Computational Linguistics and Chinese Language Processing
Vol. 8, No. 2 , August 2003, pp. 1-28

A Class-based Language Model Approach to Chinese Named Entity Identification[1]

Jian Sun*, Ming Zhou+, Jianfeng Gao+

Abstract

This paper presents a method of Chinese named entity (NE) identification using a class-based language model (LM). Our NE identification concentrates on three types of NEs, namely, personal names (PERs), location names (LOCs) and organization names (ORGs). Each type of NE is defined as a class. Our language model consists of two sub-models: (1) a set of entity models, each of which estimates the generative probability of a Chinese character string given an NE class; and (2) a contextual model, which estimates the generative probability of a class sequence. The class-based LM thus provides a statistical framework for incorporating Chinese word segmentation and NE identification in a unified way. This paper also describes methods for identifying nested NEs and NE abbreviations. Evaluation based on a test data with broad coverage shows that the proposed model achieves the performance of state-of-the-art Chinese NE identification systems.

Keywords: Named entity identification, class-based language model, contextual model, entity model

1. Introduction

Named Entity (NE) identification is the problem of detecting entity names in documents and then classifying them into corresponding categories. This is an important step in many natural language processing applications, such as information extraction (IE), question answering (QA), and machine translation (MT). A lot of researches have been carried out on English NE identification. As a result, some systems have been widely applied in practice. On the other hand, Chinese NE identification is a different task because in Chinese, there is no space to mark the boundaries of words and no clear definition of words. In addition, Chinese NE

---

[1] This work was done while the author was visiting Microsoft Research Asia.

* Beijing University of Posts&Telecommunications.
Currently is an assistant researcher in Institute of Computing Technology, Chinese Academy of Sciences. Email: sunjian@ict.ac.cn

+ Microsoft Research Asia. Beijing, 100080. Email: mingzhou@microsoft.com; jfgao@microsoft.com

identification is intertwined with word segmentation. Traditional approaches to Chinese NE identification usually employ two separate steps, namely, word segmentation and NE identification. As a result, errors in word segmentation will lead to errors in NE identification. Moreover, the identification of NE abbreviations and nested NEs has not yet been investigated thoroughly in previous works. For example, nested locations in organization names have not been discussed at the Message Understanding Conference (MUC).

In this paper, we present a method of Chinese NE identification using a class-based LM, in which the definitions of classes are extended in comparison with our previous work [Sun, Gao *et al.*, 2002]. The model consists of two sub-models: (1) a set of entity models, each of which estimates the generative probability of a Chinese character string given an NE class; and (2) a contextual model which estimates the generative probability of a class sequence. Our model thus provides a statistical framework for incorporating Chinese word segmentation and NE identification in a unified way. In the paper, we shall also describe our methods for identifying nested NEs and NE abbreviations.

The rest of this paper is organized as follows: Section 2 briefly discusses related work. Section 3 presents in detail the class-based LM for Chinese NE identification. Section 4 discusses our methods of identifying NE abbreviations. Section 5 reports experimental results. Section 6 presents conclusions and future work.

2. Related Work

Traditionally, the approaches to NE identification have been rule-based. They attempt to perform matching against a sequence of words in much the same way that a general regular expression matcher does. Some of these systems are, FACILE [Black *et al.*, 1998], IsoQuest's NetOwl [Krupha and Hausman, 1998], the LTG system [Mikheev *et al.*, 1998], the NTU system [Chen *et al.*, 1998], LaSIE [Humphreys *et al.*, 1998], the Oki system [Fukumoto *et al.*, 1998], and the Proteus system [Grishman, 1995]. However, the rule-based approaches are neither robust nor portable.

Recently, research on NE identification has focused on machine learning approaches, including the hidden Markov model [Bikel *et al.*, 1999; Miller *et al.*, 1998; Gotoh and Renals, 2000; Sun *et al.*, 2002; Zhou and Su, 2002], maximum entropy model [Borthwick, 1999], decision tree [Sekine *et al.*, 1998], transformation-based learning [Brill, 1995; Aberdeen *et al.*, 1995; Black and Vasilakopoulos, 2002], boosting [Collins, 2002; Carreras *et al.*, 2002; Tsukamoto *et al.*, 2002; Wu *et al.*, 2002], the voted perceptron [Collins, 2002], conditional Markov model [Jansche, 2002], support vector machine [McNamee and Mayfield, 2002; Takeuchi and Collier, 2002], memory-based learning [Sang, 2002] and learning approaches stacking [Florian, 2002]. Some systems, especially those for English NE identification, have

been applied to practical applications.

When it comes to the Chinese language, however, NE identification systems still cannot achieve satisfactory performance. Some representative systems include those developed in [Sun *et al.*, 1994; Chen and Lee, 1994; Chen *et al.*, 1998; Yu *et al.*, 1998; Zhang, 2001; Sun *et al.*, 2002].

We will mainly introduce two systems, namely, the rule-based NTU system for Chinese [Chen *et al.*, 1998] and the machine learning based BBN system [Bikel *et al.*, 1999], because these are representative of the two different approaches.

Generally speaking, the NTU system employs the rule-based method. It utilizes different types of information and models, including character conditions, statistic information, titles, punctuation marks, organization and location keywords, speech-act and locative verbs, cache model and n-gram model. Different kinds of NEs employ different rules. For example, one rule for identifying organization names is as follows:

OrganizationName → *CountryName OrganizationNameKeyword*

e.g. 美国 大使馆

US *Embassy*

NEs are identified in the following steps: (1) segment text into a sequence of tokens; (2) identify named persons; (3) identify named organizations; (4) identify named locations; and (5) use an n-gram model to identity named organizations/locations.

The BBN model [Bikel *et al.*, 1999], a variant of Hidden Markov Model (HMM), views NE identification as a classification problem and assigns to every word either one of the desired NE classes or the label NOT-A-NAME, meaning "none of the desired class". The HMM has a bigram LM of each NE class and other text. Another characteristic is that every word is a two-element vector consisting of the word itself and the word-feature. Given the model, the generation of words and name-classes is performed in three steps: (1) select a name-class; (2) generate the first word inside that name-class; (3) generate all the subsequent words inside the current name-class.

There have been relatively fewer attempts to deal with NE abbreviations [Chen, 1996; Sproat *et al.*, 2001]. These researches mainly investigated the recovery of acronyms and non-standard words.

In this paper, we present a method of Chinese NE identification using a class-based LM. We also describe our methods of identifying nested NEs and NE abbreviations.

3. Class-based LM Approach to NE Identification

A word-based n-gram LM is a stochastic model which predicts a word given the previous n-1

words by estimating the conditional probability $P(w_n|w_1 \ldots w_{n-1})$. A class-based LM extends the word-based LM by defining similar words as a class. It has been demonstrated to be a more effective way of dealing with the data-sparseness problem. In this study, the class-based LM is applied to integrate Chinese word segmentation and NE identification in a unified framework.

In this section, we first gives definitions of classes. Then, we describe the elements of the class-based LM, parameter estimation, and how we apply the model to NE identification.

Table 1. Definitions of Classes

Class		Explanation/Intuition	Examples
PER	FN	foreign names in transliteration	*克 林 顿* 'Clinton'
	PER1	Chinese personal name consisting only of a surname	*周 总理* 'Premier Zhou'
	PER2	Chinese personal name consisting of a surname and a one-character given name	*李 鹏* 'Li Peng'
	PER3	Chinese personal name consisting of a surname and a two-character given name	*周 恩 来* 'Zhou Enlai'
	PABB	Abbreviation of a personal name	*恩 来* 'Enlai'
LOCW[2]		Whole name of a location	*北京 市* 'Beijing City'
LABB		Abbreviation of a location name	*中 日* 关系 'Sino-Japan relation'
ORG		Organization name	*北京 邮电 大学* 'Beijing University of Posts&Telecommunications'
PT		A personal title in context (-1~1) of PER	*周 总理* 'Premier Zhou'
PV		Speech-act verb in context (-2~2) of PER	*周 总理 指出* 'Premier Zhou points out'
LK		Location keyword in a location name	*北京 市*
OK		Organization keyword in an organization name	*北京 邮电 大学*
DT		Data and time expression	*2002 年 10 月*
NU		Numerical expression	*１２亿*, *5%*
BOS		Beginning of a sentence	
EOS		End of a sentence	

[2] In the step of identifying PERs and LOCs, the classes LOCW and LABB are modeled in context ; in the step of identifying ORGs, the two classes are united into one class, LOC.

3.1 Word Classes

In this study, each kind of NE is defined as a class in our model. In practice, in order to represent different constructions for each kind of NE, we further divide each class into sub-classes. The detailed definitions of the classes are shown in Table 1. In addition, each word in a lexicon is defined as a class.

For each NE type (PER, LOC, and ORG), we define 6 tags to mark the position of the current character (word) in the entity name as shown in Table 2.

Table 2. Position Tags in NEs

Tag	Explanation	Tag in PER	Tag in LOC	Tag in ORG
B	Beginning of the NE	PB	LB	OB
E	End of the NE	PE	LE	OE
F	First character (or word) in the NE	PF	LF	OF
I	Medial character (or word) in the NE, neither initial nor final	PI	LI	OI
L	Last character (or word) in the NE	PL	LL	OL
S	Single character (or word)	PS	LS	OS

3.2 Class-based LM for Chinese NE identification

Given a Chinese character sequence $S_1^n = s_1 \cdots s_n$, in which NEs are to be identified, the identification of PERs and LOCs is the problem of find the optimal class sequence $\hat{C}_1^m = c_1 \cdots c_m$ ($m \le n$) that maximizes the conditional probability $P(C_1^m \mid S_1^n)$. This idea can be expressed by Equation (1), which gives the basic form of the class-based LM:

$$\begin{aligned} \hat{C}_1^m &= \arg \max_C P(C_1^m \mid S_1^n) \\ &= \arg \max_C P(C_1^m) \times P(S_1^n \mid C_1^m) \ . \end{aligned} \tag{1}$$

The class-based LM consists of two components: the contextual model $P(C_1^m)$ and the entity model $P(S_1^n \mid C_1^m)$. The contextual model estimates the generative probability of a class. The probability $P(C_1^m)$ can be approximated using trigram probability as shown in Equation (2):

$$P(C_1^m) \cong \prod_{i=1}^{m} P(c_i \mid c_{i-2} c_{i-1}) \tag{2}$$

The entity model $P(S_1^n \mid C_1^m)$ estimates the generative probability of a Chinese character sequence given an NE class, as shown in Equation (3):

$$\begin{aligned} & P(S_1^n \mid C_1^m) \\ & = P(s_1 \cdots s_n \mid c_1 \cdots c_m) \\ & \cong P([s_1 \cdots s_{c_1-end}] \cdots [s_{c_m-start} \cdots s_n] \mid c_1 \cdots c_m) \\ & \cong \prod_{j=1}^{m} P([s_{c_j-start} \cdots s_{c_j-end}] \mid c_j) \end{aligned} \tag{3}$$

By combining the contextual model and the entity models as in Equation (1), we obtain a statistical framework that incorporates the entity features and contextual features. The following is an example used to show how the contextual model and entity models are integrated: "*周恩来总理是我们的好总理。*" We presume that the correct result is

周 恩 来	*总理*	*是*	*我们*	*的*	*好*	*总理*	*。*
PER	PT						
Zhou Enlai	*Prime Minister*	*is*	*our*		*great*	*premier*	.

The computation of the joint probability of the two events (the input sentence and the hidden class sequence) is shown in the following equation:

$$\begin{aligned} & P(PER \mid BOS) \times P(PER\,3 \mid PER) \times P(\text{周恩来} \mid PER\,3) \\ & \times P(PT \mid BOS, PER) \times P(\text{总理} \mid PT) \\ & \times P(\text{是} \mid PER, PT) \times P(\text{我们} \mid PT, \text{是}) \times P(\text{的} \mid \text{是}, \text{我们}) \\ & \times P(\text{好} \mid \text{我们}, \text{的}) \times P(\text{总理} \mid \text{的}, \text{好}) \times P(\text{。} \mid \text{好}, \text{总理}) \times P(EOS \mid \text{总理}, \text{。}) \end{aligned}$$

where $P(\text{周恩来} \mid PER\,3)$ will be described in Section 3.3.1. It should be noted that the computations of the generative probability of the two occurrences of 总理 are different. The first one is generated as the class PT, whereas the second is generated as the common word 总理.

In Section 3.3, we will describe the entity models in detail, and in Section 3.4, we will present our model estimation approach.

3.3 Entity Models

In order to discriminate among the first, medial and last character in an NE, we design the entity models in such a way that the character (or word) position is utilized. For each kind of NE, different entity models are adopted as described below.

3.3.1 Person Model

For the class PER (including FN, PER1, PER2, and PER3), the entity model is a *character-based* trigram model. The modeling of PER3 is described in the following example.

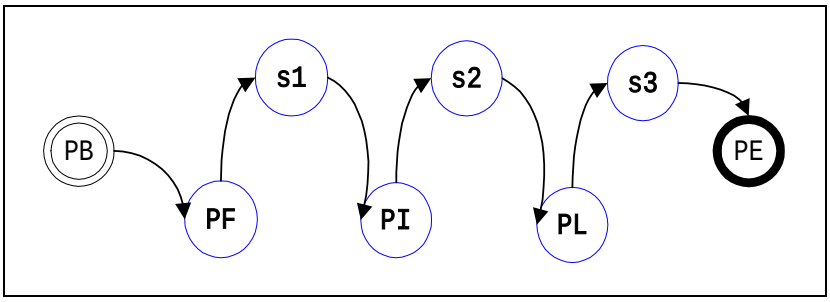


Figure 1. The generation of the sequence s_1 s_2 s_3 given the PER3 class.

As shown in Figure 1, the generative probability of the Chinese character sequence given the PER3 class is computed as follows:

$$\begin{aligned} & P(s_1 s_2 s_3 \mid c = PER\ 3) \\ & = P(PF \mid PER\ 3, PB) \times P(s_1 \mid PER\ 3, PB, PF) \\ & \times P(PI \mid PER\ 3, PF, s_1) \times P(s_2 \mid PER\ 3, s_1, PI) \\ & \times P(PL \mid PER\ 3, PI, s_2) \times P(s_3 \mid PER\ 3, s_2, PL) \\ & \times P(PE \mid PER\ 3, PL, s_3) \end{aligned} \quad (4)$$

For example, the generative probability of 周恩来 'Zhou Enlai' can be expressed as

$$\begin{aligned} & P(\text{周恩来} \mid PER\ 3) \\ & = P(PF \mid PER\ 3, PB) \times P(\text{周} \mid PER\ 3, PB, PF) \\ & \times P(PI \mid PER\ 3, PF, \text{周}) \times P(\text{恩} \mid PER\ 3, \text{周}, PI) \\ & \times P(PL \mid PER\ 3, PI, \text{恩}) \times P(\text{来} \mid PER\ 3, \text{恩}, PL) \\ & \times P(PE \mid PER\ 3, PL, \text{来}) \end{aligned}$$

The FN, PER1, and PER2 are modeled in similar ways. Each class of FN, PER1, PER2, and PER3 corresponds to an entity model for a kind of personal names. But in the contextual model, the four classes correspond to one class (PER).

3.3.2 Location Model

For the class LOCW, the entity model is a word-based trigram model. If the last word in the candidate location name is a location keyword, it can be generalized as class LK, which is also modeled in the form of a unigram. For example, the generative probability of 北京市 'Beijing City' in the location model can be expressed as:

$$
\begin{aligned}
&P(\text{北京市} \mid LOCW) \\
&= P(LF \mid LOCW, LB) \times P(\text{北京} \mid LOCW, LB, LF) \\
&\times P(LL \mid LOCW, LF, \text{北京}) \times P(LK \mid LOCW, \text{北京}, LL) \times P(\text{市} \mid LK) \\
&\times P(LE \mid LOCW, LL, LK)
\end{aligned}
$$

3.3.3 Organization Model

For the class ORG, the entity model is a class-based trigram model. Personal names and location names nested in ORG are generalized as classes PER and LOC, respectively. Thus, we can identify nested personal names and location names using the class-based model. The organization keyword in the ORG is also generalized as the OK class, which is modeled in the form of a unigram.

3.3.4 Other Models

It is obvious that personal titles and special verbs are important clues for identifying personal names (e.g., [Chen *et al.*, 1998]). In our study, personal titles and special verbs are adopted to help identify personal names by constructing a unigram model of PT and a unigram model of PV. Accordingly, the generative probability of a specific personal title w_i can be computed as

$$P(w_i \mid c = PT) \tag{5}$$

and that of a specific speech-act verb w_i can be computed as

$$P(w_i \mid c = PV) \tag{6}$$

We can also build unigram models for classes LK and OK in similar ways, respectively.

In addition, if c is a word that does not belong to the above defined classes, the generative probability is as follows:

$$P(s_{c-start} \ldots s_{c-end} \mid c) = 1 \tag{7}$$

where the Chinese character sequence $s_{c-start} \ldots s_{c-end}$ is a single word.

3.4 Model Estimation

As discussed in Section 3.2, there are two probabilities to be estimated, $P(C_1^m)$ and $P(S_1^n \mid C_1^m)$. Both of them are estimated using maximum likelihood estimation (MLE) based on the training data, which are obtained by tagging the NEs in the text using the parser

NLPWin[3]. Smoothing the MLE is essential to avoid zero probability for events that were not observed in the training data. We apply the standard techniques, in which more specific models are smoothed with progressively less specific models. The details of the back-off smoothing method we use are described in [Gao *et al.*, 2001].

In what follows, we will describe our model estimation approach. We will assume that a sample training data set has one sentence: "*周恩来总理是我们的好总理。*" The corresponding annotated training data[4] are as follows:

周 恩 来 *总 理* *是 我 们 的 好 总 理 。*
PER PT

3.4.1 Contextual Model Estimation

We extract training data for the contextual model by replacing the names in the above example with corresponding class tags, i.e., *PER PT 是 我们 的 好 总理 。*. The contextual model parameters are computed by using MLE together with back-off smoothing.

3.4.2 Entity Model Estimation

We can also obtain the training data of each entity model. For example, the PER3 list we obtained from the above example has one instance, *周 恩 来* . The corresponding training data for PER3, where position tags are introduced, are as follows:

PB PF 周 PI 恩 PL 来 PE .

The model parameters of PER3 are computed using MLE and back-off smoothing. We can also estimate other entity models in a similar way.

3.5 Decoder

The NE identification procedure is as follows: (1) identify PERs and LOCs; (2) identify ORGs based on the output of identifying PERs and LOCs. Thus, the PERs and LOCs nested in ORGs can be identified. Since the steps involved in identifying PERs and LOCs, and those involved in identifying ORGs are similar, we will only describe the former in the following.

Generally speaking, the decoding process consists of three steps: *lexical word candidate generation*, *NE candidate generation*, and *Viterbi search*. A few heuristics and NE grammars, shown in Figure 2, are used to reduce the search space when NE candidates are generated.

---

[3] The NLPWin system is a natural language processing system developed by Microsoft Research.

[4] The PV and PT are not tagged in the training data parsed by NLPWin. They are then labeled using rule-based methods.

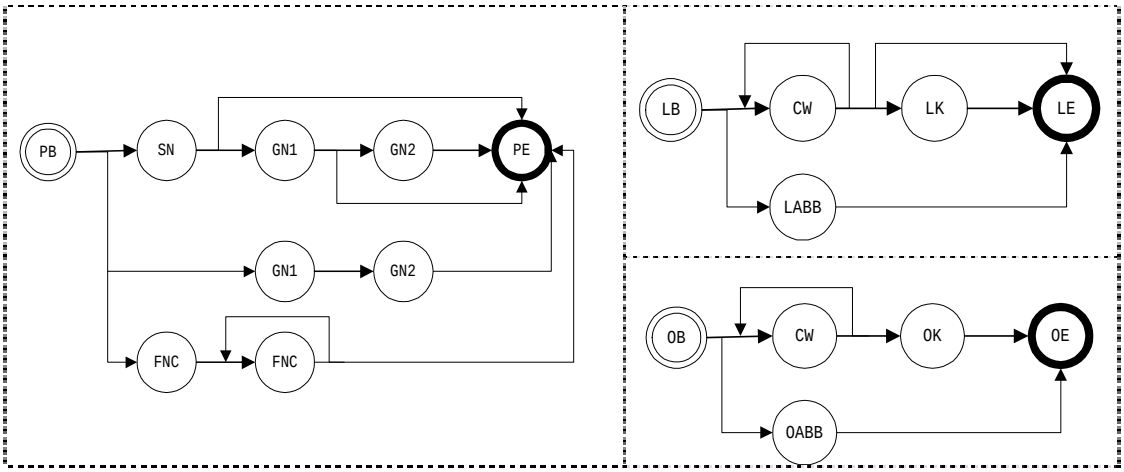


Figure 2. The grammar of PER, LOC and ORG candidates.

SN: Chinese surname; GN1: first character of a Chinese given name; GN2: second character of a Chinese given name; FNC: character of a foreign name; CW: Chinese word; LK: location keyword; LABB: abbreviation of a location name; OK: organization keyword; OABB: abbreviation of an organization name.

Given a sequence of Chinese characters, the decoding process is as follows:

Step 1:

*Lexical word candidate generatio*n. All possible word segmentations are generated according to a Chinese lexicon containing 120,050 entries. The lexicon, in which each entry does not contain the NE tags even if it is a PER, LOC or ORG, is only used for segmentation.

Step 2:

NE candidate generation. NE candidates are generated in two steps: (1) candidates are generated according to NE grammars; (2) each candidate is assigned a probability by using the corresponding entity model. Two kinds of heuristic information, namely, internal information and contextual information, are used for a more effective search. The internal information, which is used as an NE candidate trigger, includes: (1) a Chinese family name list, containing 373 entries (e.g., 周 'Zhou', 李 'Li'); (2) a transliterated name character list, containing 618 characters (e.g., 什 'shi', 顿 'dun'). The contextual information used for computing the generative probability includes: (1) a list of personal title, containing 219 entries (e.g., 总理 'premier'); (2) a list of speech-act verbs, containing 9191 entries (e.g., 指出 'point out'); (3) the left and right words of the PER.

Step 3:

Viterbi Search. Viterbi search is used to select the hypothesis with the highest probability as the best output, from which PERs and LOCs can be obtained.

For the identification of ORGs, the organization keyword list (containing 1,355 entries) is utilized both to generate candidates and to compute generative probabilities.

4. Identification of Chinese NE Abbreviations

NEs with the same meaning, which often occur more than once in a document, are likely to appear in different expressions. For example, the entity names "北京大学" (Peking university) and "北大" (an abbreviation of "北京大学") might occur in different sentences in the same document. In this case, the whole name may be identified correctly, whereas its abbreviation may not be. NE abbreviations account for about 10 percent of Chinese NEs. Therefore, identifying NE abbreviations is essential for improving the performance of Chinese NE identification. To the best of our knowledge, there has been no systematic study on this topic up to now. In this study, we applied the language model method to the task. We adopted the language model because the identification of NE abbreviations can be easily incorporated into the class-based LM framework described in Section 3. Furthermore, doing so lessens the labor required to develop rules for NE abbreviations. After a whole NE name has been identified, the procedure for identifying NE abbreviations is as follows: (1) generate all the candidates of NE abbreviations according to the corresponding generation pattern; (2) assign to each one a generative probability (or score) by using the corresponding model; (3) store the candidates in the lattice for Viterbi search.

In Sections 4.1 to 4.3, we will describe the abbreviation models applied to abbreviations of personal names, location names, and organization names, respectively.

4.1 Modeling Chinese PER Abbreviation[5]

Suppose that the whole name of PER $s_1s_2s_3$ has been identified; we generate two kinds of abbreviation candidates of personal names: s_1 and s_2s_3. The corresponding generative probabilities of these two types of candidates given PER abbreviation are computed by linearly interpolating the cache unigram model ($p_{unicache}(s_i)$) and the static entity model ($p_{static}(s_i|s_{i-1}, s_{i-2})$) as shown in Equation (8):

$$\begin{aligned} & P(s_i \mid PER \quad \text{abbr}) \\ & \cong \lambda \times P_{unicache}(s_i \mid PER) + (1-\lambda) \times P_{static}(s_i \mid s_{i-1}, s_{i-2}; PER) \end{aligned} \tag{8}$$

[5] At present, the abbreviations of transliterated personal names are not modeled.

where $\lambda \in [0,1]$ is the interpolation weight determined on the development data set. The probability $P_{static}(s_i \mid s_{i-1}, s_{i-2}; PER)$ is estimated from the training data of PER, and $P_{unicache}(s_i \mid PER)$ is estimated from the cache belonging to the PER class. At any given time during the NE identification task, the cache for a specific class contains NEs that have been identified as belonging to that class. After the abbreviation candidates are generated, they are stored in the lattice for search.

4.2 Modeling LOC Abbreviations

The LOC abbreviation (LABB) entity model is a unigram model: $P(s \mid c = LABB)$. The procedure of identifying location abbreviations can be described as follows: (1) generate LABB candidates according to the list of location abbreviations; (2) determine whether the candidates are LABB or not based on the contextual model. For example, the generative probability *P(*中日关系*)* for the sequence 中日关系 'Sino-Japan relations' is computed as follows:

$$
\begin{aligned}
&P(\text{中日关系}\ ) \\
&= P(\ LABB\ \mid BOS\ ) \times P(\text{中} \mid LABB\ ) \times P(\ LABB\ \mid BOS\ , LABB\ ) \times P(\text{日} \mid LABB\ ) \\
&\times P(\ \text{关系}\ \mid LABB\ , LABB\ ) \times P(\ EOS\ \mid LABB\ , \text{关系}\ )
\end{aligned}
$$

4.3 Empirical Modeling of ORG Abbreviations

When an organization name $A = w_1 w_2 \ldots w_N$ is recognized, all the abbreviation candidates of the organization are generated according to the patterns shown in Table 3.

Table 3. Generation Patterns[6] of Organization Abbreviations

Condition	Generation Pattern	Examples	Remark
N≥2	$s_{11}s_{21}$	北京 邮电 大学 → 北 邮	s_{ij} *denotes the jth character of the ith word of A*
	…	…	
	$s_{11}s_{21}\ldots s_{N1}$	北京 邮电 大学 → 北 邮 大	
N=2 and w_1 is not a location name	w_1	清华 大学 → 清华	w_i *denotes the ith word of A*
N=3 and w_1 is not a location name	w_1	苹果 电脑 公司 → 苹果	
	w_1w_2	苹果 电脑 公司 → 苹果 电脑	
N=3 and w_1 is a location name	w_2	北京 国安 队 → 国安	

[6] Because abbreviation formation is complex, these patterns cannot cover all cases. E.g., 中国石油天然气集团公司 abbreviated as 中石油 is not covered by our patterns.

Since there are no training data for the ORG abbreviation model, it is impossible to estimate the model parameters. We then utilize linguistic knowledge of abbreviation generation and construct a score function for the ORG abbreviation candidates. The score function is defined such that the resulting scores of the ORG abbreviation candidates are comparable to other NE candidates whose parameters (probabilities) are assigned using the probabilistic models described in Section 3.3.

The following is an example used to explain how a score is assigned. Suppose that 北京邮电大学 'Beijing University of Posts & Telecommunications' has been identified as an ORG in the previous part in the text, and that one of the ORG abbreviation candidates is 北邮. The generative probability of 北京邮电大学 *(P(北京邮电大学|ORG))* in the ORG model and that of 北邮 *P(北邮|Contextual Model)* in the contextual model can be computed. We calculate the score of 北邮 in the organization abbreviation model (denoted as *Score(北邮|ORG abbr)*) as

$$\alpha \times P(\text{北京邮电大学}|ORG) + (1-\alpha) \times P(\text{北邮}|\,contextual\,Model)),$$

where α is set to be 0.5. In addition, according to intuition, the score of 北邮 in the organization abbreviation model is larger than the probability of 北邮 in the contextual model *given that* 北京邮电大学 has been identified as an ORG, i.e.,

$$Score(\text{北邮}\,|\,ORG\,abbr) \geq P(\text{北邮}\,|\,Contextual\,Model).$$

Accordingly, a maximum function is used. Figures 3.1 and 3.2 show the state transition in the lattice of the input sequence (e.g., 北邮).

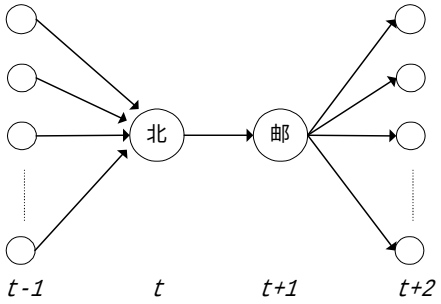


Figure 3.1. State transition in the lattice without the identification of ORG abbreviations.

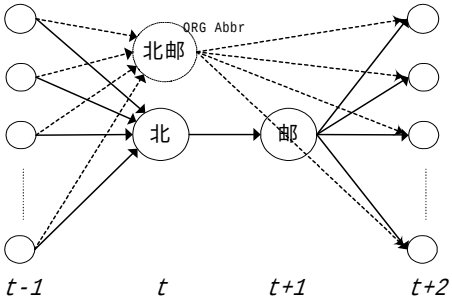


Figure 3.2. State transition in the lattice with the identification of ORG abbreviations.

To sum up, given an identified organization name $A = w_1w_2\ldots w_N$, the score of a candidate

abbreviation $J_1^{\hat{N}}$ (where $\hat{N}$ is the number of words (or characters)) is calculated as follows:

$$\begin{aligned} &Score(J_1^{\hat{N}} \mid ORG \text{ abbr}) \\ &\cong \max(P(J_1^N \mid Contextual \text{ Model}), \quad \alpha \times P(w_1 w_2 \cdots w_N \mid ORG) + (1-\alpha) \times P(J_1^N \mid Contextual \text{ Model})) \end{aligned} \tag{9}$$

where α is set to be 0.5. After the abbreviation candidates are generated, they will be added into the lattice for search.

5. Experiments

5.1 Evaluation Measures

We conducted evaluations in terms of the precision (P) and recall (R):

$$P = \frac{number \quad of \quad correctly \quad identified \quad NE}{number \quad of \quad identified \quad NE}, \tag{10}$$

$$R = \frac{number \quad of \quad correctly \quad identified \quad NE}{number \quad of \quad all \quad NE}. \tag{11}$$

There is one difference between Multilingual Entity Task (MET) evaluation and our evaluation. Nested NEs are evaluated in our system, whereas they are not in MET.

5.2 Data Sets

5.2.1 Training Data

The training corpus was taken from the People's Daily [year 1997 and year 1998]. The annotated training data set, parsed using NLPWin, contained 1,152,676 sentences (90,427k bytes). The training data set contained noises for two reasons. First, the NE guidelines used by NLPWin are slightly different from the ones we used. For example, in our output[7] of NLPWin, 北京市 (Beijing City) was tagged as <LOC>北京</LOC> 市, while 北京市 was tagged as LOC according to our guidelines. Second, there were errors in the parsing results. Therefore, we utilized 18 rules to correct the data. One of these rules is *LN LocationKeyword* → *LN*, which denotes that a location name and an adjacent location keyword are united into a location name. The following table shows some differences between parsing results and correct annotations according to our guidelines:

[7] In fact, NLPWin has many output settings.

Table 4. NLPWin parsing results and correct annotations according to our guidelines.

Examples	Corresponding English	Parsing results	Correct annotations according to our guidelines
江总书记	Secretary-General Jiang	<PER>江总书记</PER>	<PER>江</PER> 总书记
小徐	Xiao Xu	<PER>小徐<PER>	小 <PER>徐<PER>
四川 省	Sichuan Province	<LOC>四川</LOC> 省	<LOC>四川 省</LOC>
新华社	Xinhua News Agency	<LOC>新华社</LOC>	<ORG>新华 社</ORG>
联合国	The United Nations	<LOC>联合国</LOC>	<ORG>联合国</ORG>
卫生部	Ministry of Sanitation	卫生部	<ORG>卫生 部</ORG>

The statistics of the training data are shown in Table 5.

Table 5. Statistics of the Training Data.

Entity		Number of Word Tokens	
		Year 1997	Year 1998
Person	PER1	2,459	1,863
	PER2	48,404	46,141
	PER3	126,384	115,057
	FN	81,885	82,474
Locations (whole names)		376,126	354,317
Abbreviations of Locations		21,304	17,412
Organizations		122,288	125,711
Personal Titles		67,537	59,879
Speech-act Verbs		87,602	83,930
Location Keywords		49,767	53,469
Organization Keywords		115,447	117,423

5.2.2 Test Data

We developed a large open test data based on our guidelines[8]. As shown in Table 6, the data set, which was balanced in terms of domain, style and time, contained approximately half a million Chinese characters. The test set contained 11,844 sentences, 49.84% of which contain at least one NE token.

[8] One difference between our guidelines and those of MET is that nested persons and location names in organizations are tagged according to our guidelines.

Table 6. Statistics[9] of the Test Data.

ID	Domain	Number of NE Tokens			Data Size (Byte)
		PER	LOC	ORG	
1	Army	65	203	30	19k
2	Computer	62	160	134	59k
3	Culture	549	672	81	138k
4	Economy	154	824	354	108k
5	Entertainment	665	617	143	104k
6	Literature	458	715	131	96k
7	Nation	450	1195	251	101k
8	People	1134	913	400	116k
9	Politics	510	1147	214	122k
10	Science	148	206	81	60k
11	Sports	733	1194	623	114k
	Total	4928	7846	2442	1037k

Note that the open-test data set was much larger than the MET test data set (the numbers of PERs, LOCs, and ORGs were 174, 750, and 377, respectively). The numbers of abbreviations of PERs, LOCs, and ORGs in the open-test data set were 367, 729, and 475, respectively.

5.3 Baseline NLPWin Performance

We conducted a baseline experiment, which consisted of two steps: parsing the test data using NLPWin; correcting the errors according to the rules. The performance achieved is shown in Table 7.

Table 7. Baseline NLPWin Performance.

NE	P (%)	R (%)
PER	61.05	75.26
LOC	78.14	71.57
ORG	68.29	31.50
Total	70.07	66.08

[9] The statistics reported here are slightly different from those reported earlier (Sun, Gao, *et al.*, 2002) because we checked the accuracy and consistency of the test data again for our experiments.

5.4 Experimental Results

In order to investigate the contribution of the unified framework, heuristic information and the identification of NE abbreviations, the following experiments were conducted using our NE identification system:

(1) Experiments 1, 2 and 3 examined the contribution of the heuristics and unified framework.

(2) Experiments 4, 5 and 6 tested the performance of the system using our method of NE abbreviations identification.

(3) Experiment 7 compared the performance of identifying whole NEs and that of identifying NE abbreviations.

5.4.1 Experiments 1, 2 and 3: The contribution of the heuristics and unified framework

Experiment 1 was performed to examine the performance of a basic class-based model, in which no heuristic information was employed in the decoder in the unified framework. Experiment 2 examined the performance of a traditional method, which consisted of two separate steps: segmenting the sentence and recognizing NEs. In the segmentation step, we searched for the word with the maximal length in the lexicon to split the input character string[10]. Heuristic information was employed in this experiment. Experiment 3 investigated the performance of the unified framework, where the unified framework and heuristic information were adopted.

A comparison of the results of Experiment 1 and Experiment 3, which aims to show the contribution of heuristic information, is shown in Table 8. A comparison of the results of Experiment 2 and Experiment 3, which aims to show the contribution of the unified method, is shown in Table 9.

Table 8. Results of Experiment 1 and Experiment 3

NE	P (%)		R (%)	
	Exp.1[11]	Exp.3	Exp.1	Exp.3
PER	66.52	81.24	77.82	83.66
LOC	88.08	86.89	77.80	78.65
ORG	37.12	75.90	45.58	47.58
All Three	70.42	83.57	72.63	75.29

[10] Every Chinese character in the input string, which can be seen as a single character word, is also added into the segmentation lattice. We save the minimal length segmentation in the lattice so that the character-based model (for PER) can be applied.

[11] Exp.1 means the results of Experiment 1 and so on

Table 9. Results of Experiment 2 and Experiment 3

NE	P (%)		R (%)	
	Exp.2	Exp.3	Exp.2	Exp.3
PER	80.17	81.24	82.22	83.66
LOC	86.33	86.89	78.20	78.65
ORG	73.46	75.90	46.60	47.58
All Three	82.61	83.57	74.43	75.29

From Table 8, we observed that after the introduction of heuristic information, the precision of PER increased from 66.52% to 81.24%, that of ORG from 37.12% to 75.90%. We also noticed that the recall of PER from 77.82% to 83.66%, that of ORG from 45.58% to 47.58%. Therefore, the heuristic information was an important knowledge resource for recognizing NEs.

From Table 9, we find that the precision and recall of PER, LOC and ORG all improved as a result of the combining word segmentation with NE identification. For instance, the precision of PER increased from 80.17% to 81.24%, and the recall from 82.22% to 83.66%. Therefore, we can conclude that the unified framework for NE identification was a more effective method.

5.4.2 Experiments 4, 5 and 6: Performance achieved when modeling abbreviations of personal, location and organization names

In order to examine the performance of our methods of identifying NE abbreviations, Experiments 4, 5 and 6 were conducted. Experiment 4 examined the effectiveness of modeling the abbreviations of personal names. Experiment 5 incorporated modeling of the abbreviations of location names based on Experiment 4, and Experiment 6 integrated modeling of the abbreviations of organization names based on Experiment 5. The results are shown in Table 10.

Table 10. Results of Experiments 3, 4, 5 and 6.

NE	P (%)				R (%)			
	Exp.3	Exp.4	Exp.5	Exp.6	Exp.3	Exp.4	Exp.5	Exp.6
PER	81.24	79.64	79.77	79.78	**83.66**	89.31	89.31	**89.29**
LOC	86.89	87.04	85.76	86.02	**78.65**	78.61	84.91	**84.87**
ORG	75.90	75.97	75.95	76.79	**47.58**	49.50	47.71	**59.75**
All Three	83.57	82.95	82.52	82.59	75.29	77.08	80.36	82.27

It can be seen that the recall of PER, LOC and ORG showed distinct improvement. For example, the recalls increased from 83.66%, 78.65%, 47.68% to 89.31%, 84.91%, 59.75%, respectively. However, we also find that the precision of PER and LOC decreased a little (PER: from 81.24% to 79.78%; LOC: from 86.89% to 86.02%). The reason was that the precision of identifying NE abbreviations was lower than that of identifying whole NE names in general. It is difficult to decide whether a Chinese character is an NE, a single Chinese character, or a part of an ordinary word. For example, the Chinese character "中" can be an abbreviation of LOC (中国 'China'), a single Chinese character, or a part of a word (e.g., 中间 'in the middle of'). Although the precisions decreased a little, on the whole, we can conclude that the performance of NE identification improved after the models of NE abbreviations were constructed.

5.4.3 Experiment 7: Comparing the performance of identifying whole NEs and NE abbreviations

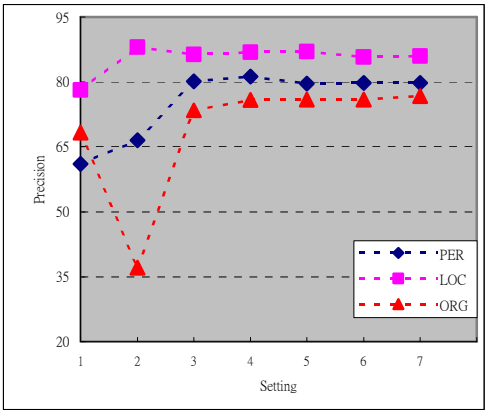
In order to compare the performance of identifying whole NE names with that of identifying NE abbreviations in more detail, we show results in Table 11. We can observe that the performance (precision and recall) of identifying NE abbreviations was about 10% lower than that of identifying whole NE names, in general.

Table 11. Results of identifying whole NEs and NE abbreviations.

NE	NE Abbreviations		Whole NEs	
	P(%)	R(%)	P(%)	R(%)
PER	61.72	78.20	81.45	90.18
LOC	67.96	71.88	88.02	86.20
ORG	78.03	65.05	76.46	58.46
All Three	68.63	71.29	84.28	83.53

5.4.4 Summary of Experiments

Figures 4 and 5 give a brief summary of the experiments in different settings.



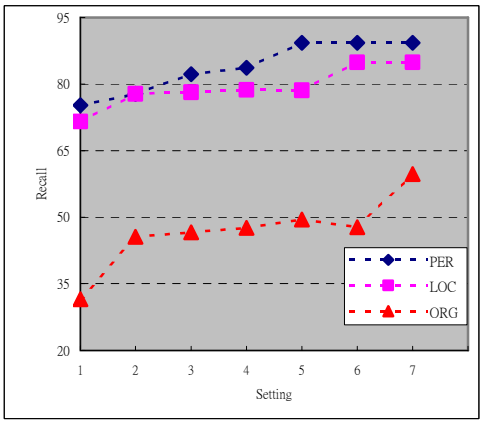


Figure 4. Precision in different settings. ***Figure 5. Recall in different settings.***

1. Results of NLPWin parsing. 2. Results of the baseline class-based model.
3. Performance of the segmentation-identification separate method.
4. Performance of integrating heuristic information and adopting the unified framework.
5. Performance of modeling for the abbreviations of personal names.
6. Performance of modeling for the abbreviations of location names.
7. Performance of modeling for the abbreviations of organization names

From these two figures, we can see that: (1) the results of the baseline class-based LM are better than those of NLPWin; (2) the distinct improvement was achieved by employing heuristic information; (3) the precision and recall rates improved when we adopted the unified framework; (4) modeling for NE abbreviations distinctly improved the recall of all NEs (as shown in Figure 5) with only a trivial decrease in precision.

5.5 Error Analysis

We classify the errors of the system into two types: Error 1 (a boundary error) and Error 2 (a class tag error) as shown in Figure 6. The distribution of these two kinds of errors is shown in Table 12.

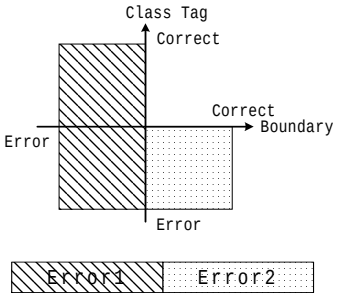


Figure 6. Two kinds of errors.

Table 12. Distribution of two kinds of errors.

NE	Error 1 (%)	Error 2 (%)
PER	87.71	12.29
LOC	96.86	3.14
ORG	97.73	2.27
All Three	93.14	6.86

From Table 12, we observe that boundary errors accounted for a large percentage of these two kinds of errors in Chinese NE identification. The errors of three kinds of NEs will be further shown in Sections 5.5.1, 5.5.2, and 5.5.3. For some errors, the solutions are given. We also indicate some cases that could not be perfectly handled in our method.

5.5.1 PER Errors

The major PER[12] errors are shown in Table 13:

Table 13. PER Errors

Cases	Identified results	Standard	Transliteration/Translation
a. Personal names that contain content word	厉 有为 高峰	厉 有为 高峰	Li Youwei Gao Feng
b. Location names that have nested personal name	胡志明 市	胡志明市	Ho Chi Minh City
c. Japanese names	藤 井 美 子	藤井 美子	Tengjing Meizi
d. Aliases of personal names	东东 娇娇	东东 娇娇	Dongdong Jiaojiao
e. Transliterated personal names and transliterated location names that cannot be distinguished	阿贾克斯 密歇根	阿贾克斯 密歇根	Ajax Michigan

We will try to deal with some of above errors in our future work. Case (b) can be handled

[12] PER LOC ORG

by adopting a nested model; Case (c) can be dealt with by constructing a model of Japanese names. Cases (a), (d), and (e) can only be partially dealt with by refining the contextual model in our framework. However, our current method does not provide a sound solution for Case (d), namely, aliases of personal names.

5.5.2 LOC Errors

LOC errors are shown in Table 14.

Table 14. LOC Errors

Cases	Identified results	Standard	Transliteration/Translation
a. Part of a sequence in LOC and the right context that can be combined into a word	深圳 市郊 布吉 河边 合浦 县城	深圳 市 郊 布吉 河 边 合浦 县 城	Suburb of Shenzhen City Buji River side Hepu county
b. Some abbreviations, which are common content words	日 (日本) 中 (中国) 港 (香港)	日 中 港	Japan China Hongkong

One reason for the errors in Case (a) was that there were noises of this kind in the training data. As for Case (b), the model of the abbreviations of location name can identify many abbreviations. However, there were a few errors of identification because location abbreviations may be common words, e.g., “中”.

5.5.3 ORG Errors

ORG errors are shown in Table 15.

Table 15. ORG Errors

Cases	Identified results	Standard	Transliteration/Translation
a. Organization names that contain other organizaiton names	联合国 维和部队	联合国 维和部队	The UN Peacekeeping Missions
	联合国 难民署	联合国 难民署	The UN Refugee Office
	新华社 澳门分社	新华社 澳门分社	Branch office of the Xinhua News Agency in Macao
b. ORGs that contain numbers, dates or English characters	八一队	八一队	August 1st Team
	六九一团	六九一团	691th Regiment
	２０世纪福克斯公司	２０世纪福克斯公司	Twentieth Century Fox
	ＮＨＫ研修中心	ＮＨＫ研修中心	NHK Research Center

Case (a) can be partly handled by refining the model of organization names. However, our system may fail to handle an instance like "新华 社 澳门 分社" because it does not have enough information to detect the right boundary of the organization name. In addition, our class-based LM cannot successfully deal with Case (b) at present.

In addition, although the language model method was adopted to identify the abbreviations of organization names, there were still some abbreviations of organization names that were not identified. One reason is that some abbreviations are not covered in the above patterns. The other reason is that the score function in Equation 9 is just an empirical formula and needs to be improved.

5.6 Evaluation with MET2 Data

We also evaluated our system (nested NEs were not numbered in this case) using the MET2 test data and compared the performance achieved with that of two public systems[13] (the NTU system and KRDL system). As shown in Table 16, our system outperformed the NTU system. Our system was also better than the KRDL system for PERs, but the performance for LOCs and ORGs was worse than that of the KRDL system. The possible reasons are: (1) Our NE definitions are slightly different from those of MET2. (2) The model is estimated using a general domain corpus, which is quite different from the domain of MET2 data. (3) An NE dictionary is not utilized in our system.

Table 16. Results using MET2 Data.

NE	**Our System**		NTU Results		Kent Ridge Digital Labs Results (KRDL)	
	P (%)	**R (%)**	P (%)	R (%)	P (%)	R (%)
PER	**77.51**	**93.10**	74	91	66	92
LOC	**86.52**	**87.20**	69	78	89	91
ORG	**88.75**	**77.25**	85	78	89	88

6. Conclusions & Future work

We have presented a method of Chinese NE identification using a class-based language model, which consists of two sub-models: a set of entity models and a contextual model. Our method provides a unified framework, in which it is easy to incorporate Chinese word segmentation

[13] Available at http://www.itl.nist.gov/iad/894.02/related_projects/muc/proceedings/ne_chinese_score_report.html.

and NE identification. As has been demonstrated, our unified method performs better than traditional methods. We have also presented our method of identifying NE abbreviations. The language model method has several advantages over rule-based ones. First, it can integrate the identification of NE abbreviations into the class-based LM. Secondly, it reduces the labor of developing rules for NE abbreviations. In addition, we have also employed a two-level ORG model so that the nested entities in organization names can be identified.

The achieved precision rates of PER, LOC, ORG on the test data were 79.78%, 86.02%, and 76.79%, respectively, and the achieved recall rates were 89.29%, 84.87%, and 59.75%, respectively.

There are several possible directions of future research. First, since we use a parser to annotate the training set, parsing errors will be an obstacle to further improvement. Therefore, we need to find an effective way to correct the mistakes and perform necessary automatic correction. Secondly, a more delicate model of ORG will be investigated to characterize the features of all kinds of organizations. Thirdly, the current method only utilizes the features in the currently processed sentence, not the global information in the text. For example, suppose that the same NE (e.g., 薄熙来) occurs twice in different sentences in a document. It is possible that the NE will be tagged PER in one sentence but not recognized in the other. This raises a question as to how to construct a model of global information. Furthermore, the model of organization name abbreviations also needs to be improved.

Acknowledgements

We would like to thank Chang-ning Huang, Andi Wu, Hang Li and other colleagues at Microsoft Research for their help. We also thank Lei Zhang for his help. In addition, we thank the three anonymous reviewers for their useful comments.

References

Aberdeen J., Day D., Hirschman L., Robinson P. and Vilain M., "MITRE: Description of the Alembic System Used for MUC-6", *Proceedings of the Sixth Message Understanding Conference*, pp. 141-155, 1995.

Black A., Taylor P. and Caley R., The Festival Speech synthesis system. http://www.cstr.ed.ac.uk/projects/festival/ , 1998.

Black W.J., Rinaldi F. and Mowatt D., "Facile: Description of the NE System Used For MUC-7", *Proceedings of 7th Message Understanding Conference*, 1998.

Black W.J. and Vasilakopoulos A., "Language Independent Named Entity Classification by modified Transformation-based Learning and by Decision Tree Induction", *The 6th Conference on Natural Language Learning*, 2002.

Borthwick. A., "A Maximum Entropy Approach to Named Entity Recognition", PhD Dissertation, 1999.

Bikel D., Schwarta R. and Weischedel R., "An algorithm that learns what's in a name", *Machine Learning Journal Special Issue on Natural Language Learning*, 34, pp. 211-231, 1999.

Brown P. F., DellaPietra V. J., deSouza P. V., Lai J. C., and Mercer R. L., "Class-based n-gram models of natural language", *Computational Linguistics*, 18(4): 467- 479, 1992.

Brill E., "Transform-based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-speech Tagging", *Computational Linguistics*, 21(4): 543-565, 1995.

Carreras X., Màrquez L. and Padró L., "Named Entity Extraction using AdaBoost", *The 6th Conference on Natural Language Learning*, 2002.

Chang J.S., Chen S. D., Zheng Y., Liu X. Z., and Ke S. J., "Large-corpus-based methods for Chinese personal name recognition", *Journal of Chinese Information Processing*, 6(3): 7–15, 1992.

Chen H.H., Ding Y.W., Tsai S.C. and Bian G.W., "Description of the NTU System Used for MET2", *Proceedings of 7th Message Understanding Conference*, 1998.

Chen H.H., Lee J.C., "The Identification of Organization Names in Chinese Texts", *Communication of Chinese and Oriental Languages Information Processing Society*, 4(2): pp. 131-142, 1994 (in Chinese).

Chen, S. F., and Goodman, J., "An empirical study of smoothing techniques for language modeling". *Computer Speech and Language*, 13: 359-394, October 1999.

Chen, Si-Qing., "The automatic identification and recovery of Chinese acronyms", *Studies in the Linguistics Sciences*, 26(1/2): 61–82. 1996.

Chinchor. N., "MUC-7 Named Entity Task Definition Version 3.5". Available by from ftp.muc.saic.com/pub/MUC/MUC7-guidelines, 1997.

Collins M., Singer Y., "Unsupervised models for named entity classification", *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999.

Collins M., "Ranking Algorithms for Named-Entity Extraction: Boosting and the Voted Perceptron", *Proceedings of the 40th Annual Meeting of the ACL*, Philadelphia, pp. 489-496, July 2002.

Florian R., "Named Entity Recognition as a House of Cards: Classifier Stacking", *The 6th Conference on Natural Language Learning*, 2002.

Fukumoto J., Shimohata M., Masui F. and Sasaki M., "Oki Electric Industry: Description of the Oki System as Used for MET-2", *Proceedings of 7th Message Understanding Conference*, 1998.

Gao J., Goodman J., Miao J., "The use of clustering techniques for language modeling – application to Asian languages", *Computational Linguistics and Chinese Language Processing*, Vol. 6, No. 1, pp 27-60.2001.

Gotoh Y., Renals S., "Information extraction from broadcast news", *Philosophical Transactions of the Royal Society of London, series A: Mathematical, Physical and Engineering Sciences*, 2000.

Grishman R., "The NYU System for MUC-6 or Where's the Syntax?", *Proceedings of the MUC-6 workshop*, Washington. November 1995.

Humphreys K., Gaizauskas R., *et al.*, Univ. of Sheffield: "Description of the LaSIE-II System as Used for MUC-7", *Proceedings of 7th Message Understanding Conference*, 1998.

Jansche M., "Named Entity Extraction with Conditional Markov Models and Classifiers", *The 6th Conference on Natural Language Learning*, 2002.

Krupka G. R., Hausman K.. "IsoQuest Inc.: Description of the NetOwlTM Extractor System as Used for MUC-7", *Proceedings of 7th Message Understanding Conference*, 1998.

Kuhn R., Mori. R.D. "A Cache-Based Natural Language Model for Speech Recognition", *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol.12. No. 6. pp 570-583, 1990.

McDonald D., "Internal and External Evidence in the Identification and Semantic Categorization of Proper Names", *Corpus Processing for Lexical Acquisition*. pp. 21-39. MIT Press. Cambridge, MA. 1996.

McNamee P. and Mayfield J., "Entity Extraction without Language-specific Resources", *The 6th Conference on Natural Language Learning*, 2002.

Mikheev A., Grover C. and Moens M., "Description of the LTG System Used for MUC-7", *Proceedings of 7th Message Understanding Conference*, 1998.

Miller S., Crystal M., *et al.*, "BBN: Description of the SIFT System as Used for MUC-7", *Proceedings of 7th Message Understanding Conference*, 1998.

Palmer D., Day D.S., "A Statistical Profile of the Named Entity Task", *Proceedings of the Fifth Conference on Applied Natural Language Processing*, Washington, D.C., March 31- April 3, 1997.

Sang E.T.K., "Memory-Based Named Entity Recognition", *The 6th Conference on Natural Language Learning*. 2002.

Sekine S., Grishman R. and Shinou H., "A decision tree method for finding and classifying names in Japanese texts", *Proceedings of the Sixth Workshop on Very Large Corpora*, Montreal, Canada, 1998.

Sproat R., Black A., Chen S., *et al.*, "Normalization of non-standard words", *Computer Speech and Language*, 15(3): 287-333, 2001.

Sproat R., Chilin Shih. "Corpus-Based Methods in Chinese Morphology and Phonology", 2001 LSA Institute, Santa Barbara.

Sun J., Gao J., Zhang L., Zhou M., Huang C., "Chinese Named Entity Identification Using Class-based Language Model". *Proceeding of the 19th International Conference on Computational Linguistics*, pp.967-973, 2002.

Sun M.S., Huang C.N., Gao H.Y., Fang J., "Identifying Chinese Names in Unrestricted Texts", *Communications of COLIPS*, Vol 4, No. 2, pp. 113-122, 1994 (in Chinese)

Takeuchi K., Collier N., "Use of Support Vector Machines in Extended Named Entity Recognition", *The 6th Conference on Natural Language Learning*, 2002.

Toole J., "A Hybrid Approach to the Identification and Expansion of Abbreviations", *RIAO'2000 Proceedings*, 2000

Tsukamoto K., Mitsuishi Y., Sassano M., "Learning with Multiple Stacking for Named Entity Recognition", *The 6th Conference on Natural Language Learning*. 2002.

Viterbi A. J., "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm", *IEEE Transactions on Information Theory*, IT(13). pp. 260-269, April 1967.

Wu D.K., Ngai G., *et al*., "Boosting for Named Entity Recognition", *The 6th Conference on Natural Language Learning*, 2002.

Yu S.H., Bai S.H. and Wu P., "Description of the Kent Ridge Digital Labs System Used for MUC-7", *Proceedings of 7th Message Understanding Conference*, 1998.

Zhang L., "Study on Chinese Proofreading Oriented Language Modeling", PhD Dissertation, 2001.

Zhou G. Su J., "Named Entity Recognition using an HMM-based Chunk Tagger", *Proceedings of the 40th Annual Meeting of the ACL*, Philadelphia, pp. 473-480, July 2000.