International Journal of Database Theory and Application
Vol.7, No.1 (2014), pp.1-10
http://dx.doi.org/10.14257/ijdta.2014.7.1.01

Named Entity Linking Based On Wikipedia

Yongbo Yu, Lizhou Zheng, Kaifang Yang and Peiquan Jin*

School of Computer Science and Technology,
University of Science and Technology of China, 230027, Hefei, China

jpq@ustc.edu.cn

Abstract

In this paper, we present the ideas and methodologies on labeling the mentioned entities with the wiki dataset. This paper presents a system for the recognition and semantic disambiguation of named entities based on information extracted from a large encyclopedic collection from Wikipedia. We focus on maximizing the similarity between the contextual information extracted from Wikipedia and the context of a document, as well as the similarity among the category tags associated with the candidate entities. Our experimental results show that the proposed methods are effective and efficient to answer complex named entities disambiguation over the Wikipedia dataset.

Keywords: *Named Entity; Disambiguation; Wikipedia*

1. Introduction

Named entity disambiguation is a quite important issue on the Web. For example, according the Google search results, the name "*Michael Jordan*" represents more than ten persons. This may introduce many serious problems in various areas such as machine translation, information retrieval, and natural language semantic analysis.

This problem can be solved by utilizing the Wikipedia dataset, as Wikipedia contains rich semi-structured information. Presently, many researchers have used Wikipedia to conduct research on Web information extraction and semantic analysis including named entity disambiguation.

In this paper, we focus on the named entity disambiguation problem, and our work is based on a dataset collected from Wikipedia. The dataset contains around 3M entity names (Wikipedia URLs) and their 40M mentions. We aim to build a system for automatically detecting mentions of entities, and to link the detected mentions to entities in the given entity file with a high accuracy.

This paper describes the system we designed and implemented for entity linking. Basically, the system consists of three steps, namely Named Entity Detection, Data Preprocessing, and Named Entity Linking. In the Named Entity Detection step, we recognize the named entities in the given test corpus with the Stanford NLP tools. In the Data Preprocessing step, according to each mention that has already been recognized, we crawl the relevant Web pages of the candidate entities and extract useful information such as contextual data and category tags. In the Named Entity Linking step, we label the detected mentions with the candidate entity based on the similarity measurement.

Our experimental results show that the proposed methods are effective and efficient to answer complex named entity disambiguation over Wikipedia dataset.

ISSN: 2005-4270 IJDTA

The rest of the paper is organized as follows. Section 2 discusses the related work. Section 3 describes the general framework of our method. Section 4 discusses the details about preprocessing and entity linking. Section 5 gives the experimental results, and the conclusions are in Section 6.

2. Related Work

Named Entity Disambiguation (NED) is partially similar to the widely-studied issue of word sense disambiguation (WSD) [9]. In WSD, we aim to identify the sense of word given in the context. WSD is often performed with respect to WordNet (index), a lexical database mapping words to synonym sets. Similarly, the goal of NED is to label a given ambiguous named entity with one of multiple canonical named entities set constructed from a knowledge base. Due to the difficulty in capturing and leveraging the semantic knowledge as humankind, NED is a task full of challenge. Most previous works mainly concerned about identifying and characterizing entity mentions, or clustering mentions within documents and across documents for the lack of a comparably comprehensive sense inventory for entities.

Fortunately, in recent years, Wikipedia and other large-scale knowledge sources contain rich structured or semi-structured semantic knowledge, which can be used as a sense inventory for disambiguation. The most relative works on NED based on Wikipedia are conducted by Bunescu and Pasca, and Cucerzan [1]. Bunescu and Pasca trained a disambiguation SVM kernel to exploit the high coverage and rich structure of knowledge [1]. Cucerzan extended the work by adding some richer features to the similarity comparison [2]. Related work has been done by Hien T. Nguyen and Tru H. Cao to overcome the shortage of training data by automatically generating an annotated corpus based on a specific ontology, they also analyze the result of different combination of features representing the named entities. It comes out that disambiguation perform best using the Wikipedia features: ET (entity title), RT (redirect page title), CAT (category label) and OL (outlink label) in combination with text features [3, 5]. Ayman Alhelbawy and Rob Gaizauskas develop a document similarity function based on the named entity mentions found in two documents instead of the common vector space model computing the cosine similarity [13].

However, in general, named entity disambiguation methods measure the similarity between the named entities using the traditional bag of word model conventionally. This model measures similarity based on only the co-occurrence statistics of terms, without considering all the semantic relations like social relatedness between named entities, and lexical relatedness (*e.g.*, acronyms, synonyms) between key terms, which cannot reflect the actual similarity between entities [7]. Recently, research focus on exploiting background knowledge to capture the various semantic relations. Xianpei Han and Jun Zhao measure the similarity more accurately by building a large-scale semantic network model from Wikipedia [4, 10, 11]. They propose a knowledge based approach to improve the disambiguation by capturing and leveraging the structural semantic information from multiple knowledge source in the follow-up work. Similar work based on the context in which mentions appear is done by Danuta Ploch and Ivo Laˇsek. Danuta Ploch employ Wikipedia relations between co-occurred entities to achieve a range of novel features [8]. Ivo Laˇsek and Peter Vojt´aˇs introduce a novel disambiguation method by analyzing the structural dependencies of recognized entities [12]. Martin Jaˇcala and Jozef Tvaroˇzek exploit existing explicit semantics to construct a disambiguation dictionary, which perform better than the traditional latent semantic analysis method [14].

There are also many other uses of Wikipedia based knowledge, such as Anna Lisa Gentile, Ziqi Zhang build a random-walk graph model to calculate semantic relatedness [6]. Ben

Hachey and Will Radford make a step forward by proposing an unsupervised approach that work over a link graph of Wikipedia articles for document mentions [9].

3. The General Framework for Named Entity Disambiguation

Figure 1 shows the general framework of Wikipedia-based named entity disambiguation. We first detect the mentions in the given test corpus. Then we perform preprocessing on each mention and generate files with specific formats. After that, we build indexes for the mention to link with the proper candidate entity.

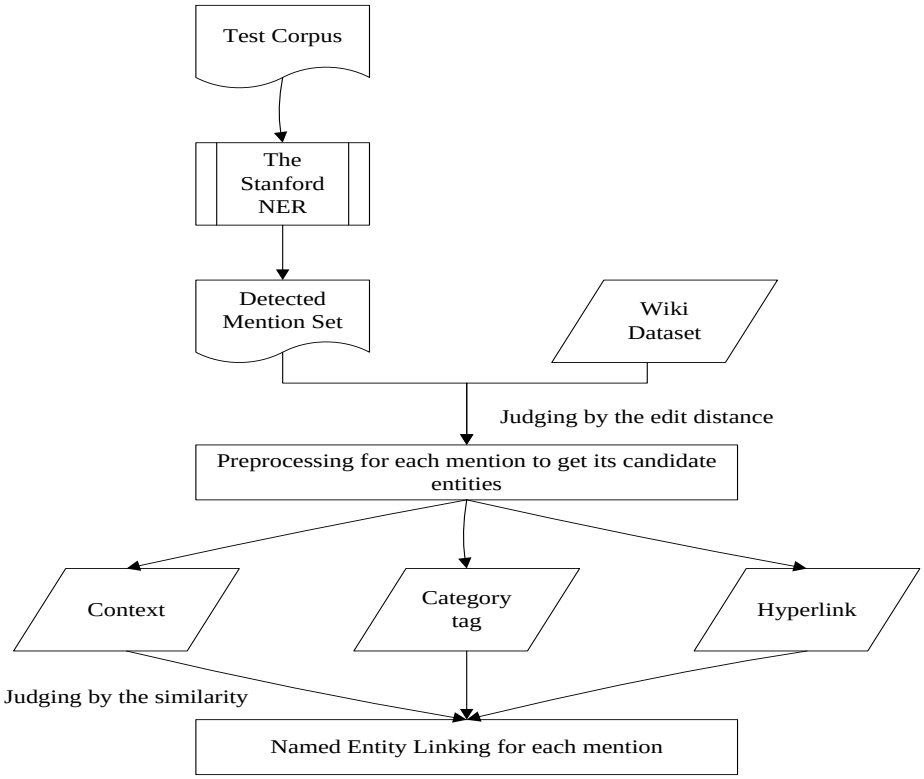

Figure 1. The framework of our method

4. Entity Detecting, Data Preprocessing, and Entity Linking

4.1 Entity Detecting and Data Processing

In our work, we use the Stanford NLP tool to detect the named entities. The Stanford NLP tool has already been demonstrated with good performance on NER (Named Entity Recognition). The mentions labeled with tags (ORGANIZATION, PERSON, LOCATION) are selected as the named entities concerned in our method.

We analyze the original Wikipedia dataset carefully, and find that it is quite time consuming to calculate the similarity between the extracted named entity and each candidate entity. So we make preprocess for each mention detected by the last step. Our preprocessing mainly contains three parts: downsizing the wiki dataset, crawling web pages and parsing the crawled webpages.

Compressing the Wikipedia Dataset. This is the preprocessing on whole given Wikipedia dataset. Its purpose is to reduce the size of the dataset. Through calculating the similarity between the mention and candidate entities, we can reduce the size of the wiki dataset remarkably. When calculating the similarity, we introduce the method of the minimum edit distance. The minimum edit distance between the candidate entity and the mention is the minimum number of single-character edits (insertion, deletion, substitution) that candidate entity required to change into the mention. We set a threshold to delete the candidate entities which are not similar to the mention. The threshold can be shown in Formula 4.1

$$\frac{\text{edit distance(mention, candidate entity)}}{\text{mention.lenth}} < 0.5 \tag{4.1}$$

The candidate entities which are higher than the threshold will be deleted from the wiki dataset.

When downsizing the wiki dataset, we find that the dataset format is shown in Figure 2.

```
James_Hanratty  James Hanratty##Hanratty
James_Hansen  James Hansen##Dr. James Hansen##Hansen##James E. Hansen‡
James_Hansen#Arrests  James Hansen was already arrested protesting coa
James_Hansen#Climate_model_development_and_projections  Hansen been at
James_Hansen#Honors_and_awards  Tides Foundation that gave Saint Hanse
James_Hansen#Who_is_responsible_for_climate_change.3F James E. Hansen
James_Hanson  James Hanson##Lord James Hanson
```

Figure 2. The Wikipedia dataset format

In Figure 2, we can see that each candidate entity is followed by one or more relative mentions. Considering these relative mentions may also affect the similarity, we also calculate the minimum edit distance between the detected mention from the test corpus and the mentions which are relative to the candidate entities, and then we take the average as the final edit distance.

After the downsizing dataset step, we can build an index from each mention to find its similar candidate entities. Table 1 is an example to show the structure form of the index for a named entity to be labelled.

Table 1. The structure of index

Mention to be labelled	*The similar candidate entities*
Frank_Potter	Frank_Porter_Patterson
	Frank_Porter_Wood
	Frank_Porter_Graham

Crawling Web Pages. This is the preprocessing based on the detected mentions. In this step we try to achieve the useful webpages of detected mentions. Last step, we have already built the index for each mention which is very convenient to craw the candidate entities URLs without visiting the whole large wiki dataset.

Parsing the Crawled Webpages. This is the preparation for the next step: Named Entity Linking. The crawled webpages contain rich information such as the contextual information, the category tags and the hyperlinks.

The contextual information of the entity URL is quite useful. If there are lots of same words and phrases in the contextual of both candidate entity and the detected mention in the

test corpus, it is more possible to label the mention with the entity than other candidate entities.

The category tags are also helpful to label the mention. Every entity has its category tags in Wikipedia. Different entities corresponding to the same mention may have different category tags. Categories separate articles into different topics, and these topics can be categorized by linking them with their parent categories. We can build a mapping structure to record the category information of each entity. The format of the mapping structure is shown in Table 2.

The hyperlinks reflect the relation among entities. When each of the two entities has a hyperlink directing to the other, the two entities may have higher similarity. We can also build a mapping table to record relative entities, as shown in Table 3.

Table 2. Mapping structure of entity

Named Entity	Category Of Entity
Victoria(Australia)	Former British colonies
	states and territories of Australia
	states and territories established in 1851
Queen Victoria	British princesses
	Monarchs of the UK
	English diarists

Table 3. Mapping table of relative entity

Named Entity	Relative Entities
Michael Jordan	Chicago Bulls
	Basketball
	1991 NBA Finals
	National Basketball Association
	1996-97 NBA Season

4.2 Entity Linking

After the preparation is done, our goal is to label the detected mentions with the candidate entity URLs. In this step, we take three factors that may have influence on the labelling in consideration: the text similarity, the context similarity, and the category similarity.

The text similarity depends a lot on the common words the two texts have. Before calculating the text similarity between the candidate and the mention, there are still some significant works we should do: we normalize the bag of words following the predefined rules as follows.

(1) Deleting special characters in some tokens, for example, normalize U.S.A to USA.

(2) Remove punctuation mark and special tokens such as commas, question mark, $, @, etc.

(3) Remove the common stop words such as a, an, the, etc.

Then we use the Vector Space Model and the Cosine Similarity. Thus, a name observation can be represented by the word vector in its context, *i.e.*, a word vector $o=\{(C_1,W_1),(C_2,W_2)\ldots(C_m, W_m)\}$, where each concept C_i represents the word appeared in the two texts that we want to compare with, and followed by its *tf-idf* weight. The cosine similarity is defined in Formula 4.2 and Formula 4.3.

$$v = (tf_1 * idf_1, tf_2 * idf_2, ..., tf_n * idf_n) \tag{4.2}$$

$$Sim(v_1, v_2) = \frac{\sum_{i=1}^{n} v_{1i} * v_{2i}}{\sqrt{\sum_{i=1}^{n} v_{1i}^2} * \sqrt{\sum_{i=1}^{n} v_{2i}^2}} \tag{4.3}$$

The same entity have different word sense in different contexts .For example, the disambiguation "*puma*" refers to the animal when with the context of "*cougar*", "*mammal*", "*felidae*", but it may refer to the "*Ford Puma*" with the context of "*ford motor company*" and "*car*". According to each detected mention *e*, we define *C*(*e*) as its context consisting of its relative words which we can get from the hyperlinks in the last step. Then we get the context agreement by calculate the degree of overlap between the context using the following Formula 4.4.

$$Con(e, D) = \frac{|C(e) \wedge C(D)|}{|C(D)|} \tag{4.4}$$

Here, |*C*| represents the size of the collection *C* in this formula. *D* refers to the test corpus.

The category tags may also show the similarity between the mention and the candidate entity. If both of them are about the same topic, they should have lots of category tags in common. For each detected mention's candidate entity *e*, we define *U*(*e*) as the category information which consist of the category tags of e. Using the Formula 4.5 as follows, we get the category similarity. In the formula, |*U*| refers to the size of the collection *U*.

$$Cat(e, D) = \frac{|U(e) \wedge U(D) - U(e)|}{|U(D)|} \tag{4.5}$$

After having got the three eigenvalues, we use the linear integration to get the final value as the factor to judge the similarity between the candidate entity and the detected mention. We choose the candidate which has the max integration value as the result to label the mention. The linear integration is shown in Formula 4.6.

$$\arg\max\{\lambda_1 * Sim(e, D) + \lambda_2 * Con(e, D) + (1 - \lambda_1 - \lambda_2) * Cat(e, D)\} \tag{4.6}$$

The parameters λ1, λ2 can be achieved by Logistic regression.

For each mention *m* to be disambiguated, we build a set of candidate entities *C*. We define the named entity linking as a ranking issue based on a hypothesis that there is a suitable function to calculate the semantic similarity between the mention *m* and the candidate entity *e* in the candidate set *C*. Here we use the feature vectors of the entities as the input of the function, and the output is the candidate entity with the highest similarity score. We use the similarity function as given in the front formula. The detailed procedure of what we described is shown in Algorithm 1. The similarity function is used in the fourth line in Algorithm 1.

Algorithm1.	Named Entity Linking based on the ranking similarity
1:	build a set of ambiguous mentions M
2:	**For** each mention $m \in M$ **Do**
3:	build a set of candidate entities of mention m: C
4:	$c \leftarrow \arg\max_{c_i \in c} Sim(Vector(c_i), Vector(m))$
5:	assign c to mention m
6:	**End For**

5. Performance Evaluation

We evaluate the performance of our system on a machine with configuration of Intel(R) Core(TM) i3-2120 CPU @ 3.30GHz and 4GB RAM. The operation system isWindows7 Ultimate and the maximum heap space of Java virtual machine is 2.5GB (JDK1.6). Since the truth label set for evaluating the results is unknown, we just select 10 files randomly to shown the coverage rate and precision rate. Table 4 shows the recognized mention numbers and the correct labeled mention numbers. Our experiments show that our method achieves a high precision of 82% and a recall of about 60%. However, the recall has a lot to do with the Stanford Entity Recognizer we used. In the test dataset, there are quite a lot of mentions are just ordinary persons that we cannot find a candidate entity in the knowledge base to label with.

Table 4. Result for randomly selected 10 files

File name	Mention recognized	Real mention number	Correct Labelled
739.txt	5	7	5
1340.txt	11	14	8
2227.txt	11	17	7
4898.txt	1	3	1
5731.txt	10	15	7
6826.txt	6	11	5
7546.txt	5	5	4
8544.txt	8	10	6
3782.txt	7	10	5
1.txt	2	2	2

The precision and recall for disambiguation by 10 files individually is shown in Figure 3. The methods we proposed perform well in labeling the location mentions and organization mentions. That is because these files contain more location and organization mentions such as “739.txt” and “1.txt”. In general, those files contain many person mentions will reach relative low recall, such as “4898.txt” and “6826.txt”.

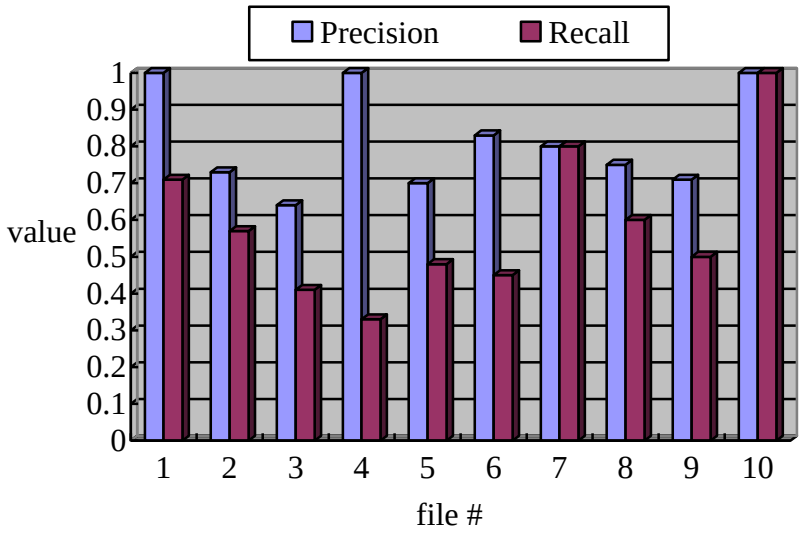

Figure 3. Precision and recall of 10 random files

6. Conclusion

In this paper, we present an approach to capture and exploit the novel features based on the Wikipedia's linking structure, which can enhance the disambiguation by exploring the explicit semantic knowledge in the Wikipedia knowledge base. Specially, we propose a semantic similarity function to measure the relatedness between the mention to be labeled and the candidate named entities. The experiment result shows that our system can achieve competitive performance over the traditional methods.

However, one problem of our system is that when dealing with the person mentions, it may not be able to find a matching entity. That is because we may not find a relevant webpage to calculate the similarity since there is not a wiki-url for the person item, especially when the mention is just a name for an ordinary person. In future work, we plan to mix up multiple knowledge sources as the knowledge base to gain more explicit information of the candidate entities, through which to strengthen the named entity disambiguation. We plan to not only capture the semantic relatedness from Wikipedia, but also use the Wordnet to achieve the linguistics relevance among the common words and obtain the social relation of the named entities from the web pages.

Acknowledgements

This work is supported by the National Science Foundation of China (no. 71273010 and 61379037), and the National Science Foundation of Anhui Province (no. 1208085MG117).

References

[1] R. Bunescu and M. Pasca, "Using Encyclopedic Knowledge for Named Entity Disambiguation", Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, (**2006**) April 3-7; Trento, Italy.

[2] S. Cucerzan, "Large-Scale Named Entity Disambiguation Based on Wikipedia Data", Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, (**2007**) June 28-30; Prague, Czech Republic.

[3] T. Nguyen and H. cao, "Named entity disambiguation on an ontology enriched by Wikipedia", Proceedings of the IEEE International Conference On Research, Innovation, & Vision for the Future Information & Communications Technol, (**2008**) July 7-11; Hochiminh City, Vietnam.
[4] X. Han and J. Zhao, "Named entity disambiguation by leveraging wikipedia semantic knowledge", Proceedings of the 18th ACM conference on Information and knowledge management, (**2009**) November 2-6; Hong Kong, China.
[5] T. Nguyen and H. cao, "Exploring Wikipedia and text features for named entity disambiguation", Proceedings of the Second international conference on Intelligent information and database systems, (**2010**) March 24-26; Hue City, Vietnam.
[6] A. L. Gentile, Z. Zhang, L. Xia and J. Iria, "Semantic relatedness approach for named entity disambiguation", Proceedings of the 6th Italian Research Conference, (**2010**) January 28-29; Padua, Italy.
[7] X. Han and J. Zhao, "Structural semantic relatedness: a knowledge-based method to named entity disambiguation", Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, (**2010**) July 11-16; Uppsala, Sweden.
[8] P. Danuta, "Exploring Entity Relations for Named Entity Disambiguation", Proceedings of The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies Student Session, (**2011**) June 19-24; Portland, Oregon, USA.
[9] B. Hachaey, W. Radford and J. R. Curran, "Graph-based named entity linking with Wikipedia", Proceedings of the 12th international conference on Web information system engineering, (**2011**) October 12-14; Sydney, Australia.
[10] X. Han, L. Sun and J. Zhao, "Collective Entity Linking in Web Text: A Graph-Based Method", Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, (**2011**) July 24-28; Beijing, China.
[11] X. Han and L. Sun, "A Generative Entity-Mention Model for Linking Entities with Knowledge Base", Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, (**2011**) June 19-24; Portland, Oregon, USA.
[12] I. Lasek and P. Vojtas, "Context Aware Named Entity Disambiguation", Proceedings of the 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology, (**2012**) December 4-7; Macau, China.
[13] A. Alhelbawy and R. Gaizauskas, "Named Entity Based Document Similarity with SVM-Based Re-ranking for Entity Linking", Proceedings of the first International Conference on Advanced Machine Learning Technologies and Applications, (**2012**) December 8-10; Cairo, Egypt.
[14] M. Jačala and J. Tvarožek, "Named entity disambiguation based on explicit semantics", Proceedings of the 38th international conference on Current Trends in Theory and Practice of Computer Science, (**2012**) January 21–27; Špindlerův Mlýn, Czech Republic.
[15] G. Rizzo and R. Troncy, "NERD: A Framework for Unifying Named Entity Recognition and Disambiguation Extraction Tools", Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, (**2012**) April 23-27; Avignon, France.

