A Joint Chinese Named Entity Recognition and Disambiguation System

Longyue Wang
Natural Language Processing & Portuguese-Chinese Machine Translation Laboratory, Department of Computer and Information Science, University of Macau, Macau S.A.R., China.
vincentwang0229@hotmail.com

Shuo Li
Natural Language Processing & Portuguese-Chinese Machine Translation Laboratory, Department of Computer and Information Science, University of Macau, Macau S.A.R., China.
leevis1987@gmail.com

Derek F. Wong
Natural Language Processing & Portuguese-Chinese Machine Translation Laboratory, Department of Computer and Information Science, University of Macau, Macau S.A.R., China.
derekfw@umac.mo

Lidia S. Chao
Natural Language Processing & Portuguese-Chinese Machine Translation Laboratory, Department of Computer and Information Science, University of Macau, Macau S.A.R., China.
lidiasc@umac.mo

Abstract

In this paper we describe an integrated approach for named entity recognition and disambiguation in Chinese. The proposed method relies on named entity recognition (NER), entity linking and document clustering models. Different from other tasks of named entities, both classification and clustering are considered in our models. After segmentation, information extraction and indexing in the preprocessing step, the test names in the documents would be judged to be common words or named entities based on hidden Markov model (HMM). And then each predicted entity should be linked to the category in the given knowledge base (KB) according to the character attributes and keywords. Finally, the named entities which have no reference in KB would be clustered into a new category based on singular value decomposition (SVD). An implementation of our presented models is described, along with experiments and evaluation results on the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing Bakeoff (Bakeoff-2012). Named entity recognition F-measure reaches up to 76.67% and named entity disambiguation F-measure up to 69.47% within the test set of 32 names.

1 Introduction

The ability to identify the named entities has been established as an important task in several areas, including topic detection and tracking, machine translation, and information retrieval (Cucerzan, 2007). NER is the first step that seeks to locate and classify atomic elements in text into predefined categories such as the names of persons, organizations, locations, etc.. Another big issue in this area is based on a factor that millions of names (especially for person names) and references appear on the Internet, which raises the problem of co-reference resolution, also called name disambiguation (Wu, 2010). Therefore, named entity recognition and disambiguation are both important in Natural Language Processing (NLP), especially in Chinese language.

Unlike Roman alphabetic languages such as English, Portuguese, etc., Chinese named entity recognition and disambiguation are more difficult due to the unavailability of morphology variations, explicit word delimiters etc.. For example, given a word "温馨 (*warm*)", it is hard to determine whether it is a common adjective or a person name. Besides, different types of named entities can use the same names. For instance, "金山 (*Gold Hill*)" can be used as the name of persons, locations and organizations. Finally, it is typical that many persons share the same name.

Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing, pages 146–151,
Tianjin, China, 20-21 DEC. 2012

It is reported that, nearly 300,000 persons have the same name of "张伟 (*Zhang Wei*)" in China.

To further investigate these issues, SIGHAN 2012 establishes a more difficult task, which can be seen as combination of related tasks in KBP (Knowledge Base Population) and WPS (Web People Search). It is divided into three parts and described as follows:

- **Named Entity Recognition**. The test names in the document should be judged to be common words or named entities.
- **Entity Linking**. Each predicted named entity should be further determined which named entity in the KB it refer to.
- **Unlinked Name Clustering**. Some predicted named entities that do not have references in the KB, should be clustered into new categories.

For these three sub-tasks, we presented a PRLC approach, which integrates with named entity **P**re-processing, **R**ecognition, **L**inking and **C**lustering modules. Word segmentation, keywords generation and character attributes extraction are ential for all the documents both in test name set and KB. And then given a test name document, recognition module will determine whether it is a name of person, place, organization or non-entity. Besides, the linking module adopts the technology of information retrieval (IR) to find the category in the indexed KB. Finally, all the unlinked documents would be classified by the named entities they refer to. Different from the traditional methods, we divided our model into four independent parts but all work together to deal with named entity recognition, linking and clustering. The word segmentation and indexing were well conducted in the preprocessing step. And both keywords and character attributes were extracted as quires. In addition, the problem is transformed from named linking to similarity calculation, where conventional IR techniques can be used. So the similarity between each document in KB and a certain query on a test name document can be evaluated to obtain best reference. Finally, an SVD-based method was adopted to group the unlinked entities by the named entities they refer to.

The paper is organized as follows. The related works are reviewed and discussed in Section 2. The proposed PRLC approach based on four models is described in Section 3 and 4. Results, discussion and comparison between different strategies are given in Section 5 followed by a conclusion and future improvements to end the paper.

2 Related Work

The issues of named entity recognition and disambiguation have been discussed from different perspectives for several decades. In this section, we briefly describe some related methods.

NER has been widely addressed by symbolic, statistical as well as hybrid approaches. Its major part in information extraction (IE) and other NLP applications has been stated and encouraged by several editions of evaluation campaigns such as MUC (Marsh and Perzanowski, 1998), the CoNLL-2003 NER shared task (Tjong Kim Sang and De Meulder, 2003) or ACE (Doddington et al., 2004), where NER systems show near-human performances for the English language. However, Chinese NER is far from mature (Wu, 2005). Recent years, a lot statistic-based methods including hidden Markov models (HMMs) (Zhou, 2002; Fu, 2005) have been applied. Comparing with rule-based NER, statistic-based methods utilize the human labeled corpus as the training set, and it doesn't require the extensive knowledge of linguistics when labeling the corpus. Carpenter (2006) presented the character language models with a good accuracy of 97.57% (precision 81.88, recall 80.97 and F-measure 81.42) in the closed track of the 3rd SIGHAN bakeoff. The results show that HMMs can perform well both in accuracy and speed.

With the development of NER, there have been some researches on combining this component with entity linking (EL). Stern et al. (2012) introduced a system based on a joint application of NER and EL, where the NER output is given to the linking component as a set of possible mentions, preserving a number of ambiguous readings. Although the system achieved a high linking accuracy (87%), it is only evaluated in French language. Regarding the Chinese person name disambiguation, Xu et al. (2010) described a system incorporating person name recognition, identity and an agglomerative hierarchical clustering. And finally his proposed method achieves encouraging recall and good overall performance for the task in the CIPS-SIGHAN 2010, which is simpler than the one we tackled.

In order to extract useful information from the descriptive documents, a method named "bags of words" is widely used to find the keywords based on Term Frequency–Inverse Document Frequency (TF-IDF) or Term Frequency (TF). Additionally, the vector space model is usually used to represent the documents and calculated the similarities (Bollegala, 2006). Although the

keyword has more relationship to the document itself instead of the information of the person, the contents of the documents in this task are mainly the description of persons. Mann and Yarowsky (2003) proposed another approach which used character attributes to build a person model and achieved a good performance.

3 Pre-processing

Different from the other languages such as English, Portuguese etc., pre-processing like word segmentation is the foundation for Chinese named entity recognition and disambiguation. In order to reduce the search space during entity linking and clustering, both keywords and character attributes are also extracted to represent the documents. We mainly completed the works as follows.

3.1 Word Segmentation

Our task is thought to be more challenging due to the need for word segmentation which could bring errors into the subsequent processes.

After years of intensive researches, Chinese word segmentation has achieved a quite high performance (Huang, 2007). Among all of them, the ICTCLAS (developed by Chinese Academy of Sciences) is currently the best one both in accuracy and speed. This Chinese lexical analysis system combines part-of-speech (POS) tagging, word segmentation and unknown word recognition.

Therefore, ICTCLAS 2007[1] is involved to deal with word segmentation and POS tagging for the documents both in the knowledge base and in the test name set. In order to make all the names segmented correctly, all the test names are collected manually as the external dictionary, Furthermore, persons often have much to do with corresponding works, books etc.. So all these segmented titles should be re-combined for further extraction.

3.2 Character Attributes Extraction

After segmentation, character attributes are extracted by some simple matching rules. According to the character categories in WPS and the contents of the documents in this task, nine kinds of attributes such as gender, political status, educational background etc. were defined to describe a person. The detailed character attributes used in our system are shown in Table 1.

No.	Attributes	Description
1	Gender	Male, female or not mentioned
2	Date	Dates of the events
3	Nation	Like Miao, Han etc.
4	Political Status	Like party members etc.
5	Educational Background	The degrees such as master, PhD etc.
6	Occupation	Name of job or titles
7	Publications	Name of books, films etc.
8	Other Names	Names of other persons, locations and organizations
9	Foreign Words	English words like names of foreigners

Table 1: Character attributes used in our system

3.3 Keywords Generation

After selecting the attributes from the documents of the test name set and KB, the keywords will be selected from the common words (not attributes). Keywords can be supplemented for some documents, which are limited with character information. Therefore, a keywords generation model was designed according the POS, TF-IDF and positions.

Based on the classical algorithm of TF-IDF (Ramos, 2003), a weight is added to obtain the words, which have more relations to the test names. Given a document collection D (e.g. test name set or KB), a word w, and an individual document $d \in D$, we calculate

$$P(w,d) = \frac{\alpha}{Dis} \times f(w,d) \times \log \frac{|D|}{f(w,D)} \quad (1)$$

where $f(w,d)$ denotes the number of times w that appears in d, $|D|$ is the size of the corpus, and $f(w,D)$ indicates the number of documents in which w appears in D. Firstly, nouns and verbs have more ability of describing than other words. In implementation, α should be set as 1 for the nouns and verbs while others as 0. Besides, the words around the entities also have more relation to the person. Therefore, the *Dis* is used to cacu-late the distance between a certain word and the closest test name. Division of *Dis* means that the words with longer distance to the test name should be less important. Finally, all the common words will be ranked by the values of $P(w,d)$

[1] ICTCLAS can be download from http://www.ictclas.org/ictclas_download.aspx

and the best *N*th words will be selected as keywords.

3.4 Query and Indexing

For the document in KB, each character attribute is indexed in respective field and all the keywords are indexed in another filed together. For the test name set, both attributes and keywords of each test name document are combined as a query for retrieving the indexed KB.

4 Proposed Approach

In addition to the pre-processing, the approach relies on three models: recognition model which judges the test name whether name entity or not; linking model which determines which named entity in the KB the test name refer to; and clustering model, which groups the same unlinked entities according by the entities they refer to. The workflow of the approach for PRLC is shown in Fig. 1.

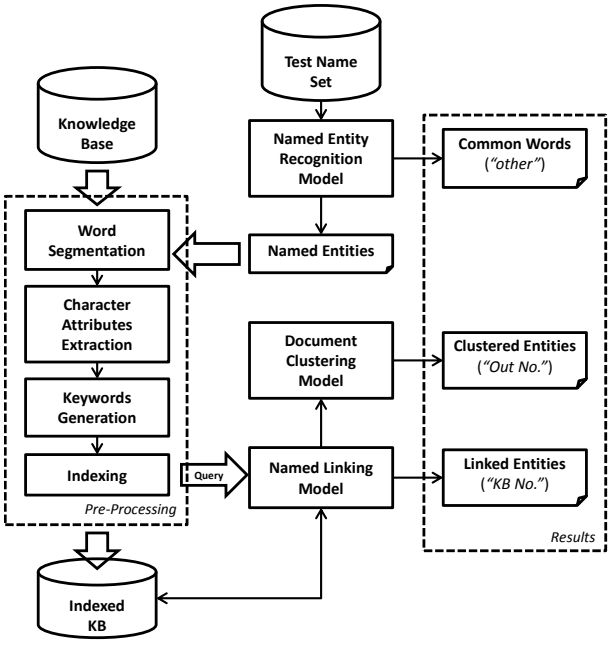

Figure 1. Approach for PRLC.

4.1 Named Entity Recognition

Proper noun of persons (PER), locations (LOC) and organizations (ORG) are included by name entities. Each sentence consists of a single character, a single space character and a tag with BIO coding scheme.

An open source NLP toolkit, LingPipe[2] was utilized to deal with the NER task, which depends on *n*-gram based character language model with the Witten-Bell smoothing (Witten et al., 1991). Regarding training phrase, the model provides a probability distribution of strings over a fixed Chinese character. The recognizer introduced an HMM interface with *n*-best decoder. The approach proposed by Carpenter (2006) was referred in the implementation of the model: the transition between tags is modeled by a maximum likelihood estimate over the training corpus. Therefore, a bounded character language model is trained to estimate the tags.

During decoding, a Chinese chunking implementation was introduced. The chunking utilizes a refinement of the standard "BIO" coding scheme (Culotta and McCallum, 2004), which means more tags were defined to label the Chinese character instead of the original tags. So the confidence estimation of Chinese characters was simplified and the probabilities will be normalized to model the joint probabilities of the Chinese character or tag (Carpenter, 2006). For example, the person's name can be generated with a tag in a person model which is built based on n-best chunker, in which each Chinese word is scored. Finally, a new output is returned with a best score by a re-scoring model.

In summary, the NER model is helpful to distinguish the name entity and none name entity. The performance of this model will be evaluated and shown in Section 5.

4.2 Entity Linking

After indexing the KB and generating the queries, the problem of entity linking is transformed into information retrieval. The core algorithm of the retrieval model is derived from the Vector Space Model (VSM). Our system takes this model to calculate the similarity between each indexed KB and the input query. The final scoring formula is given by:

$$Score(q,d) = coord(q,d) \sum_{t_{in}q} tf(t,d) \times idf(t) \times bst \times norm(t,d) \quad (2)$$

where $tf(t,d)$ is the term frequency factor for term t in document d, $idf(t)$ is the inverse document frequency of term t, while $coord(q,d)$ is frequency of all the terms in query occur in a document. bst is a weight for each term in the query. $Norm(t,d)$ encapsulates a few (indexing time)

[2] http://alias-i.com/lingpipe/web/download.html

boost and length factors, for instance, weights for each document and field. As a summary, many factors that could affect the overall score are taken into account in this model.

The model can return *N*-best candidates with the scores. In our system, only if the size of candidate set is more than 1 and the highest score is more than a threshold, the top candidate will be linked to the category in the KB. Otherwise, the test name will be treated as unlinked one.

4.3 Document Clustering

In the clustering model, a snippet-based clustering engine named Carrot2[3] was applied for the task. It can automatically organize small collections of documents (search results but not only) into thematic categories. Lingo is one of the algorithms in Carrot, which constructs a "term-document matrix" where each snippet gets a column, each word a row and the values are the frequency of that word in that snippet. It then applies a matrix factorization called singular value decomposition (SVD). All the documents of unlinked test names were group by the toolkit according to the queries.

5 Evaluation and Discussion

A number of experiments have been conducted to investigate our proposed method on different settings. In order to evaluate the performance of the recognition model, we tested it respectively with external corpus.

Measurement	Values	Average
R_{PER}	**0.8540**	0.7220
R_{LOC}	0.6823	
R_{ORG}	0.6123	
P_{PER}	**0.8868**	0.8173
P_{LOC}	0.8411	
P_{ORG}	0.6642	
F_{PER}	**0.8701**	**0.7667**
F_{LOC}	0.7534	
F_{ORG}	0.6372	

Table 2: The NER result

Two years of People's Daily (PD) corpus is used for training data, which are manual segmented and tagged with POS with high quality by Peking University. And then the test set of Microsoft Research in the 3rd SIGHAN Bakeoff was used to evaluate. The results of the person, location and organization are shown in Table 2. Although the total F-measure is only 0.7667, a large amount of test names are person name. With the high F-measure of 0.8701 in person, it fully illustrates the effectiveness of the NER model.

We also use a small test set within 6 test names, which is released by the Second CIPS-SIGHAN. The results in Table 3 show that the proposed method gives an average precision of 74.41%. However, the recall value is not ideal and the distribution is not balanced. It is unmoral that the lowest recall is 0.5925 while the highest is 0.9154. Through analyzing the data, the main reason is that the clustering model is not good enough to group the documents together based SVD. This leads to a not very high F-measure totally. The encouraging results in precision prove a good ability to distinguish categories in KB. Therefore, the technology of information retrieval using the character information or keywords is more useful to named disambiguation.

Personal Name	P	R	FB1
白雪 (*Bai Xue*)	0.8191	0.6684	0.7361
白云 (*Bai Xun*)	0.7796	0.6090	0.6839
丛林 (*Cong Lin)*	0.7024	0.7551	0.7278
杜鹃 (*Du Juan*)	0.8651	**0.5925**	0.7033
方正 (*Fang Zheng*)	0.6378	0.6051	**0.6210**
胡琴 (*Hu Qin*)	0.6604	0.9154	**0.7673**
Total	**0.7441**	0.6909	0.7165

Table 3: The result with a small test set

Finally, we evaluated our system, on the test set of 32 test names. Table 4 shows our official CIPS-SIGHAN bakeoff results. It shows the average precision, recall and FB1[4] of our system. The results show that we still can improve the clustering model to obtain a higher recall. On the whole, the presented PRLC approach is suitable to task of Chinese named entity recognition and disambiguation, but still should be improved in the future.

6 Conclusion

This article presents an integrated approach for the special task in Chinese personal name recognition and disambiguation. We divided our model into four independent parts but all work together and are easy to improve each model independently. In implementation, we combined the

[3] http://project.carrot2.org/download.html

[4] http://www.cipsc.org.cn/clp2012/task2.html

pre-processing, named entity recognition, named linking and document clustering modules into our system. Besides, the character attributes and TF-IDF keywords are both used to build person model for entity linking and clustering. Finally, we simplified the problem of named linking with the technology of information retrieval, which obtains a high precision in the task..

Precision	Recall	FB1
0.7885	0.6209	0.6947

Table 4: The official results

Acknowledgments
This work was partially supported by the Research Committee of University of Macau under grant UL019B/09-Y3/EEE/LYP01/FST, and also supported by Science and Technology Development Fund of Macau under grant 057/2009/A2.

References

Bollegala D., Matsuo Y., and Ishizuka M. 2006. Extracting key phrases to disambiguate personal name queries in web search. *Proceedings of the Workshop on How Can Computational Linguistics Improve Information Retrieval*. 17–24.

Carpenter B. 2006. Character language models for Chinese word segmentation and named entity recognition. *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*. 169–172.

Cucerzan S. 2007. Large-scale named entity disambiguation based on Wikipedia data. *Proceedings of EMNLP-CoNLL*. 6:708–716.

Culotta A. and McCallum A. 2004. Confidence estimation for information extraction. *Proceedings of HLT-NAACL 2004: Short Papers*. 109–112.

Doddington G., Mitchell A., Przybocki M., Ramshaw L., Strassel S., and Weischedel R. 2004. The automatic content extraction (ACE) program-tasks, data, and evaluation. *Proceedings of LREC*. 4:837–840.

Fu G. and Luke K.K. 2005. Chinese named entity recognition using lexicalized HMMs. *ACM SIGKDD Explorations Newsletter*. 7:19–25.

Huang C. and Zhao H. 2007. Chinese word segmentation: A decade review. *Journal of Chinese Information Processing*. 21:8–20.

Mann G.S. and Yarowsky D. 2003. Unsupervised personal name disambiguation. *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL* 2003-Volume 4. 33–40.

Marsh E. and Perzanowski D. 1998. MUC-7 evaluation of IE technology: Overview of results. *Proceedings of the Seventh Message Understanding Conference* (*MUC-7*). 20.

Ramos J. 2003. Using tf-idf to determine word relevance in document queries. *Proceedings of the First Instructional Conference on Machine Learning*.

Tjong Kim Sang E.F. and Meulder F. De. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL* 2003-Volume 4. 142–147.

Witten I.H. and Bell T.C. 1991. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *Information Theory, IEEE Transactions On*. 37:1085–1094.

Wu C., Gong L., and Zeng J. 2010. Multi-document Chinese name disambiguation based on Latent Semantic Analysis. *Fuzzy Systems and Knowledge Discovery (FSKD), 2010 Seventh International Conference On*. 5:2367–2371.

Wu Y., Zhao J., Xu B., and Yu H. 2005. Chinese named entity recognition based on multiple features. *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. 427–434.

Zhou G.D. and Su J. 2002. Named entity recognition using an HMM-based chunk tagger. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. 473–480.