Sun X, Huang DG, Song HY *et al.* Chinese new word identification: a latent discriminative model with global features. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 26(1): 14–24 Jan. 2011. DOI 10.1007/s11390-011-1107-x

Chinese New Word Identification: A Latent Discriminative Model with Global Features

Xiao Sun[1] (孙 晓), De-Gen Huang[2] (黄德根), *Senior Member*, *CCF*, Hai-Yu Song[1] (宋海玉) and Fu-Ji Ren[3] (任福继), *Member*, *IEEE*

[1] *School of Computer Science and Engineering, Dalian Nationalities University, Dalian 116600, China*

[2] *School of Computer Science and Engineering, Dalian University of Technology, Dalian 116024, China*

[3] *Department of Information Science and Intelligent Systems, Tokushima University, Tokushima 7708506, Japan*

E-mail: sunxiao@dlnu.edu.cn; huangdg@dlut.edu.cn; shy@dlnu.edu.cn; ren@is.tokushima-u.ac.jp

Received June 19, 2009; revised December 14, 2010.

Abstract Chinese new words are particularly problematic in Chinese natural language processing. With the fast development of Internet and information explosion, it is impossible to get a complete system lexicon for applications in Chinese natural language processing, as new words out of dictionaries are always being created. The procedure of new words identification and POS tagging are usually separated and the features of lexical information cannot be fully used. A latent discriminative model, which combines the strengths of Latent Dynamic Conditional Random Field (LDCRF) and semi-CRF, is proposed to detect new words together with their POS synchronously regardless of the types of new words from Chinese text without being pre-segmented. Unlike semi-CRF, in proposed latent discriminative model, LDCRF is applied to generate candidate entities, which accelerates the training speed and decreases the computational cost. The complexity of proposed hidden semi-CRF could be further adjusted by tuning the number of hidden variables and the number of candidate entities from the Nbest outputs of LDCRF model. A new-word-generating framework is proposed for model training and testing, under which the definitions and distributions of new words conform to the ones in real text. The global feature called "Global Fragment Features" for new word identification is adopted. We tested our model on the corpus from SIGHAN-6. Experimental results show that the proposed method is capable of detecting even low frequency new words together with their POS tags with satisfactory results. The proposed model performs competitively with the state-of-the-art models.

Keywords new word identification, new words POS tagging, conditional random fields, hidden semi-CRF, global fragment features

1 Introduction

In Chinese natural language processing, the occurrences of new words or the so-called unknown word identification have made this task more difficult. New words, also called Out-Of-Vocabulary (OOV) words or unknown words, are main holdbacks in Chinese natural language processing. New words cannot be segmented correctly as they are not found in the existing system basic lexicon[1-3]. With the fast development of Internet and information explosion, even the largest lexicon that we may think, will not be capable of registering all geographical names, person names, organization names, technical terms, etc. All possibilities of derivational morphology cannot be foreseen in the form of a lexicon with a fixed number of entries. Therefore, new words are sure to appear in real world applications. New words usually cause some segment fragments in Chinese word segmentation, which is the basic step in Chinese natural language processing. Recent research reported that about 60% errors in Chinese word segmentation were caused by new words[4]. These new word related errors reduce the overall precision of the system. The problems caused by existence of new words must be solved in order to increase the effectiveness of Chinese natural language processing systems. Although definitions of new words in Chinese text are not very clear, there are still some specific characteristics for new words. First, new words are generated according to basic lexicon of the system; second, new words appear in a certain period of time under specific circumstances; third, new words basically obey existing morphological rules. Furthermore, distribution of new words' POS disperses widely. The POS of new words include not only geographical names, person names and organization names, but also normal nouns, normal verbs and

Regular Paper

This work is partially supported by the Doctor Startup Fund of Liaoning Province under Grant No.20101021.

even some adjective words. However, some statistical laws for distribution of new words' POS tags can still apply. Proper machine learning methods are possible for Chinese new word identification and POS tagging in order to increase precision of Chinese word segmentation and other tasks in Chinese natural language processing.

Researchers have studied some methods to detect new words in Chinese text. Zheng[5] detects new words totally based on rules. They used the knowledge of new word constructions to build some common rule bases. Some special rule bases are also built according to the constructions of new words from the Internet. All the rules are adopted to filter the candidate strings to find new words. Yet it is hard to summarize all rules for new words as new rules appear all the time. The system based on some rules of a certain period of time will soon be out of date as time goes by. Yan[6] also proposed a rule-based method to mine Chinese new words from dynamic current corpus, and provided a means of new word identification using a modified VSM (Vector Space Model) method and new word judging based on the dynamic current corpus. The features used by VSM are represented by 1 or 0, which cannot separate the complex new words, and it is easy to introduce some noises. Chen[7] presented a primarily data-driven Chinese word segmentation system. The system consists of a new word recognizer, a base segmentation algorithm, and procedures for combining single characters, suffixes, and checking segmentation consistencies. New words recognition, combining single characters, and checking consistencies contributed the most to the improvement in precision and recall over the performance of the baseline segmentation algorithm. There are some limitations because only new words with two characters are considered. Wu[8] proposed a mechanism of new word identification in Chinese text where probabilities are used to filter candidate character strings and to assign POS tags to the selected strings in a ruled-based system. This mechanism avoids the sparse data problem of pure statistical approaches and the over-generation problem of rule-based approaches. It improves parser coverage and provides a tool for the lexical acquisition of new words. The method did not adopt some specific feature of new words. Only the new words with 2 to 4 characters can be recognized. Zou[9] presented a method for detecting new words automatically through analyzing webpages grabbed from the Internet, a large set of words and strings is built, from which new words are detected and filtered by rules. At last, new words which exist in the grabbed webpages are extracted. The system built in this way can find new words in any length and in any field. They adopted the construction features and time features of new words, but the experimental result was not good. Peng[10] regarded the process of Chinese word segmentation and new word identification as a unified step using the Conditional Random Field (CRF) model[11-12], the method only detects new words and does not assign POS tags to new words. Peng[10] proved that the character-based model performs better than the word-based model in new word identification. Li[13] presented a study of new word identification (NWI) to improve the performance of a Chinese word segmenter. In their method the distribution and types of new words are discussed empirically. In particular, they focused on new words of two surface patterns, which account for more than 80% of new words in their datasets: NW11 (two-character new word) and NW21 (a bi-character word followed with a single character). NWI is defined as a problem of binary classification. A statistical learning approach based on an SVM classifier is used. Different features for NWI are explored, including in-word probability of a character (IWP), the analogy between new words and lexicon words, anti-word list, and frequency in documents. The experiments show that these features are useful for NWI. The F-scores of NWI they achieved are 64.4% and 54.7% for NW11 and NW21, respectively. The constructions of new words are limited in their paper, which cannot be applied in real system. Asahara[14] introduced a character-based chunking for Japanese unknown word identification in Japanese text. A major advantage of this method is the ability to detect low frequency unknown words of unrestricted character type patterns. The method is built upon SVM-based chunking, using character n-gram and surrounding context of Nbest word segmentation candidates from statistical morphological analysis as features. Goh[15] proposed a hierarchical model, with multiple classifiers (same model but different feature sets and parameters) for the identification. They created one classifier for each unknown word type: numbers, time nouns, person names and others. The experimental results show that their model can get higher precision (89%) compared to that using only one classifier (86%), but the recall of new words was not satisfied. Goh[16] proposed a unified solution to detect unknown words in Chinese texts regardless of the word types such as compound words, abbreviation, person names. First, POS tagging is conducted in order to obtain an initial segmentation and POS tags for known words. Next, segmentation output from the POS tagging, which is word-based, is converted into character-based features. Finally, unknown words are detected by chunking sequences of characters. By combining the detected unknown words with the initial segmentation,

they obtained the final segmentation. They also proposed a method for guessing the part-of-speech tags of the detected unknown words using contextual and internal component features. With unknown word processing, they have improved the accuracy of Chinese word segmentation and POS tagging.

Yet there are still some limitations in all these methods. First, new word identification and POS tagging are regarded as two separate steps, which bring out the facts that the lexical features information cannot be fully considered and used. Second, these methods have not proposed a proper framework of building reasonable size of basic lexicon and new word corpus for training and testing. The number of new word is depending on the size of basic lexicon adopted in certain system. Certainly, the larger the lexicon is, the less new word occurrence in texts. One can create a lexicon from all the tagged corpus, but that will not be a proper lexicon. Furthermore, if all words in tagged corpus are used to create the lexicon, then there will be no new words in the texts for training and testing. Therefore, it is important to define the meaning of new words properly and to propose a reasonable framework for model training and testing. In previous work[1], those words that occur only once in the corpus are treated as new words in their experiment. However, some people argue that this is not really true because even low frequency words are actually words in some dictionaries but those person names even with high frequency could not be found in a lexicon. A more natural way is by building a proper basic lexicon. We can consider those words that are not in a proper basic lexicon to be new words. In this case, some words in the corpus are not found in the basic lexicon and can be marked as new words in training data for new word identification[3,15]. A new-words-generating framework is proposed for model training and testing. The characteristics of new words in the proposed framework obey the rules of new words in real text on Internet or other corpus. Hidden semi-CRF trained under such framework is flexible and could be used to detect new word in widely kinds of fields.

In order to detect new words and assign POS tags to them synchronously, here we proposed a hidden semi-CRF model, which combines the LDCRF[17-19] and the semi-CRF model[20-21]. Hidden semi-CRF model thus combines the strength of LDCRF, which could capture both extrinsic dynamics and intrinsic sub-structure, with the strength of semi-CRF, which could attach labels to the subsequences of a sentence, rather than to the tokens. The LDCRF model generates the Nbest outputs of new word boundaries, which are combined with candidate POS tags and adopted to build the candidate entities for hidden semi-CRF. In such a way, the scalability of the hidden semi-CRF is improved because the numbers of candidate entities for training and testing are significantly reduced by introducing the Nbest outputs from the LDCRF model. We could adjust the Nbest outputs to tune the degrees of pruning candidate entities. Hidden semi-CRF could detect new words together with their POS tags synchronously. The contextual word-level information and character-level information for new words could be fully used. In addition, the global information (or global features) for new words called "Global Fragment Features" (GFF) is proposed and adopted, which could obviously increase the precision of new word identification. Hidden semi-CRF costs less in computation complexity than the semi-CRF because the candidate entities are adopted from the Nbest outputs from the LDCRF, which could be further adjusted. The computation cost of unnecessary POS tagging for incorrect candidate words are avoided.

2 Hidden Semi-CRF Model

2.1 Introduction to Hidden Semi-CRF

For a given sequence X that includes new words, since both the boundaries and the POS tags of the new words are unknown, we need to segment the input sequence x (assigning BIO tags) as well as assign POS tags to the segments with new words and assigning "O" to the segments without new words. There are too many candidate segments for the sequence and candidate POS tags for the segments. If we directly adopt the semi-CRF or LDCRF to detect new words assign POS tags, the computation cost is very high. For the semi-CRF model, all the candidate segments with all candidate POS tags have to be enumerated in model training and testing[20]. Furthermore, in semi-CRF, a reasonable value of L (upper bound length of entities) has to be set for different tasks[20]. However, in the tasks of new word identification, the length of new words in the sequence might be longer than the fixed L, thus longer words cannot be detected correctly. We extended the semi-CRF method and inserted LDCRF to generate the candidate entities for semi-CRF in its model training and testing. In such a way, with the strength of the LDCRF model, we do not have to limit L in the semi-CRF model. The LDCRF approach effectively learns the substructure of an input sequence X, and outputs the boundaries of new words for the input sequence X. Furthermore, the advantage of LDCRF is that LDCRF can output Nbest label sequences and their probabilities using efficient marginalization operations[18]. We use this characteristic and combine

Nbest outputs (candidate new words) from LDCRF with possible tags to build candidate entities for hidden semi-CRF in model training and testing. We can adjust the number of the Nbest outputs from the LDCRF to control computation cost and precision of hidden semi-CRF.

In order to build the hidden semi-CRF model, we follow the original definition[20]. Let $X = \{x_0, x_1, \ldots, x_i \ldots\}$ $(0 \leqslant i \leqslant |X|)$ denote a sequence of Chinese characters that includes new words to be detected. Let $Y = \{y_0, y_1, \ldots, y_j, \ldots\}$ $(0 \leqslant i \leqslant |Y|)$ denote the output label sequence. Let $s = (s_1, \ldots, s_j, \ldots)$ denote a segmentation of X, where a segment $s_j = (t_j, u_j, y_j)$ consists of a start position t_j, an end position u_j, and a label y_j. Conceptually, a segment means that the tag y_j is given to all x_i's between t_j and u_j, inclusive. In the tasks of new word identification, this means that all the characters in a new word share the same POS tag, each of the characters out of new words has the tag "O". We assume that segments have a positive length bounded above by the pre-defined upper bound L $(1 \leqslant t_j \leqslant u_j \leqslant |s|, u_j - t_j + 1 \leqslant L)$ and completely cover the sequence x without overlapping, that is, s satisfies $t_1 = 1$, $u_{|s|} = |x|$ and $t_{j+1} = u_j + 1$ for $j = 1, \ldots, |s| - 1$. For new word identification and POS tagging, a correct segmentation of sentence "在寻找锡安的过程中 (In the process of looking for Zion)" might be s = ((0, 1, O), (2, 3, O), (4, 5, O), (6, 9, n), (10, 11, O), (12, 13, O), (13, 14, O), (15, 16, O)). We also make a restriction on the features, analogous to the usual Markovian assumption made in CRFs, and original semi-CRFs define a conditional probability of a state sequence y given an observation sequence x by:

$$p(y|x, \lambda) = \frac{1}{Z(x)} \exp\Big(\sum_j \sum_i \lambda_i f_i(s_j)\Big) \qquad (1)$$

where $f_i(y_{j-1}, y_j, x, t_j, u_j)$ is a feature function, s_j is the j-th segment in s and $Z(x)$ is the normalization factor as defined for semi-CRF, $Z(x) = \exp\big(\sum_{s(x)} \sum_j \sum_i \lambda_i f_i(y_{j-1}, y_j, x, t_j, u_j)\big)$, $s(x)$ is normalization factor that denotes all the candidate segments of x. From (1), we can see that in order to get original semi-CRF work for new word identification and POS tagging, we have to enumerate all the candidate segments with different lengths for every x and enumerate all POS tags for each candidate segment. These candidate entities made the inference of semi-CRF very expensive. So we adopted LDCRF and used Nbest output of LDCRF to generate the candidate entities for hidden semi-CRF.

We follow the original definition of LDCRF. Let the input of the LDCRF be the sequence of Chinese characters. LDCRF outputs "BIO" boundary tags for the input sequence to mark new words (or characters out of new words). Let $Path_{NBEST} = \{path^1, \ldots, path^{NBEST}\}$ denote Nbest output of LDCRF for input sequence x. *NBEST* is a predefined const denoting the number of Nbest output paths. In the task of new word identification and POS tagging, LDCRF is adopted to generate all the candidate new words first, and then candidate POS tags are assigned to candidate new words and tag "O" is assigned to single character out of new words to build candidate entities for hidden semi-CRF. Let $s_{NBEST}(x)$ denote all the candidate entities generated from LDCRF for x. We replace $s(x)$ in (1) with $s_{NBEST}(x)$ to get hidden semi-CRF:

$$\begin{aligned} & p(y|x, \lambda) \\ = & \frac{1}{\exp\Big(\sum\limits_{s_{NBEST}(x)} \sum\limits_j \sum\limits_i \lambda_i f_i(y_{j-1}, y_j, x, t_j, u_j)\Big)} \times \\ & \exp\Big(\sum_j \sum_i \lambda_i f_i(y_{j-1}, y_j, x, t_j, u_j)\Big) \end{aligned} \qquad (2)$$

where $f_i(y_{j-1}, y_j, x, t_j, u_j)$ is a feature function. We can see that through adjusting *NBEST*, we could tune the complexity of hidden semi-CRF. If we set *NBEST* to 1, then hidden semi-CRF shrinks into a two-layer linear-chain CRF. If we set *NBEST* large enough the complexity of hidden semi-CRF is still lower than semi-CRF as hidden semi-CRF does not have to enumerate all the candidate segments with different lengths. In hidden semi-CRF, obviously we do not have to limit the upper bound length of entity L.

2.2 Inference Algorithm for Hidden Semi-CRF

We revised the inference algorithm from the original semi-CRF[20] for hidden semi-CRF. The inference algorithm for the hidden semi-CRF is described as follows.

First, given the input character sequence x, we use the LDCRF model to estimate the most probable label sequence y^* (new words boundaries sequence) that maximizes the conditional model:

$$y^* = \arg\max_y P(y|x, \theta) \qquad (3)$$

where parameter values θ are learned from training examples. Assuming each class label is associated with a disjoint set of hidden states, the previous equation can be rewritten as:

$$y^* = \arg\max_y \sum_{h:\forall h_i \in H_{y_i}} P(h|x, \theta). \qquad (4)$$

To estimate label y_j^* of frame j, marginal probabilities $P(h_j = a|x, \theta)$ are computed for all possible hidden states $a \in H$. Then marginal probabilities H_{y_j} and the

label associated with the optimal set is chosen. In order to generate candidate entities for the hidden semi-CRF, first, we apply LDCRF to output Nbest label sequence for new words boundaries and their probabilities. The candidate entities for hidden semi-CRF are the segments with labels, but Nbest results from LDCRF only include the information of all candidate segments for new words without proper labels (POS tags). So we assign the possible POS tags to candidate new words to set up candidate entities for the hidden semi-CRF. Other segments (one Chinese character in each segment), which are not candidate new words, are labeled with "O". The candidate entities set generated by Nbest output of LDCRF is $S_{NBEST}(x)$, which includes all the candidate new word boundary sequences together with proper tags (POS tags or "O").

The inference algorithm for hidden semi-CRF is to get the final result for the following equation $\arg\max_{s\in S_{NBEST}(x)} P(s|x,\lambda)$. We use $F(x,s)$ to denote $\sum_j \sum_i f_i(s_j)$, use $\boldsymbol{\lambda}$ to denote weight vector for $F(x,s)$ and use $\boldsymbol{f}(s_j)$ to denote $\sum_i f_i(s_j)$, so that the former equation $\arg\max_{s\in S_{NBEST}(x)} P(s|x,\lambda)$ can be rewritten into:

$$\arg \max_{s\in S_{NBEST}(x)} \boldsymbol{\lambda} * F(x,s) = \arg \max_{s\in S_{NBEST}(x)} \boldsymbol{\lambda} \cdot \sum_j \boldsymbol{f}(s_j). \tag{5}$$

We do not have to set limitation for L, which denotes upper bound on segment length, so let $S^i(x)$ denote the set of all the partial segmentation with index starting from 1 to i. Let $V(i,y)$ denote the largest value of $F(x,s^i)$ for any segmentation $s^i \in S^i(x)$. Let $S^i_{end(k)}(x)$ denote the set of all the segments in $S^i(x)$ with the end index k $(0 \leqslant k \leqslant i)$ and $s^i_{end(k)} \in S^i_{end(k)}(x)$ is a number, which is a segment with the end index k. The recursive calculation for hidden semi-CRF can be defined as:

$$V(i,y) = \begin{cases} \max\limits_{y', s^i_{end(i)} \in S^i_{end(i)}(x)} \{V(i-len, y') + \\ \quad \boldsymbol{\lambda} \cdot \boldsymbol{f}(y, y', x, i-len, i)\}, & \text{if } (i>0), \\ 0, & \text{if } (i=0), \\ -\infty, & \text{if } (i<0), \end{cases} \tag{6}$$

where $|s^i_{end(i)}|$ denotes the length of the segment $s^i_{end(i)}$. The best segmentation then corresponds to the path traced by $\max_y V(|x|, y)$.

2.3 Parameter Estimation for Hidden Semi-CRF

We revised the original learning algorithm token from semi-CRF[20] for parameter estimation in hidden semi-CRF.

First, we train the LDCRF model in order to generate candidate entities for training hidden semi-CRF. For LDCRF model we use the following objective function to learn parameter θ:

$$L(\theta) = \sum_{i=1}^{n} \log P(y_i|x_i,\theta) - \frac{1}{2\sigma^2}\|\theta\|^2. \tag{7}$$

The first term in the equation is conditional log-likelihood of training data. The second term is the log of a Gaussian prior with variance σ^2[17].

We here adopted Limited-memory BFGS method (L-BFGS) to estimate the parameter. L-BFGS algorithm is currently the most effective optimization method for CRF parameter estimation[22]. As the parameter θ of LDCRF is already known, so it is possible for LDCRF to generate the Nbest results (boundaries for all candidate new words) from the input sequence to build candidate entities for training hidden semi-CRF. We generate candidate entities for training hidden semi-CRF in the same way as we did in inference algorithm for hidden semi-CRF: adding possible tags to the segments in Nbest results from LDCRF. The candidate entities set generated from Nbest output of LDCRF is $S_{NBEST}(x)$, which includes all the candidate new words together with their POS tags. For original semi-CRF, over a given training set $T = \{(x_l, s_l)\}_{l=1}^N$, we express log-likelihood over the training sequences as:

$$\begin{aligned} L(\lambda) &= \sum_l \log P(s_l|x_l,\lambda) \\ &= \sum_l \{F(x_l,s_l) - \log Z_\lambda(x_l)\}. \end{aligned} \tag{8}$$

In hidden semi-CRF, for a given input sequence x_l, we use $S_{NBEST}(x_l)$ to replace the set of all candidate entities, which means the candidate entities generated from Nbest result of x_l. So the equation can be rewritten as:

$$\begin{aligned} \Delta L(\lambda) &= \sum_l F(x_l,s_l) - \\ &\quad \frac{\sum\limits_{s_{NBEST}(x_l)} F(s_{NBEST}(x_l),x_l) e^{\boldsymbol{\lambda}\cdot F(s_{NBEST}(x_l),x_l)}}{Z_\lambda(x_l)} \\ &= \sum_l F(x_l,s_l) - E_{P(s_{NBEST}(x_l)|\boldsymbol{\lambda})} F(x_l, s_{NBEST}(x_l)) \end{aligned} \tag{9}$$

From (9) we can see that we only have to consider the candidate entities generated by Nbest result of LDCRF, which obviously reduce computation cost for hidden semi-CRF.

3 New-Word-Generating Framework

New words are words that do not exist in system lexicon, so it is difficult to regenerate new words for

training models. Some researchers have proposed some methods that regarded the low frequency words in the training corpus as new words[1], but the distribution and characteristic of new words under such frameworks may not confirm to the new words existing in real text. In real text, sometimes new words appear more times than known words do in specific texts. We here proposed a framework for new word training and testing. The generation of new words in real text has two factors, the first one is a system basic lexicon with proper scale, and the second is that as the time goes, newly coming texts include words that do not appear in the system basic lexicon. We analyzed the PKU corpus[23], which include all the news text of People's Daily in year 2000 classified and separated by the months to build basic lexicon of proper size for new word generation. Several consequent months of PKU corpus are used to build the basic lexicon, and then the following one month corpus is used as the corpus with new words, which are also called new word corpus. The characteristic of new words generates under such a framework is in accordance with new words in real text, so the proposed new-word-generating framework has the expansibility to adopt sundry new words.

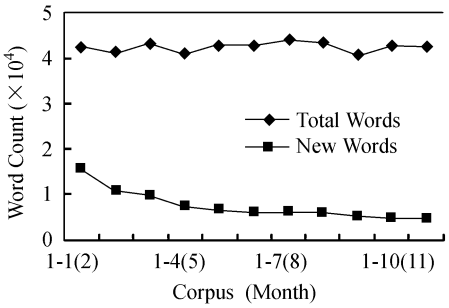


Fig.1. New words in corpus.

In order to generate the proper size of basic lexicon, we count the number of the words in all the possible consequent of corpus and new words in the following one month. The result is shown in Fig.1. In Fig.1, the horizontal axis 1-1(2) means that month of basic corpus is January, and the month of corpus including new words is February. 1-4(5) means the basic corpus is built from January to April, and the month of corpus including new words is May. The upper line means the number of total words in the corpus that includes new words. The lower line means the number of new words in the corpus that includes new words. It can be seen from Fig.1 that after half year, the percentage of new words is becoming steady. The POSs distribution of new words in June and July is shown in Table 1 (only top 10 POS tags are listed). The POS of new words is directly token from the PKU corpus and the definition of POS can be referred to [23]. We can see from Table 1 that the distribution of the POS has some statistical laws, such as the most POS tags of new words are noun(n), the top 10 of POS tags in Table 1 are almost the same. This means that the distribution and the characteristic of new words under the proposed framework are steady.

Table 1. Distribution of POSs of New Words

June			July		
POS	Count	P (%)	POS	Count	P (%)
n	4 000	41.718 8	n	4 298	46.096 1
m	1 619	16.885 7	m	1 726	18.511 4
nz	1 111	11.587 4	nz	1 106	11.861 9
ns	521	5.433 9	j	519	5.566 3
j	451	4.703 8	v	361	3.871 7
v	441	4.599 5	l	210	2.252 3
nr	353	3.681 7	i	184	1.973 4
l	210	2.190 2	nr	178	1.909 1
i	194	2.023 4	ns	113	1.211 9
t	139	1.449 7	t	107	1.147 6

We finally use the corpus from January to June to build the basic lexicon and adopt the rest corpus as new words corpus (such as the corpus of July), because after half year, both the percentage of new words and the distribution of the POS of new words are becoming steady. The basic lexicon contains 94 849 entries. Based on this basic lexicon, there are about 14.43% new words in the corpus of July, which disperse evenly in the training and testing data.

4 Features for New Word Identification and POS Tagging

4.1 Features and Templates

First, we need to define the features and templates for training and testing LDCRF model. The PKU corpus of July is divided into 80% for training and 20% for testing. Take the sentence "好享来/nz 中文/nz 网/n" (HaoXiangLai Chinese Net) in the training corpus for example, the "好享来 (HaoXiangLai)" is a new word according to the basic lexicon. To train the LDCRF model, we used the 5-tag label set described in Table 2 as the boundary labels.

Table 2. 5-Tag Label Set for Word Boundaries

Labels	Description
B	The beginning character in a word
I	Internal character in a word with more than two characters
E	The ending character in a new word
S	Single character of a new word
O	Other character in the known words

For example, "好" is labeled with "B", which means it is the beginning character in the new word; "享" is labeled with "I", which means it is the internal character in the new word; "来" labeled with "E", which means it is the ending character in the new word. The training corpus re-labeled with the 5-tag labels are adopted to train the LDCRF model. The templates of the features for the LDCRF model are listed in Table 3.

Table 3. Template of Features for LDCRF

Type	Feature	Description
Unigram	C_{-1}, C_0, C_1	Single character
Bigram	$C_{-1}C_0$, C_0C_1	The combination of two characters
Trigram	$C_{-1}C_0C_1$	The combination of three characters
GFF	$G(C_0)$	The global fragment feature of C_0
Style	$S(C_0)$	The predefined classes for the character
Seg	$M(C_0)$	The HMM segmenter for the character
Basic Dic	$B(C_0)$, $B(C_{-1}C_0)$, $B(C_0C_1)$	Whether C_0, $C_{-1}C_0$, C_0C_1 exist in words of basic dictionary or not
UW Dic	$N(C_0)$, $N(C_{-1}C_0)$, $N(C_0C_1)$	Whether C_0, $C_{-1}C_0$, C_0C_1 exist in words of UW dictionary or not

GFF is short for "Global Fragment Feature". The predefined $S(C_0)$ for the characters are five classes: Class 1 represents numbers; Class 2 represents English letters; Class 3 represents punctuations; Class 4 represents Chinese characters; Class 5 represents other characters. The basic lextion is built by the corpus from January to June. The new lexicon includes all the new words in the training corpus (we also collected some new words from the Internet). HMM (Hidden Markov Model) segmenter is built from basic lexicon using Forward Maximum Matching (FMM) method. The LDCRF is character-based, but we also imported the information from the outer lexicon as features from "Basic Dic" and "New Dic" template. The $B(C_{-1}C_0)$ means whether the character sequence $C_{-1}C_0$ exists in some word of the basic lexicon. New word lexicon is very useful for detecting new words like personal name, because the set of last name for a person is limited. The "New Dic" template is built according to new word lexicon.

In a hidden semi-CRF learner, features are no longer applied to individual words, but applied to the segment with words and POS tags. This makes it somewhat more natural to define new features, as well as providing more context[20]. Supposed that the current segment is S_0, the word in S_0 is W_0 (the characters in W_0 is C_n) and the POS tag is P_0. In Table 4, the templates and features for the hidden semi-CRF are listed.

Table 4. Templates for the Hidden Semi-CRF

Feature	Description
W_0, $G(W_0)$	Unigram features for current segment. $G(W_0)$ is GFF.
$N(W_0(C_n))$ $(1 \leqslant n \leqslant \lvert W_0\rvert)$ $N(W_0(C_0C_1))$ $N(W_0(C_0C_1C_2))$	Whether character sequence C_n, C_0C_1, $C_0C_1C_2$ in segment exist in new words dictionary or not. If so, this template outputs words and POS tags from new word dictionary
W_{-1}/W_0, W_0/W_1	The combination of two words in two segments.
$W_{-2}W_{-1}W_0$ $W_{-1}W_0W_1$ $W_0W_1W_2$	The combination of three words in three segments.
$L(W_0)$	The length of a word in current segment

4.2 Global Fragment Features for New Words

For new words, global information in context is important for identification. Certain new words appear considerable times in certain context. Although they cannot be segmented correctly, the fragments generated by new words have some disciplines that can be counted statistically. Take the new phrase "正龙拍虎 (Zhenglong took photos of tigers)" for example, after segmentation by FMM (Forward Maximum Matching) segmenter, the possible fragments for the phrase could be "正(right)/龙(dragon)/拍(pat)/虎(tiger)". The joint possibilities inside the fragments or the possibilities between the fragments and known words are lower than normal. According to this property, we could find the fragments and their counts from the Chinese text, and then use these as features in model training for new word identification. For example, "正龙拍虎" appears

Table 5. Global Fragment Features for New Word Identification and POS Tagging

GFF for LDCRF: $G(C_0)$			
C_0	Count	Length	Position
正	10	4	0
龙	10	4	1
拍	10	4	2
虎	10	4	3
GFF for Hidden Semi-CRF: $G(W_0)$			
W_0	Count	Length	Position
正龙 (Zhenglong)	10	4	0
拍 (take)	10	4	2
虎 (tiger)	10	4	3

10 times in a certain Chinese news, which are all segmented into "正/龙/拍/虎" after segmentation. We could find that "正/龙/拍/虎" is a fragment, which has the following global features: the fragment appears 10 times in context and the length of the fragment is 4 Chinese characters. These global features could be used as features for LDCRF and hidden semi-CRF, which are listed in Table 5. In Table 5, the position means the start position of the character or word in the fragment.

5 Experimental Results

5.1 Experiments on PKU Corpus

In order to get proper hidden states for LDCRF and Nbest variable for hidden semi-CRF in training and testing, we first performed the cross-validation by using 20% test corpus of July and test the overall F-score of the new word identification and POS tagging. We finally set the number of hidden states to 4 and the Nbest to 30 for LDCRF and hidden semi-CRF. In the following experiments, we will use the fixed hidden states number and Nbest number. We first tested the model on the PKU corpus using 20% test corpus of July. We evaluated the recall of new word identification by their POS tags. The results are shown in Fig.2.

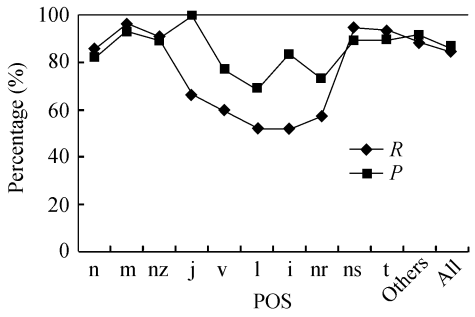


Fig.2. Distribution of detected new words by POS tags.

As there is no single standard definition for words (or new words) in Chinese, we could hardly say that the gold data is perfectly correct. Therefore, human judgment is necessary. Since there are not so many incorrectly detected new words, we have gone through all the errors to examine what kind of mistakes has been made. Surprisingly, there are quite a number of words in the error list which are said to be acceptable by human judgment. There are some abbreviations and new words in specific fields which cannot be detected and given the POS correctly. The POS tagging is not quite satisfied because we only apply hidden semi-CRF model to guess the POS tags for new words, the information of POS of the known word could not be used. We will apply hidden semi-CRF to detecting known words together with new words so that the lexical information of known words can be fully used. The precision of new words identification can be further increased. In order to test the effectiveness of our proposed "Global Fragment Features", we deleted GFF from hidden semi-CRF model (No GFF) and compare the result with the model with Global Fragment Features (with GFF). The results are listed in Table 6.

Table 6. Effectiveness of Global Fragment Features

	R (%)	P (%)	F (%)
Without GFF	85.83	86.62	86.22
With GFF	86.55	87.98	87.26

5.2 Comparisons with Other Models

In order to compare the hidden semi-CRF model with other models, we built five other Chinese new word identification models, which are listed in Table 7. We first adopted the HMM model with some rules to build the new word identification and POS tagging tools, based on the methods proposed by [7] and [24], which can be treated as the baseline. The "SVM + ME" model is based on [16]. The "CRF + ME" model is partly based on [10], and the POS tagging is still based on the method proposed by [16]. The LDCRF and the semi-CRF model are also adopted to make sure that the hidden semi-CRF is better in the fields of new word identification and POS tagging. We adopted cross validation to get the optimum parameters for all these models.

Table 7. Comparison with Other Models

	R (%)	P (%)	F (%)	T_l	T_t
HMM + Rules	75.43	70.61	72.94	1.00	1.00
SVM + ME	71.32	89.11	79.23	4.12	2.38
CRF + ME	80.59	79.61	80.10	3.83	1.59
Semi-CRF	87.01	83.89	85.42	5.14	2.43
LDCRF	85.74	84.38	85.06	4.96	2.67
HSCRF	86.55	87.98	87.26	3.89	1.48

In Table 7, in order to compare the computational cost between models, we calculated the training (T_l) and testing (T_t) corpus of each method. In order to make the comparison clear, in the training time column (T_l), we set the time of model "HMM + Rules" to 1. In the testing time column (T_t), we also set the time of model "HMM + Rules" to 1. HSCRF is short for hidden semi-CRF.

From year 2003, a competition for Chinese word segmentation and other Chinese natural language processing tasks, such as POS tagging and Chinese Named Entity Recognition, was carried out in SIGHAN workshop to compare the accuracy of various methods[25-28]. The

score of all the tasks shows distinguished increase from year 2003 to 2008[28]. In the SIGHAN-6, seven corpora are provided for the evaluation. We only consider the simplified character corpora, which are CTB corpus, NCC corpus and SXU corpus. Take the CTB corpus for example, the F-score of the OOV (out-of-vocabulary) of the close test for CTB dataset are 51.05∼77.45% points and the recalls for OOV are 52.99∼77.30%. The F-score of the OOV (out-of-vocabulary) of the open test for CTB dataset are 96.54∼65.81% points and the recalls for OOV are 59.67∼96.85%. The high score in open track may due to the reason that the open track could adopt the possible training corpora as many as the participant want. We know that if we adopt more outer corpora, the score for the OOV could be promoted further. Therefore, the OOV score in open track is not comparable. In open track, we did not re-train our model with their training materials in open track, but just used what we have on hand to run on the testing data. We made the comparison with Rank 1 in each corpus. The result is listed in Table 8.

Table 8. Results on the SIGHAN 2008 Corpus

		R_{OOV} (%)	P_{OOV} (%)	F_{OOV} (%)
CTB	Open	96.85	96.23	96.54
	Close	77.70	77.61	77.45
Our	Open	89.81	87.48	88.63
CTB	Close	82.23	80.10	81.15
NCC	Open	88.93	88.67	88.80
	Close	61.79	59.84	60.80
Our	Open	86.75	85.00	85.87
NCC	Close	71.68	70.01	70.84
SXU	Open	78.25	84.15	81.09
	Close	74.29	71.59	72.92
Our	Open	81.78	82.32	82.05
SXU	Close	76.32	75.61	75.96

The standards for segmentation (segment granularity) are different in the three corpora, which will affect the results of segmentation and POS tagging. The model trained on the corpora with large segment granularity will perform badly on the corpora with small segment granularity. In the CTB corpus, the open track results presented in SIGHAN-6 are very high, this may caused by that the CTB corpus is widely adopted by the participants in their daily work and the CTB corpus is used by the former SIGHAN. The standards of the CTB are very common and well adopted. The participants could have adopted some other CTB corpus for the open track, which may cause the high score in the open track. The other corpora are the corpora that first appear in the SIGHAN, so the results of the open track and the close track are reasonable and comparable. Our model is better than the state-of-the-art models in new word identification. We get quite satisfactory precision by using the proposed method.

6 Conclusion and Future Work

In order to detect the Chinese new words with their POSs in real text or on the Internet, we proposed a hidden semi-CRF model, which combines the strength of the LDCRF model and the semi-CRF model. By importing the LDCRF model in the hidden semi-CRF model, we do not have to enumerate all the candidate entities as what the semi-CRF does. The proposed hidden semi-CRF adopts the candidate entities generated from the Nbest results of the LDCRF, which obviously decreases the computational cost in training and testing. By virtues of CRFs, a number of correlated features for hierarchical tag sets can be incorporated which was not possible in HMMs, and influences of label bias and length bias are minimized which caused errors in MEMMs. A new-word-generating framework is proposed here to build the basic lexicon and training/testing corpus for the hidden semi-CRF model. Under such a framework, new words for training are in accordance with the characteristic of the Chinese new words in real text, so with the framework it is easy to extend and detect new words in other fields. Some global features called "Global Fragment Features" are adopted in the model training and testing. The global fragment information for new words is a very useful feature for Chinese new word identification and POS tagging, which was adopted for training and testing the hidden semi-CRF model. There exist some phenomena which cannot be analyzed only with bi-gram features in new word identification. To improve accuracy, trigram or more general n-gram features would be useful. Hidden semi-CRF has capability of handling such features. We also need a practical feature selection which effectively trades between accuracy and efficiency. We will apply the proposed hidden semi-CRF model in Chinese word segmentation and POS tagging, in such a way more context information such as known words together with their POS tags for new words will be imported and the precision of new word identification and POS tagging can be further increased. As the precision of Chinese word segmentation and POS tagging could be improved by hidden semi-CRF model, machine translation, Chinese base phrase chunking and other high level natural language processing applications could also benefit from this.

References

[1] Goh C, Asahara M, Matsumoto Y. Chinese unknown word identification using character-based tagging and chunking. In

Proc. the 41st Annual Meeting on Association for Computational Linguistics, Sapporo, Japan, Jul. 7-12, 2003, pp.197-200.

[2] Nie J, Hannan M, Jin W. Unknown word detection and segmentation of Chinese using statistical and heuristic knowledge. *Communications of COLIPS*, 1995, 5(1): 47-57.

[3] Chen C, Bai M, Chen K. Category guessing for Chinese unknown words. In *Proc. the Natural Language Processing Pacific Rim Symposium*, Phuket, Thailand, Dec. 2-4, 1997, pp.35-40.

[4] Sproat R, Shih C, Gale W, Chang N. A stochastic finite-state word-segmentation algorithm for Chinese. *Computational Linguistics*, 1996, 22(2): 377-404.

[5] Zheng J H, Li W H. A study on automatic identification for Internet new words according to word-building rule. *Journal of Shanxi University* (*Natural Science Edition*), 2002, 25(2): 115-119. (In Chinese)

[6] Yan W. New words mining from the dynamic current corpus based on VSM. In *Proc. Dictionaries and Digital Symposium*, Yantai, China, Aug. 16-20, 2004. (In Chinese)

[7] Chen A. Chinese word segmentation using minimal linguistic knowledge. In *Proc. the Second SIGHAN Workshop on Chinese Language Processing*, Sapporo, Japan, Jul. 11-12, 2003, pp.148-151.

[8] Wu A D, Jiang Z X. Statistically-enhanced new word identification in a rule-based Chinese system. In *Proc. the Second Chinese Language Processing Workshop*, Hong Kong, China, Oct. 1-8, 2000, pp.46-51.

[9] Zou G., Liu Y., Liu Q. Internet-oriented Chinese New Words Detection (in Chinese). *Journal of Chinese Information Processing*, 2004, 18: 1-9.

[10] Peng F, Feng F, McCallum A. Chinese segmentation and new word detection using conditional random fields. In *Proc. the 20th International Conference on Computational Linguistics*, Geneva, Switzerland, Aug. 23-27, 2004, pp.562-569.

[11] Lafferty J, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. the 18th Int. Conf. Machine Learning*, Williamstown, USA, Jun. 28-Jul. 1, 2001, pp.282-289.

[12] Zhao H, Kit C. Scaling conditional random fields by one-against-the-other decomposition. *Journal of Computer Science and Technology*, July, 2008, 23(4): 612-619.

[13] Li H Q, Huang C N, Gao J F, Fan X Z. The use of SVM for Chinese new word identification. In *Proc. IJCNLP 2004*, Sanya, China, Mar. 22-24, 2004, pp.723-732.

[14] Asahara M, Matsumoto Y. Japanese unknown word identification by character-based chunking. In *Proc. the 20th International Conference on Computational Linguistics*, Geneva, Switzerland, Aug. 23-27, 2004, pp.459-465.

[15] Goh C L, Asahara M, Matsumoto Y. Training multi-classifiers for Chinese unknown word detection. *Journal of Chinese Language and Computing*, 2005, 15(1): 1-12.

[16] Goh G, Asahara M, Matsumoto Y. Machine learning-based methods to Chinese unknown word detection and POS tag guessing. *Journal of Chinese Language and Computing*, 2006, 16: 185-206.

[17] Morency L, Quattoni A, Darrell T. Latent-dynamic discriminative models for continuous gesture recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, USA, Jun. 17-22, 2007, pp.1-8.

[18] Sun X, Wang H, Wang B. Predicting Chinese abbreviations from definitions: An empirical learning approach using support vector regression. *Journal of Computer Science and Technology*, 2008, 23(4): 602-611.

[19] Sun X, Huang D, Ren F. Detecting new words from Chinese text using latent semi-CRF models. *IEICE Transactions on Information and Systems*, 2010, E93-D(6): 1386-1393.

[20] Sarawagi S, Cohen W. Semi-Markov conditional random fields for information extraction. In *Proc. NIPS 2004*, Vancouver, Canada, Dec. 13-18, 2004, pp.1185-1192.

[21] Okanohara D, Miyao Y, Tsuruoka Y, Tsujii J. Improving the scalability of semi-Markov conditional random fields for named entity recognition. In *Proc. the 21st Int. Conf. Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia, Jul. 17-21, 2006, pp.465-472.

[22] Liu D, Nocedal J. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 1989, 45(3): 503-528.

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

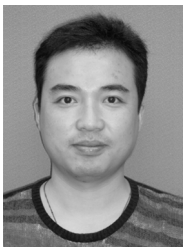
[24] Zhou G. A chunking strategy towards unknown word detection in Chinese word segmentation. In *Proc. IJCNLP 2005*, Jeju Island, Korea, Oct. 11-13, 2005, pp.530-541.

[25] Sproat R, Emerson T. The first international Chinese word segmentation bakeoff. In *Proc. the 2nd SIGHAN Workshop on Chinese Language Processing*, Sapporo, Japan, Jul. 11-12, 2003, pp.133-143.

[26] Emerson T. The second international Chinese word segmentation bakeoff. In *Proc. the 4th SIGHAN Workshop on Chinese Language Processing*, Jeju Island, Korea, Oct. 14-15, 2005, pp.123-133.

[27] Levow G A. The third international Chinese language processing bakeoff: Word segmentation and named entity recognition. In *Proc. the 5th SIGHAN Workshop on Chinese Language Processing*, Sydney, Australia, Jul. 22-23, 2006, pp.108-117.

[28] Jin G, Chen X. The fourth international Chinese language processing bakeoff: Chinese word segmentation, named entity recognition and Chinese POS tagging. In *Proc. Sixth SIGHAN Workshop on Chinese Language Processing*, Hyderabad, India, Jan. 11-12, 2008, pp.69-81.



Xiao Sun received the M.E. degree in 2004 from the Department of Computer Sciences and Engineering, Dalian University of Technology, Dalian, China. He is now working in School of Computer Science and Engineering, Dalian Nationalities University. He received his double-Ph.D. degree from Dalian University of Technology, China, and University of Tokushima in Japan. His research interests include natural language processing, machine translation, Chinese lexical analysis, and machine learning.

De-Gen Huang was born in 1965. He is a professor in the Dalian University of Technology. His main research interests include natural language processing, machine learning and machine translation. He is now working at the Department of Computer Science and Engineering, Dalian University of Technology. He is now a senior member of CCF, and an associate editor of Int. J. Advanced Intelligence.

Hai-Yu Song received the B.E. degree in computer and application in 1996, the M.E. degree in computer software and theory in 2003, both from Jilin University, China. Now he is a Ph.D. candidate in computer software and theory at Jilin University, and working in Dalian Nationalities University. His research interests include image analysis and understanding, image retrieval, data mining, and computer graphics.

Fu-Ji Ren received the B.E. degree in 1982 and M.E. degree in 1985 from the Department of Computer Sciences, Beijing University of Posts and Telecommunications, Beijing, China. He received the Ph.D. degree in 1991 from Faculty of Engineering, Hokkaido University, Japan. He worked at CSK, Japan, where he was a chief researcher of NLP. From 1994 to 2000, he was an associate professor. His research interests include natural language processing, machine translation, artificial intelligence, language understanding and communication.