expansion

As discussed above microblog-genre text has less disambiguation context, consequently, we have to expand the initial microblog posts. We use two methods to expand microblog posts, namely, keywords based and unambiguous entities based method. In particular, we first extract keywords or unambiguous entities around the mention from initial microblog posts, and then use these keywords or entities to retrieval topical related text from the given microblog post corpus or the web. In this paper, we use normalized Google distance to extract keywords.

$$NGD(x, y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x, y)}{\log M - \min\{\log f(x), \log f(y)\}} \quad (1)$$

3.5 Entity Resolution

In this section, we introduce the cascade approach for entity resolution. The main goal of the module is to link a mention with a knowledge base (*K*) entity or NIL. The flow chart of this module is shown in Figure 2.

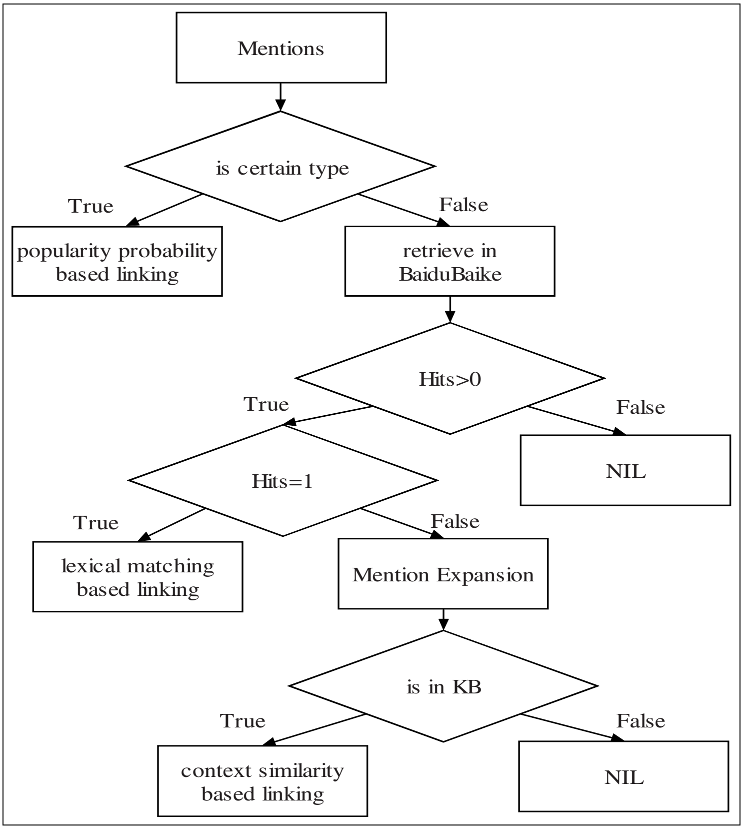

Fig. 2. Flow chart of entity resolution module

The entity resolution module contains four steps. In the first step, we adopt popularity probability to link some specific kind mentions. In microblog post, users usually talk about popular or common entities such as geographic name, teams and players names. For example, mention “北京” means the capital of China in most cases, and “热火” denotes “迈阿密热火队”, the professional basketball team based in Miami. Although for some mentions it may be a dangerous bias to prefer popular or common entities, it seemed helpful for entity linking in microblog-genre text. In the second step, we feed mentions to Baidu Baike, and retrieve the results. If the search results do not contain the mention, we think the mention does not exist in the given knowledge base, because the knowledge base is a subset of Baidu Baike encyclopedia. If we find the unique result, we use a lexical matching based method to identify the entity. In the third step, we use the mention expansion module described in section 4.3 to expand a mention, and then we search these entity candidates in the knowledge base. If no candidate exists in the KB, we assign NIL to this mention, otherwise, we adopt a contextual similarity based method to identity the entity in step four. In step four, we use a threshold to determine whether the mention should be linked with an entity or NIL. In this work, we use validation data set to tune the threshold.

4 Experiments

In this section, we report a primary experiment aimed at evaluating the proposed system MSAS.

4.1 Experimental Setup

In this experiment, we use the Chinese microblog entity linking evaluation data sets provided by Natural Language Processing and Chinese Computing Conference[1].

Table 1. The statistics of the dataset

Data Set	Microblog posts	Mentions	Topics
Training data	177	249	12
Test data	787	1249	63

4.2 Experimental Results

In this experiment, we report the experiment results of two systems. The first system MSAS1 expands the initial microblog post by retrieving both the given microblog post corpus and the web, while the second system MSAS2 does not use the web data. The threshold is assigned 0.06 and 0.03 in MSAS1 and MSAS2, respectively. Table 2 and 3 shows the experiment result of two systems. From table 2, we can see that microblog post expansion is useful in contextual similarity based disambiguation. From table 3, we can see NIL detection methods achieve both high precision and recall. Generally speaking, contextual similarity can achieve promising results, but it fails in some cases when the entity candidates are in similar domain. For example, “霸王别姬” can refer to different entity in movie and literature domain. Another example is “凯恩斯”, which could refer to the a famous economist “John Maynard Keynes” and a netizen with network name “凯恩斯” who usually publish economic content in his blogs. Some informal mentions are also hard to link, such as “CCAV”, “毛总”, “周同学”.

[1] http://tcci.ccf.org.cn/conference/2013/pages/page04_eva.html

Table 2. The overall results on micro-averaged accuracy

System	micro-averaged accuracy
MSAS1	0.9092
MSAS2	0.8995

Table 3. The In-KB linking and NIL linking result on precision, recall and F-measure

System	In-KB results			NIL results		
	Precision	Recall	F1	Precision	Recall	F1
MSAS1	0.8983	0.8812	0.8897	0.9201	0.9383	0.9291
MSAS2	0.8859	0.8670	0.8764	0.9130	0.9333	0.9231

5 Conclusion

In this paper we investigate how to link mentions in microblog posts with knowledge base entities and present a microblog semantic annotation system (MSAS). This system can automatically create links between mentions and entities in knowledge base. In particular, we first build knowledge repository by mining entity knowledge from multiple sources. Second, we develop a mention expansion model to identify all possible entities in the knowledge base. Finally, we employ a divide and conquer strategy to identity the entity in knowledge base or NIL. In addition, we test the effects of microblog context expansion method, based on topic-based retrieval and web search. Experimental results on real world datasets show promising results and demonstrate the proposed system is effective. As a future research, we plan to use more sophisticated mention normalization methods to solve informal name variant issues. For entity disambiguation, we also plan to exploit reasoning with local and global evidence to reach a collective agreement.

References

1. Meij, E., Weerkamp, W., Rijke, M.D.: Adding Semantics to Microblog Posts. In: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, pp. 563–572 (2012)
2. Guo, S., Chang, M.W., Kıcıman, E.: To Link or Not to Link? A Study on End-to-End Tweet Entity Linking. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (2013)
3. Derczynski, L., Maynard, D., Aswani, N., Bontcheva, K.: Microblog-Genre Noise and Impact on Semantic Annotation Accuracy. In: Proceedings of the 24th ACM Conference on Hypertext and Social Media, pp. 21–30 (2013)
4. Cassidy, T., Ji, H., Ratinov, L., Zubiaga, A., Huang, H.Z.: Analysis and Enhancement of Wikification for Microblogs with Context Expansion. In: Proceedings of 24th International Conference on Computational Linguistics, pp. 441–456 (2012)

5. Bontcheva, K., Rout, D.: Making Sense of Social Media Streams through Semantics: a Survey. Semantic Web Journal (2012)
6. Kiryakov, A., Popov, B., Ognyanoff, D., Manov, D., Kirilov, A., Goranov, M.: Semantic Annotation, Indexing and Retrieval. Journal of Web Semantics 1(2), 49–79 (2004)
7. 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)
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. Ferragina, P., Scaiella, U.: TAGME: On-the-fly Annotation of Short Text Fragments. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, pp. 1625–1628 (2010)
10. Mendes, P.N., Jakob, M., García-Silva, A., Bizer, C.: DBpedia Spotlight: Shedding Light on the Web of Documents. In: Proceedings of the 7th International Conference on Semantic Systems, pp. 1–8 (2011)
11. Ratinov, L., Roth, D.: Design Challenges and Misconceptions in Named Entity Recognition. In: Proceedings of the Thirteenth Conference on Computational Natural Language Learning, pp. 147–155 (2009)
12. Yosef, M.A., Hoffart, J., Bordino, I., Spaniol, M., Weikum, G.: AIDA: an Online Tool for Accurate Disambiguation of Named Entities in Text and Tables. In: Proceedings of the PVLDB 2011, pp. 1450–1453 (2011)
13. Varma, V., Bharat, V., Kovelamudi, S., Bysani, P.: GSK, S., Kumar, N. K., Reddy, K., Kumar, K., Maganti, N.: IIIT Hyderabad at TAC 2009. In: Proceedings of Text Analysis Conference, TAC (2009)
14. Suchanek, F.M., Kasneci, G., Weikum, G.: YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia. In: Proceedings of the International World Wide Web Conference, pp. 697–706 (2007)
15. 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)