Chinese Named Entity Recognition Combining a Statistical Model with Human Knowledge

Youzheng WU **Jun ZHAO** **Bo XU**
National Laboratory of Pattern Recognition
Institute of Automation Chinese Academy of Sciences
No.95 Zhongguancun East Road, 100080, Beijing, China
(yzwu, jzhao,boxu)@nlpr.ia.ac.cn

Abstract

Named Entity Recognition is one of the key techniques in the fields of natural language processing, information retrieval, question answering and so on. Unfortunately, Chinese Named Entity Recognition (NER) is more difficult for the lack of capitalization information and the uncertainty in word segmentation. In this paper, we present a hybrid algorithm which can combine a class-based statistical model with various types of human knowledge very well. In order to avoid data sparseness problem, we employ a back-off model and《同义词词林/TONG YI CI CI LIN》, a Chinese thesaurus, to smooth the parameters in the model. The F-measure of person names, location names, and organization names on the newswire test data for the 1999 IEER evaluation in Mandarin is 86.84%, 84.40% and 76.22% respectively.

1 Introduction

The NER task was first introduced as Message Understanding Conference (MUC) subtask in 1995 (MUC-6). Named Entities were defined as entity names (organizations, persons and locations), temporal expressions (dates and times) and number expressions (monetary values and percentages). Compared with the entity name recognition, the recognition of temporal and number expressions is simpler. So, our research focuses on the recognition of person, location and organization names.

The Multilingual NE task first started in 1995(MET-1), including Chinese, Japanese, and Spanish in that year, and continued for Chinese, Japanese in 1998(MET-2). Compared with English NER, Chinese NER is more difficult. We think the main differences between Chinese NER and English NER lie in:

First, unlike English, Chinese lacks the capitalization information that plays an important role in signaling named entities.

Second, there is no space between words in Chinese, and we have to segment the text before NER. However, the errors in word segmentation will affect the result of NER.

Third, Different types of named entities have different structures, especially for abbreviative entities. Therefore, a single unified model can't capture all the types of entities. Typical structures of Chinese person name (CN), location name (LN) and organization name (ON) are as follows:

CN--><surname> <given name>

LN--><name part>* <a salient word>

ON-->{[person name] [organization name] [place name] [kernel name] }* [organization type] <a salient word>

Here <>* means repeating one or several times. {}* means selecting at least one of items.

Fourth, there are few openly available resources for Chinese NER. Thus we have to resort to the algorithm that doesn't rely on large NER-tagged text corpus.

Based on the above analysis, we present a hybrid algorithm that incorporating various types

of human knowledge into a statistical model. The innovative points of our paper are as follows.

First, the hybrid algorithm can make the best use of existing limited resources to develop an effective NER system. These resources include one-month's Chinese People's Daily tagged with NER tags by Peking University (which contains about two-million Chinese characters) and various types of human knowledge.

Second, in order to compensate for the lack of labeled corpus, we use several types of human knowledge, such as《同义词词林/TONG YI CI CI LIN》 [Mei.J.J, et al. 1983], a general location names list, the list of the salient words in location name, the list of the salient words in organization names, a Chinese surnames list, the list of Chinese characters that could be included in transliterated person names, and so on.

Third, we emphasize that human knowledge and statistical information should be combined very well. For example, a general LN list and a general famous ON list are used in our system. However, we only accept words in the lists as entity **candidates** with a probability. Whether it is a LN or ON depends on the context. This is different from other systems which accept them as a LN or ON once the system meets them. More details refer to section 4.

This paper will be organized as follows. Section 2 is the background of NER. Section 3 describes the class-based statistical baseline Chinese NER model. Section 4 describes different types of human knowledge for different named entities recognitions and how to combine them with a statistical model organically in details. Section 5 is the evaluation and section 6 is the conclusion.

2 Backgroud

The researches on English NER have made impressive achievement. The best NER system [Mikheev, et al. 1999] in MUC7 achieved 95% precision and 92% recall. Recent methods for English NER focus on machine-learning algorithms such as DL-CoTrain, CoBoost [Collins and Singer 1999], HMM [Daniel M. Bikel 1997], maximum entropy model [Borthwick, et al, 1999] and so on.

However, Chinese NER is still at its immature phase. Typical Chinese NER systems are as follows.

NTU system [Hsin-His Chen, et al. 1997] relied on a statistical model when recognizing person names, but rules when recognizing location and organization names. In the formal run of MET-2, the total F-measure is 79.61%. As a result, they may miss the person names whose probability is lower than the threshold, the location and organization names may also be missed for those which don't accord with the rules.

[Yu et al. 1998] uses both a contextual model and a morphological model. However, their system requires information of POS tags, semantic tags and NE lists. The system obtains 86.38% F-measure.

[CHUA et al. 2000] employs a combination of template-based rules supplemented by the default-exception trees and decision tree that obtains over 91% F-measure on MET-2 test data. It also uses HowNet [Dong & Dong 2000] to cluster semantically related words.

[Jian Sun, 2002] presents a class-based language model for Chinese NER which achieves 81.79% F-measure on MET-2 test set and 78.75% F-measure on IEER test data. However, the model heavily depends on statistical information, and must be trained on large labeled corpus.

For Chinese NER, we can't achieve satisfactory performance if we use only a statistical model or handcrafted heuristic rules. Therefore, we have to resort to the algorithm that can incorporate human knowledge into a statistical model.

In the following sections, we will introduce a statistical Chinese NER model first, and then incorporate various types of human knowledge into the statistical model in order to show the power of human knowledge for Chinese NER.

3 The Baseline Class-based Statistical Model

We regard NER as a tagging problem. Given a sequence of Chinese string $W = w_1 w_2 \cdots w_n$, the task of NER is to find the most likely sequence of class sequence $C^* = c_1 c_2 \cdots c_m (m <= n)$ that maximizes the probability $P(C|W)$. We use Bayes' Rule to rewrite $P(C|W)$ as equation (3.1):

$$P(C|W) = \frac{P(C,W)}{P(W)} = \frac{P(W|C) \times P(C)}{P(W)} \qquad (3.1)$$

So, the class-based baseline model can be expressed as equation (3.2).

$$C^* = \arg\max_C (P(W \mid C) \times P(C))$$
$$= \arg\max_C (P(w_1 w_2 \cdots w_n \mid c_1 c_2 \cdots c_m) \times P(c_1 c_2 \cdots c_m)) \quad (3.2)$$
$$\approx \arg\max_C \left(\prod_{i=1}^{m} P(w_{i1} \cdots w_{ij} \mid c_i) \times P(c_i \mid c_{i-1}) \right)$$

We call $P(C)$ as the contextual model and $P(W \mid C)$ as the morphological model. Formally, we can regard such a class-based statistical model as HMM. The classes used in our model are shown in Table 1, where $|V|$ means the size of vocabulary used for word segmentation.

Class	Description
PN	Person Name
LN	Location Name
ON	Organization Name
TM	Time Name
NM	Number Name
Other	One word is on Class
Total	$\|V\| + 5$

Table 1 Classes used in our model

3.1 Contextual Model

Due to our small-sized labeled corpus, we use a statistical bi-gram language model as the contextual model. This model can be described as equation (3.3).

$$P(C) \cong \prod_{i=1}^{i=m} P(c_i \mid c_{i-1}) \quad (3.3)$$

Theoretically, trigram is more powerful for NER than bi-gram, however when training corpus is small, trigram can't work effectively. Using bi-gram model, we still need $(|V|+5)^2$ transmission probabilities, some of which can't be observed in our small-sized labeled corpus and some of which are unauthentic. That is, data sparseness is still serious. We will explain how to resolve data sparseness problem in details in section 3 and 4.

3.2 Morphological Model

Recognition of Person Names

The model of person names recognition (including Chinese person names abbreviated to CN and Transliterated person names abbreviated to TN) is a character-based tri-states unigram model.

In principle, Chinese person name is composed of a surname (including single-character surname like "吴/wu" and double-character surname like"欧阳/Ouyang") and a given name (one or two characters like "鹏/peng" or "友政/youzheng"). So we divide Chinese name words into three parts as the surname (surCN), the middle name (midCN) and the end name (endCN), which means the probability of a specific character used in different position in person names isn't equal. For example,

$$P(\text{吴/wu} \mid c_j = \text{surCN}) \neq P(\text{吴/wu} \mid c_j = \text{secCN}) \neq P(\text{吴/wu} \mid c_j = \text{endCN}) \quad (3.4)$$

The model for three-character-CN recognition is described as equation (3.5).

$$P(w_{j1} w_{j2} w_{j3} \mid c_j = CN) \cong P(w_{j1} \mid c_j = surCN) \times P(w_{j2} \mid c_j = midCN) \times P(w_{j3} \mid c_j = endCN) \quad (3.5)$$

The model for two-character-CN recognition is described as equation (3.6).

$$P(w_{j1} w_{j2} \mid c_j = CN) \cong P(w_{j1} \mid surCN) \times P(w_{j2} \mid endCN) \quad (3.6)$$

where $P(w_{j1} w_{j2} w_{j3} \mid c_j = CN)$ means the probability of emitting the candidate person name $w_{j1} w_{j2} w_{j3}$ under the state of CN.

For TN, we don't divide transliterated name words into several different parts. That is, the probability of a word used in different position in TN is same. The model is as follows.

$$P(w_{j1} w_{j2} \cdots w_{jk} \mid c_j = TN) \cong \prod_{i=1}^{i=k} P(w_{ji} \mid c_j = TN) \quad (3.7)$$

Must be mentioned is that all these probabilities are estimated from labeled corpus using maximum likelihood estimation.

Recognition of Location Names

For location names recognition, we use a word-based bi-state unigram model, and divide words used in the location name into two parts: location-end-words (LE) and non-location-end words (NLE). That means the probability of the word used in the end position of location name is different from that of in other position.

The model for location name recognition is shown in equation (3.8).

$$P(w_{j1} w_{j2} \cdots w_{jk} \mid c_j = LN) \cong \prod_{i=1}^{i=k-1} P(w_{ji} \mid c_j = NLE) \times P(w_{jk} \mid c_j = LE) \quad (3.8)$$

The parameters in equation (3.8) are also estimated from labeled training corpus.

Recognition of Organization Names

For the model of organization names recognition, we use bi-state unigram that is similar to the location morphological model shown as equation (3.9):

$$P(w_{j1}w_{j2}\cdots w_{jk} \mid c_j = ON) = \prod_{i=1}^{i=k-1} P(w_{ji} \mid c_j = OE) \times P(w_{jk} \mid c_j = NOE) \quad (3.9)$$

where OE means the word used in the end position of organization name, while NOE is not.

The parameters in equation (3.9) are also estimated from the labeled training corpus.

Back-off Models to Smooth

Data sparseness problem still exists. As some parameters were never observed in trained corpus, the model will back off to a less-powerful model. We employ escape probability to smooth the statistical model [Teahan, et al. 1999].

An escape probability is the probability that a previously unseen character will occur. There is no theoretical basis for choosing the escape probability optimally. Here we estimate the escape probability in a particular context as:

$$\lambda = \frac{0.5d}{n} \quad (3.10)$$

The probability of a word c_i that has occurred c times in that context c_{i-1} is:

$$P(c_i \mid c_{i-1}) = \frac{c - 0.5}{n} \quad (3.11)$$

While the probability of a word that has never occurred in that context is:

$$P(c_i \mid c_{i-1}) = \lambda \times P(c_i) \quad (3.12)$$

where n is the number of times that context has appeared and d is the number of different symbols that have directly followed it.

As a example, if we observe the bi-gram "A B" once in training corpus and “A C" three times, and nowhere else did we see the word "A", then $P(C \mid A) = \frac{3-0.5}{1+3}$, while the escape probability $\lambda = \frac{0.5 \times 2}{1+3}$ and unseen transition probability of $P(D \mid A) = \lambda \times P(D)$.

The Evaluation for the Baseline

The baseline model was evaluated in terms of precision (P), recall (R) and F-measure (F) metrics.

$$P = \frac{number\ of\ correct\ responses}{number\ of\ responses}$$

$$R = \frac{number\ of\ correct\ responses}{number\ of\ all\ NE}$$

$$F = \frac{(\beta^2 + 1) \times P \times R}{(\beta \times P) + R} \quad (3.13)$$

where β is a weighted constant often set to 1.

We test the baseline system on the newswire test data for the 1999 IEER evaluation in Mandarin (http://www.nist.gov/speech/tests/ie-r/er_99/er_ 99.htm). Table 2 in section 4 summarizes the result of baseline model.

	Precision	Recall	F-measure
PN	80.23%	89.55%	84.63%
LN	45.05%	66.96%	53.86%
ON	42.98%	61.45%	50.58%
Total	52.61%	71.53%	60.63%

Table 2 The Performance of The Baseline

4 The Hybrid Model Incorporating Human Knowledge into the Baseline

From table 1, we find that the performance of the above statistical baseline model isn’t satisfactory. The problems mainly lie in:

① Data sparseness is still serious though we only use bi-gram contextual model, unigram morphological model and smooth the parameters with a back-off model.

② In order to recognize the named entities, we have to estimate the probability of every word in text as named entities. Thus redundant candidates not only enlarge search space but also result in many unpredictable errors.

③ Abbreviative named entities especially organization abbreviation can’t be resolved by the baseline model. Because abbreviations have weak statistical regularities, so can’t be captured by such a baseline model.

We try to resolve these problems by incorporating human knowledge. In fact, human being usually uses prior knowledge when recognizing named entities. In this section, we introduce the human knowledge that is used for NER and the method of how to incorporate them into the baseline model.

Given a sequence of Chinese characters, the recognition process after combined with human

knowledge consists of the five steps shown in Figure1.

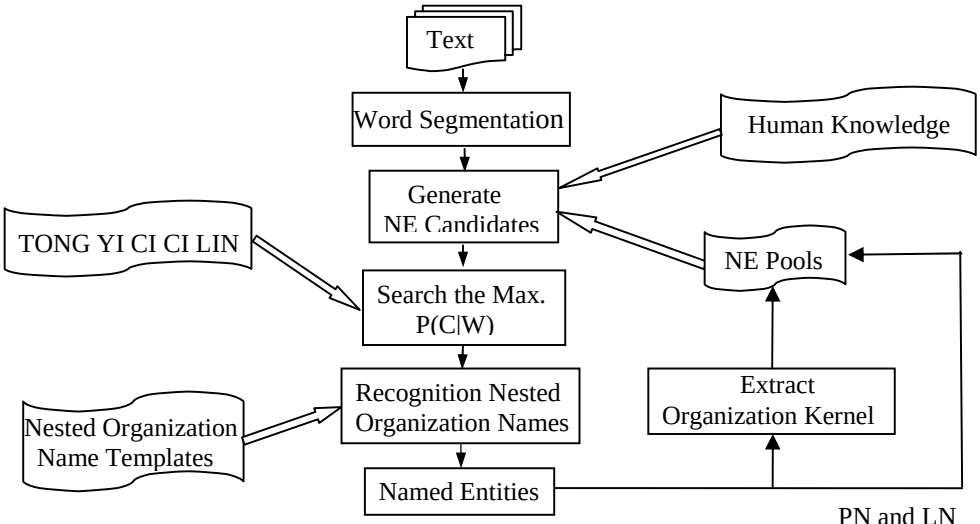

Figure 1 Recognition Process of the Hybrid Model

4.1 Incorporate Knowledge for Person Name Recognition

Chinese person names are composed of a surname and a given name. Usually the characters used for Chinese person names are limited. [Maosong Sun, Changning Huang, 1994] presents 365 most high frequently used surnames cover 99% Chinese surnames. 1141 most high frequently used characters cover 99% Chinese given names. Similarly the characters used for transliterated names are also limited. We extract about 476 transliterated characters from the training corpus.

The following is the human knowledge used for person name recognition and the method of how to incorporate them into the baseline.

① **A Chinese single and plural surname list:** Only those characters in the surname list can trigger person name recognition.

② **A list of person title list:** Only when the current character belongs to the surname list and the next word is in the title list, candidates are accepted.

③ **A transliterated character list:** Only those consecutive characters in the transliterated character list form a candidate transliterated name.

④ Person name can't span any punctuation and the length of CN can't exceed 8 characters while the length of TN is unrestrained.

All these knowledge are used for restricting search space.

4.2 Incorporate Knowledge for Location Name Recognition

A complete location name is composed of the name part and a salient word. For the location name "北京市/Beijing City", the name part is "北京/Beijing" and the salient word is "市/city". Unfortunately, the salient word is omitted in many occasions. So it is unfeasible to trigger LN recognition only depending on the salient words in location name. In order to improve the precision and recall of LN recognition, we use the following human knowledge. The method of incorporating them is also explained.

① **A general location name list:** The list includes the names of Chinese provinces and counties, foreign country and its capitals, some famous geographical names and foreign cities. If the current word is in the list, we accept it as a **candidate** LN.

② **A location salient word list:** If the word w_i belongs to the list, 2~6 words before the salient word are accepted as candidate LNs.

③ **A general word list** (such as verbs and prepositions) which usually is followed by a location name, such as "在/at", "去/go". If the word w_i is in the list, 2~6 words following it are accepted as candidate LNs.

④ **An abbreviative location name list:** If the current word is in the list, we accept it as a candidate LN such as "中/China", "美/America".

⑤ **Coordinate LN recognition:** If w_{i-2} is a candidate LN and w_{i-1} is "、"(a punctuation denoting coordinate relation), LN recognition is triggered at the position of word w_i.

⑥ Location name can't span punctuations and its length couldn't exceed 6 words.

Knowledge ①, ②, ③, ⑤, ⑥ can restrict search space while knowledge ④ deals with abbreviative location name.

4.3 Incorporate Knowledge for Organization Name Recognition

The organization names recognition is the most difficult task. The reasons lie in nested ONs and abbreviative ONs especially.

Nested ON means there are one or more location names, person names and/or organization names embedded in organization name. Typical structure of ON has been given in section 1. We can capture most of the nested organization names by several ON templates mentioned in the following section.

Abbreviative ONs include continuous and discrete abbreviation which omits some words in the full name. Take "上海华联超市股份有限公司" as example, abbreviative ON of it may omit LN "上海/Shanghai", organization types like"超市/supermarket", "股份/stock", "有限/limited", and salient word like "公司/company" from full names but usually remains organization kernel "华联/Hualian". Table 3 lists some examples of abbreviative ONs.

Continuous Abbreviation	上海华联超市 股份有限公司 Shanghai Hualian Co.,Ltd	上海华联 Shanghai Hualian
	清华大学 Tsinghua niversity	清华 Tsinghua
Discrete Abbreviation	上海证券交易所 Shanghai Stock Exchange	上证 Shanghai Stock
	北京大学 Peking University	北大 Bei Da

Table 3 Nest Organization Full Names and Its Abbreviative Names

So it is important to extract organization kernel from the full name in order to recognize abbreviative ON like "上海华联". Moreover, an organization's abbreviative names usually occur after its' full name, unless it is a well-known organization. So this strategy for abbreviation organization name recognition is effective.

The following is the human knowledge used for ON recognition and the method of how to incorporate them.

① **An organization salient word (OrgSws) list:** If the current word w_i is in OrgSws list, 2~6 words before OrgSw are accepted as the candidate ONs.

② **A general famous organization name list:** If the current word is in the list, we accept it as a **candidate** ON such as "国务院/ State Department", "联合国/ U.N. ".

③ **An organization names template list:** We mainly use organization name templates to recognize the nested ONs. Some of these templates are as follows:

ON-->LN D* OrgSw
ON-->PN D* OrgSw
ON-->ON OrgSw

D means words used in the middle of organization names. *D** means repeating zero or more times. This component runs in the end stage of recognition process shown in Figure 1.

④ **An organization type list:** The list is used to extract organization kernels from recognized ONs. We have a pool which memorizes ONs recognized in current paragraph and its kernel. If the current word belongs to organization kernel in pool, we accept it as a candidate ON. The idea is effective especially in financial domain which contains many stocks such as"上海华联/Shanghai Hualian", "长江科技/Changjiang Technology".

Knowledge ①, ②, ③ restrict search space while knowledge ④ deals with abbreviative organization name.

4.4 Semantic Similarity Computation for Data Sparseness

《同义词词林/TONG YI CI CI LIN》classifies the words in terms of semantic similarity. Here we use it to resolve data sparseness problem. If current transmission probability doesn't exist, we resort to its synonym transmission. In statistical sense, synonym transmissions are approximate. Take an example, the probability of $P(A|B)$ doesn't exist, but there has $P(C|B)$, meanwhile, the word A and C are thesaurus according to 《同义词词林/TONG

YI CI CI LIN》, then we use *P(C|B)* to replace *P(A|B)*.

5 Results of Evaluation

We also test our hybrid model on IEER-99 neswire test data. The performance is shown in Table 4.

	Precision	Recall	F-measure
PN	83.30%	92.28%	87.56%
LN	88.31%	84.69%	86.47%
ON	84.49%	71.08%	77.21%
Total	86.09%	83.18%	84.61%

Table 4 The Performance of the Hybrid Model

Comparing Table 1 with 4, we find that the performance of the hybrid model increases remarkably. More specifically, the precision and the recall of PNs increase from 80.23% to 83.30% and from 89.55% to 92.28% respectively. The precision and recall of LNs increase from 45.05% to 82.18% and from 66.96% to 86.74% respectively. The precision and recall of ONs increase from 42.98% to 80.86% and from 61.45% to 72.09% respectively. The reason that the improvement of PNs is slighter than that of ONs and LNs is that the statistical information estimated from labeled corpus for PNs is good enough but not for LNs and ONs.

Must be mentioned is that, in our evaluation, only NEs with both correct boundary and correct type label are considered as the correct recognitions, which is a little different from other evaluation systems.

We also test our system on data set of sport, finance, news and entertainment domains. These test data are downloaded from Internet shown in Table 4.

Domain	Number of NE			File size
	PN	LN	ON	
Sport(S)	954	510	609	91K
Finance(F)	212	406	461	80K
News(N)	526	961	437	76K
Entertainment(E)	1016	511	133	100K
Total	2708	2388	1640	247K

Table 4 Statistic of Multi-field Test Data

The results are shown in Table 5.

		Precision	Recall	F-measure
PN	**S**	80.17%	91.10%	85.28%
	F	61.35%	94.34%	74.35%
	N	88.66%	83.27%	85.88%
	E	82.20%	82.28%	82.24%
LN	**S**	82.90%	81.76%	82.33%
	F	83.72%	81.03%	82.35%
	N	91.95%	91.56%	91.75%
	E	81.64%	87.87%	84.64%
ON	**S**	73.43%	67.16%	70.15%
	F	65.88%	60.30%	62.97%
	N	92.52%	84.70%	88.44%
	E	78.30%	62.41%	69.46%
Total		81.01%	81.24%	81.12%

Table 5 Results on different domain

Table 5 shows that the performance on financial domain is much lower. The reason is that, in financial domain, there are many stock names which are the abbreviation of organization names. Moreover, organization full name never appear in the text. So the system can't recognize them as an organization name. However, on many occasions, they are recognized as person names. As a result, the precision of PNs declines, meanwhile, the precision and recall of ONs can't be high.

Based on the above analysis, we find that the main sources of errors in our system are as follows.

First, we still have not found a good strategy for the abbreviation location names and organization names. Because abbreviative LNs and ONs sometimes appear before full LN, sometimes not, so the pool strategy can't work well.

Second, some famous organization names that always appear in the shape of abbreviation can't be recognized as ON because the full name never appear such as 高通/GaoTong, 新浪/Xinlang. However, these ONs are often recognized as PNs. Such errors are especially serious in finance domain shown Table 5.

Third, many words can't be found in 《同义词词林/TONG YI CI CI LIN》.

6 Conclusions

Chinese NER is a more difficult task than English NER. Though many approaches have been tried, the result is still not satisfactory. In this paper, we present a hybrid algorithm of incorporating human

knowledge into statistical model. Thus we only need a relative small-sized labeled corpus (one-month's Chinese People's Daily tagged with NER tags at Peking University) and human knowledge, but can achieve better performance. The main contribution of this paper is putting forward an approach which can make up for the limitation of using the statistical model or human knowledge purely by combining them organically.

Our lab was mainly devoted to cross-language information processing and its application. So in the future we will shift our algorithm to other languages. And fine-tune to a specific domain such as sports.

ACKNOWLEDGEMENT

This paper is supported by the National "973" project G1998030501A-06 and the Natural Science Foundation of China 60272041.

References

Jian Sun, et al. 2002. *Chinese Named Entity Identification Using Class-based Language Model.* Proceedings of the 19th International Conference on Computational Linguistics

Hsin-His Chen, et al. 1997. *Description of the NTU System Used for MET2*. Proceedings of the Seventh Message Understanding Conference

Tat-Seng Chua, et al. 2002. *Learning Pattern Rules for Chinese Named Entity Extraction*. Proceedings of AAAI'02

W.J.Teahan, et al. 1999. *A Compression-based Algorithm for Chinese Word Segmentation*. Computational Linguistic 26(2000) 375-393

Maosong Sun, et al. 1994. *Identifying Chinese Names in Unrestricted Texts*. Journal of Chinese Information Processing. 1994,8(2)

Collins, Singer. 1999. *Unsupervised Models for Named Entity Classification*. Proceedings of 1999 Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora

Daniel M. Bikel, et al. 1997. *Nymble: a High-Performance Learning Name-finder*. Proceedings of ANLP-97, page 194-201, 1997

Yu et al. 1998. *Description of the Kent Ridge Digital Labs System Used for MUC-7*. Proceedings of the Seventh Message Understanding Conference

Silviu Cucerzan, David Yarowsky. 1999. *Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence*. Proceedings 1999 Joint SIGDAT Conference on EMNLP and VLC

Peter F.Brown, et al. 1992. *Class-Based n-gram Model of Natural Language*. 1992 Association for Computational Linguistics

A.Mikheev, M.Moens, and C.Grover. 1999. *Named entity recognition without gazetteers*. Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics. Bergen, Norway

Borthwich. A. 1999. *A Maximum Entropy Approach to Named Entity Recognition*. PhD Dissertation

Dong & Dong. 2000. *Hownet. At: http://www.keenage.com*

Yu.S.W. 1999. *The Specification and Manual of Chinese Word Segmentation and Part of Speech Tagging*. At: http://www.icl.pku.edu.cn/Introduction/corpustagging. htm

Mei.J.J, et al. 1983. *《同义词词林/TONG YI CI CI LIN》*. Shanghai CISHU Press