Chinese Named Entity Recognition using Lexicalized HMMs

Guohong Fu
Department of Linguistics, The University of Hong Kong
Pokfulam Road, Hong Kong
ghfu@hkucc.hku.hk

Kang-Kwong Luke
Department of Linguistics, The University of Hong Kong
Pokfulam Road, Hong Kong
kkluke@hkusua.hku.hk

ABSTRACT

This paper presents a lexicalized HMM-based approach to Chinese named entity recognition (NER). To tackle the problem of unknown words, we unify unknown word identification and NER as a single tagging task on a sequence of known words. To do this, we first employ a known-word bigram-based model to segment a sentence into a sequence of known words, and then apply the uniformly lexicalized HMMs to assign each known word a proper hybrid tag that indicates its pattern in forming an entity and the category of the formed entity. Our system is able to integrate both the internal formation patterns and the surrounding contextual clues for NER under the framework of HMMs. As a result, the performance of the system can be improved without losing its efficiency in training and tagging. We have tested our system using different public corpora. The results show that lexicalized HMMs can substantially improve NER performance over standard HMMs. The results also indicate that character-based tagging (viz. the tagging based on pure single-character words) is comparable to and can even outperform the relevant known-word based tagging when a lexicalization technique is applied.

Keywords

Chinese named entity recognition, lexicalized hidden Markov models, known word tagging, character tagging.

1. INTRODUCTION

The goal of named entity recognition (NER) is to recognize phrases in a document that indicate the names of persons, organizations, locations, times or quantities. As an important subtask of information extraction and text mining, NER has been attracting more and more attention in the NLP community. It has now become a shared task of a number of conferences or projects, such as the Multilingual Entity Task (MET) at the Message Understanding Conferences (MUCs), the language-independent NER task at CoNLL-2002 and CoNLL-2003, and the 1999 DARPA-TIDES Information Extraction-Entity Recognition (IEER-99) technology evaluation project.

Current research on NER has focused on machine learning approaches, including hidden Markov models (HMMs) [1][2][3], maximum entropy (ME) [4], transformation-based error-driven learning (TBL) [5], and support vector machines (SVMs) [6]. In comparison with rule-based methods, machine-learning approaches are more adaptive and robust. However, it is still a challenge for most of them to keep a balance between capacity and computational cost [7]. While a HMM-based tagger has proven to be very speedy in training and tagging [8], it usually achieves relatively lower tagging accuracy for it only takes into account the context of the category tags, and no contextual word information, which sometimes gives strong evidence for NER. On the contrary, some learning methods such as ME and SVMs are capable of combining much richer lexical information in a straightforward way. However, they usually need much more time in training and tagging, which will become a serious problem in processing a large amount of data or some on-line applications like text mining. In order to address these problems, some recent work suggested the use of lexicalization techniques to enhance the standard HMMs [8][9][10]. Their experiments demonstrated that their systems could be improved without increasing much computational cost in training and processing.

Recently, a number of methods have been reported for Chinese NER. Sun *et al*. proposed a class-based language model approach to Chinese NER [11]. In their work, they used different models to identify different types of NEs in Chinese text, including a character-based trigram for *person*, a word-based model for *location* and a more complicated model for *organization*. Further, in [12], Wu *et al*. modified the class-based language model approach by incorporating human knowledge, particularly semantic information. Zhang *et al*. put forward a stochastic role model to recognize Chinese NEs [13]. In this work, they defined a set of roles about component tokens within a Chinese NE and the relevant contexts. Their experiments showed that the role-based model was effective for different NEs. More recently, Chen *et al.* proposed a smoothing maximum entropy model for Chinese nominal entity tagging [14]. They suggested that simple semantic features extracted from a dictionary help improve the performance of the model in NER, especially when the training data is not sufficient. Guo *et al.* presented a robust risk minimization (RRM) classification method to Chinese NER, which was able to incorporate the advantages of character-based and word-based models [15]. Their experiments have also demonstrated that local Chinese characters, Chinese word segmentation information, the surrounding context and part-of-speech (POS) are the most informative features that have significant impacts on the performance of NER.

Although much progress has been made in the literature, it is still a big challenge to develop a high-performance NER system for Chinese due to the language-specific issues in Chinese. Unlike other languages such as English and Spanish, there are no explicit delimiters to indicate word boundaries in a plain Chinese text. Word segmentation is therefore an essential step to many Chinese processing tasks. The second issue concerns unknown words in open-ended documents. Most current systems need a dictionary to guide their analysis. However, no dictionary could be complete. While a predefined dictionary may cover most words in use, there are many other words in open-ended documents, such as proper nouns and domain-specific terms that cannot be exhaustively listed. On the other hand, unknown word

identification (UWI) is still a difficult problem for unknown words are constructed freely and dynamically in Chinese. Furthermore, it is not easy to explore word-internal cues and contextual information for NER from an open set of unknown words. Finally, there is less exterior information in plain Chinese texts, such as capitalization in English to help identify entity names and unknown words.

In this paper, we propose a lexicalized HMM approach to Chinese NER. In order to address the problem of unknown words, we unify Chinese UWI and NER, and reformulate them as a single tagging process on a sequence of known words (viz. lexicon words that are listed in the system lexicon). To do this, we develop a two-stage NER system for Chinese. Given a sentence, a known word bigram model is first applied to segment it into a meaningful sequence of known words. Then, a lexicalized HMM tagger is used to assign each known word a proper hybrid tag that indicates its pattern in forming an entity and the category of the formed entity. In comparison with previous methods, our system is able to explore three types of features, i.e. entity-internal formation patterns, contextual word evidence and contextual category information, and combine them for NER under the framework of HMMs. As a consequence, the system's performance can be improved without losing its efficiency in training and processing.

The rest of this paper is organized as follows: In section 2, we discuss how to reformulate Chinese NER as a tagging problem on a sequence of known words. In section 3, we present a bigram model for known word segmentation. In section 4, we describe in detail a lexicalized HMM-based tagger for Chinese NER. We report in section 5 our experimental results and give our conclusions on this work in section 6.

2. NER AS KNOWN WORD TAGGING

2.1 Categorization of Entities

Named entity types	Abbreviated SGML tags	
Person	<PER>	</PER>
Chinese personal names	<CPN>	</CPN>
Transliterated personal names	<TPN>	</TPN>
Location	<LOC>	</LOC>
Organization	<ORG>	</ORG>
Other names	<ONR>	</ONR>
Date	<DTE>	</DTE>
Time	<TME>	</TME>
Duration	<DUR>	</DUR>
Money	<MNY>	</MNY>
Measure	<MSR>	</MSR>
Percent	<PCT>	</PCT>
Cardinal	<CRD>	</CRD>
Other numbers	<ONU>	</ONU>

Table 1 Categories of named entities in Chinese

In our work, we use the same named entity tag set as defined in the IEER-99 Mandarin named entity task. [1]As shown in Table 1, this task specifies twelve different types of NEs for Chinese. These entity categories are further encoded using twelve different abbreviated SGML tags. To show the different formation rules between Chinese personal names and transliterated personal names, we subdivide the class *personal name* (PER) into two groups, namely *Chinese personal name* (CPN) and *transliterated personal name* (TPN). In addition to NEs, our system will also assign each common word in the input sentence a proper POS tag. For convenience, we adopt the Peking University POS tag-set, which contains 48 different POS tags [16].

[1] The detail of IEER-99 Mandarin named entity task is available at http://www.nist.gov/speech/tests/ie-er/er_99/er_99.htm.

2.2 Patterns of Known Words in NER

In general, a named entity can be composed of one known word or several known words. In other words, a known word may present itself as an independent entity or a component of an entity after NER. Similar to UWI [17], a known word *w* may take one of the following four patterns to present itself during NER: (1) *w* is an independent named entity; (2) *w* is the beginning component of a named entity; (3) *w* is at the middle of a named entity; (4) *w* is at the end of a named entity. In our work, we use four tags *ISE*, *BOE*, *MOE* and *EOE* to denote the above four patterns respectively.

Other than common segmented words in a sentence, we consider known words to be the basic units or components within a named entity, because: Firstly, any Chinese unknown word or entity name is actually a combination of known words if the system dictionary covers all Chinese characters. It is therefore very convenient to handle word-internal clues for NER based on known words. Secondly, tagging based on known words is more general and actually contains two major notions for NER: the character-level model and the common known-word model. In fact, the character-level model discussed in [18][19][20] is a special form of the known word model, in which the system dictionary only consists of single-character words. For this reason, we also refer to this model as pure single-character word model. Thirdly, UWI and NER can be unified as a single tagging task on a sequence of known words. Moreover, a Chinese sentence can be segmented into a sequence of known words with accuracy using the known-word n-grams [17].

Obviously, a segmented named entity in a sentence can be represented as a sequence of known words together with their pattern tags. For example, the segmented string "温家宝/总理/" (*wen1jia1bao3 zong3li3*, *Premier Wen Jiabao*) is equivalent to "<BOE>温</BOE><MOE>家</MOE><EOE>宝</EOE><ISE>总理</ISE>".

In other words, the boundary of an entity name will be determined if all its components are assigned a proper pattern tag. At this point, the identification of NEs can be viewed as a process of assigning each known word in the input an appropriate pattern tag that indicates its position in an entity. For example, a known word will be tagged with *ISE* if it is an independent entity name. Similarly, a known word will be labeled with *BOE*, *MOE* or *EOE* respectively if it is a beginning, middle or end component of a named entity.

2.3 NER as Known Word Tagging

However, a full named entity task involves identifying and classifying NEs in documents. To do this, we define a hybrid tag set by merging the category tags defined in section 2.1 and the pattern tags defined in section 2.2. In our work, a hybrid tag has

a format as follows: $t_C - t_P$. Where, t_C denotes the category tag of a named entity, and t_P denotes the pattern tag of a known word within the named entity.

Thus, a NE-tagged sentence can be fully reformulated as a sequence of known words together with their hybrid tags. Given an entity name $E = w_1 w_2 \cdots w_n$, it is normally tagged as $< t_C > E < /t_C >$ after NER. Under our new formulation, this standard format is represented as follows:

$$< t_C\text{-}t_{P1} > w_1 < /t_C\text{-}t_{P1} > \cdots < t_C\text{-}t_{Pn} > w_n < /t_C\text{-}t_{Pn} > \quad (1)$$

Where, $w_i (1 \le i \le n)$ stands for a known word within the named entity E, t_C stands for the category tag of the named entity E; $t_{Pi} (1 \le i \le n)$ denotes the pattern tag of the known word w_i. In this formulation, each known word in an entity should have the same category tag as the entity.

(a) 中国国家主席胡锦涛同北朝鲜领导人金正日举行会谈。

(b) *Chinese President Hu Jintao held talks with North Korean leader Kim Jong-Il*

(c) *zhong1guo2 guo2jia1 zhu3xi2 hu2jin3tao1 tong2 bei3chao2xian3 ling3dao3ren2 jin1zheng4ri4 ju3xing2 hui4tan2.*

(d) 中国/国家/主席/胡/锦/涛/同/北朝鲜/领导人/金/正/日/举行/会谈/。/

(e) <LOC-ISE>中国</LOC-ISE><n-ISE>国家</n-ISE> <n-ISE>主 席</n-ISE> <CPN-BOE>胡 </CPN-BOE> <CPN-MOE>锦 </CPN-MOE> <CPN-EOE>涛 </CPN-EOE> <p-ISE>同</p-ISE> <LOC-ISE>北朝鲜</LOC-ISE> <n-ISE>领导人</n-ISE> <CPN-BOE>金</CPN-BOE> <CPN-MOE>正 </CPN–MOE> <CPN-EOE>日 </CPN-EOE <v-ISE>举 行 </v-ISE> <vn-ISE>会谈</vn-ISE> <w-ISE>。</w-ISE>

(f) <LOC>中国</LOC> <n>国家</n> <n>主席</n> <CPN>胡锦涛</CPN> 同 <LOC>北朝鲜</LOC> <n>领导人 </n> <CPN>金正日 </CPN> <v>举行 </v> <vn>会谈 </vn> <w>。</w>

Figure 1 Representing a NE-tagged sentence as a sequence of known words together with their hybrid-tags

Figure 1 gives an example of different representations of NEs in the Chinese sentence "中国国家主席胡锦涛同北朝鲜领导人金正日举行会谈。". Where, (a) is the original sentence, and the next three sequences, i.e. (b), (c) and (d), are respectively the English translation, the transcription in Chinese Phonetic Alphabet and the segmentation of known words for this sentence. As can be seen from this figure, the standard NE tagged string (f) can be equivalently converted to a sequence of known words and their hybrid tags, as shown in (e).

3. KNOWN WORD SEGMENTATION

The goal of known word segmentation is to segment a sequence of characters into a meaningful sequence of known words. In a sense, known word segmentation is actually a process of disambiguation. In our system, we apply known word bigram language models to resolve word boundary ambiguities in known word segmentation.

Given a Chinese character string $C = c_1 c_2 \cdots c_m$, there may be multiple candidate known word sequences $\{W = w_1 w_2 \cdots w_n\}$ according to a given system lexicon. Known word bigram segmentation aims to find the most appropriate segmentation $\hat{W} = w_1 w_2 \cdots w_n$ that maximizes the conditional probability $P(W \mid C)$, i.e.

$$\hat{W} = \arg\max_W P(W \mid C) \approx \arg\max_W \prod_{t=1}^{n} P(w_t \mid w_{t-1}) \quad (2)$$

where $P(w_i \mid w_{i-1})$ denotes the known word bigram probability, which can be estimated from a segmented corpus using maximum likelihood estimation (MLE). It should be noted that all unknown words in the training corpus must be decomposed to a sequence of known words before counting known word bigrams. For simplicity, we employ the maximum match technique [21] to perform this conversion. To resolve the issue of data sparseness in MLE, we apply the linear interpolation technique to smooth the estimated word bigram probabilities.

4. LEXICALIZED HMM TAGER

4.1 Lexicalized HMMs

At present, two types of lexicalization techniques are used to improve HMM-based taggers, i.e. the uniformly lexicalized HMMs [9] and the selectively lexicalized HMMs [8][10]. In view of the convenience in implementation, we employ the uniformly lexicalized models to perform the tagging of known words for Chinese NER.

Given a sequence of known words $W = w_1 w_2 \cdots w_n$, the task of the tagger for Chinese NER is to find an appropriate sequence of hybrid tags $\hat{T} = t_1 t_2 \cdots t_n$ that maximizes the conditional probability $P(T \mid W)$, namely

$$\hat{T} = \arg\max_T P(T \mid W) = \arg\max_T \frac{P(W \mid T) P(T)}{P(W)} \quad (3)$$

Since the probability $P(W)$ remains fixed for all candidate tag sequences, we can disregard it. Thus, we have a general statistical model for Chinese NER as follows

$$\begin{aligned} \hat{T} &= \arg\max_T P(W \mid T) P(T) = \arg\max_T P(W, T) \\ &= \arg\max_T \prod_{i=1}^{n} P(w_i \mid w_1 \cdots w_{i-1}, t_1 \cdots t_i) P(t_i \mid w_1 \cdots w_{i-1}, t_1 \cdots t_{i-1}) \end{aligned} \quad (4)$$

In theory, the general model in Equation (4) can provide the tagging system with a powerful capacity of disambiguation. However, this general model is not computable in practice for it involves too many parameters. Generally, two types of approximations are employed to simplify the general model.

The first approximation is based on the independent hypothesis used in standard HMMs: The appearance of current word w_i depends only on current tag t_i during tagging, and the

assignment of current tag t_i depends only on its previous K $(1 \le K \le i-1)$ tags $t_{i-K} \cdots t_{i-1}$. Based on these assumptions, the general model in Equation (4) can be rewritten as

$$\hat{T} = \arg\max_{T} \prod_{i=1}^{n} P(w_i \mid t_i) P(t_i \mid t_{i-K} \cdots t_{i-1}) \quad (5)$$

Where, $P(w_i \mid t_i)$ denotes the so-called lexical probability; and $P(t_i \mid t_{i-K} \cdots t_{i-1})$ denotes the contextual tag probability. In view of the problem of data sparseness, we use the first-order HMMs in our system, i.e. $P(t_i \mid t_{i-K} \cdots t_{i-1}) \approx P(t_i \mid t_{i-1})$.

The second approximation follows the notion of the lexicalization technique, where two main hypotheses are made: The appearance of current word w_i is assumed to depend not only on the current tag t_i and the previous $I(1 \le I \le i-1)$ tags $t_{i-I} \cdots t_{i-1}$ but also the previous $J(1 \le J \le i-1)$ words $w_{i-J} \cdots w_{i-1}$; The assignment of current tag t_i is supposed to depend both on its previous $K(1 \le K \le i-1)$ words $w_{i-K} \cdots w_{i-1}$ and $L(1 \le L \le i-1)$ tags $t_{i-L} \cdots t_{i-1}$. Thus,

$$\hat{T} = \arg\max_{T} \prod_{i=1}^{n} \begin{pmatrix} P(w_i \mid w_{i-J} \cdots w_{i-1}, t_{i-I} \cdots t_{i-1} t_i) \\ \times P(t_i \mid w_{i-K} \cdots w_{i-1}, t_{i-L} \cdots t_{i-1}) \end{pmatrix} \quad (6)$$

Equation (6) gives a general form of the uniformly lexicalized HMMs for Chinese NER. With a view to the issue of data sparseness, we set $I = 0$ and $J = K = L = 1$.

By comparison, the uniform lexicalization technique is able to handle richer contextual information for the assignment of tags to known words, including both contextual words and contextual tags under the framework of HMMs. Consequently, the accuracy of the named entity recognizer can be improved without losing its efficiency in training and tagging.

If a large NE-tagged corpus is available, the parameters in Equation (5) and (6) can be easily estimated using the MLE technique. However, MLE will yield zero probabilities for any cases that are not observed in the training data. To solve this problem, we employ the linear interpolation smoothing technique to smooth higher-order models with their relevant lower-order models, or to smooth the lexicalized parameters using the related non-lexicalized probabilities, namely

$$\begin{cases} P'(w_i \mid w_{i-1}, t_i) = \lambda P(w_i \mid w_{i-1}, t_i) + (1-\lambda) P(w_i \mid t_i) \\ P'(t_i \mid w_{i-1}, t_{i-1}) = \mu P(t_i \mid w_{i-1}, t_{i-1}) + (1-\mu) P(t_i \mid t_{i-1}) \end{cases} \quad (7)$$

where λ and μ denote the interpolation coefficients.

4.2 Lattice-Based Tagging

Based on the above models, the tagging algorithm aims at finding the most probable sequence of hybrid tags for a given sequence of known words. In our implementation, we employ the classical Viterbi algorithm to perform this task, which works in three major steps as follows:

(1) The generation of candidate tags: This step aims to generate a lattice of candidate hybrid tags for a sequence of known words produced by the known word segmenter. As discussed above, a hybrid tag of a known word involves a category tag and a pattern tag. Given a known word, it may take one of the four patterns defined in Section 2.2 to present itself in a segmented word or entity. All the four pattern tags are therefore its eligible candidates. As for its category tag candidates, they can be constructed by looking up the system dictionary and the lexical probability library. The candidate hybrid tags of a known word are a combination of its candidate category tags and its candidate pattern tags. All these candidates are stored in a lattice structure.

(2) The decoding of the best tag sequence: In this step, the well-known Viterbi algorithm is employed to score all candidate tags with the proposed language models, and then search the best path through the lattice that has the maximal score. This path contains the best sequence of tags for the known word string.

(3) The conversion of the results: The direct output of our tagger has the same format as shown in formula (1). For evaluation purposes, we further convert it to the standard representation by merging the consecutive known words into entities in terms of their patterns.

4.3 Inconsistent Tagging

Our system may yield two types of inconsistent tagging, namely pattern inconsistency and class inconsistency.

Pattern inconsistency arises when two adjacent known words are assigned inconsistent pattern tags such as “ISE:MOE” or “ISE:EOE”. It has been shown that the inconsistent pattern tagging hardly exerts any influence on the results in word segmentation [20]. In practice, entity boundary detection is very similar to word segmentation. This suggests by analogy that the inconsistency in pattern tagging has no effects on the identification of entity boundaries. For this reason, we do nothing to the inconsistent patterns during the result conversion.

Category inconsistency means that two adjacent known words are labeled with different category-tags while at the same time, they are assigned the pattern tags that indicate they should appear in the same word or named entity. For example, the Chinese personal name 张晓华 (Zhang Xiaohua) might be inconsistently tagged as <CPN-BOE> 张 </CPN-BOE><Vg-MOW> 晓 </Vg-MOW><CPN-EOE> 华 </CPN-EOE>. In this case, the system cannot make its decision in choosing a category tag for the personal name 张 晓 华 (Zhang Xiaohua). According to our intuition, the end component may be more informative in classifying Chinese NEs. Furthermore, few inconsistent category-tags can occur in the results because they usually have lower probabilities, and will be accordingly blocked by the decoder. Therefore, we resolve these inconsistent categories just by assuming the categories of ending components to be that of the relevant NEs or unknown words.

5. EXPERIMENTS

To evaluate our approach, we conducted a number of experiments on our system using the public PFR corpus, the IEER-99 newswire data and the MET2 data. This section reports the results of these experiments.

5.1 Experimental Measures

We evaluate our system in terms of *recall* (R), *precision* (P) and *F-measure* (F). Here, recall (R) is defined as the number of correctly recognized NEs divided by the total number of NEs in

the manually annotated corpus, and precision (P) is defined as the number of correctly recognized NEs divided by the total number of NEs recognized by the system. In our evaluation, a recognized entity is correct if and only if both its boundary and its category are the same as the manual annotations in the data for testing. As shown in Equation (8), F-measure is a weighted harmonic mean of precision and recall.

$$F_\beta = \frac{(\beta^2 + 1) \times P \times R}{\beta^2 \times P + R} \quad (8)$$

Where β is the weighing coefficient. In our experiments, we use the balanced F-score (viz. $F_\beta = 1$) to evaluate the overall performance of our system because it is not clear whether recall or precision is more important in evaluating a NE recognizer.

5.2 Experimental Data

As shown in Table 2, we use three types of corpora in our experiments: The PKU corpus is a manually tagged corpus containing one month of news texts from *the People's Daily* (January 1998) [16]. In this work, we further annotate this corpus with the entity tags defined in Table 1 mainly under the guidance of the IEER-99 Mandarin NE Task Definition (version 1.2). Moreover, we divide it into two parts: 90% for training and 10% for testing. The IEER-99 newswire test data is originally used for the IEER evaluation sponsored by the National Institute of Standard and Technology. The third corpus is the MET2 test data, which is originally used for Chinese NER evaluation at the Second Multilingual Entity Task (MET2). In our experiments, we use the later two corpora as the data for an open comparison evaluation.

Entity Category	PKU Corpus		IEER-99 Test Data	MET2 Test Data
	Training	Testing		
CPN	12,861	1,462	489	174
TPN	2,919	333		
LOC	23,626	2,428	1,026	750
ORG	15,228	1,709	497	377
Total	58,707	5,932	2,012	1,301

Table 2 Experimental corpora

5.3 Experimental Results

In order to examine the effectiveness of our system, we conducted a number of experiments using the corpora in Table 2. In particular, we intended to examine the following three issues through these experiments:

(1) In principle, lexicalized HMMs should be more powerful than standard HMMs in the tagging for Chinese NER because lexicalized HMMs can handle richer contextual information for tagging, in particular the contextual lexical information. Consequently, our first aim is to examine how the use of the lexicalization technique affects the performance of our system.

(2) In practice, the formulation of NER as a tagging task on a sequence of known words involves two different models: the word-level model (viz. the common known word model) and the character-level model (viz. the pure single-character known word model). There are some arguments in the community of NER about whether a word model or a character model is better. For this reason, our second intention is to investigate whether the word-level mode or the character-level model is more effective for Chinese NER.

(3) The third motivation of our experiments is to compare our system with other public systems for Chinese NER.

For comparison purpose, we concentrate our evaluation on the three major groups of NEs, i.e. personal names (PER, including Chinese personal names (CPN) and transliterated personal names (TPN), organization names (ORG) and location names (LOC). The experimental results are presented below.

Methods	Entity	R (%)	P (%)	$F_{\beta=1}$ (%)
Character based tagging with standard HMMs	CPN	79.41	77.97	78.69
	TPN	67.27	51.58	62.05
	LOC	52.22	67.66	58.95
	ORG	26.02	46.05	33.25
Character based tagging with lexicalized HMMs	CPN	91.24	92.45	91.84
	TPN	89.19	90.27	89.73
	LOC	85.01	87.11	86.05
	ORG	80.92	82.39	81.65
Known-word based tagging with standard HMMs	CPN	87.89	82.64	85.18
	TPN	84.38	68.70	75.74
	LOC	76.44	78.11	77.27
	ORG	66.47	71.09	68.70
Known-word based tagging with lexicalized HMMs	CPN	91.72	89.88	90.79
	TPN	92.49	90.06	91.26
	LOC	88.67	85.64	87.13
	ORG	84.55	82.67	83.60

Table 3 Results for the evaluation of different models using the PKU corpus

Table 3 shows the results of the experiments on the NE-tagged PKU test corpus.

Systems	Entity	R (%)	P (%)	$F_{\beta=1}$ (%)
Sun *et al.* [11]	PER	84.43	79.38	81.83
	LOC	80.18	79.09	79.63
	ORG	62.30	88.03	72.96
Wu *et al.* [12]	PER	92.28	83.30	87.56
	LOC	84.69	88.31	86.47
	ORG	71.08	86.09	84.61
Character-based tagging with lexicalized HMMs	PER	86.71	89.26	87.97
	LOC	80.66	84.72	82.64
	ORG	76.07	74.63	75.34
Known-word based tagging with lexicalized HMMs	PER	87.73	87.37	87.55
	LOC	82.03	82.84	82.43
	ORG	71.15	70.40	70.78

Table 4 Results for the evaluation using the IEER-99 data

Table 4 gives the results of the evaluation using the IEER-99 test data. In this evaluation, two other public systems, i.e. the system

developed by Sun *et al*. [11] and the system by Wu *et al*. [12] are shown for comparison.

Systems	Entity	R (%)	P (%)	$F_{\beta=1}$ (%)
The KRDL system [22]	PER	92	66	76.7
	LOC	91	89	90.0
	ORG	88	89	88.5
The NTU system [23]	PER	91	74	81.6
	LOC	78	69	73.2
	ORG	78	85	81.3
Character-based tagging with lexicalized HMMs	PER	89.66	69.03	78.00
	LOC	81.78	73.13	77.21
	ORG	74.01	67.72	70.72
Known-word based tagging with lexicalized HMMs	PER	92.53	64.92	76.30
	LOC	80.72	72.78	76.54
	ORG	73.47	66.27	69.69

Table 5 Results for the evaluation using the MET-2 data

Table 5 lists the results of the evaluation using the MET2 data. For comparison purpose, we also list the corresponding results of two public systems, i.e. the NTU (National Taiwan University) System and the KRDL (Kent Ridge Digital Labs) system.

From these results, we can draw the following conclusions:

(1) As can be seen in Table 3, the lexicalized HMMs significantly outperform the standard HMMs for all types of NEs under investigation. This indicates that the use of lexicalization technique leads to the improvement of accuracy in NER.

(2) The character model can yield results that are comparable to or better than the word-level model with the lexicalization technique, for all test data. However, the character model performs worse than the word-level model without the lexicalization technique.

(3) It can be observed that the proposed lexicalized HMM approaches are effective for most Chinese NEs. However, it achieves the relatively lower performance for entities like ORG. The reason may be that organization names usually have more complicated structures that are possibly beyond the current models. Moreover, some types of organization names are not clearly specified in the IEER-99 named entity task, it is therefore difficult to perform consistent annotation on them.

(4) As shown in Table 4, our methods, whether the word level model or the character-level model, perform better than the system in [11] as a whole. However, they perform worse than the system of Wu *et al*.[12] except for personal names. The reason may be that Wu *et al*.[12] have integrated some additional human knowledge, in particular semantic features in their system.

(5) By comparing Table 3, 4 and 5, we can see that our system yields worse results for MET2 data than for IEER-99 data or the NE-tagged PFR corpus. An intensive error analysis shows that wrongly recognized entities mainly result from three causes: the problem of data sparseness, the inconsistent tagging between the training data and the MET2 data, and some complicated NEs such as nested organization names that are beyond the sequence models.

6. CONCLUSIONS

In this paper, we have presented a lexicalized HMM-based approach to Chinese NER. In particular, we formalize Chinese NER as a tagging task on a sequence of known words. We have also developed a two-stage NER system for Chinese, which consists of two major modules: a segmenter using known-word bigrams and a tagger using lexicalized HMMs. In this way, both the internal entity formation clues and the surrounding contextual information, in particular the contextual lexical information, are explored and combined to recognize different types of NEs in Chinese documents. The experimental results on different public corpora show that the NER performance can be significantly enhanced using lexicalization techniques. The results also indicate that character-level tagging (viz. the pure single-character word models) are comparable to and may even outperform known-word based tagging when a lexicalized method is applied.

While our system has achieved a promising performance, there is still much to be done to improve it. First, our current tagger is a purely statistical system; it will inevitably suffer from the problem of data sparseness, particularly in open-domain applications. Secondly, our system usually fails to yield correct results for some complicated NEs such as nested organization names. For future work, we intend to explore some domain-adaptive techniques and heuristic information to enhance our system.

7. ACKNOWLEDGMENTS

We would like to thank the Institute of Computational Linguistics, Peking University, for their lexicon and corpus, and the *U.S. National Institute of Standard and Technology* for their Mandarin named entity tag set and corpus. We also would like to thank the reviewers for their helpful and valuable comments.

8. REFERENCES

[1] Bikel, D. M. Bikel, Schwartz, R., and Weischedel, R.M. An algorithm that learns what's in a name. Machine Learning, 34, 1-3 (1999), 211-231.

[2] Zhou, G.D., and Su J. Named entity recognition using an HMM-based chunk tagger. in Proceedings of the 40th Annual Meeting of the Association for Compuational Linguistics (ACL-2002), Philadelphia, USA, 2002, 473-480.

[3] Cohen, W.W., and Sarawagi, S. Exploiting dictionaries in named entity extraction: Combining semi-Markov extraction processes and data integration methods. in Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA, USA, 2004, 89-98.

[4] Borthwick, A. A maximum entropy approach to named entity recognition. Ph.D. Thesis, New York University, 1999.

[5] Brill, E. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. Computational Linguistics, 21, 4 (1995), 543-565.

[6] Isozaki, H., and Kazawa, H. Efficient support vector classifiers for named entity recognition. in Proceedings of

the 19th International Conference on Computational Linguistics (COLING 2002), Taipei, 2002, 953-959.

[7] Nakagawa, T. Kudo, T., and Matsumoto, Y. Revision learning and its application to part-of-speech tagging. in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002), Philadelphia, USA, 2002, 497-504.

[8] Pla, F., and Molina, A: Improving part-of-speech tagging using lexicalized HMMs. Natural Language Engineering, 10, 2 (2004), 167-189.

[9] Lee, S.-Z., Tsujii, J., and Rim, H.-C. Lexicalized hidden Markov models for part-of-speech tagging. In Proceeding of The 18th Conference on Computational Linguistics (COLING 2000), Saarbruken, Germany, 2000, 481-487.

[10] Molina, A., and Pla, P. Shallow parsing using specialized HMMs. Journal of Machine Learning Research, 2 (2002), 595-613.

[11] Sun, J., Gao, J., Zhang, L., Zhou, M., and Huang, C. Chinese named entity identification using class-based language model. in Proceedings of the 19th International Conference on Computational Linguistics (COLING-2002), Taipei, 2002, 967-973.

[12] Wu, Y., Zhao, J., and Xu, B. Chinese named entity recognition combining a statistical model with human knowledge. In Proceedings of the ACL 2003 Workshop on Multilingual and Mixed-language Named Entity Recognition, Sapporo, Japan, 2003, 65-72.

[13] Zhang, H.-P., Liu, Q., Yu, H.-K., Cheng, Y.-Q., and Bai, S. Chinese named entity recognition using role model. Computational Linguistics and Chinese Language Processing, 8, 2 (2003), 29-60.

[14] Chen, J., Xue, N., and Palmer, M. Using smoothing maximum entropy model for Chinese nominal entity tagging. In Proceedings of the First International Joint Conference on Natural Language Processing, Sanya, Hainan Island, China, 2004, 123-128.

[15] Guo, H., Jiang, J., Hu, G., and Zhang, T. Chinese named entity recognition based on multilevel linguistic features. In Proceedings of the First International Joint Conference on Natural Language Processing, Sanya, Hainan Island, China, 2004, 294-301.

[16] Yu, S., Duan, H., Zhu, S., Swen, B., and Chang, B. Specification for corpus processing at Peking University: Word segmentation, POS tagging and phonetic notation. Journal of Chinese Language and Computing, 13, 2 (2003), 121-158.

[17] Fu, G., and Luke, K.-K. Chinese unknown word identification using class-based LM. Lecture Notes in Computer Science (IJCNLP 2004), 3248 (2005), 704-713.

[18] Klein, D., Smarr, J., Nguyen, H., and Manning, C.D. Named entity recognition with character-level models. in Proceedings of the Seventh Conference on Natural Language Learning (CoNLL-2003), Edmonton, Canada, 2003, 180-183.

[19] Jing, H., Florian, R., Luo, X., Zhang, T., Ittycheriah, A. How to get a Chinese name (entity): Segmentation and combination issues. in Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP-2003), Sapporo, Japan, 2003, 200-207.

[20] Xue, N. Chinese word segmentation as character tagging. Computational Linguistics and Chinese Language Processing, 8, 1 (2003), 29-48.

[21] Liang, N. CDWS --- A written Chinese automatic word segmentation system. Journal of Chinese Information Processing, 1, 2 (1987), 44-52.

[22] Yu, S., Bai, S., and Wu, P. Description of the Kent Ridge Digital Labs system used for MUC-7. in The Seventh Message Understanding Conference Proceedings (MUC-7), Washington, D.C., USA, 1998,

[23] Chen, H.-H., Ding, Y.-W., Tsai, S.-C., and Bian, G.-W. Description of the NTU system used for MET2. in the Seventh Message Understanding Conference Proceedings (MUC-7), Washington, D.C., USA, 1998,

ABOUT THE AUTHORS:

Guohong Fu is now a post-doctoral fellow in the Department of Linguistics, the University of Hong Kong. He received his Ph.D. degree in computer science from Harbin Institute of Technology. His research interests mainly include Chinese information processing, natural language processing, and machine learning.

Kang-Kwong Luke is now a senior lecturer and serves as the head of the Department of Linguistics at the University of Hong Kong. He received his Ph.D. degree from the York University in 1988. His current research interests include Chinese grammar, Chinese dialects, discourse and conversation, computational linguistics, corpus linguistics, etc.