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 $P(w_n|w_1...w_{n-1})$. 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:

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

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:

$$\begin{aligned}
&P(PER \mid BOS) \times P(PER3 \mid PER) \times P(\text{周恩来} \mid PER3) \\
&\times P(PT \mid BOS, PER) \times P(\text{总理} \mid PT) \\
&\times P(\text{是} \mid PER, PT) \times P(\text{我们} \mid PT, \text{是}) \times P(\text{的} \mid \text{是}, \text{我们}) \\
&\times P(\text{好} \mid \text{我们}, \text{的}) \times P(\text{总理} \mid \text{的}, \text{好}) \times P(\text{。} \mid \text{好}, \text{总理}) \times P(EOS \mid \text{总理}, \text{。})
\end{aligned}$$

where $P(\text{周恩来} \mid PER3)$ 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.

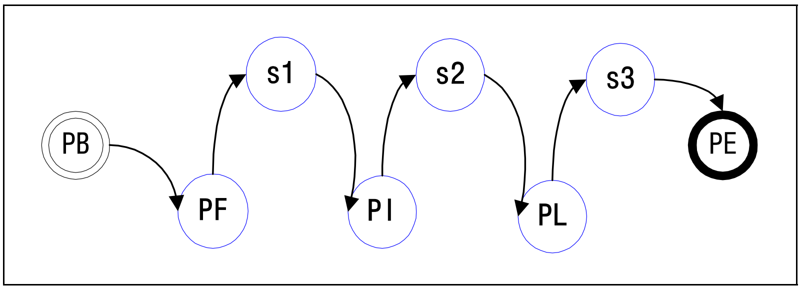


Figure 1. The generation of the sequence* s₁ s₂ s₃ *given the PER3 class.

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

$$\begin{aligned} & P(s_1 s_2 s_3 \mid c = PER3) \\ & = P(PF \mid PER3, PB) \times P(s_1 \mid PER3, PB, PF) \\ & \times P(PI \mid PER3, PF, s_1) \times P(s_2 \mid PER3, s_1, PI) \\ & \times P(PL \mid PER3, PI, s_2) \times P(s_3 \mid PER3, s_2, PL) \\ & \times P(PE \mid PER3, PL, s_3) \end{aligned} \tag{4}$$

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

$$\begin{aligned} & P(\text{周恩来} \mid PER3) \\ & = P(PF \mid PER3, PB) \times P(\text{周} \mid PER3, PB, PF) \\ & \times P(PI \mid PER3, PF, \text{周}) \times P(\text{恩} \mid PER3, \text{周}, PI) \\ & \times P(PL \mid PER3, PI, \text{恩}) \times P(\text{来} \mid PER3, \text{恩}, PL) \\ & \times P(PE \mid PER3, PL, \text{来}) \end{aligned}$$

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:

$\underset{\text{PER}}{\underline{\textit{周 恩 来}}}$ $\underset{\text{PT}}{\underline{\textit{总 理}}}$ *是 我 们 的 好 总 理 。*

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.

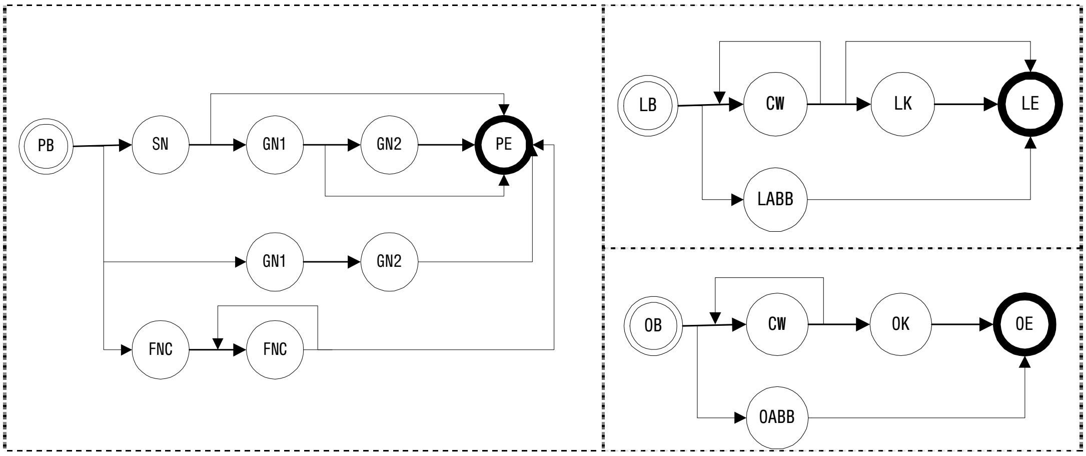


Figure 2. The grammar of PER, LOC and ORG candidates.

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):

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

[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	$s_{11}s_{21}$ … $s_{11}s_{21}\ldots s_{N1}$	北京 邮电 大学 → 北 邮 … 北京 邮电 大学 → 北 邮 大	*s_{ij} denotes the jth character of the ith word of A*
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 w_1w_2	苹果 电脑 公司 → 苹果 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

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

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.,

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

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).

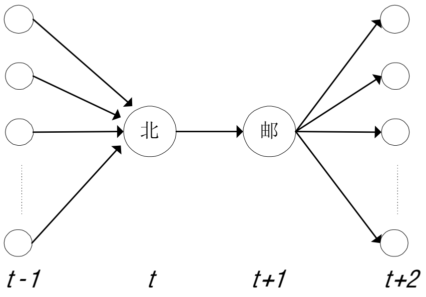


Figure 3.1. State transition in the lattice without the identification of ORG abbreviations.

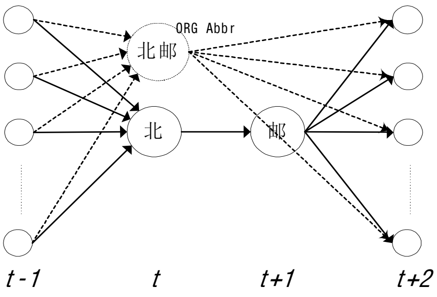


Figure 3.2. State transition in the lattice with the identification of ORG abbreviations.

To sum up, given an identified organization name $A = w_1 w_2 \ldots w_N$, 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
		PER	LOC	ORG	(Byte)
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

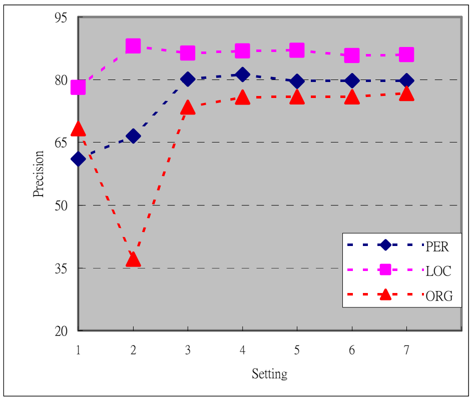
In order to compare the performance of identifying whole NE names with that of identifying NE abbreviations in more detail, we show results in Table 11. We can observe that the performance (precision and recall) of identifying NE abbreviations was about 10% lower than that of identifying whole NE names,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



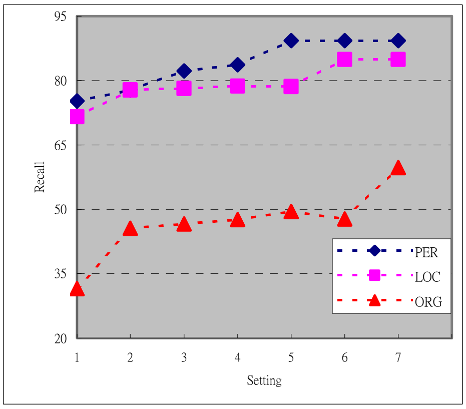


Figure 4. Precision in different settings. ***Figure 5. Recall in different settings.***

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.

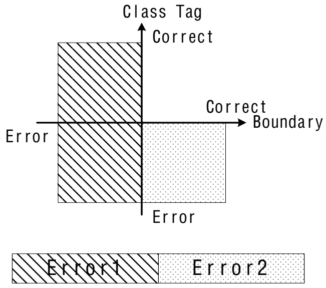


Figure 6. Two kinds of errors.

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.

Computational Linguistics and Chinese Language Processing
Vol. 8, No. 2 , August 2003, pp. 29-60

Chinese Named Entity Recognition Using Role Model[1]

Hua-Ping ZHANG[*], Qun LIU[*+], Hong-Kui YU[*],

Xue-Qi CHENG[*], Shuo BAI[*]

Abstract

This paper presents a stochastic model to tackle the problem of Chinese named entity recognition. In this research, we unify component tokens of named entity and their contexts into a generalized role set, which is like part-of-speech (POS). The probabilities of role emission and transition are acquired after machine learning on a role-labeled data set, which is transformed from a hand-corrected corpus after word segmentation and POS tagging are performed. Given an original string, role Viterbi tagging is employed on tokens segmented in the initial process. Then named entities are identified and classified through maximum matching on the best role sequence. In addition, named entity recognition using role model is incorporated along with the unified class-based bigram model for word segmentation. Thus, named entity candidates can be further selected in the final process of Chinese lexical analysis. Various evaluations conducted using one

[1] This research is supported by the national 973 fundamental research program under grants number G1998030507-4 and G1998030510 and the ICT Youth Fund under contract number 20026180-23.
Hua-Ping Zhang (Kevin Zhang): born in February, 1978, a PhD candidate in the Institute of Computing Technology (ICT), Chinese Academy of Sciences. His research interests include computational linguistics, Chinese natural language processing and information extraction.
Qun Liu: born in October 1966, an associate professor at ICT and a PhD candidate at Peking University. His research interests include machine translation, computational linguistics and Chinese natural language processing.
Hong-KuiYu: born in November 1978, a visiting student at ICT from Beijing University of Chemical Technology. His research interests include natural language processing and named entity extraction.
Xue-Qi Cheng: born in 1971, an associate professor and director of the software division of ICT. His research fields include computational linguistics, network and information security.
Shuo Bai: born in March 1956, a professor, PhD supervisor and principal scientist of the software division of ICT. His research fields include computational linguistics, network and information security.

[*] Software Division, Institute of Computing Technology, The Chinese Academy of Sciences, Beijing, P.R. China, 100080 Email: zhanghp@ software.ict.ac.cn

[+] Institute of Computational Linguistics, Peking University, Beijing, P.R. China, 100871

month of news from the People's Daily and MET-2 data set demonstrate that the role modeled can achieve competitive performance in Chinese named entity recognition. We then survey the relationship between named entity recognition and Chinese lexical analysis via experiments on a 1,105,611-word corpus using comparative cases. It was found that: on one hand, Chinese named entity recognition substantially contributes to the performance of lexical analysis; on the other hand, the subsequent process of word segmentation greatly improves the precision of Chinese named entity recognition. We have applied the role model to named entity identification in our Chinese lexical analysis system, ICTCLAS, which is free software and available at the Open Platform of Chinese NLP (www.nlp.org.cn). ICTCLAS ranked first with 97.58% in word segmentation precision in a recent official evaluation, which was held by the National 973 Fundamental Research Program of China.

Keywords: Chinese named entity recognition, word segmentation, role model, ICTCLAS

1. Introduction

Named entities (NE) are broadly distributed in original texts from many domains, especially politics, sports, and economics. NE can answer for us many questions like "who", "where", "when", "what", "how much", and "how long". NE recognition (NER) is an essential process widely required in natural language understanding and many other text-based applications, such as question answering, information retrieval, and information extraction.

NER is also an important subtask of the Multilingual Entity Task (MET), which was established in the spring of 1996 and run in conjunction with the Message Understanding Conference (MUC). The entities defined in MET are divided into three categories: entities [organizations (ORG), persons (PER), locations (LOC)], times (dates and times), and quantities (monetary values and percentages) [N.A.Chinchor, 1998]. As for NE in Chinese, we further divide PER into two sub-classes: Chinese PER and transliterated PER on the basis of their distinct features. Similarly, LOC is split into Chinese LOC and transliterated LOC. In this work, we only focus on those more difficult but commonly used categories: PER, LOC and ORG. Other NE such as times (TIME) and quantities (QUAN), in a border sense, can be recognized simply via finite state automata.

Chinese NER has not been researched intensively till now, while English NER has received much attention. Because of the inherent difference between the two languages, Chinese NER is more complicated and difficult. Approaches that are successfully applied in English cannot be simply extended to cope with the problems of Chinese NER. Unlike Western languages such as English and Spanish, there are no delimiters to mark word boundaries and no explicit definitions of words in Chinese. Generally speaking, Chinese NER has two sub-tasks: locating the string of NE and identifying its category. NER is an intermediate step in Chinese word segmentation, and token sequences greatly influence the process of NER. Take "孙家正在工作" (pronunciation: "sun jia zheng zai gong zuo") as an example. "孙家正"(Sun Jia-Zheng) in "孙家正/在/工作/" (Sun Jia-Zheng is working) can be recognized as a Chinese PER, and "孙家" is also an ORG in "孙家/正在/工作/"(The Sun family is working). Here, "孙家正在" contains some ambiguous cases: "孙家正"(Sun Jia-Zheng, a PER name), "孙家" (the Sun family, an ORG name), and "正在" (just now, a common word). Such problems are caused by Chinese character strings without word segmentation, and they are hard to solve in the process of NER. Sun *et al.* [2002] points out that "Chinese NE identification and word segmentation are interactional in nature."

In this paper, we present a unified statistical approach, namely, a role model, to recognize Chinese NE. Here, roles are defined as some special token classes, including an NE component and its neighboring and remote contexts. The probabilities of role emission and transition in the NER model are trained on modified corpus, whose tags are converted from POS to roles according to the definition. To some extent, roles are POS-like tags. As in POS tagging, we can tag the global optimal role sequence to obtain tokens using the Viterbi algorithm. NE candidates can be recognized through pattern matching on the role sequence, not the original string or token sequence. NE candidates with credible probability are, furthermore, added into a class-based bigram model for Chinese word segmentation. In the generalized frame, any out-of-vocabulary NE is handled in the same way as known words listed in the segmentation lexicon. And improper NE candidates are eliminated if they fail in compete with other words, while correctly recognized NE are further confirmed in comparison with other cases. Thus, Chinese word segmentation improves the precision of NER. Moreover, NER using the role model optimizes the segmentation result, especially in unknown words identification. A survey on the relationship between NER and word segmentation supports this conclusion. NER evaluation was conducted on a large corpus from MET-2 and the People's Daily. The precisions of PER, LOC, ORG on the 1,105,611-word news corpus were 94.90%, 79.75% and 76.06%, respectively; and the recall rate were is 95.88%, 95.23% and 89.76%, respectively.

This paper is organized as follows: Section 2 overviews problems in Chinese NER, and the next section details our approach using the role model. The class-based segmentation model integrated with NE candidates is described in Section 4. Section 5 presents a comparison between the role model and previous works. An NER evaluation and survey of segmentation and NER is reported in Section 6. The last section gives our conclusions.

2. Problems in Chinese NER

NE appear frequently in real texts. After surveying a Chinese news corpus with 7,198,387 words from the People's Daily (Jan.1-Jun.30, 1998), we found that the percentage of NE was 10.58%. The distributions of various NE is given in Table 1.

Table 1. Distributions of NE in a Chinese news corpus from the People's Daily (Jan.1-Jun.30, 1998).

NE	Frequency	Percentage in NE (%)	Percentage in corpus (%)
Chinese PER	97,522	12.49	1.35
Transliterated PER	24,219	3.10	0.34
PER	121,741	15.59	1.69
Chinese LOC	157,083	20.11	2.18
Transliterated LOC	27,921	3.57	0.39
LOC	185,004	23.69	2.57
ORG	78,689	10.07	1.09
TIME	127,545	16.33	1.77
QUAN	268,063	34.43	3.72
Total	781,042	100.00	10.85

As mentioned above, Chinese sentences are made up of character strings, not word sequences. A single sentence often has many different tokenizations. In order to reduce the complexity and be more specific, it would be better to conduct NER on tokens after word segmentation rather than on an original sentence. However, word segmentation cannot achieve good performance without unknown word detection in the process of NER. Due to this a problem, Chinese NER has special difficulties.

Firstly, an NE component may be a known word inside the vocabulary; such as "王国"(*kingdom*) in the PER "王国维" (*Wang Guo-Wei*) or "联想"(*to associate*) in the ORG "北京联想集团"(*Beijing Legend Group*). It's difficult to make decisions between common words and parts of NE. As far as we know, this has not been considered previously. Thus, NE containing known words are very likely to be missed in the final recognition results.

The second problem is ambiguity, and it is almost impossible to be solved only in NER. Ambiguities in NER can be categorized into segmentation and classification ambiguities. "孙家正在工作" (pronunciation: "sun jia zheng zai gong zuo"), presented in the Introduction, has segmentation ambiguity: "孙家正/在"(Sun Jia-Zheng is at …) and "孙家/正在" (The Sun family is doing something). Classification ambiguity means that an NE may be have one more class even if its position in the string is properly located. For instance, in the sentence "吕梁的特点是穷"(The characteristic of Lv Liang is poverty), it is not difficult to detect the NE "吕梁"(Lv Liang). However, we cannot judge whether this NE is a Chinese PER name or a Chinese LOC name while considering the single sentence without any additional information, Moreover, NE tends to stick to its neighboring contexts. There are also two types: head components of NE binding with their left neighboring tokens and those tail binding with their right tokens. This greatly increases the complexity of Chinese NER and word segmentation. In Figure 1, “内塔尼亚胡”(*Netanyahu*) in “克林顿对内塔尼亚胡说”(pronunciation: “ke lin dun dui nei ta ni ya hu shuo”) is a transliterated PER. However its left token “对”(to) sticks to the head component “内”(Inside) and forms a common word “对内”(to one’s own side) ; similarly, the tail component “胡”(to) and right neighbor “说”(to say) become a common word, “胡说” (nonsense). Therefore, the most possible segmentation result would not be “克林顿/对/内塔尼亚胡/说”(*Clinton* said to *Netanyahu*) but “克林顿/对内/塔尼亚/胡说”(*Clinton* points to his own side and *Tanya* talks nonsense.), and then not “内塔尼亚胡”(*Netanyahu*) but “塔尼亚”(*Tanya*) would be recognized as a PER. We can draw the conclusion that such a problem not only reduces the recall rate of Chinese NER, but also influences the segmentation of normal neighboring words like “对”(to) and “说”(to say). Appendix I provides more Chinese PER cases that were extracted from our corpus.

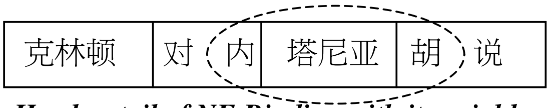


Figure 1: Head or tail of NE Binding with its neighbours.

1. Words within a solid square are tokens.
2. “内塔尼亚胡”(Netanyahu) inside the dashed ellipse is a PER, and its head and tail stick to their neighbouring tokens.

3. Role model for Chinese NER

Considering the problems encountered in NER, we will introduce a role model to unify all possible NE and sentences. Our motivation is to classify similar tokens into some role categories according to their linguistic features, to assign a corresponding role to each token automatically, and to then perform NER based on the role sequence.

3.1 What Are Roles Like?

Given a sentence like “孔泉说，江泽民主席今年访美期间向布什总统发出了邀请”(*Kong Quan* said that President *Jiang Ze-Min* had invited President *Bush* while visiting the USA), the tokenization result without considering NER would be “孔/泉/说/，/江/泽/民/主席/今年/访/美/期间/向/布/什/总统/发出/了/邀请”(shown in Figure 2a). Here “孔泉”(Kong Quan) and “江泽民”(Jiang Ze-Min) are Chinese PERs, while “美”(USA) is an LOC and “布什”(Bush) is a transliterated PER.

孔泉	说	，	***江 泽民***	主席	今年	访	***美***	期间	向	***布什***	总统	发出	了	邀请

Figure 2a: Token sequence without detecting Chinese NE, which is in bold type and italics.

(*Kong Quan* said that President *Jiang Ze-Min* had invited President *Bush* while visiting the USA).

When we consider the generation of NE, we find that different tokens play different roles in sentences. Here, the term “role” is referred to a generalized class of tokens with similar functions in forming a NE and its context. For instance, “曾” (pronunciation: “zeng”) and “张” (pronunciation: “zhang”) can both act as common Chinese surnames, while both “说”(to speak) and “主席”(chairman) may be right neighboring tokens following PER names. Relevant roles for the above example are explained in Figure 2b.

Tokens	**Role played in the token sequence**

孔(pronunciation: “kong”);
江(pronunciation: “jiang”)
	Surname of Chinese NER

泉(pronunciation: “quan”)	Given name with a single Hanzi (Chinese character)
泽(pronunciation: “ze”)	Head character of 2-Hanzi given name
民(pronunciation: “min”)	Tail character of 2-Hanzi given name

布(pronunciation: “bu”);
什(pronunciation: “shi”)
	Component of transliterated PER

说(say);主席(chairman); 总统(president)	Right neighboring token following PER
，(comma); 向(toward)	Left neighboring token in front of PER
美(USA)	Component of LOC
访(visit)	Left neighboring token in front of LOC
期间(period)	Right neighboring token following LOC
今年(this year); 发出(put forward);了 (have); 邀请(invite)	Remote context, which distance is more than one word. from NE

Figure 2b: Relevant roles of various tokens in

“孔/泉/说/，/江/泽/民/主席/今年/访/美/期间/向/布/什/总统/发出/了/邀请”
(*Kong Quan* said that President *Jiang Ze-Min* had invited President *Bush* while visiting the USA).

If NE is identified in a sentence, it is easy to extract the roles listed above through simple analysis on NE and other tokens. On the other hand, if we get the role sequence, can NE be identified properly? The answer to this question is clearly yes. Take a token-role segment like “孔/ Surname 泉/Given-name 说/context ，/context 江/Surname 泽/first component of given-name 民/second component of given-name 主席/context” as an example. If we either know that “江”(pronunciation: “jiang”) is a surname while “泽”(pronunciation: “ze”) and “民”

(pronunciation: “min”) are components of the given name, or if we know that “，”(comma) and “主席”(chairman) are its left and right neighbours, then “江泽民”(Jiang Ze-Min) can be identified as a PER. Similarly, “孔泉”(Kong Quan) and “布什”(Bush) can be recognized as PERs , and at the same time, “美”(an abbreviation of USA in Chinese) can be picked up as an LOC..

In other words, the NER problem can be solved with the correct role sequence on tokens, and many intricate character strings can be avoided. However, the question when applying the role model to NER is: “How can we define roles and assign roles to the tokens automatically?”

3.2 What Roles Are Defined?

To some extent, a role is POS-like, and a role set can be viewed as a token tag collection. However, a POS tag is defined according to the part-of-speech of a word, while a role is defined based purely on linguistic features from the point of view of the NER. Similarly, like a POS tag, a role is a collection of similar tokens, and a token has one or more roles. In the Chinese PER role set shown in Table 2a, the role SS includes almost 900 single-Hanzi (Chinese character) surnames and 60 double-Hanzi surnames. Meanwhile, the token “曾”(pronunciation “ceng” or “zeng”) can play role SS in the sequence “*曾*/菲/小姐”(Ms. Zeng *Fei*),play role GS in “记者/唐/师/*曾*”(Reporter Tang Shi-*Ceng*), play role NF in “胡锦涛*曾*视察西柏坡”(Hu Jin-Tao *has* surveyed Xi Bai Po), and also play some other roles.

If the size of a role set is too large, NER will suffer severely from the problem of data sparseness. Therefore, we do not attempt to set up a general role set for all NE categories. In order to reduce complexity, we build a specific role model using its own role set for each NE category. In another words, we apply the role model to PER, LOC, and ORG, respectively. Their role models are customized and trained individually. Finally, different recognized NE is all added into our unified class-based segmentation frame, which selects the global optimal result among all possible candidates.

The role set for Chinese PER, Chinese LOC, ORG, transliterated PER, and transliterated LOC are defined in Table 2a, Table 2b, Table 2c, Table 2d, and Table 2e, respectively. Considering the possible segmentation ambiguity mentioned in Section 2, we introduce some special roles, such as LH and TR, in Chinese PER. Such roles indicate that the token should be split into two halves before NER. Such a policy can improve NER recall. The process will be demonstrated in detail in the following section.

For the sake of clarity and to avoid loss of generality, we will focus our discussion mainly on Chinese PER entities. The problems and techniques discussed below are applicable to other entities.

Table 2a. Role set for Chinese PER.

Roles	Significance	Examples
SS	Surname.	*欧阳*/修 (*Ouyang* Xiu)
GH	Head component of 2-character given name	张/*华*/平/先生(Mr. Zhang *Hua*-Ping)
GT	Tail component of 2-character given name	张/华/*平*/先生(Mr. Zhang Hua-*Ping*)
GS	Given name with a single Chinese character	曾/*菲*/小姐(Ms. Zeng *Fei*)
PR	Prefix in the name	*老*/刘(*Old* Liu)、*小*/李(*Little* Li)
SU		王/总(*President* Wang)、曾/氏(*Ms* Zeng)
NI	Neighboring token in front of NE	*来到*/于/洪/洋/的/家 (Come to Yu Hong-Yang's house)
NF	Neighboring token following NE	新华社/记者/黄/文/*摄* (*Photographed* by Huang Wen from the Xinhua News Agency)
NB	Tokens between two NE.	编剧/邵/钧/林/*和*/稽/道/青/说 (Editor Shao Jun-Lin *and* Ji Dao-Qin said)
LH	Words formed by its left neighbor and head of NE.	现任/主席/*为何*/鲁/丽/。/ (Current chair *is He* Lu-Li.) * "*is He*" in Chinese forms word "why"
TR	Words formed by tail of NE and its right neighbor.	龚/学/*平等*/领导/ (Gong Xue-*Ping and other* leaders) * "*Ping and other*" forms the word "equality"
WH	Words formed by surname and GH (List in item 2)	*王国*/维 (*Wang Guo*-Wei) * "*Wang Guo* " in Chinese forms word "kingdom"
WS	Words formed by a surname and GS (List in item 3)	*高峰*(*Gao Feng*) *"*Gao Feng*" in Chinese forms the word "high ridge"
WG	Words formed by GH and GT	张*朝阳*(Zhang *Zhao-Yang*) *"*Zhao-Yang*" in Chinese forms the term "rising sun"
RC	Remote context, except for roles listed above.	*全国*/*人民*/*深切*/缅怀/邓/小/平/(The whole nation memorialized Mr. Deng Xiao-Ping)

Table 2b. Role set for Chinese LOC.

Roles	Significance	Examples
LH	Location head component	*石*/河/子/乡/ (*Shi* He Zi Village)
LM	Location middle component	石/*河*/子/乡/ (Shi *He* Zi Village)
LT	Location tail component	石/河/*子*/乡/ (Shi He Zi Village)
SU		海/淀*区*(Hai Dian *district*)
NI	Neighboring token in front of NE	我/*来到*/中/关/园(I *came* to Zong Guan Garden.)
NF	Neighboring token following NE	波/阳/县/*是*/我/的/老家

NB	Tokens between two NE	刘家村/*和*/下岸村/相邻(Liu Jia village *and* Xia An village are neighboring villages.)
RC	Remote context, except roles listed above.	波/阳/县/是/我/*的*/*老家*(Bo Yang county is my home)

Table 2c. Role set for ORG.

Roles	**Significance**	**Examples**
TO	Tail component of ORG	中央/人民/广播/*电台*/(China Central Broadcasting *Station*)
OO	Other component of ORG	*中央*/*人民*/*广播*/电台/(*China* *Central* *Broadcasting* Station)
NI	Neighboring token in front of NE	*通过*/中央/人民/广播/电台/(*via* China Central Broadcasting Station)
NF	Neighboring token following NE	/中央/电视台/*是*/国办的(China Central TV Station *is* run by the state)
NB	Tokens between two NE.	中国/国际/广播/电台/*和*/中央/电视台/(China Central Broadcasting Station and CCTV)
RC	Remote context, except for the roles listed above.	*1998 年*/*来临*/*之际*/(At the forthcoming of the year of 1998)

Table 2d. Role set for transliterated PER.

Roles	**Significance**	**Examples**
TH	Heading component of transliterated PER	*尼*/古/拉/斯/· /凯/奇(“ni” in “Nicolas Cage”)
TM	Middle component of transliterated PER	尼/*古*/*拉*/*斯*/· /*凯*/奇(“colas ca” in “Nicolas Cage”)
TT	Tail component of transliterated PER	尼古拉斯· 凯*奇*(“ge” in “Nicolas Cage”)
NI	Neighboring token in front of NE	*会见*/蒙/帕/蒂/· /梅/拉/费(meet)
NF	Neighboring token following NE	蒙/帕/蒂/· /梅/拉/费/*表示*(figure)
NB	Tokens between two NE.	里/根/*与*/南/茜/是/患难/夫妻(and)
TS	Tokens needed split	铁/木/尔/· /达/瓦/买/*提高*/度/评价/了(“Ti” is a tail component of a transliterated PER, and “Gao” or “highly” is a neighboring token; *提高* or “Ti Gao” forms a common word: “enhance”.)
RC	Remote context, except for the roles listed above.	里/根/与/南/茜/是/*患难*(adversity)/*夫妻* *(couple)*

Table 2e. Role set for Transliterated LOC.

Roles	Significance	Examples
TH	Heading component of transliterated LOC	*喀*布尔(“Ka” in Kabul)
TM	Middle component of transliterated LOC	喀*布*尔(“Bu” in Kabul)
TT	Tail component of transliterated LOC	喀布*尔*(“l” in Kabul)
NI	Neighboring token in front of NE	*到达（arrive）*喀布尔
NF	Neighboring token following NE	喀布尔*位于*（*locate*）
NB	Tokens between two NE.	喀布尔*和*(*and*)坎大哈

3.3 Role corpus

Since a role is self-defined and very different from a POS or other tag set, there is no special corpus that meets our requirement. How can we prepare the role corpus and extract role statistical information from it? Our strategy is to modify an available corpus by converting the POS tags to roles automatically.

We use a six-month news corpus from the *People's Daily*. It was all manually checked after word segmentation and POS tagging were performed. The work was done at the Institute of Computational Linguistics, Peking University (PKU). It is a high-quality corpus and widely used for Chinese language processing. The POS standard used in the corpus is defined in PKU, and we call it the PKU-POS set. Figure 3a shows a segment of our corpus labelled PKU-POS. Though PKU-POS is refined, it is implicit and not large enough for Chinese NER. In Figure 3a, the Chinese PER “黄振中”(Huang Zhen-Zhong) is split into the surname“黄”(Huang) and given name“振中”(Zhen-Zhong), but both of them are assigned the same tag, “nr”. In addition, there are no tags to distinguish transliterated PERs or LOCs from Chinese ones. Moreover, some NE abbreviations are not tagged with the right NE category, but with an abbreviation label, “j”. Here, “淮”(abbreviation for “淮河” or “Huai He River”) is a Chinese LOC and should be tagged with the location label “ns”.

Based on the PKU-POS, we made some modifications and added some finer labels for Chinese NE. Then, we built up our own modified POS set called ICTPOS (Institute of Computing Technology, part-of-speech set). In ICTPOS, we used the label “nf” to tag a surname and the label “nl” to tag a given name. In addition, we also separated each transliterated PER and transliterated LOC from each “nr” (PER) and “ns”(LOC), and tagged them with “tr” and “ts”, respectively. In the final step, we replaced each ambiguous label “j” with its NE category. Besides the NE changes, labels for different punctuations were added, too. The final version of ICTPOS contains 99 POS tags, and it is more useful for the NER task. Also, the modified corpus with ICTPOS labels is better in terms of quality after hand correcting. Figure 3b shows the equivalent segment with ICTPOS.

Next, we converted our corpus labelled with ICTPOS into a role corpus. The conversion procedure included the following steps:

(1) Extract the sequence of words and their POS.

(2) According to the POS, locate the particular NE category under consideration. Here, we only locate words labelled ‘nf’ or ‘nl’ when considering Chinese PER.

(3) Convert the POS of the NE’s components, their neighbours, and remote contexts into corresponding roles in that role set of the particular category.

Figures 3c and 3d show the corresponding training data after label conversion from ICTPOS tags to roles of Chinese PER and Chinese LOC, respectively. What we should point out is that the PER role corpus is totally different from the LOC corpus and other ones. For instance, the first pronoun word “本报”(this newspaper) in the PER role corpus is just a remote context, while it is a left neighboring context before “蚌埠”(Feng Pu) when LOC roles are applied. Though we use the same symbol “NI” to tag NE left neighboring tokens in both Figures 3c and 3d, it has different meanings. The first is for Chinese PER left tokens, and the other is for LOC. In a word, each NE category has its own role definition, its own training corpus, and its own role parameters though they all make use of the role model.

19980101-02-009-002/m 本报/r **蚌埠/ns** １月/t １日/t 电/n 记者/n **黄/nr 振中/nr** 、/w **白/nr** **剑峰/nr** 报道/v ：/w 新年/t 的/u 钟声/n 刚刚/d 敲响/v ，/we 千/m 里/q **淮河/ns** 传来/v 喜讯/n ：/w 沿/p **淮/j** 工业/n 污染源/n 实现/v 达标/v 排放/v ，/w 削减/v 污染/v 负荷/n ４０％/m 以上/f ，/we **淮河/ns** 治/v 污/Ng 第一/m 战役/n 告捷/v 。/w

Figure 3a: A segment of a corpus labeled with PKU-POS.

(Translation: 19980101-02-009-002 Jan. 1, reporters Huang Zhen-Zhong and Bai Jian-Feng from Feng Pu reporting: Since the bell for the New Year just rang, good news spread over the thousands miles Huai He river. The pollution source from industry near the Huai River achieved the standard with reducing pollution by over 40%. The first step in Huai River decontamination has been accomplished.)

19980101-02-009-002/m 本报/r **蚌埠/ns** １月/t １日/t 电/n 记者/n **黄/nf 振中/nl** 、/we **白/nf 剑峰/nl** 报道/v ：/we 新年/t 的/uj 钟声/n 刚刚/d 敲响/v ，/we 千/m 里/q **淮河/ns** 传来/v 喜讯/n ：/we 沿/p **淮/ns** 工业/n 污染源/n 实现/v 达标/v 排放/v ，/we 削减/v 污染/v 负荷/n ４０％/m 以上/f ，/we **淮河/ns** 治/v 污/Ng 第一/m 战役/n 告捷/v 。/we

Figure 3b: The segment from our corpus labeled with our modified POS.

19980101-02-009-002/RC 本报/RC 蚌埠/RC １月/RC １日/RC 电/RC **记者/NI 黄/SS 振/GH 中/GT 、/NM 白/SS 剑/GH 峰/GT 报道/NF** ：/RC 新年/RC 的/RC 钟声/RC 刚刚/RC 敲响/RC ，/RC 千/RC 里/RC 淮河/RC 传来/RC 喜讯/RC ：/RC 沿/RC 淮/RC 工业/RC 污染源/RC 实现/RC 达标/RC 排放/RC ，/RC 削减/RC 污染/RC 负荷/RC ４０％/RC 以上/RC ，/RC 淮河/RC 治/RC 污/RC 第一/RC 战役/RC 告捷/RC 。/RC

Figure 3c: The corresponding corpus labeled with Chinese PER roles.

19980101-02-009-002/RC **本报/NI 蚌/LH 埠/LT １月/NF** １日/RC 电/RC 记者/RC 黄/RC 振中/RC 、/RC 白/RC 剑峰/RC 报道/RC ：/RC 新年/RC 的/RC 钟声/RC 刚刚/RC 敲响/RC ，/RC 千/RC **里/NI 淮/LH 河/LT 传来/NF** 喜讯/RC ：/RC **沿/NI 淮/LH 工业/NF** 污染源/RC 实现/RC 达标/RC 排放/RC ，/RC 削减/RC 污染/RC 负荷/RC ４０％/RC 以上/RC **，/NI 淮/LH 河/LT 治/NF** 污/RC 第一/RC 战役/RC 告捷/RC 。/RC

Figure 3d: The corresponding corpus labeled with Chinese LOC roles.

3.4 Role tagging using the Viterbi Algorithm

Next, we prepared the role set and role corpus. Then, we could return to the key problem described in Section 3.1. That is: Given a token sequence, how can we tag a proper role sequence automatically?

Similar to POS tagging, we use the Viterbi algorithm [Rabiner and Juang, 1989] to select a global optimal role result from all the role sequences. The methodology and its calculation are given below:

Suppose that T is the token sequence after tokenization, R is the role sequence for T, and $R^{\#}$ is the best choice with the maximum probability. That is,

$T=(t_1, t_2, \dots, t_m)$,

$R=(r_1, r_2, \dots, r_m)$, m>0,

$$R^{\#}=\arg\max_{R} P(R|T) \qquad \text{E1}$$

According to the Bayes' Theorem, we can get

$$P(R|T)=P(R)P(T|R)/P(T) \qquad \text{E2}$$

For a particular token sequence, P(T) is a constant. Therefore, we can get E3 based on E1 and E2:

$$R^{\#}=\arg\max_{R} P(R)P(T|R) \qquad \text{E3}$$

We may consider T as the observation sequence and R as the state sequence hidden behind the observation. Next we use the Hidden Markov Model [Rabiner and Juang, 1986] to tackle a typical problem:

$P(R)\ P(T|R) \approx \prod_{i=1}^{m} p(t_i \mid r_i)\, p(r_i \mid r_{i-1})$, where r_0 is the beginning of a sentence;

$$\therefore R^{\#} \approx \arg\max_{R} \prod_{i=1}^{m} p(t_i \mid r_i)\, p(r_i \mid r_{i-1}) \qquad \text{E4}$$

For convenience, we often use the negative log probability instead of the proper form. That is,

$$R^{\#} \approx \arg\min_{R} \sum_{i=1}^{m} \{-\ln p(t_i \mid r_i) - \ln p(r_i \mid r_{i-1})\} \qquad \text{E5}$$

Finally, role tagging is done by as solving E5 using Viterbi algorithm.

Next, we will use the sentence “张华平等着你”(Zhang Hua-Ping is waiting for you) to explain the global optimal selection process. After tokenization is performed using any approach, the most probable token sequence will be “张/华/平等/着/你”. Here, “平”(pronunciation “ping”) is separated from the PER name “张华平” (Zhang Hua-Ping) and forms a token “平等”(equality) while it sticks to “等”(pronunciation “deng”). In Figure 4, we illustrate the process of role tagging with Viterbi selection on tokens sequence “张/华/平等/着/你”. Here, the best role result $R^{\#}$ is “张/SS 华/GH 平等/TR 着/RC 你/RC” based on Vitebi selection.

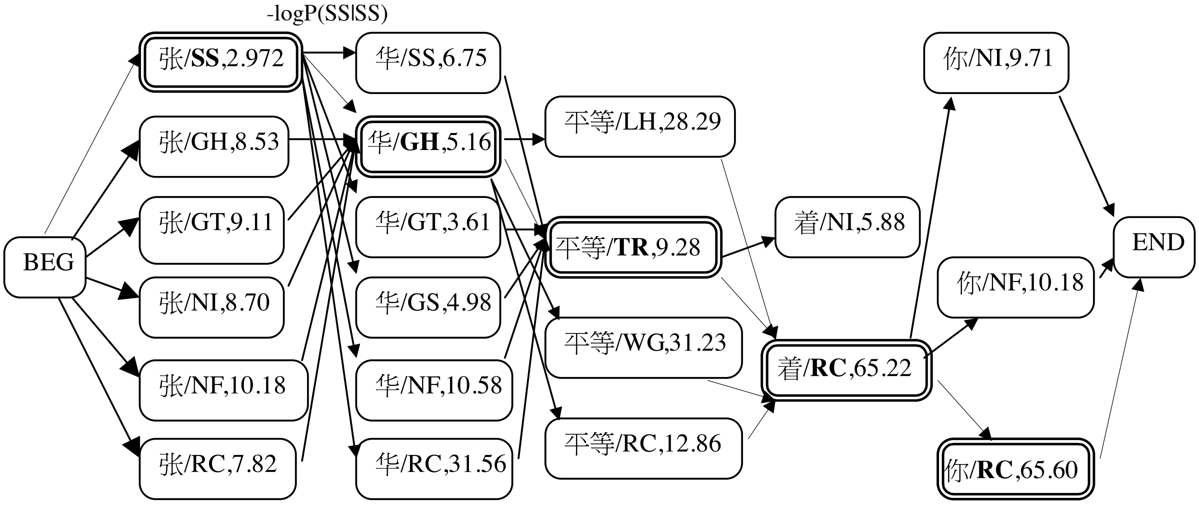


Figure 4: Role selection using the Viterbi algorithm.

Notes:

1. The data shown in each square are organized as follows: Token t_i /role r_i, -logP(t_i | r_i).
2. The value on the directed edges in the figure is –logP(r_i | r_{i-1}). Here, we do not paint all the possible edges for simplicity.
3. The double-edged squares are the best choices after Viterbi selection.

3.5 Training the Role model

In E5, $p(t_i \mid r_i)$ is the emission probability of token t_i given its role r_i, while $p(r_i \mid r_{i-1})$ is the role transitive probability from the previous role r_{i-1} to the current one r_i. They are estimated with maximum likelihood as follows:

$$p(t_i \mid r_i) \approx C(t_i, r_i)/C(r_i) \qquad \text{E6}$$

, where C(t_i, r_i) is the count of token t_i with role r_i, and C(r_i) is the count of role r_i;

$$p(r_i \mid r_{i-1}) \approx C(r_{i-1}, r_i)/C(r_{i-1}) \qquad \text{E7}$$

, where C(r_{i-1},r_i) is the count of role r_{i-1} followed by role r_i.

C(t_i, r_i), C(r_i) and C(r_{i-1},r_i) can be easily calculated based on our roles corpus during the process of role model training. In Figure 3c, C(“黄”,SS), C(“白”,SS), C(SS) ,C(NI, SS) and C(NM,SS) are 1,1,2,1 and 1, respectively.

3.6 The probability that an NE is recognized correctly

A recognized NE may be correct or incorrect. The result is uncertain and it is essential to quantify the uncertainty with a reliable probability measure. The probability that an NE is recognized correctly is the essential basis for our further processing, such as improving the performance of NER by filtering some results with lower probability. Suppose *N* is the NE, and that its type is *T*. *N* consists of the token sequence (t_i t_{i+1} …. t_{i+k}), and its roles are (r_i r_{i+1} …. r_{i+k}). Then, we can estimate the possibility as follows:

$$P(N|T) \approx \prod_{j=0}^{k} p(t_{i+j} \mid r_{i+j}) \times \prod_{j=1}^{k} p(r_{i+j} \mid r_{i+j-1}) \qquad \text{E8}$$

For the previous Chinese PER “张华平”(Zhang Hua-Ping), we can compute P(张华平 |Chinese PER) using the following equation:

P(张华平|Chinese PER)=p(SS|NI)xp(张|SS)xp(GH|SS)xp(华|GH)xp(GT|GH)xp(平|GT).

3.7 The Work Flow of Chinese NER

After the role model is trained, Chinese NE can be recognized in an original sentence through the steps listed below:

(1) Tokenization on a sentence. In our work, we use a tokenization method called the “Model of Chinese Word Rough Segmentation Based on N-Shortest-Paths Method” [Zhang and Liu, 2002]. It aims to produce the top N results as required and to enhance the recall rates of right tokens.

(2) Tag token sequences with roles using the Viterbi algorithm. Get the role sequence $R^{\#}$ with the maximum possibility.

(3) In $R^{\#}$, split the tokens whose roles are “LH” or “TR”. These roles indicate that the internal components stick to their contexts. Suppose R^{*} is the final role sequence.

(4) NE recognized after maximum matching on R^{*} with the particular NE templates. Templates of Chinese PER are shown in Table 3.

(5) Computing the possibilities of NE candidates using formula E8.

Table 3. Chinese PER Templates

No	Roles Templates	Examples
1	SS+SS+ GH+ GT	香港立法会/* 主席/* **范/SS 徐/SS 丽/GH 泰/GT** (Council chair Fan Xu Li-tai)
2	SS+ GH+ GT	**张/SS 华/GH 平/GT** 先生/* (Mr. Zhang Hua-Ping)
3	SS+ GS	**曾/SS 菲/GS** 表示/* (Zeng fei expressed…)
4	SS +WG	**张/SS 朝阳/WG** (Zhang Zhao-Yang; Zhao-yang is a common word meaning “morning sun” in English)
5	WG	宝玉**/WG** 回到/*了/* 怡香院/* (Bao-Yu went back to Yi-Xiang yard, Bao-Yu is a common word meaning “Jade” in English)
6	GH+ GT	**华/GH 平/GT** 先生/* (Mr. Hua-Ping)
7	PR+ SS	老/ PR 刘**/SS**(Old Liu); 小**/PR** 李**/SS**(Little Li)
……		

Note: “*” in the examples indicates any role.

We will continue our demonstration with the previous example “张华平等着你”.After Viterbi tagging, its optimal role sequence $R^{\#}$ is “张/SS 华/GH 平等/TR 着/RC 你/RC”. The role “RC” forces us to split the token “平等”(equality) into two parts: “平”(pronunciation: “ping”) and “等”(etc.). Then, the modified role result R^{*} will be “张/SS 华/GH 平/GT 等/NF 着/RC 你/RC”. Through maximum pattern matching using the Chinese PER patterns listed in Table 3, we find that the second template “SS+ GH+ GT” can be applied. Therefore, the token sequence “张/SS 华/GH 平/GT” is located, and the string “张华平” is recognized as a common

Chinese PER name.

4. Class-based Segmentation Model Integrated into NER

In section 3-2, we emphasized that each NE category uses an independent role model. Each NE candidate is the global optimum result in its role model. However, it has not competed with other models, and all the different models have not been combined together. One problem is as follows: If a word is recognized as a location name by the LOC role model, and as an ORG, PER or even a common word by another, which one should we choose in the end? Another problem is as follows: Although Chinese NER using role models can achieve higher recall rates than previous approaches (the recall rate of Chinese PER is nearly 100%), the precision result is not satisfactory because some NE candidates are common words or belong to other categories.

Here, we use a class-based word segmentation model that is integrated into NER. In the generalized segmentation frame, NE candidates from various role models can compete with common words and themselves.

Given a word w_i, a word class c_i is defined as shown in Figure 5a. Suppose |LEX| is the lexicon size; then, the size of the word classes is |LEX|+9. In Figure 5b, we show the corresponding class sample based on Figure 3b.

$$c_i = \begin{cases} w_i & \text{if } w_i \text{ is listed in the segmentation lexicon;} \\ \text{Chinese PER} & \text{if } w_i \text{ is an unlisted}^{*} \text{ Chinese PER;} \\ \text{Transliterated PER} & \text{if } w_i \text{ is an unlisted transliterated PER;} \\ \text{Chinese LOC} & \text{if } w_i \text{ is an unlisted Chinese LOC;} \\ \text{TIME} & \text{if } w_i \text{ is an unlisted time expression;} \\ \text{QUAN} & \text{if } w_i \text{ is an unlisted numeric expression;} \\ \text{STR} & \text{if } w_i \text{ is an unlisted symbol string;} \\ \text{BEG} & \text{if } w_i \text{ is beginning of a sentence} \\ \text{END} & \text{if } w_i \text{ is ending of a sentence} \\ \text{OTHER} & \text{otherwise.} \end{cases}$$

* "unlisted" means outside the segmentation lexicon.

Figure 5a: Class Definition of word $\mathbf{w}_\mathrm{i}$

[QUAN] 本报/r **[Chinese LOC] [TIME] [TIME]** 电/n 记者/n **[Chinese PER]** 、/we **[Chinese PER]** 报道/v ：/we 新年/t 的/uj 钟声/n 刚刚/d 敲响/v ，/we 千/m 里/q **[Chinese LOC]** 传来/v 喜讯/n ：/we 沿/p **[Chinese LOC]** 工业/n 污染源/n 实现/v 达标/v 排放/v ，/we 削减/v 污染/v 负荷/n **[QUAN]**/m 以上/f ，/we **[Chinese LOC]** 治/v 污/Ng 第一/m 战役/n 告捷/v 。/we

Figure 5b: The corresponding class corpus.

Let W be the word sequence, let C be its class sequence, and let $W^{\#}$ be the segmentation result with the maximum likelihood. Then, we can get a class-based word segmentation model integrated into unknown Chinese NE. That is,

$$\begin{aligned} W^{\#} &= \arg\max_{W} P(W) \\ &= \arg\max_{W} P(W/C)P(C). \end{aligned}$$

After introducing a class-based bigram model, we can get

$$W^{\#} \approx \arg\max_{w_1 w_2 \ldots w_m} \prod_{i=1}^{m} p'(w_i \mid c_i) p(c_i \mid c_{i-1}) \text{ , where } c_0 \text{ is the begin of a sentence} \qquad \text{E9}$$

Based on the class definition, we can compute $p'(w_i/c_i)$ using the following formula:

$$p'(w_i/c_i) = \begin{cases} \text{estimated using E8; if } w_i \text{ is an unknown Chinese NE} \\ 1; \quad \text{otherwise} \end{cases}$$

Another factor $p(c_i/c_{i-1})$ in E9 indicates the transitive probability from one class to another. It can be extracted from corpus as shown in Figure 5b. The training of word classes is similar that of role models, thus we skip the detail.

If there are no unknown Chinese NE, the class approach will back off to a word-based language model. All in all, the class-based approach is an extension of the word-based language model. One difference is that class-based segmentation covers unknown NE besides common words. With this strategy, it not only the segmentation performance, but also the precision of Chinese NER is improved. For the sentence “张华平等着你” shown in Figure 6, both “张华” and “张华平” can be identified as Chinese PERs. It is very difficult to make decision between the two candidates solely in NER. In our work, we do not attempt to make such a choice in a earlier step; we add the two possible NE candidates to the class-based segmentation model. When the ambiguous candidates compete with each other in the unified frame, the segmentation result “张华平/等着/你” will defeat “张华/平等/着/你” because of its much higher probability.

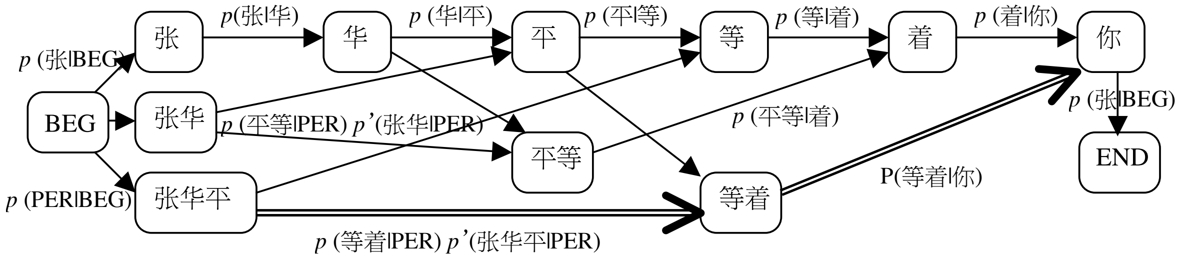


Figure 6:Demonstration of segmentation on "张华平等着你" using the class-based approach.

Note: "张华平"(Zhang Hua-Ping) and "张华" are NE candidates from role models.

5. Comparison with Previous Works

Since MET came into existence, NER has received increasing attention, especially in research on written and spoken English. Some systems have been put into practice. The approaches tend to involve statistics mixed with rules, such as the hidden Markov model (HMM), the expectation maximum, transformation-based learning, etc. [Zhou and Su, 2002; Bikel *et al.* 1997; Borthwick *et al.* 1999]. Besides making use of a corpus with labels, Andrei *et al.* [1999] proposed another statistical method without Gazetteers.

Historically, much work has been done on Chinese NER, but the research is still in its early stages. Previous solutions can be broadly categorized into rule-based approaches [Luo, 2001; Ji, 2001; Song, 1993; Tan, 1999], statistics-based ones [Zhang *et al.* 2002; Sun *et al.* 2002; Sun, 1993] and approaches that are a combination of both [Ye, 2002, Lv *et al.* 2001]. Compared with our approach using the role model, previous works have some disadvantages. First of all, many researchers used handcrafted rules, which are mostly summarized by linguists through painful study on large corpuses and huge NE libraries [Luo, 2001]. This is time-consuming, expensive and inflexible. The NE categories are diverse, and the number of words for each category is huge. With the rapid development of the Internet, this situation is becoming more and more serious. Therefore, it is very difficult to summarize simple yet thorough rules for NE components and contexts. However, in the role model, the mapping from roles to entities is done based on by simple rules. Secondly, the recognition process in previous approaches could not be activated until some "indicator" tokens were scanned in. For instance, possible surnames or titles often trigger personal name recognition on the following 2 or more characters. In the case of place name recognition, postfixes such as "县"(county) and "市"(city) activate recognition on previous characters. Furthermore, this trigger mechanism cannot resolve the ambiguity. For example, the unknown word "方林山" (Fang Lin Shan) may be a personal name, "方/林山"(Fang Linshan), or a place name, "方林/

山”(Fanglin Mountain). What’s more, previous approaches tended to work only on monosyllabic tokens, which are obvious fragments after tokenization [Luo, 2001; Lv *et al*. 2001]. This risks losing those NE that lack explicit features. On the other hand, the role model tries to select possible NE candidates based on the whole token sequence and then select the most promising ones using Viterbi tagging. Last but not least, to the best of our knowledge, some statistical works only focus on the frequency of characters or tokens in NE and their common contexts. Thus, it is harder to compute a reliable probability for a recognized NE. Unlike the role-based approach, previous works could not satisfy other requirements, such as NE candidate filtering and statistical lexical analysis.

In one sense, BBN' s name finder IdentiFinder [F. Kubala *et al*. 1998] is very close to our role model. Both the role model and IdentiFinder extract NE using a hidden Markov Model, which is also trained on a corpus. In addition, the authors claim that it can perform NER in multilingual languages, including Chinese. Now, we will explain how IdentiFinder and the role model differ.

(1) IdentiFinder uses general name-classes, which include all kinds of NE and Not-A-Names, while we build a different instance for each NE category with the same role model. As explained in Section 3, a general name-class will suffer from data sparseness. The role model does not require a large-scale corpus because we can transform the same corpuses into different role corpus, from which role probabilities can be extracted.

(2) IdentiFinder is applied to token sequences, but Chinese sentences are made up of character strings. It is impossible to apply the name-class HMM to Chinese original texts. Even if it is applied after tokenization, there is no more consideration on unification between tokenization and NER. Here, tokenization becomes an independent preprocessing step for Chinese NER.

(3) The name-classes used in IdentiFinder seem too simple for Chinese, a complicated language. IdentiFinder has only 10 classes: PER, ORG, five other named entities, Not-A-Name, start-of-sentences and end-of sentence. However, just for PER recognition, we use 16 roles to differentiate various tokens, such as component, left and right neighboring contexts and other helpful ones. Actually, they boost the recall rate of Chinese NER.

All in all, IdentiFinder have the similar motivation as we described here, and it successfully solves the problem of English NER. Nevertheless, much work must still be done to extend its approach to Chinese NER.

6. Experiments and Discussion

6.1 Evaluation Metric

As is commonly done, we conducted experiments on precision (P), recall (R) and the F-measure (F). The last term, F, is defined as a weighted combination of precision and recall. That is,

$$P = \frac{\text{number of correctly recognized NE}}{\text{number of recognized NE}} \qquad \text{E10}$$

$$R = \frac{\text{number of correctly recognized NE}}{\text{number of all NE}} \qquad \text{E11}$$

$$F = \frac{R \times P \times (1 + \beta^2)}{R + P \times \beta^2} \qquad \text{E12}$$

In E12, β is the relative weight of precision and recall. Here, Supposed that precision and recall are equally weighted, and we assign 1 to β, namely F-1 value.

In order to compare with other evaluation results, we only provide the result of PER(including Chinese PER and transliterated PER) and LOC (including Chinese LOC and transliterated LOC) although Chinese NE and transliterated ones are recognized with the different instances of role model.

6.2 Training Data Set

As far as we known, the traditional evaluation approach is to prepare a collection of sentences that include some special NE and to then perform NER on the collection. Those sentences that do not contain specific NE are not considered. In our experiments, we used a realistic corpus and did no filtering. The precision rates we obtained may be lower than but closer to the realistic linguistic environment than those obtained in previous tests.

We used the news corpus from January as the test data with 1,105,611 words and used the other five months as the training set. The ratio between the training and testing data was about 5:1. The testing corpus was obtained from the homepage of the Institute of Computational Linguistics at www.icl.pku.edu.cn at no cost since it was for non-commercial use. In the training of the role model, we did not used any heuristic information (such as the length of name, the particular features of characters used, etc.) or special NE libraries, such as person name collections or location suffix collections. It was purely a statistical process.

6.3 NER Evaluation Experiments

In a broad sense, automatic recognition of known Chinese NE depends more on the lexicon than on the NER approach. If the size of the NE collection in the segmentation lexicon is large enough, Chinese NER will back to the problems of word segmentation and disambiguation. Undoubtedly, it is easier than a pure NER. Therefore, evaluation of unlisted NE, which is outside the lexicon, can reflect the actual performance of NER method. It approach will be more objective, informative and useful. Here, we will report our results both for unlisted and listed NE. In order to evaluate the function of class-based segmentation, we also give some contrast testing. We conducted the five NER evaluation experiments listed in Table 4.

Table 4. Different evaluation experiments.

ID	Testing Set	Unlisted* NE or listed ones?	Class-based segmentation applied?
Exp1	PKU corpus	Considering only unlisted NE	No
Exp2	PKU corpus	Both	No
Exp3	PKU corpus	Considering only unlisted NE	Yes
Exp4	PKU corpus	Both	Yes
Exp5	MET2 testing data	Considering only unlisted NE	Yes

* "Unlisted" means outside the segmentation lexicon

The PKU corpus is January 1998 news from the People's Daily.

6.3.1 Exp1: individual NER conducted on unlisted names using a specific role model

Exp1 includes 3 sub-experiments: personal name recognition with the PER role model, LOC recognition with its own model, and ORG. In Exp1, we evaluate the performance only on unlisted NE. The performance achieved is reported in Table 5.

Table 5. Performance achieved in Exp1.

NE	Total Num	Recognized	Correct	P (%)	R (%)	F (%)
PER	17,051	29,991	15,880	56.85	93.13	70.61
LOC	4,903	12,711	3,538	27.83	72.16	51.84
ORG	9,065	9,832	6,125	62.30	67.58	64.83

6.3.2 Exp2: Individual NER conducted on all names using a specific role model

The only differences between Exp1 and Exp2 were that Exp2 ignored the segmentation lexicon, and that the performance in Exp2 is evaluated on both unlisted and listed NE. Comparing Table 5 and Table 6, we find that the NER results were better when listed NE were added. We can also draw the conclusion that location items in the lexicon contribute more to LOC recognition than to the LOC role model.

Table 6. Performance achieved in Exp2.

NE	Total Num	Recognized	Correct	P (%)	R (%)	F (%)
PER	19,556	32,406	18,915	58.37	96.72	72.80
LOC	22,476	30,239	22,366	67.54	99.51	80.55
ORG	10,811	11,483	7,776	67.72	71.93	69.77

6.3.3 Exp3 and Exp4: Introducing Class-based Segmentation Model

Exp1 and Exp2 are conducted on PER, LOC and ORG candidates with their individual role models. They were not integrated into a complete frame. In Exp3 and Exp4, we used the class-based segmentation model to further filter all the NE candidates. As we explained in the Section 4, common words and recognized NE from various role models could be added to the class-based segmentation model. After they competed with each other, either the optimal segmentation or the NER result would be selected. From Table 7, it can be concluded that the word segmentation model greatly improved the performance of Chinese NER.

We also found an interesting phenomenon in that unlisted PER recognition was a little better than recognition of all personal names. The main reason was that unlisted PER recognition could achieve a good recall rate, but some listed PERs could not be recalled because of ambiguity. For instance, "江泽民主张…" (Jiang Ze-Min proposed ..) would produce the wrong tokenization result "江/泽/民主/张" while the role model failed because "江泽民"(Jiang Ze-Min) was listed in the segmentation lexicon. On the other hand, if "江泽民"(Jiang Ze-Min) was not listed in the core lexicon, then "民主" (democracy) would be tagged with role "TR", and the token would be split before recognition. We provide more examples in Appendix II.

Table 7**. **Performance achieved in Exp3 and Exp4.

NE	Unlisted NE in Exp3			All NE in Exp4		
	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)
PER	95.18	96.50	95.84	95.49	95.66	95.57
LOC	71.83	74.67	73.23	92.64	95.38	93.99
ORG	66.06	81.76	73.08	75.83	88.39	81.63

6.3.4 Exp5: Evaluation of the MET2 Data

We conducted an evaluation experiment, Exp5, on the MET2 test data. The results for unlisted NE are shown in Table 8. Compared with the PKU standard, the MET2 data have some slight differences in terms of NE definitions. For example, in the PKU corpus, "新华社"(Xinhua News Agency) is not treated as an ORG but as an abbreviation. "酒泉卫星发射中心"(Jiu Quan Satellite Emission Center) is viewed as an LOC in MET-2, but as an ORG according to our definition. Therefore, the performance of NER for MET2 was not as good as that for the

PKU corpus.

Table 8. Performance achieved in Exp 5.

NE	Total Num	Recognized	Correct	P (%)	R (%)	F (%)
PER	162	231	150	64.94	92.59	76.34
LOC	751	882	661	74.94	88.02	80.96
ORG	378	366	313	85.52	82.80	84.14

6.4 A survey of on the relationship between NER and Chinese lexical analysis

A good tokenization or lexical analysis approach provides a specific basis for role tagging; meanwhile, correctly recognized NE will modify the token sequence and improve the performance of the Chinese lexical analyser.

Next, we will survey the relationship between NER and Chinese lexical analysis based on a group of contrast experiments. On a 4MB news corpus, we conducted four experiments:

1) BASE: Chinese lexical analysis without any NER;

2) +PER: Adding the PER role model to BASE;

3) +LOC: Adding the LOC role model to +PER;

4) +ORG: Adding the ORG role model to +LOC.

Table 9. A survey of on the relationship between NER and Chinese lexical analysis.

CASE	PER F-1 (%)	LOC F-1 (%)	ORG F-1 (%)	SEG	TAG1(%)	TAG2(%)
BASE	27.86	83.67	51.13	96.55	93.92	91.72
+PER	95.40	83.84	53.14	97.96	95.34	93.09
+LOC	95.50	85.50	52.76	98.05	95.44	93.18
+ORG	95.57	93.99	81.63	98.38	95.76	93.52

Note:

1) PER F-1: F-1 rate of PER recognition;
LOC F-1: F-1 rate of LOC recognition;
ORG F-1: F-1 rate of ORG recognition;

2) SEG=#of correctly segmented words/ #of words;

3) TAG1=#of correctly tagged 24-tag POS/#of words;

4) TAG2=#of correctly tagged 48-tag finer POS/#of words.

Table 9 shows the performance achieved in the four experiments. Based on these results, we draw the following conclusions:

Firstly, each role model contributes to Chinese lexical analysis. For instance, SEG increases from 96.55% to 97.96% after the PER role model is added. If all the role models are integrated, ICTCLAS achieves 98.38% SEG, 95.76% TAG1, and 93.52% TAG2.

Secondly, the preceding role model benefits from the succeeding one. We can find that after ORGs are recognized, Org F-1 increase by 25.91%; furthermore, the performance of PER and LOC also improve. It can be inferred that the ORG role model not only solves its own problem, but also helps exclude improper PER or LOC candidates in the segmentation model. Similarly, the LOC model aids PER recognition, too. Take "刘庄的水很甜"(The water in Liu village is sweet) as an example, here, "刘庄"(Liu village) is very likely to be incorrectly recognized as a personal name. However, it will be recognized as a location name after HMM is added for location recognition.

6.2 Official evaluation of our lexical analyser ICTCLAS

We have developed our Chinese lexical analyser ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System). ICTCLAS applies the role model to recognize unlisted NE names. We also integrate class-based word segmentation into the whole ICTCLAS frame. The full source code and documents of ICTCLAS are available at no cost for non-commercial use. Researchers and technical users are welcome to download ICTCLAS from the Open Platform of Chinese NLP (www.nlp.org.cn).

On July 6, 2002, ICTCLAS participated in the official evaluation, which was held by the National 973 Fundamental Research Program in China. The testing set consisted of 800KB of original news from six different domains. ICTCLAS achieved 97.58% in segmentation precision and ranked at the top. This proved that ICTCLAS is one of the best lexical analysers, and we are convinced that the role model is suitable for Chinese NER. Detailed information about the evaluation is given in Table 10.

Table 10. Official evaluation results for ICTCLAS.

Domain	Words	SEG	TAG1	RTAG
Sport	33,348	97.01%	86.77%	89.31%
International	59,683	97.51%	88.55%	90.78%
Literature	20,524	96.40%	87.47%	90.59%
Law	14,668	98.44%	85.26%	86.59%
Theoretic	55,225	98.12%	87.29%	88.91%
Economics	24,765	97.80%	86.25%	88.16%
Total:	208,213	97.58%	87.32%	89.42%

Note:

1) RTAG=TAG1/SEG*100%

2) The results related to POS are not comparable because our tag set is greatly different from their definition.

6.5 Discussion

Our approach is merely corpus-based. It is well known that, in any usual corpus, NE is sparsely distributed. If we depend solely on the corpus, the problem of sparseness inevitably be encountered. But by fine-tuning our system, we can alleviate this problem through some modifications described below:

Firstly, lexical knowledge from linguists can be incorporated into the system. This does not mean that we fall back to rule-based approaches. We just need some general and heuristic rules about NE formation to reduce some errors. As for Chinese PER recognition, there are several strict restrictions, such as the length of names and the order of surnames and given names.

Secondly, we can produce one more tokenization result. In this way, we can improve the recall rate at the expense of the precision rate. Precision can be improved in the class-based segmentation model. In this work, we only use the best tokenization result. We have tried rough word segmentation based on the N-Shortest-Paths method [Zhang and Liu, 2002]. When the top 3 token sequences are considered, the recall and precision of NER in ICTCLAS can be significantly enhanced.

Lastly, we can add some huge NE libraries besides the corpus. As is well known, it is easier and cheaper to get a personal name library or other special NE libraries than a segmented and tagged corpus. We can extract more precise component roles from NE libraries and then merge these data into the contextual roles obtained from the original corpus.

7. Conclusion

The main contributions of this study are as follows:

(1) We have propose the use of self-defined roles based on to linguistic features in named entity recognition. The roles consist of NE components, their neighboring tokens and remote contexts. Then, NER can be performed more easily on role sequences than on original character strings or token sequences.

(2) Different roles are integrated into a unified model, which is trained through an HMM. With emission and transitive probabilities, the global optimal role sequence is tagged through Viterbi selection.

(3) A class-based bigram word segmentation model has been presented. The segmentation frame can adopt common words and NE candidates from different role models. Then, the final segmentation result can be selected following competition among possible choices. Therefore, promising NE candidates can be reserved and others filtered out.

(4) Lastly, we have surveyed the relationship between Chinese NER and lexical

analysis. It has been shown that the role model can enhance the performance of lexical analysis after NE are successfully recalled, while class-based word segmentation can improve the NER precision rate.

We have conducted various experiments to evaluate the performance of Chinese NER on the PKU corpus and MET-2 data. F-1 measurement of recognizing PER, LOC, ORG on the 1,105,611-word PKU corpus were 95.57%, 93.99%, and 81.63%, respectively.

In our future work, we will build a finely tuned role model by adding more linguistic knowledge into the role set, more tokenization results as further candidates, and more heuristic information for NE filtering.

Acknowledgements

The authors wish to thank Prof. Shiwen Yu of the Institute of Computational Linguistics, Peking University, for the corpus mentioned in section 3.2 and Gang Zou for his wonderful work in the evaluation of named entity recognition. We also acknowledge our debt to our colleagues: Associate Professor Wang Bin, Dr. Jian Sun, Hao Zhang, Ji-Feng Li, Guo-Dong Ding, Dan Deng, and De-Yi Xiong. Kevin Zhang especially thanks his graceful girl friend Feifei for her encouragement during this research. We also thank the three anonymous reviewers for their elaborate and helpful comments.

References

Andrei M., Marc M. and Claire G., “Named Entity Recognition using an HMM-based Chunk Tagger”, *Proc. of EACL* '99.

Bikel D., Schwarta R., Weischedel. R. “An algorithm that learns what’s in a name”. *Machine Learning* 34, 1997, pp. 211-231

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

Chen X. H. “One-for-all Solution for Unknown Word in Chinese Segmentation”. *Application of Language and Character*, 3. 1999

F. Kubala, R. Schwartz, R. Stone, and R. Weischedel, “Named entity extraction from speech”, in *Proceedings of DARPA Broadcast News Transcription and Understanding Workshop*, (Lansdowne, VA), February 1998.

L. R.Rabiner, “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition”. *Proceedings of IEEE* 77(2): pp.257-286, 1989.

L.R. Rabiner and B.H. Juang, “An Introduction to Hidden Markov Models”. *IEEE ASSP Mag.*, pp.4-166.

Luo H. and Ji Z. “Inverse Name Frequency Model and Rules Based on Chinese Name Identifying”. In *Natural Language Understanding and Machine Translation*, C. N. Huang & P. Zhang, ed., Tsinghua Univ. Press, Beijing, China, Jun. 1986, pp. 123-128.

Luo Z. and Song R. "Integrated and Fast Recognition of Proper Noun in Modern Chinese Word Segmentation". *Proceedings of International Conference on Chinese Computing*, 2001, Singapore, pp. 323-328.

Lv Y.J., Zhao T. J. "Levelled Unknown Chinese Words Resolution by Dynamic Programming". *Journal of Chinese Information Processing*. 2001,15, 1, pp. 28-33.

N.A. Chinchor , "MUC-7 Named Entity Task Definition". In *Proceedings of the Seventh Message Understanding Conference*, 1998

Song R., "Person Name Recognition Method Based on Corpus and Rule". In *Computational Language Research and Development*, L. W. Chen & Q. Yuan, ed., Beijing Institute of Linguistic Press.1993

Sun H. L., "A content chunk parser for unrestricted Chinese text", PhD Dissertation, 2001, pp 22-35

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

Sun M.S. "English Transliteration Automatic Recognition". In *Computational Language Research and Development*, L. W. Chen & Q. Yuan, ed., Beijing Institute of Linguistic Press.1993.

Tan H. Y. "Chinese Place Automatic Recognition Research". In *Proceedings of Computational Language*, C. N. Huang & Z.D. Dong, ed., Tsinghua Univ. Press, Beijing, China. 1999

Ye S.R, Chua T.S., Liu J. M., "An Agent-based Approach to Chinese Named Entity Recognition", *Proc. of the 19th International Conference on Computational Linguistics*, Taipei, Aug. 2002. pp 1149-1155

ZHANG Hua-Ping, LIU Qun, "Model of Chinese Words Rough Segmentation Based on N-Shortest-Paths Method". *Journal of Chinese Information Processing*. Feb. 2002, 16, 5, pp.1-7.

ZHANG Hua-Ping, LIU Qun, "Automatic Recognition of Chinese Person based on Roles Taging". *Chinese Journal of Computer*, 2003(To be published).

ZHANG Hua-Ping, LIU Qun, "Automatic Recognition of Chinese Person based on Roles Taging". *Proc. of 7th Graduate Conference on Computer Science in Chinese Academy of Sciences*. Si Chuan, July, 2002.

ZHANG Hua-Ping, LIU Qun, Zhang Hao and Cheng Xue-Qi, "Automatic Recognition of Chinese Unknown Words Recognition". Proc. of COLING 2002 workshop on SIGHAN, Aug. 2002 pp.71-77.

Zhou G. D., Su J., "Named Entity Recognition using an HMM-based Chunk Tagger", *Proc. of the 40th ACL*, Philadelphia, July 2002, pp. 473-480.

Appendices

Appendix I. Cases that head or tail of Chinese PER binds with the neighboring tokens

(Cases illustrated with the format: Known words: left neighbor/Chinese PER/right neighbor)

波长(wave length)： 。/陈昌波(Chen Chang-Bo)/长大成人(grow up)

长安(Chang’An: an olden city of China)： 会长(chairman)/安士伟(An Shi-Wei)/代表 (present)

长长(long)： 局长(director general)/长孙(Zhang Sun)/介绍(introduce)

长发(long hair)： 会长(chairman)/钱伟长(Qian Wei-Chang)/发(deliver)

长江(the Changjiang River)： 院长(dean)/江泽慧(Jiang Ze-Hui)/指出(point out)

长孙(surname: “Zhang Sun”)： 队长(captain)/孙雯(Sun Wen)/门前(in front of goal)

长项(one's strong suit)： 局长(director general)/项怀诚(Xiang Huai-Cheng)/的(‘s)

超生(over birth)： 和(and)/邓颖超(Deng Ying-Chao)/生前(before one's death)

陈说(state)： 。/小陈(Xiao Chen)/说(say)

成都(ChengDu: a city of China)： ，/童志成(Tong Zhi-Cheng)/都(all)

成为(become)： 选举(elect)/李玉成(Li Yu-Cheng)/为(become)

成心(deliberately)： ，/童志成(Tong Zhi-Ch)/心中(in one’s heart)

初等(primary)： 主席(chairman)/董寅初(Dong Yin-Chu)/等(etc)

慈和(kindly)： 中郎将(general)/太史慈(Taishi Ci)/和(and)

到时(on time)： 到(go to)/时传祥(Shi Chuan-Xiang)/老伴(old partner)

东家(master)： 在(at)/赵孝东(Zhao Xiao-Dong)/家(home)

队章(discipline)： 河北队(Hebei team)/章钟(Zhang Zhong)/、

对白(dialogue)： 对(toward)/白晓燕(Bai Xiao-yan)/绑架(kidnap)

方向(direction)： /邓朴方(Deng Pu-Fang)/向(toward)

高手(expert)： 交到(hand in)/张洪高(Zhang Hong-Gao)/手上(keep)

光明(sunshine)： ，/苏洪光(Su Hong-Guang)/明白(understand)

光能(energy of light)： ，/苏洪光(Su Hong-Guang)/能(can)

国都(capital)： ，/邱娥国(Qiu Er-Guo)/都(all)

好在(thank to)： 总裁(president)/刘永好(Liu Yong-Hao)/在(at)

家门(the gate of a house)： 大家(everybody)/门文元(Men Wen-Yuan)/任(occupy)

家史(family tree)： 家(home)/史德才(Shi De-Cai)/一家(household)

健在(be still living and in good health)： /褚时健(Chu Shi-Jian)/在(at)

老是(always)： 侯老(Hou Lao)/是(is)

老总(master)： 许老(Xu Lao)/总是(always)

林中(in woods)： 繁荣(thrive)/李清林(Li Qing-Lin)/中共(Chinese Communist)

明说(say directly)： 主编(editor in chief)/周明(Zhou Ming)/说(say)

平等(equality)： 主席(chairman)/吴修平(Wu Xiu-Ping)/等(etc)

平和(gentle)： 向(toward)/小平(Xiao-Ping)/和(and)

平行(parallel)： 向(toward)/小平(Xiao-Ping)/行礼(salute)

谦和(modesty)：吴学谦(Wu Xue-Qian)/和(and)

前程(future)： 前(front)/程增强(Cheng Zeng-Qiang)/（

前身(preexistence)：魏光前(Wei Guang-Qian)/身(body)

请安(pay respects to)： 请(invite)/安金鹏(An Jin-Peng)/寒假(winter vacation)

若是(if)： 。/吕赫若(Lv He-Ruo)/是(is)

商周(“Shang” dynasty and “Zhou” dynasty)： 台商(Taiwan trader)/周荣顺(Zhou Rong-Shun)/先生(mister)

生就(be born with)： 主任(director)/徐寅生(Xu Yin-Sheng)/就(toward)

生来(be born with)： 对于(toward)/吕建生(Lv Jian-Sheng)/来说(toward)

帅才(person with marshal’s ability)： /刘帅(Liu Shuai)/才(just)

水上(aquatic)： /李长水(Li Chang-Shui)/上任(take a post)

为何(why)： 为(wei)/何鲁丽(He Lu-Li)/。

文中(in the text)： 主任(director)/陈振文(Chen Zhen-Wen)/中(middle)

西站(west station)： 发现(see)/张海西(Zhang Hai-Xi)/站(stand)

学说(theory)： 逄新学(Pang Xin-Xue)/说(say)

一等(first class)： 、/陆定一(Lu Ding-Yi)/等(etc)

怡和(mellowness)： 、/张怡(Zhang Yi)/和(and)

永不(never)： 责备(accuse)/仲永(Zhong-Rong)/不(no)

有关(about)： 有(has)/关天培(Guan Tian-Pei)/的(‘s)

远在(far away)： 会长(chairman)/齐怀远(Qi Huai-Ruan)/在(at)

在理(reasonable)： 在(at)/理琪(Li Qi)/司令员(chief of staff)

照说(ordinarily)： 学生(student)/毛照(Mao Zhao)/说(say)

正品(quality goods)： 《/朱乃正(Zhu Nai-Zheng)/品艺录(note)

正在(in process of)： 部长(minister)/孙家正(Sun Jia-Zheng)/在(at)

之和(summation)： 会长(chairman)/朱穆之(Zhu Mu-Zhi)/和(and)

中和(counteract)： 院士(academician)/吴咸中(Wu Xian-Zhong)/和(and)

主张(affirmation)： 业主(owner)/张洪芳(Zhang Hong-Fang)/被(by)

子孙(offspring)： 子(son)/孙占海(Sun Zhan-Hai)/是(is)

Appendix II. Some error samples in ICTCLAS (Missing or error NE is italic and underlined)

1. [LOC: *龙(dragon)/n 胜(defeat)/v 镇(town)/n*] [LOC: *勒(rein in)/v 黄村(Huang village)/ns*] 村主任(village director)/n [PER: 梁/nf 光林/nl](Liang Guang-Lin)
   Translation: Liang Guang-Lin, the village director of Long-Sheng town Le-Huang village.
2. [ORG: 湘潭市/ns 中级/b 人民法院/l](XiangTan city intermediate people's court)nt 裁定(judge)/vn [ORG: *湖(lake)/n 南方(South)/s*] 按(according to)/p 21.6%/m 的(of)/u 比例(proportion)/n 赔偿(compensate)/v [ORG: *河南(He Nan)/ns 方(just)/d*] 38 万(380,000)/m 元(Yuan)/q ，/w [ORG*:河(river)/n 南方(South)/s*] 不(don't)/d 同意(agree)/v ，/w 而(but)/c [ORG: *湖南(HuNan)/ns 方(Fang)/nl*] *则(Ze)/nl* 认为(consider)/v 应(ought)/v 按(according to)/p 法律(law)/n 裁定(judge)/vn 办(transact)/v 。/w
   Translation: XiangTan intermediate people's court sentence HuNan compensate HeNan 380,000 Yuan (21.6%), HeNan disgree while HuNan think it ought to judge by law.
3. 向(toward)/p 站(stand)/v 长*江(Chang Jiang river)/ns 秀忱(Xiu-Chen)/nr* （/w 右(right)/f 二(two)/m ）/w 赠送(present)/v 锦旗(silk flag)/n 。/w
   Translation: Donate silk flag towards stationmaster Jiang Xiu-Chen (the second from right)
4. 据(according)/p [ORG: 新华社(Xin Hua She)/nt 南京(NanJing)/ns] １月(Jan)/t ６日(6th)/t 电(telegram)/n （/wf *范(Fan)/nf 春(Chun)/nl 生于(born)/v 力(power)/n* ）
   Translation: According to the report of Xin-Hua She from NanJing, Jan, 6th (Fan Chun-Sheng, Yu Li)
5. 五十(fifty)/m 年(year)/q 前(before)/f 的(of)/u *周(Zhou)/nf 公*之(Gong-Zhi)/nl 与(and)/p 红岩(Hong Ran)/nz ，/w
   Translation: The Zhou Gong and Hong Ran of fifty years ago
6. 子翼(Zi-Yi)/nl 望(look over)/v 着(at)/u *孟(Meng)/nf 德远(De-Yuan)/nl* 去(leave)/v 的(of)/u 背影(a view of sb.'s back)/n ，/w
   Translation: Zi-Yi look over Meng De-Yuan's fading view of back
7. *刘家庄(Liu Jia Zhang)/ns 村(village)/n* 的(of)/u 农民(countrymen)/n 美事不断

(happiness after happiness)/l

Translation: The peasants in Liu-Jia-Zhang village enjoy happiness after happiness.

8. 图(photo)/n 为(is)/p 大河乡(Da He Xiang)/ns *水乡(Shui Xiang)/n 村(village)/n* ６５(65)/m 岁(age)/q 的(of)/u 席(Xi)/nr 星顺(Xing-Shun)/nr 领到(draw)/v 油毡(felt)/n

Translation: in the photo is Xi Xing-Shun, a 65 years man of Da-He Xiang Shui-Xiang village, receiving the Rou felt.

Computational Linguistics and Chinese Language Processing
Vol. 8, No. 2, August 2003, pp. 61-76

Building A Chinese WordNet Via Class-Based Translation Model

Jason S. Chang[*], Tracy Lin[+], Geeng-Neng You[**], Thomas C. Chuang[++], Ching-Ting Hsieh[***]

Abstract

Semantic lexicons are indispensable to research in lexical semantics and word sense disambiguation (WSD). For the study of WSD for English text, researchers have been using different kinds of lexicographic resources, including machine readable dictionaries (MRDs), machine readable thesauri, and bilingual corpora. In recent years, WordNet has become the most widely used resource for the study of WSD and lexical semantics in general. This paper describes the Class-Based Translation Model and its application in assigning translations to nominal senses in WordNet in order to build a prototype Chinese WordNet. Experiments and evaluations show that the proposed approach can potentially be adopted to speed up the construction of WordNet for Chinese and other languages.

1. Introduction

WordNet has received widespread interest since its introduction in 1990 [Miller 1990]. As a large-scale semantic lexical database, WordNet covers a large vocabulary, similar to a typical college dictionary, but its information is organized differently. The synonymous word senses are grouped into so-called synsets. Noun senses are further organized into a deep IS-A hierarchy. The database also contains many semantic relations, including hypernyms, hyponyms, holonyms, meronyms, etc. WordNet has been applied in a wide range of studies on

[*] Department of Computer Science, National Tsing Hua University
101, Sec. 2, Kuang Fu Road, Hsinchu, Taiwan, ROC E-mail: jschang@cs.nthu.edu.tw
[+] Department of Communication Engineering, National Chiao Tung University
1001, University Road, Hsinchu, Taiwan, ROC E-mail: tracylin@mail.nctu.edu.tw
[**] Department of Information Manangement, National Taichung Institute of Technology
San Ming Road, Taichung, Taiwan, ROC E-mail: gny@mail.ntit.edu.tw
[++] Dept of Computer Science, Van Nung Institute of Technology
1 Van-Nung Road, Chung-Li, Taiwan, ROC E-mail: tomchuang@cc.vit.edu.tw
[***] Panasonic Taiwan Laboratories Co., Ltd. (PTL) E-Mail: chingting@ptl.com.tw

such topics as word sense disambiguation [Towell and Voothees, 1998; Mihalcea and Moldovan, 1999], information retrieval [Pasca and Harabagiu, 2001], and computer-assisted language learning [Wible and Liu, 2001].

Thus, there is a universally shared interest in the construction of WordNet in different languages. However, constructing a WordNet for a new language is a formidable task. To exploit the resources of WordNet for other languages, researchers have begun to study ways of speeding up the construction of WordNet for many European languages [Vossen, Diez-Orzas, and Peters, 1997]. One of many ways to build a WordNet for a language other than English is to associate WordNet senses with appropriate translations. Many researchers have proposed using existing monolingual and bilingual Machine Readable Dictionaries (MRD) with an emphasis on nouns [Daude, Padro & Rigau, 1999]. Very little study has been done on using corpora or on covering other parts of speech, including adjectives, verbs, and adverbs. In this paper, we describe a new method for automating the process of constructing Chinese WordNet. The method was developed specifically for nouns and is capable of assigning Chinese translations to some 20,000 nominal synsets in WordNet.

The rest of this paper is divided into four sections. The next section provides the background on using a bilingual dictionary to build a Chinese WordNet and semantic concordance. Section 3 describes a class-based translation model for assigning translations to WordNet senses. Section 4 describes the experimental setup and results. A conclusion is provided in Section 5 along with directions of future work.

2. From Bilingual MRD and Corpus to Bilingual Semantic Database

In this section, we describe the proposed method for automating the construction process of a Chinese WordNet. We have experimented to find the simplest way of attaching an appropriate translation to each WordNet sense under a Class-Based Translation Model. The translation candidates are taken from a bilingual word list or Machine Readable Dictionaries (MRDs). We will use an example to show the idea, and a formal description will follow in Section 3.

Table 1. *Words in the same conceptual class that often share common Chinese characters in their translations.*

Code (set title)	Hyponyms	Chinese translation
fish (aquatic vertebrate)	carp	鯉魚
fish (aquatic vertebrate)	catfish	鯰魚
fish (aquatic vertebrate)	eel	鰻魚
complex (building)	factory	工廠
complex (building)	cannery	罐頭工廠
complex (building)	mill	製造廠
speech (communication)	discussion	討論;議論

speech (communication)	argument	論據;論點;爭論
speech (communication)	debate	辯論

Let us consider the example of assigning appropriate translations for the nominal senses of "plant" in WordNet 1.7.1. The noun "plant" in WordNet has four senses:

1. plant, works, industrial plant (buildings for carrying on industrial labor);
2. plant, flora, plant life (a living organism lacking the power of locomotion);
3. plant (something planted secretly for discovery by another person);
4. plant (an actor situated in the audience whose acting is rehearsed but seems spontaneous to the audience).

The following translations are listed for the noun "plant" in the Longman Dictionary of Contemporary English (English-Chinese Edition) [Longman Group 1992]:

1. 植物, 2. 設備, 3. 機器, 4. 工廠, 5. 內線人, and 6. 栽的贓 .

For words such as "plant" with multiple senses and translations, the question arises: Which translation goes with which synset? We make the following observations that are crucial to the solution of the problem:

1. Each nominal synset has a chain of hypernyms which give ever more general concepts of the word sense. For instance, *plant*-1 is a *building complex*, which in turn is a *structure* and so on and so forth, while *plant*-2 can be generalized as *a life form*.
2. The hyponyms of a certain top concept in WordNet form a set of semantically related word senses.
3. Semantically related senses tend to have surface realization in Chinese with shared characters.

For instance, *building complex* spawns the hyponyms *factory*, *mill*, *assembly plant*, *cannery*, *foundry, maquiladora*, etc., all of which realize in Chinese using the characters "廠" or "工廠." Therefore, we can say that there is a high probability that senses which are direct or indirect hyponyms of *building complex* share the Chinese characters "工" and "廠" in their Chinese translations. Therefore, it is clear that one can determine that *plant*-1, a hyponym of *building complex*, should have "工廠" instead of "植物" as its translation. See Table 1 for more examples. That intuition can be expanded into a systematic way of assigning the most appropriate translation to a given word sense. Figure 1 shows how the method works for four senses of *plant*.

In the following, we will consider the task of assigning the most appropriate translation to *plant*-1, the first sense of the noun "plant." First, the system looks up "plant" in the Translation Table (T Table) for candidate translations of *plant*-1:

(*plant*, 植物), (*plant*, 機器), (*plant*, 設備), (*plant*, 工廠), (*plant*, 內線人), (*plant*, 栽的贓).

Next, the semantic class *g* to which *plant*-1 belongs is determined by consulting the Semantic Class Table (SC Table). In this study we use some 1,145 top hypernyms *h* to represent the class of word senses that are direct or transitive hyponyms of *h*. The path designator of *h* in WordNet is used to represent the class. The hypernyms are chosen to correspond roughly to the division of sets of words in the Longman Lexicon of Contemporary English (LLOCE) [McArthur 1992]. Table 2 provides examples of classes related to *plant* and their class codes.

Table 2. *Words in four classes related to the noun* ***plant****.*

English	WN sense	Class Code	Words in the Class
Plant	1	N001004003030	factory, mill, assembly plant, …
Plant	2	N001001005	flora, plant life, …
Plant	3	N001001015008	thought, idea, …
Plant	4	N001001003001001	producer, supernatural, …
Plant	4	N001003001002001	announcer, conceiver, …

For instance, *plant*-1 belongs to the class *g* represented by the WordNet synset (*structure*, *construction*):

g = N001004003030.

Subsequently, the system evaluates the probabilities of each translation conditioned on the semantic class *g*:

P(“植物” | N001004003030),

P(“機器” | N001004003030),

P(“設備” | N001004003030),

P(“工廠” | N001004003030),

P(“內線人” | N001004003030),

P(“栽的贓” | N001004003030).

These probabilities are not evaluated directly. The system takes apart the characters in a translation and looks up P(*u* | *g*), the probabilities for each translation character *u* conditioned on *g*:

P(“植” | **N001004003030**) = 0.000025,

P(“物” | **N001004003030**) = 0.000025,

P(“機” | **N001004003030**) = 0.00278,

P(“器” | **N001004003030**) = 0.00278,

P(“設” | **N001004003030**) = 0.00306,

P(“備” | **N001004003030**) = 0.00075,

P(“工” | **N001004003030**) = 0.00711,

P(“廠” | **N001004003030**) = 0.01689,

P(“內” | **N001004003030**) = 0.00152,

P(“線” | **N001004003030**) = 0.00152,

P(“人” | **N001004003030**) = 0.00152,

P(“栽” | **N001004003030**) = 0.00152,

P(“的” | **N001004003030**) = 0.00152,

P(“贓” | **N001004003030**) = 0.00152.

Note that to deal with lookup failure, a smoothing probability is given (0.000025, derived using the Good-Turing method). By using a statistical estimate based on simple linear interpolation, we can get

P(“工廠” | *plant*-1) ≈ P (“工廠” | **N001004003030**)

$$\approx \frac{1}{2} P(\text{“工”} \mid \textbf{N001004003030}) + \frac{1}{2} P(\text{“廠”} \mid \textbf{N001004003030})$$

$$= \frac{1}{2} (0.0178+0.0073) = 0.0124.$$

Similarly, we have

P(“植物” | **N001004003030**) = 0.0013,

P(“機器” | **N001004003030**) = 0.0023,

P(“設備” | **N001004003030**) = 0.0028,

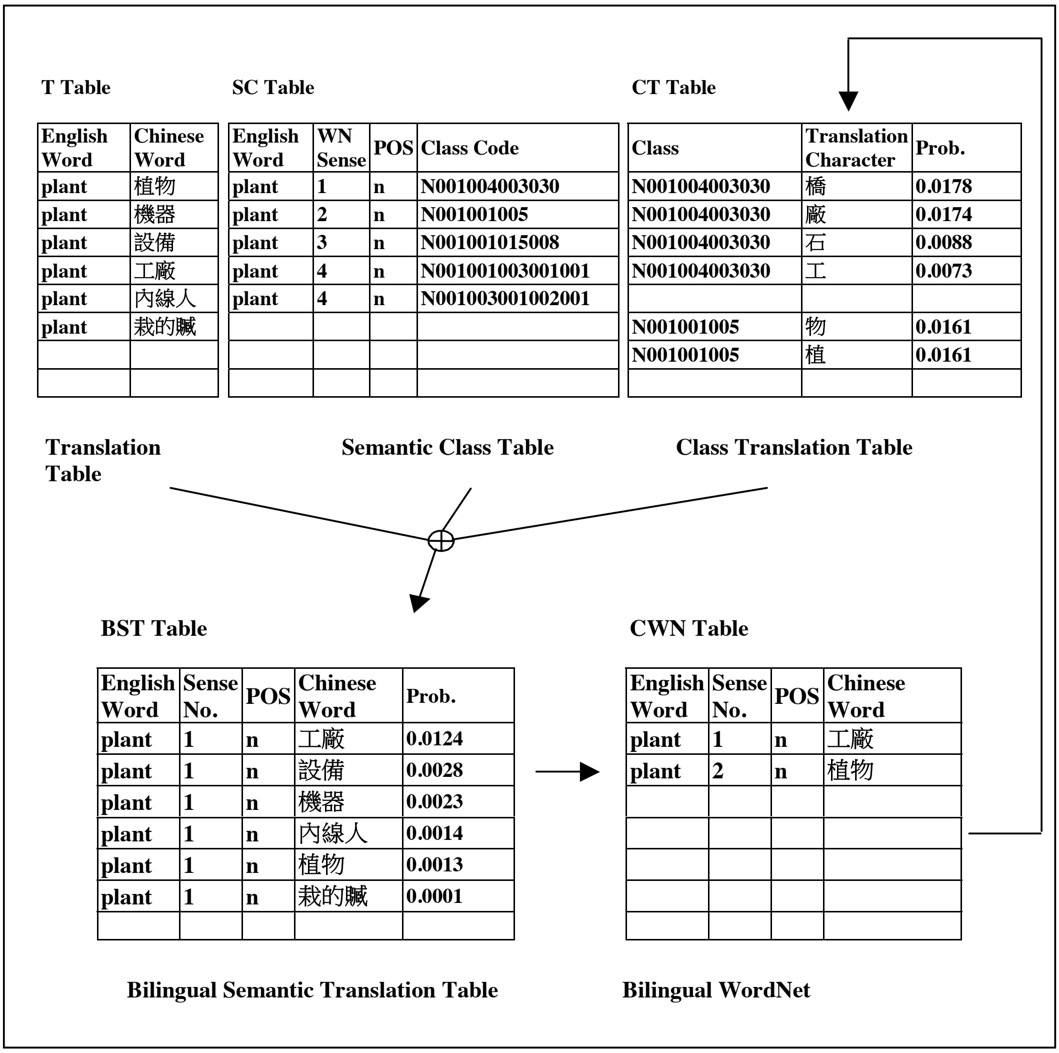
P(“內線人” | **N001004003030**) = 0.0014,

P(“栽的贓” | **N001004003030**) = 0.0001.

Finally, by choosing the translation with the highest probabilistic value for *g*, we can get an entry for Chinese WordNet (CWN Table):

(plant, 工廠, n, 1, “buildings for carrying on industrial labor”)

After we get the correct translation of *plant*-1 and many other word senses in *g*, we will be able to re-estimate the class-based translation probability for *g* and produce a new CT Table. However, the reader may wonder how we can get the initial CT Table. This dilemma can be resolved by adopting an iterative algorithm that establishes an initial CT Table and makes revision until the values in the CT Table converge. More details will be provided in Section 3.



T Table

English Word	Chinese Word
plant	植物
plant	機器
plant	設備
plant	工廠
plant	內線人
plant	栽的贓

Translation Table

SC Table

English Word	WN Sense	POS	Class Code
plant	1	n	N001004003030
plant	2	n	N001001005
plant	3	n	N001001015008
plant	4	n	N001001003001001
plant	4	n	N001003001002001

Semantic Class Table

CT Table

Class	Translation Character	Prob.
N001004003030	橋	0.0178
N001004003030	廠	0.0174
N001004003030	石	0.0088
N001004003030	工	0.0073
N001001005	物	0.0161
N001001005	植	0.0161

Class Translation Table

BST Table

English Word	Sense No.	POS	Chinese Word	Prob.
plant	1	n	工廠	0.0124
plant	1	n	設備	0.0028
plant	1	n	機器	0.0023
plant	1	n	內線人	0.0014
plant	1	n	植物	0.0013
plant	1	n	栽的贓	0.0001

Bilingual Semantic Translation Table

CWN Table

English Word	Sense No.	POS	Chinese Word
plant	1	n	工廠
plant	2	n	植物

Bilingual WordNet

Fig. 1 *Using CBTM to build Chinese WordNet. This example shows how the first sense of plant receives an appropriate translation via the Class-Based Translation Model and how the model can be trained iteratively.*

3. The Class-Based Translation Model

In this section, we will formally describe the proposed class-based translation model, how it can be trained, and how it can be applied to the task of assigning appropriate translations to different word senses. Given E_k, the kth sense of an English word E in the WordNet, the probability of its Chinese translation is denoted as $P(\ C \mid E_k)$. Therefore, the best Chinese translation C^* is

$$C^*(E_{\mathrm{k}}) \cong \arg\max_{C \in T(E)} P(C \mid E_{\mathrm{k}}) \quad , \tag{1}$$

where *T*(*X*) is the set of Chinese translations of sense X listed in a bilingual dictionary.

Based on our observation that semantically related senses tend to be realized in Chinese using shared Chinese characters, we tie together the probability functions of translation words in the same semantic class and use the class-based probability as an approximation. Thus, we have

$$P(C \mid E_{\mathrm{k}}) \cong P(C \mid g) \ , \tag{2}$$

where $g = g(E_{\mathrm{k}})$ is the semantic class containing E_{k}.

The probability of *P*(*C*|g) can be estimated using the Expectation and Maximization Algorithm as follows:

(Initialization) $P(C \mid E_{\mathrm{k}}) = \dfrac{1}{m}$, $m = \mid T(E) \mid$ and $C \in T(E)$; (3)

(Maximization) $P(C \mid g) = \dfrac{\sum_{E,k,i} P(C_{\mathrm{i}} \mid E_{\mathrm{k}}) I(C = C_{\mathrm{i}}) I(E_{\mathrm{k}} \in g)}{\sum_{E,k,i} P(C_{\mathrm{i}} \mid E_{\mathrm{k}}) I(E_{\mathrm{k}} \in g)}$, (4)

where C_{i} = the *i*th translation of E_{k} in $T(E_{\mathrm{k}})$,

$I(x) = 1$ if *x* is true and 0 otherwise;

(Expectation) $P_1(C \mid E_{\mathrm{k}}) = P(C \mid g)$, (5)

where $g = g\ (E_{\mathrm{k}})$ is the class that contains E_{k} ;

(Normalization) $P(C \mid E_{\mathrm{k}}) = \dfrac{P_1(C \mid E_{\mathrm{k}})}{\sum_{D \in T(E_{\mathrm{k}})} P_1(D \mid E_{\mathrm{k}})}$. (6)

In order to avoid the problem of data sparseness, *P*(*C*|*g*) is estimated indirectly via the unigrams and bigrams in *C*. We also weigh the contribution of each unigram and bigram to avoid the domination of a particular character in the semantic class. Therefore, we rewrite Equations 4 and 5 as follows:

(Maximization) $P_u(u \mid g) = \dfrac{\sum_{E,k,i,j} \frac{1}{m} I(E_{\mathrm{k}} \in g) I(u = u_{\mathrm{i,j}}) P(u_{\mathrm{i,j}} \mid E_{\mathrm{k}})}{\sum_{E,k,i,j} \frac{1}{m} I(E_{\mathrm{k}} \in g) P(u_{\mathrm{i,j}} \mid E_{\mathrm{k}})}$, (4a)

where $u_{\mathrm{i,j}}$ = the *j*th unigram of the *i*th translation in $T(E_{\mathrm{k}})$,

m = the number of characters in the *i*th translation in $T(E_{\mathrm{k}})$,

$$P_b(b \mid g) = \frac{\sum_{E,k,i,j} \frac{1}{m-1} I(E_k \in g) I(b = b_{i,j}) P(b_{i,j} \mid E_k)}{\sum_{E,k,i,j} \frac{1}{m-1} I(E_k \in g) P(b_{i,j} \mid E_k)}, \tag{4b}$$

where $b_{i,j}$ = the jth overlapping bigram of the ith translation in $T(E_k)$;

(Expectation) $$P_1(C \mid E_k) \cong P(C \mid g) \cong \sum_{i=1}^{m} \frac{P_u(u_i \mid g)}{m} \quad \text{(unigram)}, \tag{5a}$$

$$P_1(C \mid E_k) \cong P(C \mid g) \cong \sum_{i=1}^{m} \frac{P_u(u_i \mid g)}{2m} + \sum_{i=1}^{m-1} \frac{P_b(b_i \mid g)}{2(m-1)} \quad \text{(+bigram)}, \tag{5b}$$

where u_i is a unigram, b_i is an on overlapping bigram of C, and m is the number of characters in C .

For instance, assume that we have the first sense *trunk*-1 of the word *trunk* in WordNet and the translations in LDOCE as follows:

trunk-1 (the main stem of a tree; usually covered with bark; the bole is usually the part that is commercially useful for lumber),

Translations of *trunk* — 大皮箱, 大衣箱, 樹幹, and 象鼻 .

Initially, the probabilities of each translation for *trunk*-1 are as follows:

P(大皮箱 | *trunk*-1) = 1/4, *P*(大衣箱 | *trunk*-1) = 1/4,

P(樹幹 | *trunk*-1) = 1/4, *P*(象鼻 | *trunk*-1) = 1/4.

Table 3 shows the words in the semantic class N001004001018013014 (stalk, stem), containing *trunk*-1 and relevant translations. Following Equations 4a and 4b, we took the unigrams and overlapping bigrams from these translations to calculate the probability of unigram and bigram translations for (stalk, stem). Although initially irrelevant translations such as bulb-電燈泡(light bulb) can not be excluded, after one iteration of the maximization step, the noise is suppressed substantially, and the top ranking translations shown in Tables 4 and 5 seem to be the "genus" terms of the class. For instance, the top ranking unigrams for N001004001018013014 include莖 (stem), 枝 (branch), 條 (branch), 根 (stump) 樹 (tree) 幹 (trunk) etc. Similarly, the top ranking bigrams include 球莖 (bulb), 樹枝 (branch), 柳條 (willow branch), and 樹幹 (trunk). All indicate the general concepts of the class.

With the unigram translation probability *P*(*u* | *g*), one can apply Equations 5a and 6 to proceed with the Expectation Step and calculate the probability of each translation candidate for a word sense as shown in Example 1:

Example 1.

P_1(樹幹|*trunk*-1)=1/2*(*P*(樹|N001004001018013014)+*P*(幹| N001004001018013014))
=1/2*(0.0145+0.0103) = **0.0124**,

P_1(象鼻|*trunk*-1) =1/2*(*P*(象|N001004001018013014)+*P*(鼻 |N001004001018013014))
=1/2* (0.00054+0.00054) = 0.00054,

P_1(大皮箱|*trunk*-1) =1/3*(*P*(大|N001004001018013014)+*P*(皮|N001004001018013014)
+ *P*(箱 |N001004001018013014)) ,
=1/3*(0.0074+0.00036+0.00072) = 0.00283,

P_1(大衣箱|*trunk*-1) =1/3*(*P*(大|N001004001018013014)+*P*(衣|N001004001018013014)
+ *P*(箱 | N001004001018013014))
=1/3*(0.0074 + 0.00043 + 0.00072) = 0.00285

P (樹幹 | *trunk*-1) = 0.0124/(0.0124+0.00054+0.00283+0.00285) = 0.665950591,

P (象鼻 | *trunk*-1) = 0.0124/(0.0124+0.00054+0.00283+0.00285) = 0.0290010741,

P (大皮箱 | *trunk*-1) = 0.0124/(0.0124+0.00054+0.00283+0.00285) = 0.1519871106,

P (大衣箱 | *trunk*-1) = 0.0124/(0.0124+0.00054+0.00283+0.00285) = 0.1530612245.

Using simple linear interpolation of translation unigrams and bigrams (Equation 5b), the probability of each translation candidate for a word sense can be calculated as shown in Example 2:

Example 2.

P_1(樹幹 | *trunk*-1) = 1/2 * {1/2 * (*P*(樹 | N001004001018013014)
+*P*(幹 | N001004001018013014))
+*P*(樹幹 | N001004001018013014) }
= 1/2 * (0.0124 + 0.0145) = **0.01345**,

P_1(象鼻 | *trunk*-1) = 1/2 * {1/2 * (*P*(象 | N001004001018013014)
+*P*(鼻 | N001004001018013014))
+*P*(象鼻 | N001004001018013014) }
= 1/2 * (0.00054 + 0.00107) = 0.000805,

P_1(大皮箱 | *trunk*-1) = 1/2 * {1/3 * (*P*(大 | N001004001018013014)
+ *P*(皮 | N001004001018013014))
+ *P*(箱 | N001004001018013014)}
+ 1/2 * (*P*(大皮 | N001004001018013014)
+*P*(皮箱 | N001004001018013014)) }
= 1/2 * (0.00283 + 0.00054) = 0.001685,

P_1(大衣箱 | *trunk*-1) = 1/2 * {1/3 * (*P*(大 | N001004001018013014)
+ *P*(衣 | N001004001018013014))
+ *P*(箱 | N001004001018013014) }
+ 1/2 * (*P*(大衣 | N001004001018013014)

$+P($ 衣箱 $\mid$ N001004001018013014 $)\,)\,\}$

$= 1/2 * (0.00285 + 0.00054) = 0.001695$

$P\,($樹幹$\mid trunk\text{-}1) = 0.01345/(0.01345+0.000805+0.001685+0.001695)= 0.76268783669,$

$P\,($象鼻$\mid trunk\text{-}1) = 0.000805/(0.01345+0.000805+0.001685+0.001695)$

$= 0.045647859371,$

$P\,($大皮箱$\mid trunk\text{-}1) = 0.001685/(0.01345+0.000805+0.001685+0.001695)$

$= 0.095548624894,$

$P\,($大衣箱$\mid trunk\text{-}1) = 0.001695/(0.01345+0.000805+0.001685+0.001695)$

$= 0.096115679047.$

Table 3. *Words and their translations in the semantic class N001004001018013014*

English *E*	WN sense *k*	G(E_k)	Chinese Translation
Beanstalk	1	N001004001018013014	豆莖
Bole	2	N001004001018013014	樹幹
Branch	2	N001004001018013014	分枝
Branch	2	N001004001018013014	部門
Branch	2	N001004001018013014	樹枝
Brier	2	N001004001018013014	荊棘
Bulb	1	N001004001018013014	球莖狀物
Bulb	1	N001004001018013014	電燈泡
Cane	2	N001004001018013014	籐條
Cutting	2	N001004001018013014	剪報
Cutting	2	N001004001018013014	插枝
Stick	2	N001004001018013014	小樹枝
Stick	2	N001004001018013014	手杖
Stem	2	N001004001018013014	家系
Stem	2	N001004001018013014	幹

Table 4. *Probabilities of each unigram for the semantic class containing* trunk-*1, etc.*

Unigram (u)	Semantic Class Code (g)	*P*(*u* | *g*)
莖	N001004001018013014	0.0706
枝	N001004001018013014	0.0274
豆	N001004001018013014	0.0216
條	N001004001018013014	0.0162
樹	N001004001018013014	0.0145
根	N001004001018013014	0.0134

幹	N001004001018013014	0.0103
籐	N001004001018013014	0.0080

Table 5. *Probabilities of each bigram for the semantic class containing* trunk-*1, etc.*

Bigram (b)	**Semantic Class Code (g)**	$P(b \mid g)$
球莖	N001004001018013014	0.0287
柳條	N001004001018013014	0.0269
樹幹	N001004001018013014	0.0145
樹枝	N001004001018013014	0.0144
嫩枝	N001004001018013014	0.0134
…	…………………………	…

Both examples show that the class-based translation model produces reasonable probabilistic values. The examples also show that for *trunk*-1, the linear interpolation method gives a higher probabilistic value for the correct translation “樹幹” than the unigram-based approach does (0.76268783669 vs. 0.665950591). In this case, linear interpolation is a better parameter estimation scheme. Our experiments showed, in general, that combining both unigrams and bigrams does lead to better overall performance.

4. Experiments

We carried out two experiments to see how well CBTM can be applied to assign appropriate translations to nominal senses in WordNet. In the first experiment, the translation probability was estimated using Chinese character unigrams, while in the second experiment, both unigrams and bigrams were used. The linguistic resources used in the experiments included:

1. **WordNet 1.6**: WordNet contains approximately 116,317 nominal word senses organized into approximately 57,559 word meanings (synsets).
2. **Longman English-Chinese Dictionary of Contemporary English (LDOCE E-C)**: LDOCE is a learner’s dictionary with 55,000 entries. Each word sense contains information, such as a definition, the part-of-speech, examples, and so on. In our method, we take advantage of its wide coverage of frequently used senses and corresponding Chinese translations. In the experiments, we tried to restrict the translations to lexicalized words rather than descriptive phrases. We set a limit on the length of a translation: nine Chinese characters or less. Many of the nominal entries in WordNet are not covered by learner dictionaries; therefore, the experiments focused on those senses for which Chinese translations are available in LDOCE.
3. **Longman Lexicon of Contemporary English (LLOCE)**: LLOCE is a bilingual

taxonomy, which brings together words with related meanings and lists them in topical/semantic classes with definitions, examples, and illustrations.

The three tables shown in Figure 1 were generated in the course of the experiments:

1. The Translation Table has 44,726 entries and was easily constructed by extracting Chinese translations from LDOCE E-C [Proctor 1988].
2. We obtained the Sense Class Table by finding the common hypernyms of sets of words in LLOCE. 1,145 classes were used in the experiments.
3. The Class Translation Table was constructed using the EM algorithm based on the T Table and SC Table. The CT Table contains 155,512 entries.

Table 6 shows the results of using CBTM and Equation 1 to find the best translations for a word sense. We are concerned with the coverage of word senses in average text. In that sense, the translation of *plant*-3 is incorrect, but this error is not very significant, since this word sense is used infrequently. We chose the WordNet semantic concordance, SEMCOR, as our testing corpus. There are 13,494 distinct nominal word senses in SEMCOR. After the translation probability calculation step, our results covered 10,314 word senses in SEMCOR; thus, the coverage rate was 76.43%.

Table 6. *The results and appropriate translations for each sense of the English word.*

English	WN sense	Chinese Translation	Appropriate Chinese Translation
Plant	1	工廠	工廠
Plant	2	植物	植物
Plant	3	內線人	栽的贓
Plant	4	內線人	內線人
Spur	1	鼓勵	鼓勵
Spur	2	激勵	刺，針
Spur	4	馬刺	馬刺
Spur	5	支線	支線
Bank	1	銀行	銀行
Bank	2	邊坡	沙洲
Bank	3	庫	庫，儲存所
Scale	1	記數法或基準	記數法或基準
Scale	2	比例	規模
Scale	3	比例	比例
Scale	5	脫下的乾燥皮屑	脫下的乾燥皮屑
Scale	6	音階	音階

To see how well the model assigns translations to WordNet senses appearing in average text, we randomly selected 500 noun instances from SEMCOR as our test data. There were 410 distinct words. Only 75 words had a unique sense in WordNet. There were 77 words with two senses in WordNet, while 70 words had three senses in WordNet, and so on. The average degree of sense ambiguity was 4.2.

Table 7. *The degree of ambiguity and number of words in the test data with different degree of ambiguity.*

Degree of ambiguity **# of senses in WordNet**	**# of word types in the test data**	**Examples**
1	75	aptitude, controversy, regret
2	77	camera, fluid, saloon
3	70	drain, manner, triviality
4	51	confusion, fountain, lesson
5	35	isolation, pressure, spur
6	25	blood, creation, seat
7	28	column, growth, mind
8	9	contact, hall. program
9	7	body, company, track
10	8	bank, change, front
>10	25	control, corner, deaft

Among our 500 test data, 280 entries were the first sense, while 112 entries were the second sense. Over half of the words had the meaning of the first sense. Therefore, the first sense was most frequently used. Therefore, it was found to be more important to get the first and the second senses right. We manually gave each word sense an appropriate Chinese translation whenever one was available from LDOCE. From these translations, we found the following:

1. There were 491 word senses for which corresponding translations were available from LDOCE.
2. There were 5 word senses for which no relevant translations could be found in LDOCE due to the limited coverage of this learner's dictionary. Those word senses and relevant translations included assignment-2 (轉讓), marriage-3 (婚禮), snowball-1(繡球莢), prime-1(質數), and program-7 (政綱).
3. There were 4 words, that have no translations due to the particular cross-referencing scheme of LDOCE. Under this scheme, some nouns in LDOCE are not directly given a definition and translation, but rather a pointer to a more frequently used spelling. For instance, "groom" is given a pointer to "BRIDEGROOM" rather than the relevant definition and translation ("新郎").

In the first experiment, we started out by ranking the relevant translations for each noun sense using the class-based translation model. If two translations had the same probabilistic value, we gave them the same rank. For instance, Table 8 shows that the top 1 translation for *plant*-1 was "工廠."

Table 8. *The rank of each translation corresponding to each word sense. (plant-2, 栽的贓) and (plant-2, 設備) have the same probability and rank.*

English	Semantic class	WN sense	Chinese Translation	Probability	Rank
Plant	N001004003030 (structure)	1	工廠	0.012372	1
Plant	N001004003030 (structure)	1	設備	0.002823	2
Plant	N001004003030 (structure)	1	機器	0.002270	3
Plant	N001004003030 (structure)	1	內線人	0.001375	4
Plant	N001004003030 (structure)	1	植物	0.001278	5
Plant	N001004003030 (structure)	1	栽的贓	0.000130	6
Plant	N001001005 (flora)	2	植物	0.016084	1
Plant	N001001005 (flora)	2	機器	0.002623	2
Plant	N001001005 (flora)	2	工廠	0.000874	3
Plant	N001001005 (flora)	2	設備	0.000525	4
Plant	N001001005 (flora)	2	栽的贓	0.000525	4
Plant	N001001005 (flora)	2	內線人	0.000360	5

Table 9. *The recall rate in the first experiment*

The number of top-ranking translations	Correct Entries (Total entries =500)	Recall rate (unigram)	Recall rate (unigram+bigram)
Top 1	344	68.8%	70.2%
Top 2	408	81.6%	83.2%
Top 3	441	88.2%	89.0%
Top 4	449	89.8%	91.4%
Top 5	462	92.4%	93.2%

We used the same method to evaluate the recall rate in the second experiment, where both unigrams and bigrams were used. The experimental results show a slight improvement over the results obtained using only unigrams.

In these experiments, we estimated the translation probability based on unigrams and bigrams. The evaluation results confirm our observation that we can exploit shared characters in translations of semantically related senses to obtain relevant translations. We evaluated the experimental results based on whether the Top 1 to Top 5 translations covered all appropriate translations. If we selected the Top 1 translation in the first experiment as the most appropriate translation, there were 344 correct entries, and the recall rate was 68.8%. The Top 2 translations covered 408 correct entries, and the recall rate was 81.6%. Table 9 shows the recall rate with regard to the number of top-ranking translations used for the purpose of evaluation.

5. Conclusion

In this paper, a statistical class-based translation model for the semi-automatic construction of a Chinese WordNet has been proposed. Our approach is based on selecting the appropriate Chinese translation for each word sense in WordNet. We observe that a set of semantically related words tend to share some Chinese characters in their Chinese translations. We propose to rely on the knowledge base of a Class Based Translation Model derived from statistical analysis of the relationship between semantic classes in WordNet and translations in the bilingual version of the Longman Dictionary of Contemporary English (LDOCE). We carried out two experiments that show that CBTM is effective in speeding up the construction of a Chinese WordNet.

The first experiment was based on the translation probability of unigrams, and the second was based on both unigrams and bigrams. Experimental results show that the method produces a Chinese WordNet covering 76.43% of the nominal senses in SEMCOR, which implies that a high percentage of the word senses can be effectively handled. Among our 500 testing cases, the recall rate was around 70%, 80% and 90%, respectively, when the Top 1, Top 2, and Top 3 translations were evaluated. The recall rate when using both unigrams and bigrams was slightly higher than that when using only unigrams. Our results can be used to assist the manual editing of word sense translations.

A number of interesting future directions present themselves. First, obviously, there is potential for combining two or more methods to get even better results in connecting WordNet senses with translations. Second, although nouns are most important for information retrieval, other parts of speech are important for other applications. We plan to extend the method to verbs, adjectives and adverbs. Third, the translations in a machine readable dictionary are at times not very well lexicalized. The translations in a bilingual corpus cauld be used to improve the degree of lexicalization.

Acknowledgement

This study was partially supported by grants from the National Science Council (NSC 90-2411-H-007-033-MC) and the MOE (project EX 91-E-FA06-4-4).

References

Daudé, J., L. Padró and G. Rigau, “Mapping Multilingual Hierarchies using Relaxation Labelling,” *Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999

Daudé, J., L. Padró and G. Rigau, "Mapping WordNets using Structural Information," *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, 2000.

McArthur, T., "Longman Lexicon of Contemporary English," Longman Group (Far East) Ltd., Hong Kong, 1992.

Mihalcea, R. and D. Moldovan., "A method for Word Sense Disambiguation of unrestricted text," *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 1999, pp. 152-158.

Miller, G., "Five papers on WordNet," *International Journal of Lexicography,* 3(4), 1990.

Pasca, M. and S. Harabagiu, "The Informative Role of WordNet in Open-Domain Question Answering," in *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, June 2001, Carnegie Mellon University, Pittsburgh PA, pp. 138-143.

Proctor, P., "Longman English-Chinese Dictionary of Contemporary English," Longman Group (Far East) Ltd., Hong Kong, 1988.

Towell, G. and E. Voothees, "Disambiguating Highly Ambiguous Words," *Computational Linguistics*, 24(1) 1998, pp. 125-146.

Vossen, P., P. Diez-Orzas and W. Peters, "The Multilingual Design of the EuroWordNet Database," *Processing of the IJCAI-97 workshop Multilingual Ontologies for NLP Applications,* 1997.

Wible, D. and A. Liu, "A syntax-lexical semantics interface analysis of collocation errors," *PacSLRF* 2001.

Computational Linguistics and Chinese Language Processing
Vol. 8, No. 2, August 2003, pp. 77-92

Preparatory Work on Automatic Extraction of Bilingual Multi-Word Units from Parallel Corpora

Boxing Chen[*], Limin Du[*]

Abstract

Automatic extraction of bilingual Multi-Word Units is an important subject of research in the automatic bilingual corpus alignment field. There are many cases of single source words corresponding to target multi-word units. This paper presents an algorithm for the automatic alignment of single source words and target multi-word units from a sentence-aligned parallel spoken language corpus. On the other hand, the output can be also used to extract bilingual multi-word units. The problem with previous approaches is that the retrieval results mainly depend on the identification of suitable Bi-grams to initiate the iterative process. To extract multi-word units, this algorithm utilizes the normalized association score difference of multi target words corresponding to the same single source word, and then utilizes the average association score to align the single source words and target multi-word units. The algorithm is based on the Local Bests algorithm supplemented by two heuristic strategies: excluding words in a stop-list and preferring longer multi-word units.

Key words: bilingual alignment; multiword unit; translation lexicon; average association score; normalized association score difference;

1. Introduction

1.1 The Background of Automatic Extraction of Bilingual Multi-Word Units

In the natural language processing field, which includes machine translation, machine assistant translation, bilingual lexicon compilation, terminology, information retrieval, natural language generation, second language teaching etc., the automatic extraction of bilingual multi-word units (steady collocations, multi-word phrases, multi-word terms etc.) is an

[*] Center for Speech Interaction Technology Research, Institute of Acoustics, Chinese Academy of Sciences
Address: 17 Zhongguancun Rd. Beijing 100080, China
E-mail: {chenbx , dulm}@iis.ac.cn

important aspect of the automatic alignment of bilingual corpus technology. Since the 1980's, the technique of automatic alignment of a bilingual corpus has undergone great improvement; and during the mid- and late-1990's, many researchers began to research the automatic construction of a bilingual translation lexicon [Fung 1995; Wu *et al.* 1995; Hiemstra 1996; Melamed 1996 etc.] Their works have focused on the alignment of single words. At the same time, the extraction of multi-word units in singular languages has been also studied. Church utilized mutual information to evaluate the degree of association between two words [Church 1990]; hence, mutual information has played an important role in multi-word unit extraction research, and it is used most often with this technology by means of a statistical method. Many researchers [Smadja 1993; Nagao *et al.* 1994; Kita *et al.* 1994; Zhou *et al.* 1995; Shimohata *et al.* 1997; Yamamoto *et al.* 1998] have utilized mutual information (or the transformation of mutual information) as an important parameter to extract multi-word units. The shortcoming of these methods is that low frequency multi-word units are easy to eliminate, and the output of extraction mainly depends on the verification of suitable Bi-grams when the iterative algorithm initiates.

Automatic extraction of bilingual multi-word units is based on the automatic extraction of bilingual word and multi-word units in singular languages. Research in this field has also proceeded [Smadja *et al.* 1996; Haruno *et al.* 1996; Melamed 1997 etc], but the problem with this approach is that it relies on statistical methods more than the characteristics of the language per se and is mainly limited to the extraction of noun phrases.

Because of the above problems and the fact that Chinese-English corpuses are commonly small, we provide an algorithm that uses the average association score and normalized association score difference. We also apply the Local Bests algorithm, stopword filtration and longer unit preference methods to extract Chinese or English multi-word units.

1.2 The Object of Our Research

In research on the results produced by single-English-word to single-Chinese-word alignment, we have found an interesting phenomenon: During the phase of Chinese word segmentation, if the translation of an English word ("*A*") comprises of several Chinese words ("*BCD*"), the mutual information and the t-score for each "*B-A, C-A, D-A*" mapping are both very high and close to each other. Thus, we can use the average association score and the normalized association score difference to extract the translation equivalent pairs of single-English-word to multiple-Chinese-word mappings.

For example, when names and professional terms are translated, "*Patterson*" is translated as "佩特逊," which includes three entries in a Chinese dictionary ("佩," "特," and "逊"); "*Internet*" is translated as "因特网," which includes three entries in a Chinese dictionary

(“因,” “特,” and “网”). Furthermore, the same situation occurs with some non-professional terms. For example, “*my*” is translated as “我的.” Also, the same rule applies to Chinese-English translation. For example, “不三不四” is translated as “*get funny,*” and “放肆” as “*get fresh*.”

Therefore, the research presented in this paper is focused on single-source-word to multi-target-word-unit alignment. The alignment of bilingual multi-word units will be the focus of our future research.

2. Algorithm

The method we use to align single source words with target multi-word units from a parallel corpus can be divided into the following steps (we use the mutual information and t-score as the association score):

(1) Word segmentation:

We do word segmentation first because Chinese has no word delimiters.

(2) Calculating the co-occurrence frequency:

If a word pair appears once in an aligned bilingual sentence pair, one co-occurrence is counted.

(3) Computing the association score of single word pairs:

We calculate the mutual information and t-score of the source words and their co-occurrence target words.

(4) Calculating the average association score and normalized association score:

We calculate the average mutual information and normalized mutual information difference, and the average t-score and normalized t-score difference of every source word and its co-occurrence target words’ N-gram (N: 2-7, since most phrases have of 2-6 words).

(5) The Local Bests algorithm:

We utilize the Local Bests algorithm to eliminate non-local best target multi-word units.

(6) Stop-word list filtration:

Some words cannot be used as the first or the last word of a multi-word unit, so we use the stop-word list to filter these multi-word units.

(7) Bigger association score preference:

After the above filtration, from among the remaining multi-word units, we choose N items with the maximal average mutual information and average t-score as the candidate target translation.

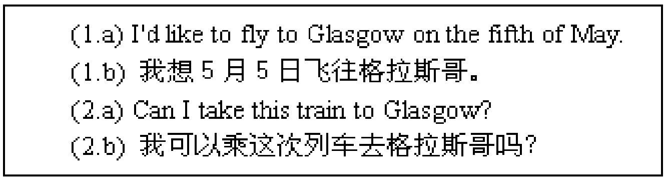
(8) Longer unit preference:

We extract multi-word units but not words, so if the longer word string C_1 entirely contains another shorter word string C_2, then string C_1 is taken as the translation of the source word.

(9) Lexicon classification:

According to the above four parameters, we classify the lexicons into four levels of translation lexicons.

We will use "*Glasgow*: 格拉斯哥," which appears in the corpus as shown in Figure 1, as an example to explain the whole process.



(1.a) I'd like to fly to Glasgow on the fifth of May.

(1.b) 我想5月5日飞往格拉斯哥。

(2.a) Can I take this train to Glasgow?

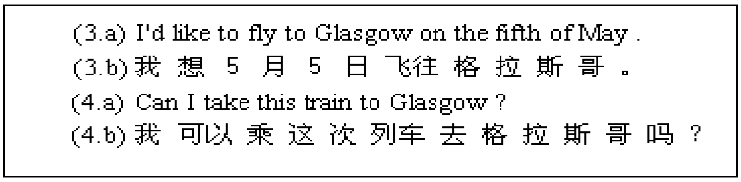
(2.b) 我可以乘这次列车去格拉斯哥吗?

Figure 1. Sentence Example.

The reasons why we choose "*Glasgow*" are: (1) the occurrence frequency of "*Glasgow*" is quite low, only two times, which is easily ignored by the previous algorithm; (2) the Chinese translation of "*Glasgow*" is unique, so the correct extraction of this lemma can prove the accuracy of our algorithm; (3) "*Glasgow*" contains four single-character words, and it will be found later that our algorithm is more effective with multi-word units made up of two words, so here we use "*Glasgow*" to prove that our algorithm is also effective with multi-word units made up of more than two words.

2.1 Chinese Word Segmentation

We used the "maximum probability word segmentation method" [Chen 1999] and *The Grammatical Knowledge-base of Contemporary Chinese* published by Peking University [Yu 1998]. The idea behind this method is: first find out all the possible words in the input Chinese string on a vocabulary basis and then find out all the possible segmentation paths, from which we can find the best path (with the maximal probability) as the output. We randomly sampled 1000 sentences to check: if we did not take "un-listed words that are divided" as an error, then the precision rate was 98.88%; but if it was being taken as an error, the precision rate was 88.74%. The unlisted words in DECC1.0 (Daily English-Chinese Corpus) were mainly the Chinese translations of foreign personal names and place names. The main focus of our research here was the aggregation of single Chinese characters that are produced through segmentation. The results of word segmentation are shown in Figure 2:



(3.a) I'd like to fly to Glasgow on the fifth of May .
(3.b) 我 想 5 月 5 日 飞往 格 拉 斯 哥 。
(4.a) Can I take this train to Glasgow ?
(4.b) 我 可以 乘 这 次 列车 去 格 拉 斯 哥 吗 ？

Figure 2. Word Segmentation Results.

2.2 Calculate the Co-occurrence Frequency

There were many translation sentence pairs in the corpus. For each possible word pair in these translation sentence pairs, the higher the probability of appearance it had, the higher the probability it had of being the correct translation word pair. We built a co-occurrence model to count the number of appearances: it was counted as a co-occurrence each time the word pair appears in a sentence pair. The reasons are as follows: First, the length of a sentence in spoken language is usually shorter than that in a written language; for example, in the corpus DECC1.0, the average length of English sentences is 7.07 words, and the average length of Chinese sentences is 6.87 words and expressions. Secondly, the corresponding sense units of English-Chinese sentence pairs in spoken language are not always aligned in terms of position, as shown in Figure 3.

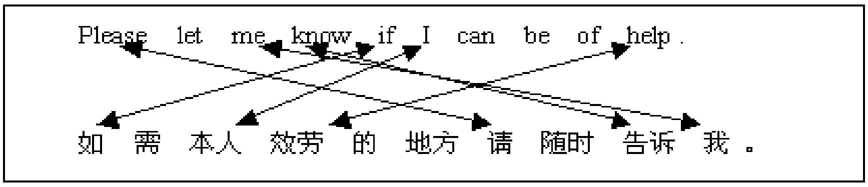

Please let me know if I can be of help .
如 需 本人 效劳 的 地方 请 随时 告诉 我 。

Figure 3. Example of Word Alignment.

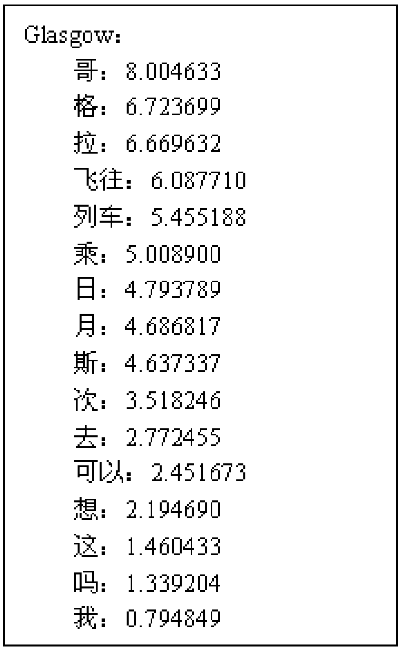
2.3 Calculate the Mutual Information and T-Score

Having calculated the word pair’s co-occurrence frequency and the frequency of every word, we use formulas (1) and (2) to calculate the mutual information $MI(S,T)$ and t-score $t(S,T)$ of any source word and its single target word. As for the association verifying score [Fung 1995], the higher the t-score, the higher the degree of association between S and T:

$$MI(S,T) = \log \frac{\Pr(S,T)}{\Pr(S)\Pr(T)}, \tag{1}$$

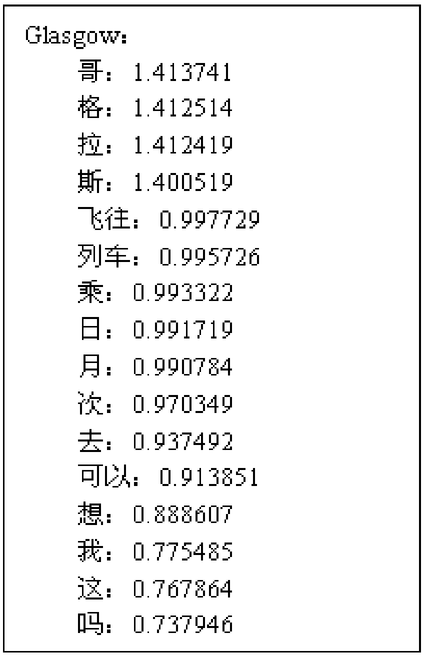
$$t(S,T) \approx \frac{\Pr(S,T) - \Pr(S)\Pr(T)}{\sqrt{\frac{1}{N}\Pr(S,T)}} . \qquad (2)$$

Here, N is the total number of sentence pairs in the corpus, S is the source word, T is the target word, and *Pr(.)* is the probability of the source word or target word. For the "*Glasgow*" example, the outcome of Formula (1) is shown in Figure 4, and the outcome of Formula (2) is shown in Figure 5.



```
Glasgow:
    哥: 8.004633
    格: 6.723699
    拉: 6.669632
    飞往: 6.087710
    列车: 5.455188
    乘: 5.008900
    日: 4.793789
    月: 4.686817
    斯: 4.637337
    次: 3.518246
    去: 2.772455
    可以: 2.451673
    想: 2.194690
    这: 1.460433
    吗: 1.339204
    我: 0.794849
```

Figure 4. Mutual Information Score



```
Glasgow:
    哥: 1.413741
    格: 1.412514
    拉: 1.412419
    斯: 1.400519
    飞往: 0.997729
    列车: 0.995726
    乘: 0.993322
    日: 0.991719
    月: 0.990784
    次: 0.970349
    去: 0.937492
    可以: 0.913851
    想: 0.888607
    我: 0.775485
    这: 0.767864
    吗: 0.737946
```

Figure 5. T-Score.

2.4 Calculate the Average Association Score and its Normalized Difference

The Average Association Score (AAS) is the average association score of the source word and every word in the target language N-gram. It can measure the association degree between the source language and target language. The Normalized Difference (ND) is the normalized difference for the association score of the source word and every word in the target language N-gram. It can measure the internal association of the target multiword units. Therefore, we use the AAS and ND to build the association model of the single source word and target multiword units. We compute the average mutual information, normalized mutual information difference, average t-score, and normalized t-score difference of the consecutive Chinese word string N-gram (N: 2-7), which co-occurs with "*Glasgow*." Vintar's research indicated that the length of 95% of English phrases and Slavic phrases is between 2-6 words [Vintar *et al*. 2001], and from our experience, we can conclude that Chinese multiword units of more than 6 words are also very rare. To reduce the complexity of calculation, we only consider multiword units with 6 words or less. Suppose a Chinese word string C (chunk) is expressed by the following symbols:

$$C = W_1 W_2 ... W_i ... W_n \, . \tag{3}$$

Then the formulae of AMI (Average Mutual Information), MID (Mutual Information Difference), AT (Average T-score) and TD (T-score Difference) are as follows:

$$AMI\,(C,T) = \frac{1}{n}\sum_{i=1}^{n} MI\,(W_i, T)\,, \tag{4}$$

$$MID\,(C,T) = \frac{1}{n \times AMI\,(C,T)}\sum_{i=1}^{n} |\, MI\,(W_i, T) - AMI\,(C,T)\,|\,, \tag{5}$$

$$AT\,(C,T) = \frac{1}{n}\sum_{i=1}^{n} t(W_i, T)\,, \tag{6}$$

$$TD(C,T) = \frac{1}{n \times AT\,(C,T)}\sum_{i=1}^{n} |\, t(W_i, T) - AT\,(C,T)\,|\,. \tag{7}$$

Here, t(.) is the t-score, MI(.) is the mutual information, T is the target word. The results obtained using formulae (4)-(7) are shown in Table 1. (There were 108 outputs from each parameter; we chose only 16 that were connected with the correct answer "*Glasgow*" and could be used to explain the algorithm.)

2.5 Local Bests Algorithm

Currently, the algorithms for extracting multiword units are mainly based on setting a global threshold for some association score (mutual information, entropy, mutual expectation etc.), and if only the association score of the checked word string is bigger or smaller than that threshold, then the word string is considered to be a multiword unit. However, the threshold method has many limitations because the threshold will change with the type of language, the size of the corpus, and the difference of the selected association score, and because of the threshold cannot be easily chosen.

The Local Bests algorithm [Silva *et al*. 1999] is a more robust, flexible and finely tuned approach to the extraction of multiword units, which is based on the local context, rather than on the use of global threshold methods. If a word string (n-gram) is a multiword unit, there should be stronger internal association, and the association score will be high. Also, as a local structure, a multiword unit can show the best association in a local context. Thus, when we find the association score of a word string that is high in a local context, we may consider it as a phrase. For example, there is a strong internal association within the Bi-gram <*ice, cream*>, i.e., between the words *ice* and *cream*. On the other hand, one cannot say that there is a strong internal association within the Bi-gram <*the, in*>. Therefore, let us suppose that there is a function S(.) that can measure the internal association of each n-gram.

Let Ω_{n-1} be the set of all the (n-1)-grams contained in the N-gram word string C (Chunk), and let Ω_{n+1} be the set of all the (n+1)-grams containing this N-gram word string C. Suppose the bigger the association score S(.), the better the result. The Local Bests algorithm can be described as follows:

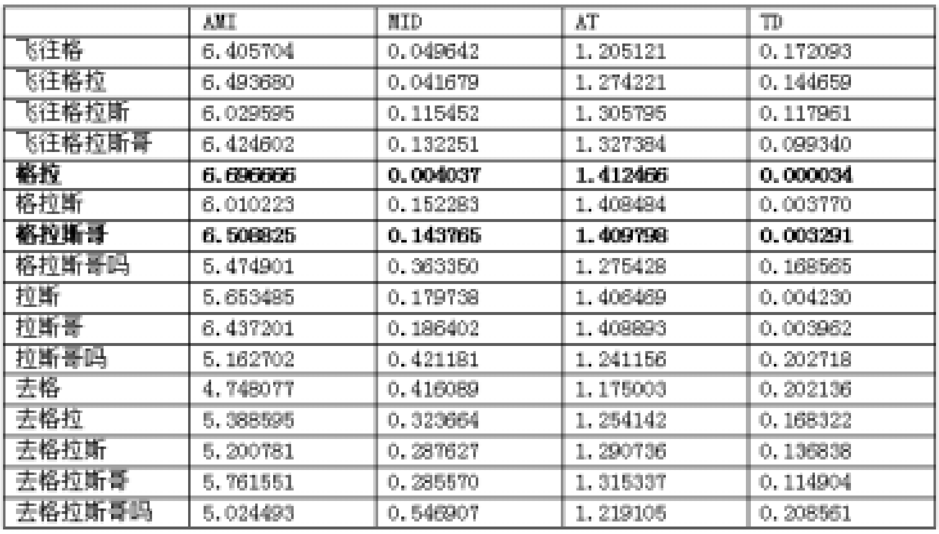
Algorithm 1. Local Bests Algorithm

```
$\forall x \in \Omega_{n-1}$ , $\forall y \in \Omega_{n+1}$ if
(length(C) = 2 and S(C) > S(y)) or
(length(C) > 2 and S(x) $\leq$ S(C) and S(C) > S(y))
then word string C is a multiword unit.
```

Here, S(.) is the internal association score of the Multi-Word Units, and length (C) is the number of words included in C.

In our algorithm, it is better if AMI and AT are bigger, and if MID and TD are smaller; every n-gram of the local best co-occurring with "*Glasgow*" is shown in boldface in Table 1. As we can see in the table, the normalized mutual Information difference of "格拉斯哥" is not a global best score, but it is a local best score, so we may exclude this Multi-Word Unit if we use the global threshold but not the local best algorithm.

Table 1. AMI, MID, AT and TD of Chinese N-gram (N=2~7) co-occurring with "Glasgow."



	AMI	MID	AT	TD
飞往格	6.405704	0.049642	1.205121	0.172093
飞往格拉	6.493680	0.041679	1.274221	0.144659
飞往格拉斯	6.029595	0.115452	1.305795	0.117961
飞往格拉斯哥	6.424602	0.132251	1.327384	0.099340
格拉	**6.696666**	**0.004037**	**1.412466**	**0.000034**
格拉斯	6.010223	0.152283	1.408484	0.003770
格拉斯哥	**6.508825**	**0.143765**	**1.409798**	**0.003291**
格拉斯哥吗	5.474901	0.363350	1.275428	0.168565
拉斯	5.653485	0.179738	1.406469	0.004230
拉斯哥	6.437201	0.186402	1.408893	0.003962
拉斯哥吗	5.162702	0.421181	1.241156	0.202718
去格	4.748077	0.416089	1.175003	0.202136
去格拉	5.388595	0.323664	1.254142	0.168322
去格拉斯	5.200781	0.287627	1.290736	0.136838
去格拉斯哥	5.761551	0.285570	1.315337	0.114904
去格拉斯哥吗	5.024493	0.546907	1.219105	0.208561

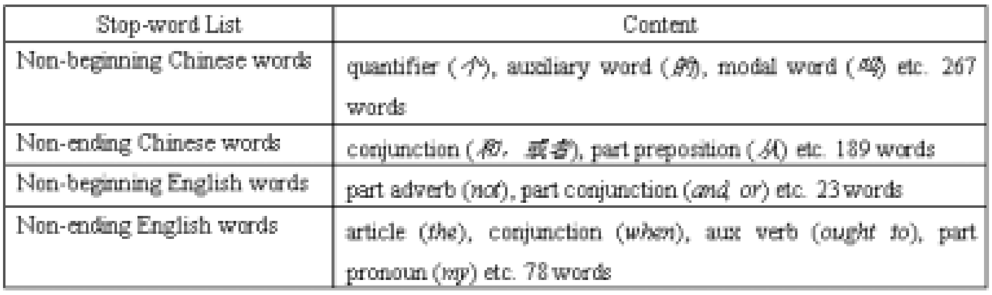
There are still two main problems with using the Local Bests algorithm to extract multiword units: (1) A fraction of the extracted multiword units are not correct, such as “的传球” and “没法把,” with improper words at the beginning or the end of a multiword unit; the same is true with English multiword units, such as “*and, or*” appearing at the beginning of a multiword unit, and “*the, may, if*” at the end of a multiword unit. (2) For a source word, several multiword units are extracted, but not all of them are correct translations.

We utilize a stop-word list to solve the first problem, and the methods based on the association score best and longer unit preference are used to solve the second.

2.6 Stop-word List Filtration

A stop-word is a word that cannot be used at the beginning or the end of a multiword unit. By analyzing the parts of speech and the characteristics of specific words arrangements, we manually create four types of stop-word lists: non-beginning and non-ending Chinese words, and non-beginning and non-ending English words. Samples of lists are shown in Table 2.

Table 2. Stopword List.



Stop-word List	Content
Non-beginning Chinese words	quantifier (个), auxiliary word (的), modal word (吗) etc. 267 words
Non-ending Chinese words	conjunction (和、或者), part preposition (从) etc. 189 words
Non-beginning English words	part adverb (*not*), part conjunction (*and, or*) etc. 23 words
Non-ending English words	article (*the*), conjunction (*when*), aux verb (*ought to*), part pronoun (*my*) etc. 78 words

Using the stop-word lists to filter multiword units, we can the first problem mentioned above.

2.7 Association Score Best Filtration

The association score (mutual information and t-score) is a measure used to judge whether the source word and the target multiword unit are translations of each other, so if a source word corresponds to several target multiword units, then the target multiword unit with a higher association score is more likely to be a translation of this source word. Then we can choose from among the remaining multiword units after two filtrations and take N items with the maximal average mutual information and average t-score as the candidate target translations. According to the results of sample tests, after local bests filtration, the association score of the correct target translation is usually among the best three scores, so we assume that N equals 3.

2.8 Longer Units Preference

A short unit is more likely to be a word [Tanapong *et al*. 2000], but for the following reasons, we apply the Longer Units Preference: (1) Our algorithm determines that the multiword units of two words, especially the two words of the maximal association score with the source word, have the higher average association score and the lower association score difference. For example we can see that "格拉" is better than "格拉斯哥" based on four parameters. (2) We extract multiword units but not words, and if a longer word string has the local best result, then this word string is a comparatively steady structure. Therefore, if a longer words string C_1 entirely contains another shorter word string C_2, then string C_1 is taken as the translation of the source word. This method might choose Multi-Word Units that are longer than necessary, a situation we call "translation units expansion," but it is useful for the extraction of bilingual Multi-Word Units, and it is can be used in the phase of bilingual Multi-Word Unit extraction.

2.9 Lexicon Classification

Thus, the work of extracting a multiword unit translation of every source word is basically accomplished. There are four parameters used in the algorithm. The Average Association Score can measure the association degree between the source language and target language. The Normalized Difference can measure the internal association of the target multiword units. If a pair of bilingual word strings can match more parameters after Local Best and N-bests association score filtering, then it must have higher probability of being correct. Based on the four parameters, four bilingual lexicons are constructed, and they can be subjected to the merge application or intersection application according to different application requirements. We calculate four outcome tables using Formulae (4), (5), (6) and (7), each of them based on a certain measure. Then we pick translation word pairs from those four tables to form five lexicons. The 1st level lexicon composed of word pairs which has appeared only once in the tables; the 2nd level lexicon composed of word pairs which has appeared twice in the tables; and the same rule applies to the 3rd and 4th level lexicons. The higher level one word pair belongs to, the more precision it has. The 0th lexicon is a union of the other four lexicons; that is, any word pairs that have appeared in the tables go into the 0th lexicon. If a source word has several target entries, we calculate the co-occurrence frequency of every entry with the source word in the corpus and then normalize the probability of every entry.

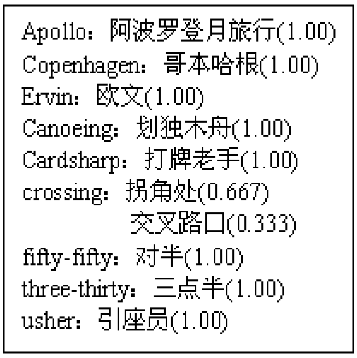
3. Results and Analysis

3.1 Bilingual Corpus

The bilingual corpus we used was DECC1.0, which consists mostly of daily life dialogues, including 14,974 aligned bilingual sentence pairs and a total of 1,039,183 bytes. In this corpus, there are 7,491 English word types and 7,344 Chinese word types.

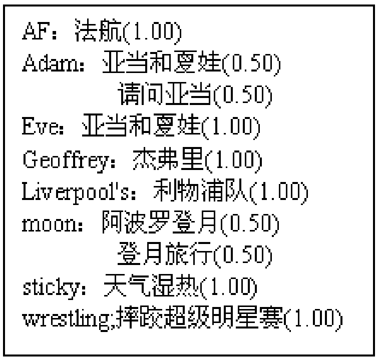
3.2 Lexicon Evaluation

Taking English as the source language and Chinese as the target language, we provide an example of the 4th level lexicon and the 0th level lexicon in Figures 6 and Figure 7.



Apollo：阿波罗登月旅行(1.00)
Copenhagen：哥本哈根(1.00)
Ervin：欧文(1.00)
Canoeing：划独木舟(1.00)
Cardsharp：打牌老手(1.00)
crossing：拐角处(0.667)
交叉路口(0.333)
fifty-fifty：对半(1.00)
three-thirty：三点半(1.00)
usher：引座员(1.00)

Figure 6. 4th level lexicon.



AF：法航(1.00)
Adam：亚当和夏娃(0.50)
请问亚当(0.50)
Eve：亚当和夏娃(1.00)
Geoffrey：杰弗里(1.00)
Liverpool's：利物浦队(1.00)
moon：阿波罗登月(0.50)
登月旅行(0.50)
sticky：天气湿热(1.00)
wrestling;摔跤超级明星赛(1.00)

Figure 7. 0th level lexicon.

There is no uniform method for calculating the precision of translation lexicons, so we take the following approach: the corpus is the measure – if and only if the lexicon entry has an exact match in the corpus, it is taken as correct. For example, the meaning of "*fifty-fifty*" in the English-Chinese dictionary is "平分为二的，对半地，平分为二分地," and in the corpus the corresponding translation of "*fifty-fifty*" is "对半," so we consider that the translation "*fifty-fifty*: 对半" in Figure 6 is correct, but in Figure 7, "*Adam*: 亚当和夏娃" is considered to be incorrect because in the corpus, the pair is "*Adam*: 亚当." The recall rate is the number of English words in each lexicon divided by the number of all the English words in the whole corpus.

The F-measure is an important parameter for balancing precision and recall [Langlais *et al*. 1998]. Table 3 shows the precision, recall and F-measure results of the English-Chinese, Chinese-English 0~4 level lexicons. For lexicons that had more than 200 entries, we randomly chose 200 entries from each of them; for those that had less than 200 entries, we used all the entries for calculation:

$$F = 2\frac{recall \times precision}{recall + precision}. \qquad (8)$$

Table 3. Precision and recall results of all levels of lexicons.

i^{th} level lexicons	precision（%）	Recall（%）	F-measure
0th E-C	41.394	98.63	0.583
1st E-C	23.535	84.22	0.368
2nd E-C	52.388	31.56	0.394
3rd E-C	78.323	5.18	0.097
4th E-C	94.900	1.36	0.027
0th C-E	38.266	96.94	0.549
1st C-E	18.943	82.58	0.308
2nd C-E	47.564	29.92	0.367
3rd C-E	75.092	7.54	0.137
4th C-E	88.293	2.83	0.055

"E-C" lexicons take the single-English-word as the source language and the multi-Chinese-word unit as the target language, and vice versa.

3.3 Analysis of the Result

By analyzing the precision and recall results, and the lemmas of all levels of lexicons, we reached the following conclusions:

(1) There are many lemmas satisfying one qualification (viz. the 1st level lexicon). Almost every English word and Chinese word and expression has at least one target word string satisfying the local best and other qualifications, but the precision of the 1st level lexicon is very low. This shows that (1) depending on a single qualification is not sufficient to construct a bilingual lexicon with high precision, and that (2) not every source word has a corresponding target phrase.

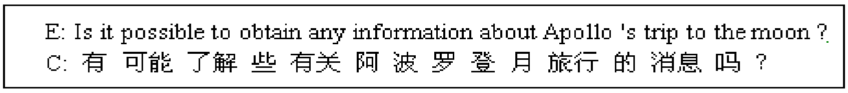
(2) Compared with the 1st level lexicon, the precision of the 2nd level lexicon is greatly increased. According to the sketchy statistics, the two qualifications satisfied by most of the correct portion of the 2nd level lexicon are mutual information and t-score, which shows that for a certain parameter (mutual information or t-score), simultaneously using the difference and average value can improve the results greatly.

(3) Compared with the 2nd level lexicon, the precision of the 3rd level lexicon is also greatly increased and recall is decreased, which shows that after one parameter has been satisfied, if a qualification of another parameter can be also satisfied, then the translation is very likely to be correct. In similar works, many other researchers needed to consider multiple parameters, and the selection of parameters was very important. From early works on word alignment and our current work on phrase extraction, we find that a combination of mutual information and t-score provides a reliable measure.

(4) Only a little manual collation work is needed to make the 4th level lexicon practical. The English-Chinese 4th level lexicon has only 98 lemmas, which, except for some common phrases with high appearance frequency, are mainly personal names, place names and specialized terms; and all of these terms have low appearance frequency, many occurring only once. This shows that for the extraction of low frequency phrases, our algorithm also is good.

(5) The higher the lexicon's level, the lower its recall rate. This shows that the cases of single source words corresponding to a target word string are comparatively few. On the other hand, it shows that our corpus is too small. If the corpus could be increased, the result would be better.

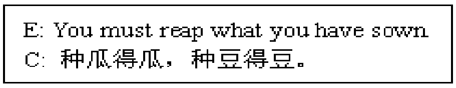
(6) There are cases of "translation unit expansion" in all levels of lexicons; for example, in the 4th level lexicon for "*Apollo:*阿波罗登月旅行," "*Apollo*" corresponds to "阿波罗," but there is only one sentence pair in which "*Apollo*" appears in the whole corpus (Figure 8). In addition, "阿波罗登月旅行" exists as a sense unit, so according to the longer units preference method, our algorithm selected "阿波罗登月旅行." It should be made clear that, although "*Apollo*: 阿波罗登月旅行" is an incorrect lemma, it provides a basis for constructing a translation lexicon in which the source language and the target language are both multi-word phrases. Especially in the 0th level lexicon, we can see that the two translations of "*moon*" are "阿波罗登月" and "登月旅行," from which, using a certain algorithm, we can extract the correct phrase "*Apollo's trip to the moon:* 阿波罗登月旅行," and this will be the focus of our future research.



E: Is it possible to obtain any information about Apollo 's trip to the moon ?
C: 有 可能 了解 些 有关 阿 波 罗 登 月 旅行 的 消息 吗 ？

Figure 8 Sentence pair in a corpus with "Apollo."

(7) Another fact that affects the precision is that the corpus we used contains 171 bilingual proverbs, and such sentence pairs can rarely be translated word for word, as demonstrated by the example shown in Figure 9.



E: You must reap what you have sown.
C: 种瓜得瓜，种豆得豆。

Figure 9. Bilingual proverb.

4. Conclusion and Future Research

4.1 Conclusion

Because there are many cases of single source words corresponding to target multi-word units, for example, English personal names and place names, we have provided an algorithm for the automatic alignment of single source words and target multi-word units from a sentence-aligned parallel spoken language corpus, which makes a translation lexicon more practical. It will be of great help for machine translation, especially Chinese-English translation. On the other hand, the outputs can also be used to extract bilingual multi-word units. Compared with other similar researches, this algorithm differs in the following ways:

(1) It utilizes the normalized association score difference as the criterion for extracting phrases.

(2) It simultaneously uses the Local Bests algorithm, stop-word filtration, and the longer units preference method to extract phrases.

(3) Classify lexicon. Different levels of lexicons can be applied to obtain practical translation lexicons or can be used as the basis for further research.

Mutual information has been used in many other similar researches, but these processes are mainly based on algorithms of iterating the Bi-gram calculation, and the retrieval results mostly depend on the identification of suitable Bi-grams for the initiation of the iterative process. Errors can accumulate during the iteration process, thus greatly affecting the precision of multi-word phrase extraction [Dias *et al.* 2000]. Our algorithm solves this problem by calculating the normalized association score difference of the target words corresponding to the same source word. The use of t-score increases the precision of the phrase translation lexicon, and the classification of the lexicon reduces the number of the incorrect entries in the high level lexicon effectively, which makes the translation lexicon more practical.

4.2 Future Research Plan

Currently, “translation unit expansion” is a common problem, and we shall utilize the outcome to extract bilingual multi-word units in our future research.

Reference:

Chen, X.H. “Automatic Analysis of Contemporary Chinese Using Visual C++,” Beijing: Beijing Language and Culture University Press, (It's published in Chinese) 1999, pp.97-103.

Church, K.W. and P. Hanks. "Word Association Norms, Mutual Information & Lexicography." *Computational Linguistics,* 16(1) 1990, pp.22-29.

Dias, G., Guilloré,S. and Pereira L.J.G. "Normalization of Association Measures for Multiword Lexical Unit Extraction." *International Conference on Artificial and Computational Intelligence for Decision Control and Automation in Engineering and Industrial Applications (ACIDCA'2000).* Monastir, Tunisia, 2000, pp. 207-216.

Fung P. "A Pattern Matching Method for Finding Noun and Proper Noun Translations from Noisy Parallel Corpora." *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, Boston, USA. 1995, pp. 236-243.

Haruno M., Ikehara S. and Yamazaki T. "Learning Bilingual Collocations by Word-level Sorting. *COLING96*. 1996, pp. 525~530.

Hiemstra, D. "Using Statistical Methods to Create a Bilingual Dictionary." *Master's Thesis*, University of Twente. 1996.

Kita, K., Kato, Y., Omoto T. and Yano Y. "A Comparative Study of Automatic Extraction of Collocation from Corpora: Mutual Information vs. Cost Criteria." *Journal of Natural Language Processing*, 1 (1), 1994, pp. 21-33.

Langlais P., Simard M. and Véronis J. "Methods and Practical Issues in Evaluating Alignment Techniques." *Proceedings of COLING-ACL*, 1998, Montréal, Canada, pp. 711-717.

Melamed I. D. "Automatic Construction of Clean Broad-Coverage Translation Lexicons." *Conference of the Association for Machine Translation in Americas,* Montreal, Canada. 1996, pp. 125-134.

Melamed I. D. "Automatic Discovery of Non-Compositional Compounds in Parallel Data." *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing*. Providence, RI. USA. 1997, pp. 97-108.

Nagao, M. and Mori, S. "A New Method of N-gram Statistics for Large Number of n and Automatic Extraction of Words and Phrases from Large Text Data of Japanese." *Proceedings of the 15th International Conference on Computational Linguistics*. 1994. pp.611-615.

Sayori Shimohata, Toshiyuki Sugio and Junji Nagata "Retrieving Collocations by Co-occurrences and Word Order Constraints." *35th Conference of the Association for Computational Linguistics (ACL'97)*, Madrid, 1997, pp. 476-481.

Silva J.F., Dias G., Guillor S. and Lopes J.G.P. "Using Localmaxs Algorithm for Extraction of Contiguous and Non-contiguous Multiword Lexical Units." *9th Portuguese Conference in Artificial Intelligence, Lecture Notes, Springer-Verlag*, Universidade de Evora, Evora, Portugal, 1999, pp. 113-132.

Smadja, F. "Retrieving Collocations from Text: Xtract." *Computational Linguistics*, 1993. Vol.19, No.1. pp. 143-177.

Smadja F., McKeown K.R. and Hatzivassiloglou V. "Translation Collocations for Bilingual Lexicons: a Statistical Approach." *Computational Linguistics* 1996, 22(1), pp. 1~38.

Tanapong Potipiti, Virach Sornlertlamvanich and Thatsanee Charoenporn. "Towards Building a Corpus-based Dictionary for Non-word-boundary Language." *Workshop on Terminology Resources and Computation, Workshop Proceedings of the Second International Conference on Language Resources and Evaluation (LREC2000)*, Athens, Greece. 2000, pp. 82-86.

Vintar, Spela. "Using Parallel Corpora for Translation-Oriented Term Extraction." *Babel Journal*, John Benjamins Publishing. 2001.

Wu, D. and Xia, X. "Large-Scale Automatic Extraction of an English-Chinese Translation Lexicon." *Machine Translation* (4). 1995, pp. 285-313.

Yamamoto, M. and Church, K.W. "Using suffix arrays to compute term frequency and document frequency for all substrings in a corpus." *Proceedings of the 6th Workshop on Very Large Corpora,* Montreal, Canada, 1998, pp.28-37.

Yu, S.W. The Grammatical Knowledge-base of Contemporary Chinese, Beijing: Tsinghua University Press, (It's published in Chinese) 1998.

Zhou, J. and Dapkus, P. "Automatic Suggestion of Significant Terms for a Predefined Topic." *Proceedings of the Third Workshop on Very Large Corpora*. Cambridge1995, pp.131-147.

Computational Linguistics and Chinese Language Processing
Vol. 8 , No. 2 , August 2003, pp. 93-108

從詞網出發的中文複合名詞的語意表達

柯淑津*

摘要

WordNet 提供豐富的詞彙語意資訊，因此對於自然語言處理相關研究有很大的幫助。但是由於 Princeton WordNet 的語意資訊僅以英文的形式呈現，為了能讓 WordNet 所蘊含的豐富資源也能應用到中文相關處理，我們試圖利用雙語字典等多項已存在的資源做為橋樑，希望能將英文 WordNet 的豐富資源自動引介到中文。但是，在我們觀察這些連結英文 WordNet 與雙語字典所產生的初步結果後，發現由於語言之間的藩籬以及雙語字典的目標語詞彙大都偏向於解釋等多種原因，使得英文同義詞集(Synset)所對應到的中文翻譯，常是一些不具結構性的中文複合詞、片語、甚至是一長串的句子，而不是獨立的中文詞彙。這樣的現象與中文詞網應以詞彙為基本元件的要求相違背。因此，本研究將針對這種現象作進一步的處理。

本文的主要目標有下列兩項：首先，自中文複合詞找出最能代表其意義的中心詞彙，及若干個特徵詞彙。其次，將這些詞彙進一步以語意概念形式表達出來。第一個部分，我們透過語法結構分析來完成。至於，第二個部分，詞彙的語意我們透過知網的概念特徵來表示。當然，在中文詞彙轉為詞義概念的部分，是存在著歧義現象的。辨識語意歧義的方法，我們除了用到詞彙的詞性之外，還透過 WordNet 的上位關係來降低歧義度。我們以名詞部分進行實驗，實驗結果顯示在語意標示方面，可達到 93.5%的應用率以及 93.8%的正確率。

Abstract

WordNet provides plenty of lexical meaning; therefore, it is very helpful in natural language processing research. Each lexical meaning in Princeton WordNet is presented in English. In this work, we attempt to use a bilingual dictionary as the backbone to automatically map English WordNet to a Chinese form. However, we encounter many barriers between the two different languages when we observe the preliminary result for the linkage between English WordNet and the bilingual dictionary. This mapping causes the Chinese translation of the English synonym

* 東吳大學資訊科學系 E-mail: ksj@cis.scu.edu.tw

collection (Synset) to correspond to unstructured Chinese compound words, phrases, and even long string sentence instead of independent Chinese lexical words. This phenomenon violates the aim of Chinese WordNet to take the lexical word as the basic component. Therefore, this research will perform further processing to study this phenomenon.

The objectives of this paper are as follows: First, we will discover core lexical words and characteristic words from Chinese compound words. Next, those lexical words will be expressed by means of conceptual representations. For the core lexical words, we use grammar structure analysis to locate such words. For characteristic words, we use sememes in HowNet to represent their lexical meanings. Certainly, there exists a problem of ambiguity when Chinese lexical words are translated into their lexical meanings. To resolve this problem, we use lexical parts-of-speech and hypernyms of WordNet to reduce the lexical ambiguity. We experimented on nouns, and the experimental results show that sense disambiguation could achieve a 93.8% applicability rate and a 93.5% correct rate.

1. 簡介

自然語言處理的研究與應用，隨著語言資源的快速增加，有愈來愈蓬勃的現象。這些語言資源包括：辭典、詞彙資料庫、語料庫等等，而其中相當引人注目的就是分類辭典，例如：GUM, CYC, ONTOS, MICROKOSMOS, EDR 和 WordNet [Gomez, 1998]。這些分類字典中，各自有不同的特徵，有些是專為某個特殊範疇設計，有些則是不限文體；它們的排列方式也各自有所不同，可能是根據詞彙關係(Lexical Relation)，也可能根據概念關係(Conceptual Relation)來排列。在這些分類詞典中，WordNet [Miller, 1990; Fellbaum, 1998]擁有最寬廣的應用空間，已然形成一種標準[Farreres, Rigau and Rodriguez, 1998]。因此，自 WordNet 推出之後，便被廣泛地應用在許多的相關研究中，像是文件檢索[Gonzalo *et al*., 1998; Mandala, Tokunaga and Tanaka, 1998]，機器翻譯[Knight and Luk, 1994]，文件生成[Jing, 1998]，影像檢索[Aslandogam *et al*., 1997]等等。WordNet 的成功，引發許多非英語系的國家，建置自己語言版本 WordNet 的構想，並且有不少計畫已開始實際進行。例如包含多種歐洲語言的 EuroWordNet 已經完成[Atserias *et al*., 1997; Farreres, Rigau, and Rodriguez, 1998]。另外，韓語版本以及日語版本的 WordNet 建構計畫也都正積極進行中[Lee, Lee and Yun, 2000]。

WordNet 提供了豐富的語意相關資訊，因此對於自然語言處理相關研究有很大的幫助。但是由於 Princeton WordNet 的語意資訊僅以英文形式呈現，為了能讓 WordNet 所蘊含的豐富資源也能應用到中文相關處理，我們試圖利用雙語字典等多項已存在的資源做為橋樑，希望能將英文 WordNet 的豐富資源自動引介到中文。但是，在我們觀察這些連結英文 WordNet 與雙語字典所產生的結果後，發現由於語言間的藩籬以及雙語字典的目

標語詞大都偏向於解釋等多項原因，使得英文同義詞集(Synset)所對應到的中文翻譯，常是一些不具結構性的中文複合詞、片語、甚至是一長串的句子，而不是獨立的中文詞彙。因此，本研究對這種現象作進一步的處理，透過提供詞義的字典，將這些中文翻譯轉化成語意概念，使得連結資料可應用於語言相關處理。

本文的目標在將詞網的中文翻譯轉化成語意概念表示，主要的工作分為下列兩項：第一項工作為自中文複合詞中找出語意中心詞彙，而第二項工作則是將中文詞彙轉化成詞義概念表達。當詞網的中文翻譯本身就是詞彙時，我們只需進行第二項工作，透過知網[董振東、董強, 2002]的概念定義，將詞彙表達成適當的詞義概念。若中文翻譯不是詞彙，而是複合詞句時，我們需先作第一項處理，找出複合詞句的語意中心詞彙，再進行第二項工作，將中心詞彙進一步以語意概念形式表達出來。當然，在中文詞彙轉為詞義概念的部分，是存在著歧義現象的。辨識語意歧義的方法，我們除了用到詞彙的詞性之外，還透過 WordNet 的上位關係來降低歧義度。

本文第二節介紹相關研究，第三節對資料進行一些觀察，第四節與第五節分別提出標示中心詞彙與概念特徵的方法。實驗設計及結果在第六節，最後是結論以及未來研究方向。

2. 相關研究

自然語言處理研究，需要豐富的詞彙知識與語意關係作為基礎。這些重要的研究資源除了透過統計技巧，由語料庫中獲得以外[Gale, Church and Yarowsky, 1992; Yarowsky, 1992, 1995; Resnik, 1993; Dagan and Itai, 1994; Luk, 1995; Ng and Lee, 1996; Riloff and Jones, 1999]，還可粹取自機讀字典[Guthrie *et al*., 1991; Slator, 1991; Li, Szpakowicz and Matwin, 1995; Chen and Chang, 1998, Yang and Ker, 2002]。

近來，有許多學者著力於建構含語意訊息的中英雙語資源，他們認為這些資源對於機器翻譯以及多語資訊檢索系統都有很大的幫助[Chang, Ker and Chen, 1998; Chen and Chang, 1998; Chen and Lin, 2000; Chen, Lin and Lin, 2000; Dorr *et al*., 2000; Carpuat *et al*., 2002; Wang, 2002]。其中，他們所使用的資源各自不一，有的是連結 WordNet 與同義詞詞林[Chen and Lin, 2000]，有的將 WordNet 與 HowNet 進行對應[Dorr *et al*., 2000; Carpuat *et al*., 2002]，也有的研究群將一般機讀字典與分類字典進行連結，設法由詞彙得到分類資訊[Chang, Ker and Chen, 1998; Chen and Chang, 1998]。

3. 觀察

透過觀察，我們發現不同的字典資源，雖收錄的詞彙、語意訊息各自有所差異，但當它們在表達同一詞彙所具有的相同語意時，常會存在著某些共通現象。這些共通現象包含有：中文複合詞之中心詞彙與上位詞之翻譯共用詞素，以及共用定義詞彙等等。以下，我們先介紹經連結後的資料，並觀察它們所包含的訊息，以及探討這些訊息如何使用於辨識中心詞彙或是進行歧義詞的詞義辨識工作。

3.1 連結上中文翻譯的WordNet資料

為了能讓 WordNet 所蘊含的豐富資源也能應用到中文的相關處理，在先前的研究我們利用雙語字典等多項已存在的電子資源做為橋樑，將 WordNet 的同義詞集連結上適當的中文翻譯。表 1 是部分的連結例子，其中第一個欄位是構成這個同義詞集的英文詞彙。第二個欄位是它們在 WordNet 的定義，而第三個欄位就是經過連結雙語字典後所得的中文翻譯。這些翻譯中像是「光線」、「大樓」以及「秋天」等都屬於中文詞彙，但是也有一些中文複合詞，例如：「一壘安打」、「扁桃腺切除術」以及「西洋棋騎士」等。這樣的現象與中文詞網應以詞彙為基本元件的要求相違背。因此，本研究將對這種現象作進一步的處理。

表 1 連結 WordNet 同義詞集與中文翻譯之例子。

WordNet 同義詞集	WordNet 定義	中文翻譯
building, edifice	a structure that has a roof and walls and stands more or less permanently in one place	大廈, 大樓, 建築物
beam, beam of light, light beam, ray, ray of light, shaft, shaft of light	a column of light (as from a beacon)	光束, 光線
clay, mud	water soaked soil; soft wet earth	泥, 泥巴, 泥漿
autumn, fall	the season when the leaves fall from the trees	秋, 秋天, 秋季
single	a base hit on which the batter stops safely at first base	一壘安打
tonsillectomy	surgical removal of the palatine tonsils; commonly performed along with adenoidectomy	扁桃腺切除術，扁桃體切除術
knight, horse	a chessman in the shape of a horses head; can move two squares horizontally and one vertically (or vice versa)	西洋棋騎士

3.2 中心詞彙與上位翻譯詞彙共用詞素

在 WordNet 中上位詞代表一種泛稱。在一個同義詞集的定義與其上位詞集的定義常會有共用詞素的情形。這種原本出現於英文定義上的現象似乎也延續至它們的中文翻譯。這種情形對於中文複合詞翻譯尤其明顯。而且，如果我們將這些中文複合詞進行斷詞處理後，更可發現與其上位詞的中文翻譯擁有最多共用詞素的詞彙往往就是該中文複合詞的中心詞彙。例如，表 2 的中文複合詞「扁桃腺切除術」經斷詞處理後得到「扁桃腺」與「切除術」兩個詞彙。其中，「切除術」與該同義詞集的上位詞之中文翻譯「切除」擁

有較多的共用詞素。同時，「扁桃腺切除術」所談的主要是「切除術」，因此「扁桃腺切除術」的中心詞彙是「切除術」。同樣的情形也發生在表 2 的其他例子，如：「曲線球」、「一壘安打」、「西洋棋騎士」等中文複合詞，它們的中心詞彙分別是「球」、「安打」以及「西洋棋」。這些中心詞彙與其所屬的上位詞翻譯皆含有較多的共用詞素。

3.3 詞彙之概念特徵辨識

詞彙往往含有多個詞義，詞義的辨識是自然語言處理的核心工作。在此小節中，我們觀察連結上 WordNet 同義詞集的中文翻譯所含有的訊息，以便瞭解如何透過訊息處理，將這些中文翻譯轉化成知網的概念特徵。

表 2 同義詞集之中文複合詞翻譯及其中心詞彙。

同義詞集	同義詞集定義	中文翻譯	上位詞詞集	上位詞中文翻譯
testate, testator	a person who makes a will	遺囑 人	mortal, individual, person, somebody, soul, someone, human	人
screwball	a pitch with reverse spin that curves toward the side of the plate from which it was thrown	曲線 球, 內曲線 球	pitch, delivery	投球
single	a base hit on which the batter stops safely at first base	一壘 安打	bingle, safety, base hit	安打
tonsillectomy	surgical removal of the palatine tonsils; commonly performed along with adenoidectomy	扁桃腺 切除術	ablation, cutting out, excision, extirpation	切除
plumbing, plumbery	the occupation of a plumber (installing and repairing pipes and fixtures for water or gas or sewage in a building)	鉛管 業	craft, trade	職業
knight, horse	a chessman in the shape of a horses head; can move two squares horizontally and one vertically (or vice versa)	西洋棋 騎士	chessman, chess piece	棋子

註：標示底線者為中文翻譯之語意中心詞彙

3.3.1 單義詞部分

有些詞彙本身的語意很確定，不具有歧義性。這些詞彙在不同的資源中，雖可能會存在

不同的解釋用語。但是，它們大多是指同一事物，因此，我們直接讓它們對應。如：表1的同義詞集{autumn, fall}所對應的中文翻譯「秋, 秋天, 秋季」，經查詢知網的概念定義後，所得到的都是「time|時間, autumn|秋」(如表3所示)。另外，同義詞集{clay, mud}的中文翻譯「泥, 泥巴, 泥漿」對應知網後所得的也都是一致的概念定義「stone|土石」。像表3列出的這些單義詞彙，我們可以直接將它們標上唯一的概念特徵。

表3 部分中文單義詞及其在知網的概念定義。

中文詞彙	知網的概念定義
秋	time|時間,autumn|秋
秋天	time|時間,autumn|秋
秋季	time|時間,autumn|秋
秋海棠	FlowerGrass|花草
光束	lights|光
光線	lights|光
泥	stone|土石
泥巴	stone|土石
泥漿	stone|土石

3.3.2 歧義詞部分

對於歧義的中文詞彙，在知網中會存在一個以上的概念定義，例如「分號」與「目標」這兩個詞彙在知網中各自有兩個概念定義。如表4所示，「分號」可能是標點符號「；」，也可能代表商場上的分支機構。至於，表5是這兩個中文歧義詞當為WordNet同義詞集的中文翻譯之例子。其中，同義詞集{semicolon}因為它的英文詞彙與知網中具{symbol|符號}概念之詞條的英文詞彙相同。因此，我們可以將此中文翻譯「分號」標上{symbol|符號}概念。但是，這樣的作法對於歧義詞彙「目標」卻是行不通的。比對表4及表5的內容之後，我們可發現WordNet同義詞集{aim, object, objective, target}與「目標」在知網中的兩項定義皆有共同的英文詞彙。其中，與概念定義{purpose|目的}擁有相同的英文詞彙'aim'以及'objective'，而與概念定義{tool|用具,#weapon|武器, $AimAt|定向, $firing|射擊}也同時擁有相同的英文詞彙'target'。另外，從表6所呈現的上位詞資訊中，我們也可以發現同義詞集{aim, object, objective, target}的上位詞中文翻譯「目的」與{purpose|目的}概念中的概念名稱完全相同。綜合上述資訊，我們可以判定同義詞集同義詞集{aim, object, objective, target}的正確知網概念為{purpose|目的}。

表 4　「分號」與「目標」在知網的概念定義與其英文詞彙。

中文詞彙	知網的概念定義	英文詞彙
分號	InstitutePlace|場所, branch|支, commercial|商	branch
分號	symbol|符號	semicolon
目標	purpose|目的	aim, goal, objective
目標	tool|用具,#weapon|武器, $AimAt|定向, $firing|射擊	target

表 5　幾個以「分號」與「目標」為中文翻譯的同義詞集及其定義。

WordNet 同義詞集	WordNet 定義	中文翻譯
semicolon	a punctuation mark (;) used to connect independent clauses; indicates a closer relation than does a period	分號
butt, target	an object set up for a marksman or archer to aim at	目標
aim, object, objective, target	the goal intended to be attained (and which is believed to be attainable)	目標

表 6　表 5 的同義詞集所含的上位詞及其相關訊息。

WordNet 同義詞集	WordNet 上位詞同義詞集	WordNet 上位詞定義	上位詞 中文翻譯
butt, target	sports equipment	equipment needed to participate in a particular sport	體育 裝備
aim, object, objective, target	end, goal	the state of affairs that a plan is intended to achieve and that (when achieved) terminates behavior intended to achieve it	目的, 終極

4. 判定中文複合名詞的中心詞彙方法

對於中文複合詞，無法以直接連結字典的方式來進行語意標示。我們分兩個階段完成所需要的處理。首先，要作的就是區分它的中心詞彙與概念特徵，也就是要找出它的主要特徵詞彙與次要特徵詞彙。然後，再標示主要特徵詞彙之語意，以主要特徵詞彙之語意代表整個複合詞之主要語意。透過先前的觀察，我們發現同義詞集的中文翻譯複合詞與其上位詞之中文翻譯常擁有相同的詞素。在此，我們先將詞彙逐字拆解成詞素，以所得

的詞素集合來代表該詞彙。再透過 Dice 係數[Dice, 1945]估算兩個詞素集的相似度，當作兩個中文詞彙的相似度。針對給定的兩個中文詞彙 W_1, W_2，它們的相似度計算公式 Sim(W_1, W_2)如公式一所示。

$$Sim(W_1, W_2) = \frac{2 \times |S(W_1) \cap S(W_2)|}{|S(W_1)| + |S(W_2)|} \qquad \text{（公式一）}$$

其中， W_i ：中文詞彙，
$S(W_i)$ ：將中文詞彙 W_i 拆解成詞素，所得的詞素集合，
$|S(W_i)|$ ：詞素集合 $S(W_i)$的長度。

5. 詞彙與詞義概念的對應方法

5.1 對應方法設計

在本節中，我們設法將 WordNet 同義詞集所連結的中文翻譯詞彙，對應到知網的詞義概念。透過觀察，我們知道單義詞的詞義概念直接標示即可。因此，以下僅就岐義詞部分討論它們與詞義概念的對應方法。

對於一詞多義的詞彙，我們需有一套辨識歧義的方法，來為詞彙找出最適當的詞義對應。透過第三節的觀察，我們發現不同的資源在定義同個詞彙的相同語意時，常常會使用共同的中英文詞彙。這種現象有時發生在中文翻譯，有時出現的卻是使用相同的概念語詞。

對於一個 WordNet 的同義詞集 S，以及它已標示的 n 個中文翻譯詞彙 W_1, W_2, …, W_n,，假設 W_i 在知網中有 k 個不同的概念定義 $D_{i1}, D_{i2}, \ldots, D_{ik}$，我們定義集合 E_{ij} 表示詞彙 W_i 在定義 D_{ij} 中對應的英文詞彙所形成的集合。我們令同義詞集 S 與 E_{ij} 交集元素最多的定義 D_{ij} 為 S 的概念定義，即 CDEF(S)，如公式二所示。

$$\text{CDEF}(S) = \arg\max_{i,j} |S \cap E_{i,j}|, \quad \forall\ i = 1, \ldots, n,\ j = 1, \ldots, k_i \qquad \text{(公式二)}$$

其中， n ：同義詞集 S 已標示的中文翻譯詞彙數，
k_i ：同義詞集 S 的第 i 個中文翻譯詞彙 W_i 在知網中的概念定義數，
$E_{i,j}$ ：表示詞彙 W_i 在定義 D_{ij} 中對應的英文詞彙所形成的集合。

5.2 對應例子

下面我們以例子說明將 WordNet 同義詞集透過中文翻譯設定其詞義概念對應的方法。表 7 給出兩個同義詞集例子，其中，同義詞集{campaign, cause, crusade, drive, effort, movement}有兩個中文翻譯詞彙「運動」（稱 W_1）與「活動」（稱 W_2）。由表 8 可看到詞彙「運動」在知網中有三個詞義概念，分別為「fact|事情,function|活動,politics|政」（稱 D_{11}）、「fact|事情,exercise|鍛練,sport|體育」（稱 D_{12}）以及「fact|事情,AlterLocation|變

空間位置」(稱 D_{13})，至於詞彙「活動」在知網中僅有一個詞義概念 D_{21} 為「fact|事情,generic|統稱」。四個詞義概念所對應之英文詞彙分別為{campaign, drive, movement}, {athletics, exercise, sports}, {motion, movement}以及{activity, maneuve}，它們與同義詞集之交集分別為{campaign, drive, movement}、空集合、{movement}以及空集合，因此，同義詞集{campaign, cause, crusade, drive, effort, movement}之詞義概念為 D_{11}「fact|事情,function|活動,politics|政」。同樣的方法，我們可為同義詞集{flank, wing}標上詞義概念「place|地方,edge|邊,military|軍」。

表 7 為同義詞集對應詞義概念方法的例子。

WordNet 同義詞集	WordNet 定義	中文翻譯
{campaign, cause, crusade, drive, effort, movement}	a series of actions advancing a principle or tending toward a particular end	運動, 活動
{flank, wing}	the side of military or naval formation	翼

表 8 表 7 之中文詞彙出現在知網的定義及其英文詞彙。

中文詞彙	知網對應英文詞彙	知網詞義概念
運動	{campaign, drive, movement}	fact|事情,function|活動,politics|政
	{athletics, exercise, sports}	fact|事情,exercise|鍛練,sport|體育
	{motion, movement}	fact|事情,AlterLocation|變空間位置
活動	{activity, maneuve}	fact|事情,generic|統稱
翼	{wing}	part|部件,%artifact|人工物,wing|翅
	{flank, wing}	place|地方,edge|邊,military|軍
	{wing}	part|部件,%bird|禽,wing|翅,*fly|飛

註： 標示底線者，表示與同義詞集共用詞彙。

6. 實驗

6.1 實驗資料

我們將實驗分為由中文複合名詞標示出中心詞彙及中文詞彙語意概念標示兩個部分。在效能部分，本研究採用正確率(Correctness)及應用率(Applicability)進行評估。實驗過程中能標示出語意概念的同義詞集個數比對於參與比對的總同義詞集數，稱為應用率，至於，正確率定義為標示結果的正確比率。

本研究以標示中文翻譯的英文 1.6 版 WordNet 同義詞集進行實驗，實驗資料來自於標示中文翻譯的 53753 個名詞詞網同義詞集，經與知網所含詞彙比對後，其中 13628 個同義詞集的中文翻譯出現在知網中。至於含單義中文翻譯詞彙的同義詞集則有 12231 個，佔總同義詞集數的 22.8%，實驗資料分佈情形如表 9 所示。我們將中文翻譯含知網詞彙的同義詞集直接進行標示語意概念實驗，其餘的同義詞集資料則進行中心詞彙標示工作。

表 9　實驗資料之分佈統計。

標示中文翻譯的同義詞集數	53753
中文翻譯含知網詞彙的同義詞集數	13628
中文翻譯含單義知網詞彙的同義詞集數	12231
中文翻譯全為多義知網詞彙的同義詞集數	1397

6.2 標示語意概念實驗

6.2.1實驗設定與結果

在標示中文詞彙語意概念部分，我們先將中文詞彙依其在知網的定義，分為單義詞與歧義詞兩個部分，再進行不同之處理。若為單義詞彙，則詞彙語意直接連結知網的定義。若為歧義詞彙，則依第 5 節之公式做出適當之語意連結。實驗結果在 1397 個不含單義中文翻譯詞彙的同義詞集中，有 546 個同義詞集可成功產生語意連結。因此，在語意連結之實驗共可產生 12777 筆連結，比對於參與語意標示實驗的 13628 個同義詞集，應用率為 93.8%(如表 10 所示)。為評估連結之正確率，作者自產生之 12777 筆連結資料中隨機選取 630 筆，經人工比對後計有 589 筆為正確之連結，正確率為 93.5%（詳見表 10）。

表 10　詞彙語意標示實驗所得結果。

進行語意連結的同義詞集數	13628
產生語意連結的同義詞集數	12777
應用率	93.8%
人工比對資料筆數	630
正確連結資料筆數	589
正確率	93.5%

6.2.2 實驗結果討論與錯誤分析

由實驗結果，我們發現有些同義詞集無法成功的產生語意連結，其原因主要有下列幾種：

1. 詞網的同義詞集與知網兩項資源未出現共用詞彙，2.多個概念定義與同義詞集擁有相同的交集個數，3. 不同詞義的中、英文詞彙皆擁有共用詞彙。

本文所提方法以共用詞彙作為相同詞義的徵象。但若兩份資源在同一詞義的詞彙表達上，作了不同的詞彙選擇，我們目前無法做出對應。這種情形共有 663 個同義詞集，比對於所有無法產生連結的 851 個同義詞集共佔 77.9%。

當多個概念定義與同義詞集擁有相同的交集個數時，我們現行的方法無法判斷出最合適的對應概念。例如表 11 一所呈現的同義詞集{pumpkin}及{Japanese plum, loquat}，即屬於這種情形。同義詞集{pumpkin}與其兩個語意概念「vegetable|蔬菜」及「part|部件,%vegetable|蔬菜,embryo|胚,$eat|吃」都同時擁有共同詞彙{pumpkin}，而且，這兩個概念應該都屬於可接受的連結。另外，詞彙「枇杷」應屬於單義詞彙，但是在知網上出現兩個定義，並且此兩份概念定義所對應的英文詞彙完全相同。若是將同義詞集{Japanese plum, loquat}對應至「fruit|水果」或「tree|樹」，應該都算是正確的對應。

通常不同詞義的原始語在翻譯至目標語時會有不同的詞彙選擇，例如，「river bank」與「money bank」在翻譯至中文時會分別使用不同的詞彙「河岸」及「銀行」。但是，表 11 的「鳶」在「tool|用具,*WhileAway|消閑」及「bird|禽」兩個很不同的語意概念中，它們的英文詞彙卻都是「kite」，這種情形，只以對應詞彙本身作為辨識歧義的依據，顯然是不夠的。

表 11 *無法成功產生語意連結的同義詞集例子。*

同義詞集	WordNet 定義	中文翻譯	知網對應英文詞彙	知網詞義概念
{pumpkin}	usually large pulpy deep-yellow round fruit of the squash family maturing in late summer or early autumn	南瓜	{pumpkin, cushaw}	vegetable|蔬菜
			{pumpkin, cushaw}	part|部件, %vegetable|蔬菜, embryo|胚,$eat|吃
{Japanese plum, loquat}	yellow olive-sized semitropical fruit with a large free stone and relatively little flesh; used for jellies	枇杷	{loquat}	fruit|水果
			{loquat}	tree|樹
{kite}	any of several small graceful hawks of the family Accipitridae having long pointed wings and feeding on insects and small animals	鳶	{kite}	tool|用具, *WhileAway|消閑
			{kite}	bird|禽

註： 標示底線者，表示與同義詞集共用詞彙。

6.3 標示語意中心詞彙實驗

我們先將複合名詞進行斷詞拆解成詞彙，在 39228 個中文複合詞中，所拆解出的詞彙組合數分佈如表 12 所示，以含兩個詞彙之複合詞最多，佔 56.9%。對於由兩個詞彙組成的複合名詞，經第 4 節的詞彙相似度量公式，選出語意中心詞彙，實驗結果在詞彙組合數為 2 的部分，複合詞之中心詞彙來自第一個詞彙的有 2120 筆，來自第二個詞彙的有 10481 筆，而無法成功辨識出中心詞彙的同義詞集共有 9727 筆（見表 13）。

接著，我們將標上中心詞彙的同義詞集資料依第 6.2 節所述方法進行語意概念標示，結果有 5756 個同義詞集可標上語意概念。因此，比對於參與語意標示實驗的 11439 個同義詞集，應用率為 50.3%。為評估連結之正確率，我們自連結資料中隨機選取 288 筆，經人工比對後計有 261 筆為正確之連結，正確率為 90.6 %。實驗結果與 6.2 節比較正確率差異不大，但是應用率卻降低不少，主要原因為中文複合詞與其語意中心詞彙在對應之英文詞彙上，較難一致。若是考慮將對英文詞彙的一致要求，轉至概念上的一致性，應可提升應用率。

表 12 參與中心詞彙標示實驗資料之複合名詞之詞彙組合數統計。

	詞彙組合數		
	2	3	4
複合詞數	22330	9742	4173
同義詞集總數	19323	8994	3991

表 13 複合名詞中心詞彙標示之實驗結果。

		詞彙組合數		
		2	3	4
複合詞數		22330	9742	4173
語意中心詞彙位置	無法判定	9729	3931	1482
	詞彙一	2120	599	240
	詞彙二	10481	790	207
	詞彙三	------	4422	299
	詞彙四	------	------	1945
應用率		56.4%	59.7%	64.5%
正確率		84.8%	75.2%	70.0%

7. 結論

本研究提出一套方法，結合比對概念語詞、詞彙本身、以及上位關係的詞彙內容等技巧，將已標示中文翻譯的 WordNet 同義詞集對應上知網概念定義。並且，對於中文複合詞找出主要特徵所在。實驗顯示有不錯的結果。在未來的研究擬將詞彙相似度比對由詞彙擴充至概念，克服同義詞問題，以提升應用率。

致謝

本研究獲得國科會編號 NSC 90-2213-E-031-005 計畫之補助，特此致謝。

参考資料

Aslandogam, Y. A., C. Their, C. T. Yu, J. Zou and N. Rishe, "Using Semantic Contents and WordNet in Image Retrieval," In *Proceedings of the 20th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, Philadelphia, 1997, pp. 286-295.

Atserias, J., S., Climent, X., Farreres, G. Rigau and H. Rodriguez, "Combining Multiple Methods for the Automatic Construction of Multilingual WordNets," In *Proceedings of International Conference of Recent Advances in Natural Language Processing* (*RANLP'97*), Tzigov Chark, Blgaria, 1997.

Carpuat, M., G. Ngai, P. Fung and K. W. Church, "Creating a Bilingual Ontology: A Corpus-Based Approach for Aligning WordNet and HowNet," In *Proceedings of the 1st International Conference on Global WordNet*, Mysore, India, 2002.

Chang, J. S., S. J. Ker and M. H. Chen, "Taxonomy and Lexical Semantics - from the Perspective of Machine Readable Dictionary," In *Proceedings of 3rd Conference of the Association for Machine Translation in the Americas* (*AMTA'98*), 1998, pp. 199-212.

Chen, J. N. and J. S. Chang, "TopSense: A Topical Sense Clustering Method based on Information Retrieval Techniques on Machine Readable Resources," *Special Issue on Word Sense Disambiguation, Computational Linguistics*, 24(1), 1998, pp. 61-95.

Chen, Hsin-Hsi, Chi-Ching Lin and Wen-Cheng Lin, "Construction of a Chinese-English WordNet and Its Application to CLIR," In *Proceedings of 5th International Workshop on Information Retrieval with Asian Languages*, Hong Kong, 2000, pp. 189-196.

Chen, Hsin-Hsi and Chi-Ching Lin, "Sense-Tagging Chinese Corpus," In *Proceedings of 2nd Chinese Language Processing Workshop*, Hong Kong, 2000, pp. 7-14.

Dagan, I. and A. Itai, "Word Sense Disambiguation Using a Second Language Monolingual Corpus," *Computational Linguistics*, 20(4), 1994, pp. 563-596.

Dice, L. R., "Measure of the Amount of Ecologic Association between Species," *Journal of Ecolog*, 26, 1945, pp. 297-302.

Dorr, B. J., G-A Levow, D. Lin and S. Thomas, “Chinese-English Semantic Resource Construction,” In *Proceedings of 2nd International Conference on Language Resources and Evaluation,* (*LREC 2000*), Athens, Greece, 2000, pp. 757-760.

Farreres, X., G. Rigau and H., Rodriguez, “Using WordNet for Building WordNets,” In *Proceedings of the Workshop of Usage of WordNet in NLPS*, COLING-ACL’98, 1998, pp. 65-72.

Fellbaum, C. ed., *WordNet: An Electronic Lexical Database*, MIT Press, May 1998.

Gale, W. A., K. W. Church and D. Yarowsky, “Using Bilingual Materials to Develop Word Sense Disambiguation Methods,” In *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation*, 1992, pp. 101-112.

Gomez, F., “Linking WordNet Verb Classes to Semantic Interpretation,” In *Proceedings of the Workshop of Usage of WordNet in NLPS, COLING-ACL’98*, 1998, pp. 58-64.

Gonzalo, J., F. Verdejo, I. Chugur and J. Cigarran, “Indexing with WordNet Synsets can Improve Text Retrieval,” In *Proceedings of the Workshop of Usage of WordNet in NLPS, COLING-ACL’98*, 1998, pp. 38-44.

Guthrie, J., L. Guthrie, Y. Wilks and H. Aidinejad, “Subject-Dependent Co-Occurrence and Word Sense Disambiguation,” In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, 1991, pp. 146-152.

Jing, H., “Usage of WordNet in Natural Language Generation,” In *Proceedings of the Workshop of Usage of WordNet in NLPS*, COLING-ACL’98, 1998, pp. 128-134.

Knight, K. and S. K. Luk, “Building a Large-scale Knowledge Base for Machine Translation,” In *Proceedings of The Twelfth National Conference on Artificial Intelligence*, 1994, pp. 773-778.

Lee, C., G. Lee and S. J. Yun, “Automatic WordNet Mapping using Word Sense Disambiguation,” In *Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora,* 2000, pp. 142-147.

Li, X., S. Szpakowicz and S. Matwin, “A WordNet-Based Algorithm for Word Semantic Sense Disambiguation,” In *Proceedings of the 14th International Joint Conference on Artificial Intelligence IJCAL-95*, Montreal, Canada, 1995.

Luk, A. K., “Statistical Sense Disambiguation with Relatively Small Corpora using Dictionary Definitions,” In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, 1995, pp. 181-188.

Mandala, R., T. Tokunaga and H. Tanaka, “The use of WordNet in Information Retrieval,” In *Proceedings of the Workshop of Usage of WordNet in NLPS*, *COLING-ACL’98*, 1998, pp. 31-37.

Miller, G. A., “Five papers on WordNet,” *International Journal of Lexicography*, 3(4) 1990.

Ng, H. T. and H. B. Lee, “Integrating Multiple Knowledge Sources to Disambiguate Word Sense: an Exemplar-Based Approach,” In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, Santa Cruz, 1996, pp. 40-47.

Resnik, P., "Selection and Information: A Class-Based Approach to Lexical Relationships," *Doctoral Dissertation, Department of Computer and Information Science, University of Pennsylvania*, 1993.

Riloff, E. and R. Jones, "Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping," In *Proceedings of 16th National Conference on Artificial Intelligence*, 1999, pp. 474-479.

Slator, B., "Using Context for Sense Preference," In Zernik (ed.) *Lexical Acquisition: Exploiting on-line Resources to Build a Lexicon,* Lawrence Erlbaum, Hillsdale, NJ, 1991.

Wang, Chi-Yung, "Knowledge-based Sense Pruning using the HowNet: An Alternative to Word Sense Disambiguation," *Thesis of Hong Kong University of Science and Technology*, *Computer Science*, 2002.

Yang, C. and S. J. Ker, "Considerations of Linking WordNet with MRD," In *Proceedings of the 19th International Conference on Computational Linguistics*, 2002, pp. 1121-1127.

Yarowsky, D., "Unsupervised Word Sense Disambiguation Rivalling Supervised Methods," In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, 1995, pp. 189-196.

Yarowsky, D., "Word-Sense Disambiguation using Statistical Models of Roget's Categories Trained on Large Corpora," In *Proceedings of the 14th International Conference on Computational Linguistics*, Nantes, France, 1992, pp. 454-460.

董振東、董強：知網, 2000, http://www.keenage.com/.