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.

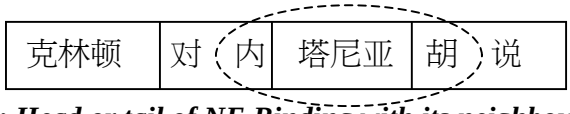

Figure 1: Head or tail of NE Binding with its neighbours.

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/*琳*/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,

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

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

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

According to the Bayes' Theorem, we can get

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

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

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

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;

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

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

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

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.

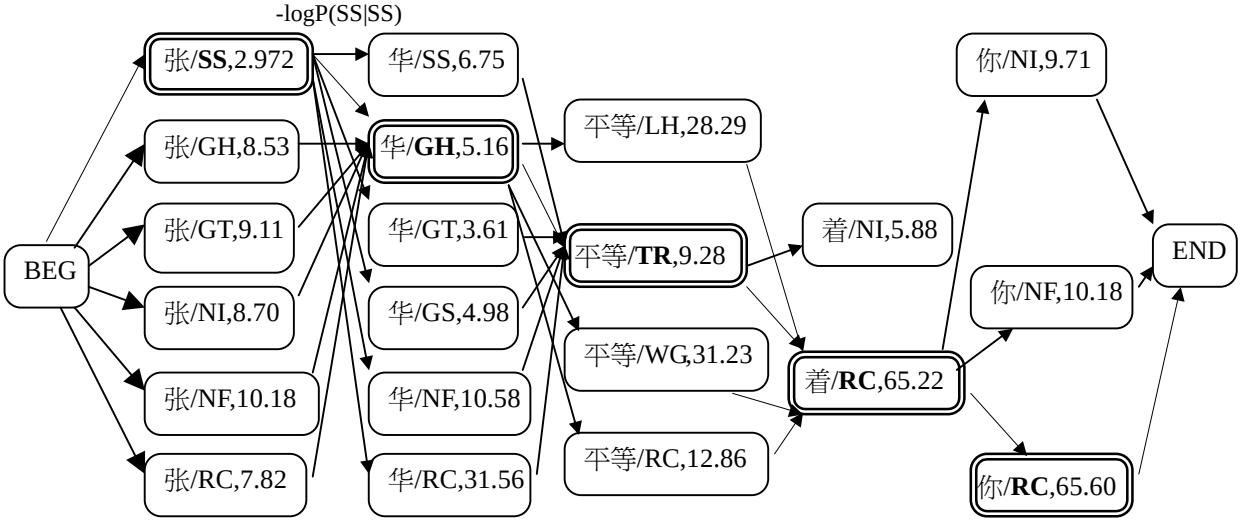

Figure 4: Role selection using the Viterbi algorithm.

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:

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

, 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;

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

, 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:

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

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)×p(张|SS)×p(GH|SS)×p(华|GH)×p(GT|GH)×p(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 $\mathbf{w}_i$

[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,

$$W^{\#} = \arg\max_{W} P(W)$$
$$= \arg\max_{W} P(W|C)P(C).$$

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

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

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.

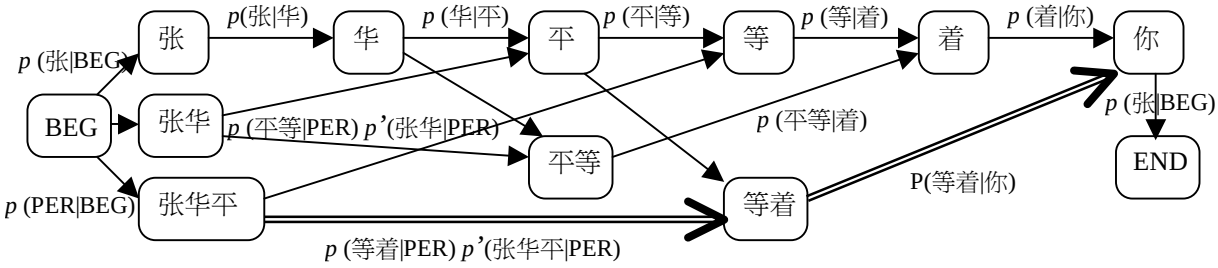

Figure 6:Demonstration of segmentation on "张华平等着你" using the class-based approach.

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/

㠭”(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,

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

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

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

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: *勤(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.