Attribute based Chinese Named Entity Recognition and Disambiguation

Wei Han, Guang Liu, Yuzhao Mao, Zhenni Huang
School of Computer,
Beijing University of Posts and Telecommunications,
Beijing, 100876 China
{hanw,liug,maoyz}@bupt.edu.cn, liangsi07@gmail.com

Abstract

In this paper, we briefly report our system for Chinese Named Entity Recognition and Disambiguation task in CIPS-SIGHAN joint conference. We first present a method to extract different types of target person attributes from text documents with multiple techniques. Then we use these attributes to disambiguate different entities. Finally a classifier is used to distinguish entities in the knowledge base, and a cluster to recognize entities out of the knowledge base.

1 Introduction

Named Entities are meaningful units in texts. The ability to identify the named entities (such as people and locations) especially person name has long been an important task in natural language processing and text mining. And it is of great significance in the field of Web information extraction, machine translation, information retrieval, etc.

Generally speaking, a particular occurrence of a name string is insufficient to uniquely identify the corresponding entity. This is due to the fact that, in natural language, the same name string can refer to more than one entity. For example "George Bush" can refer to the former president of United States, or the real estate developer. In web search, 15-21% of the queries contain person names (11-17% of the queries are composed of a person name in web search, with additional terms and 4% are identified simply as person names). So it will be greatly improved to identify the entity that corresponds to a particular occurrence of a name string in the text document for many applications.

And it is especially important and challenging in Chinese. As there are less morphology variations than many other languages, it is challenging to distinguish common words from named entities in Chinese such as 高明 (brilliant), a common adjective and also a common person name. In addition, different types of named entities can use the same names and many persons may share the same name. For this reason, SIGHAN 2012 proposed the task, Named Entity Recognition and Disambiguation in Chinese.

Similar tasks have been explored previously. The KBP task and WePS task are public evaluation campaigns for entity disambiguation, providing annotated datasets for training and testing. During these tasks, it was noticed that attributes (such as birthday, occupation, affiliation, nationality, birth place, relatives, etc.) are very important clues for disambiguation. In fact, every person has his own attributes, and we believe that it is the right direction to study such problem. So in this work, we introduce an entity disambiguation system based on attribute extraction for the Named Entity Recognition and Disambiguation in Chinese task.

The overview of our system is as follows. We split this task into five parts: preprocessing, attribute extraction, similarity measures and document clustering, document classification and remained document clustering.

The remainder of this paper is organized as follows. Section 2 explains our task and describes related work, respectively. Section 3 explains our framework. Section 4 evaluates our framework with a dataset. Section 5 summarizes our work.

2 Named Entity Recognition and Disambiguation Task

2.1 Task definition

The formal definition is described in a web page, available at the following URL.
http://www.cipsc.org.cn/clp2012/task2.html

In the Named Entity Recognition and Disambiguation Task, given a query that consists of a name string-which can be a person (PER), organization (ORG), location (LOC) or just common words- and a background knowledge base, the system is required to provide the ID of the KB entry to which the name refers; or OTHER if it is not an entity, or OUT if there is no such KB entry. In addition, the system is required to cluster together documents referring to the same entity not present in the KB and provide a unique ID for each cluster.

For example, the knowledge base is as follows:

```
<?xml version="1.0" encoding="UTF-8" ?>
- <EntityList name="雷雨">
- <Entity id="01">
<text>重庆市黔江区太极乡党委副书记、乡长。主持政府全
面工作，主管财政、金融、审计、统计、非公有制经济、城乡
统筹、乡镇企业、招商引资、烤烟、蚕桑工作。</text>
</Entity>
<Entity id="02">
<text>四川省蒲江县教育局党组书记、局长。主持县教育局
全面工作。主管教育督导、计财、基建和教仪电教等工作。
</text>
</Entity>
- <Entity id="03">
<text>女，1975 年 8 月生，回族，广西南宁人，中共党员，
1997 年 7 月广西师范大学汉语言专业毕业，2006 年获教育
硕士学位，中学中级教师，1997 年 7 月进入桂林中学任教语
文至今。</text>
</Entity>
</EntityList>
```

Given a set of documents containing the targeted name, we should give the corresponding results. For example, the document about the middle school teacher should be linked to the KB entry 03; and the document which has no corresponding KB entries should be clustered into a cluster with a unique ID such as "Out_01"; the document that describe the weather such as "雷雨天气" should be marked as "Other".

2.2 Related Work

Personal name ambiguity is so common in the web that most previous disambiguation systems choose to work on personal name disambiguation. The related task has been addressed by several researchers starting from Bagga and Baldwin in 1998. They first selected tokens from local context as features to tackle the problem of cross-document co-reference by comparing, for any pair of entities in two documents, the word vectors built from all the sentences containing mentions of the targeted entities. Niu et al. (2004) extended Bagga's method by presenting an algorithm that uses information extraction results in addition to co-occurring words. Mann and Yarowsky (2003) proposed a bottom-up agglomerative clustering algorithm based on extracting local biographical information as features.

Bekkerman and McCallum (2005) focused on social network to find the documents that refer to a particular person using two methods: one based on the link structure and the other used agglomerative/conglomerate double clustering. But their scenario focuses on simultaneously disambiguating an existing social network of people who are known to be connected. Bunescu et al. (2006) used the category information from Wikipedia to disambiguate names. However, due to the limitation of the coverage of the Wikipedia entries of people, this method cannot be applied to resolve the people who are not famous enough to be included in Wikipedia.

Ying Chen et al. (2009) used a Web 1T 5-gram corpus released by Google to extract additional features for clustering. Masaki Ikeda et al. (2009) proposed a two-stage clustering algorithm. In the first stage, reliable features such as named entities are used to find documents that refer to the same person. Then some new features are extracted from the clustered documents and bootstrapping algorithm is used in the second stage.

3 Methodology

In this section, we present our proposed named entity disambiguation approach, which consists of five main steps. The overview of our approach will be provided first, followed by detailed steps.

1. First, the given documents are processed to decide if the name string in the document is an entity.
2. Then, both the documents and the texts in KB entries are converted into an attribute vector based on the attributes extracted from the text.
3. After that, the similarity score between KB entries and documents containing the same name string is calculated through their attribute vectors as well as the similarity score between each document. And the based on these score, some of documents are clustered.
4. Then, a classifier is trained to classify the remained documents.
5. Finally, remained documents referred to the same entity are clustered.

3.1 Preprocessing

As not all the documents containing the name string are about an entity, they may just an adjective, an adverb or something else. And through the dataset, we found most of the documents that contain the targeted string but not an entity are collocation commonly used in the Web data. For example, to the name string 高明，it is often used as an adjective such as “手段高明”. These documents need to be filtered out. So we first use word segmentation and part-of-speech tagging tools to process the given dataset. We use a Web 1T 5-gram corpus released by Google to calculate the most frequent word collocations containing the targeted name. For each document, if the name string is used in the collocation we got, it is very likely to refer to a non-entity. Using these word collocations as well as part-of-speech results and some simple but efficient rules, we are able to mark those documents as other. And these documents will not be processed in the following steps.

3.2 Attribute Extraction

In order to extract the attributes, the first challenge is to define what “the attributes of people” are. These have to be general enough to cover most people, meaningful and useful for disambiguation. We first looked at the attributes used in the WePS task and then took an empirical approach to define them; we extracted possible attributes from the training set and web pages and created a set of attributes which are frequent and important enough for the evaluation. We looked at the documents from the SIGHAN corpus, and found many kinds of attributes very useful and meaningful. Finally we made up 19 attribute classes, as shown in Table 1.

Attribute Class	Examples of Attribute Value
外文名	Christina
别名	小丽
性别	男
机构	黄海医院
出生日期	1987 年 3 月
血型	A 型
星座	狮子座
身高	190cm
出生地	北京市海淀区
民族	苗族
作品	大秦帝国
国籍	美国
政治面貌	党员
关系	张三
学校	北京邮电大学
公司名	某某集团公司
现居地	北京
学历	硕士研究生
职业	记者

Table 1: Definition of 19 attributes of Person

We extract attribute candidates by using processing pipelines with multiple techniques including traditional NER, regular expression patterns, gazetteer-based matching, and manually constructed rules and so on.

First, we extract the attributes based on bootstrapping method which is a machine learning method that automatically gather information. With some seed words and patterns, we can get a lot of attribute extraction template. The implement procedure is as follows:

1. get attribute value from new pattern;
2. calculate the score of attribute value;
3. put the top 5 attribute values into the attribute value dictionary;
4. get the context of the new attribute value and make it a candidate template;
5. calculate the score of pattern;
6. Put the top 3 patterns into the pattern dictionary.

We use some texts from web pages as the training set and repeat 10 times to get patterns.

The score of value and pattern is calculated as follows:

$$\mathrm{score}(\mathrm{value_i}) = \frac{\mathrm{R\_pattern}}{\mathrm{ALL\_pattern}} * \log_2(\mathrm{R_{pattern}})(1)$$

$$\mathrm{score}(\mathrm{pattern_i}) = \frac{\mathrm{R\_value}}{\mathrm{ALL\_value}} * \log_2(\mathrm{R\_value})(2)$$

Then we use some dictionaries to match some attributes such as job. And use NER tools to get the attributes like relatives. Finally, we use hownet to extend some synonyms.

As these methods we used are no good enough that some documents we can’t extract the attributes or they may not contain any attributes we defined at all, so we can’t only use the attributes to finish the task. So we first use the attributes to get some results in step 3. Then remained documents are processed in step 4 and 5 with some other techniques to finish the task.

3.3 Similarity measures and document clustering

We can see that different attributes have different influence on the disambiguation task. For example, the job and date of birth attribute are obvious more important and useful than the nationality attribute.

To assign weights to the attributes those indicate their contribution in resolving the person name's identity, we utilized information gain method. It is an algorithm that measures the discrimination performance. Information gain value of an attribute can be expressed as the desired reduction in the entropy of the attribute partition data sets caused.

The information gain formula is as follows:

$$\mathrm{Gain}(S,A) \equiv \mathrm{Entropy}(S) - \sum_{v \in \mathrm{Value}(A)} \frac{|S_v|}{|S|} \mathrm{Entropy}\ (S_v) \qquad (3)$$

Attribute Class	Weights of attribute
外文名	0.323
别名	0.677
性别	0.842
机构	0.922
出生日期	0.988
血型	0.226
星座	0.420
身高	0.644
出生地	0.990
民族	0.659
作品	0.655
国籍	0.385
政治面貌	0.792
关系	0.512
学校	0.950
公司名	0.994
现居地	1
学历	0.821
职业	0.908

Table2: The weights of 19 attributes of Person

The similarity is calculated based on these weights. If the value on a certain attribute is the same, then the weight of that attribute is added to a score called right score. If it's not same, then the weight of that attribute is added to a score called wrong score.

We first calculate the similarity between each document and the corresponding KB entries. Then we clustered the documents based on these similarities. If the right score and wrong score is in the threshold, we link it to the corresponding KB entry. In order to ensure the correctness of these results, we manually annotate some of the documents which are very ambiguous according to the similarity score.

3.4 Classification

After the previous steps, we've already got some documents linked to their corresponding KB entry or some clustered with a unique ID that is out if the KB entry. For the remained documents, it's hard to get the result only through their attributes. So we trained a classifier using the results from previous steps as training set.

We use SVM tools to train the classifier and tf-idf as the feature. If the score is beyond the threshold we set, we would link it to the corresponding KB entry. Otherwise, the documents would be considered as out of the KB entry and be processed in the following step.

3.5 Clustering

The remained documents are all regarded as out of the KB entry. All features are represented in vector space model. Every document is modeled as a vertex in the vector space. So every document can be seen as a feature vector. Before clustering, the similarity between documents is computed by cosine value of the angle between feature vectors. We cluster these documents into a cluster with a unique ID. Till now, all the documents have their own labels.

4 Evaluation

The dataset for Chinese Named Entity Recognition and Disambiguation task contains training data and testing data. The training data contains 16 names. Every name folder contains 50-300 articles. The testing data contains 32 names. The thresholds we used are obtained from the training data.

The evaluation method is based on precision, recall and F-measure. The overall precision and recall for all test names are calculated as follows (the set of all the test names are notated as N, each name is represented as n in N)

$$\mathrm{Pre} = \frac{\sum_n \mathrm{Pre}(n)}{|N|}$$

$$\mathrm{Rec} = \frac{\sum_t \mathrm{Rec}(t)}{|N|}$$

$$\mathrm{F} = \frac{2 * \mathrm{Pre} * \mathrm{Rec}}{\mathrm{Pre} + \mathrm{Rec}}$$

Precision	Recall	F-measure
67.18	85.62	75.29

Table 3: Official Results

The official results show that our method performs not very well, the precision score is a little low. That is because the method we used relies on the performance of the third step which has impact on the following results.

5 Conclusion

In this paper, we report our named entity recognition and disambiguation system and a framework which integrates AE approaches.

In the future, we will attempt to use better methods to improve the performance of the attribute extraction. And consider how to combine the disambiguation part and the AE part to complement each other.

References

Javier Artiles, Julio Gonzalo, and Satoshi Sekine. The semeval-2007 weps evaluation: Establishing a benchmark for the web people search task. In the Fourth International Workshop on Semantic Evaluations (SemEval-2007). ACL, June 2007.

Bagga, Amit. & Baldwin, Breck. (1998). Entity-based cross-document co-referencing using the vector space model. In Proceedings of the 17th international conference on computational linguistics.

C. Niu, W. Li, and R. K. Srihari. 2004. Weakly Supervised Learning for Cross-document Person Name Disambiguation Supported by Information Extraction. In Proceedings of ACL 2004.

Gideon S. Mann and David Yarowsky. Unsupervised personal name disambiguation. In HLT-NAACL, pages 33–40, May 2003.

Ron Bekkerman and Andrew McCallum. Disambiguating web appearances of people in a social network. In WWW, pages 463–470, May 2005.

Lan, M., Zhang, Y.Z., Lu, Y., Su, J. & Tan. C.L. (2009). Which who are they? People attribute extraction and disambiguation in web search results. In 18th WWW Conference 2nd Web People Search Evaluation Workshop (WePS 2009).

Minkov, E., Wang, R. & Cohen, W. (2005). Extracting personal names from emails: applying named entity recognition to informal text. In Proceedings of HLT/EMNLP .

Rao, Delip., Garera, Nikesh & Yarowsky, David (2007). JHU1: An unsupervised approach to person name disambiguation using web snippets. In Proceedings of semeval 2007, association for computational linguistics .

Watanabe, K., Bollegala, D., Matsuo, Y. & Ishizuka, M. (2009). A two-step approach to extracting attributes for people on the web in web search results. In 18th www conference 2nd web people search evaluation workshop (WePS 2009),.

Xianpei Han and Jun Zhao. 2009. CASIANED:Web Personal Name Disambiguation Based on Professional Categorization. In 2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference.

Xianpei Han and Le Sun. 2011. A Generative Entity-Mention Model for Linking Entities with Knowledge Base. Proc. ACL2011.

Heng Ji and Ralph Grishman. 2011. Knowledge Base Population: Successful Approaches and Challenges. Proc. ACL2011.

Heng Ji, Ralph Grishman, Hoa Trang Dang, Kira Griffitt and Joe Ellis. 2010. An Overview of the TAC2010 Knowledge Base Population Track. Proc. Text Analytics Conference (TAC2010).

Minoru Yoshida, Masaki Ikeda, Shingo Ono, Issei Sato, and Hiroshi Nakagawa. Person name disambiguation on the web by two-stage clustering. In WWW, April 2009.

Minoru Yoshida, Masaki Ikeda, Shingo Ono, Issei Sato, and Hiroshi Nakagawa. Person name disambiguation by bootstrapping. In SIGIR, July 2010.

Bunescu, R., & Pas, M. (n.d.). Using Encyclopedic Knowledge for Named Entity Disambiguation, In EACL, pages 17–24, April 2006.