IEICE TRANS. INF. & SYST., VOL.E93–D, NO.6 JUNE 2010

PAPER Special Section on Info-Plosion

Detecting New Words from Chinese Text Using Latent Semi-CRF Models

Xiao SUN[†,††,†††a)], ***Member***, **Degen HUANG**[††b)], ***Nonmember***, ***and*** **Fuji REN**[†††c)], ***Member***

SUMMARY Chinese new words and their part-of-speech (POS) are particularly problematic in Chinese natural language processing. With the fast development of internet and information technology, it is impossible to get a complete system dictionary for Chinese natural language processing, as new words out of the basic system dictionary are always being created. A latent semi-CRF model, which combines the strengths of LDCRF (Latent-Dynamic Conditional Random Field) and semi-CRF, is proposed to detect the new words together with their POS synchronously regardless of the types of the new words from the Chinese text without being pre-segmented. Unlike the original semi-CRF, the LDCRF is applied to generate the candidate entities for training and testing the latent semi-CRF, which accelerates the training speed and decreases the computation cost. The complexity of the latent semi-CRF could be further adjusted by tuning the number of hidden variables in LDCRF and the number of the candidate entities from the Nbest outputs of the LDCRF. A new-words-generating framework is proposed for model training and testing, under which the definitions and distributions of the new words conform to the ones existing in real text. Specific features called "Global Fragment Information" for new word detection and POS tagging are adopted in the model training and testing. The experimental results show that the proposed method is capable of detecting even low frequency new words together with their POS tags. The proposed model is found to be performing competitively with the state-of-the-art models presented.

key words: *natural language processing, new word detection, new words POS tagging, conditional random fields, latent-dynamic CRF, semi-CRF, latent semi-CRF*

1. Introduction

During the process of Chinese natural language processing, the occurrences of new words have made this task more difficult. The new words in Chinese text, which are also called out-of-vocabulary (OOV) words or new words, are main holdbacks in the tasks of Chinese natural language processing. The new words cannot be segmented correctly as they are not found in the existing system basic dictionary [1]–[3]. In Chinese natural language processing, with the fast development of internet and the information explosion, even the largest dictionary 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 dictionary with a fixed number of entries. Therefore, the new words are sure to appear in real text during the procedure of real world Chinese natural language processing. The 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 the existing new words [4]. These errors in Chinese word segmentation will reduce the overall precision of the system. The problems caused by the existence of the new words must be solved in order to increase the effectiveness of Chinese natural language processing system. Although the definitions of the new words in Chinese text are not very clear, there are still some specific characteristics of the new words. First, the new words generate according to the basic dictionary of the system; Second, new words appear in a certain period of time under specific circumstance; Third, the new words basically obey the existing morphological rules. Furthermore, the distribution of the POS of new words disperses widely. The POS of new words not only include the geographical names, person names and organization names, but also include the normal nouns, normal verbs and even some adjective words. However, there still exist some statistical laws for the distribution of the new words' POS tags. Therefore, proper machine learning methods are necessary to be proposed for Chinese new word detection and POS tagging in order to increase the precision of Chinese word segmentation and other tasks in Chinese natural language processing.

Researchers have studied different methods to detect the new words in Chinese text. Yet there are still some limitations in all these methods. First, the new word detection and the POS tagging are regarded as two separate steps, which bring out the facts that the lexical features information can not be fully considered and used. Second, these methods haven't proposed a proper framework of building reasonable size of basic dictionary and new words corpus for training and testing. The number of new words is very much depending on the size of the basic dictionary adopted in certain system. Certainly, the larger the dictionary is, the less the new words occurrence in the texts. One can create a dictionary from all the tagged corpus but that will not be a proper dictionary. Furthermore, if all words in the tagged corpus are used to create the dictionary, then there will be no new words in the texts for training and testing. Therefore,

Manuscript received September 4, 2009.
Manuscript revised November 24, 2009.

[†]The author is with the School of Computer Science & Technology, Dalian Nationality University, China.
[††]The authors are with the Department of Computer Science & Engineering, Dalian University of Technology, China.
[†††]The authors are with the Department of Information Science & Intelligent Systems, Tokushima University, Tokushima-shi, 770–8506 Japan.
a) E-mail: sunxiao@dlnu.edu.cn
b) E-mail: huangdg@dlut.edu.cn
c) E-mail: ren@is.tokushima-u.ac.jp
DOI: 10.1587/transinf.E93.D.1386

it is important to define the meaning of new words properly and 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 dictionary. A more natural way is by building a proper basic dictionary. We can consider those words that are not in a proper basic dictionary to be new words. In this case, some words in the corpus are not found in the basic dictionary and can be marked as new words in training data for new word detection [5]–[7]. As far as we know, the definitions of words are different by institutions, such as Peking University Corpus, Penn Chinese Tree Bank and Taiwan Sinica Corpus in SIGHAN corpus. Therefore, the dictionary and the tagged corpus used must be consistent. We choose the tagged corpus provided by Peking University (PKU) to build the basic dictionary and the corpus containing new words for training and testing. From our survey of the PKU corpus, we found that the half year of the corpus is the perfect size for building basic dictionary and the percentage of new words in the new words corpus of one month (out of the corpus adopted to build the basic dictionary) is 14% or so. According to our survey, 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 the new words in real text on the internet or other corpus. The latent semi-CRF trained under such framework is flexible and could be used to detect new words in widely kinds of fields.

In order to detect the new words and assign POS tags to them synchronously, here we proposed a latent semi-CRF model, which combines the LDCRF (Latent-Dynamic Conditional Random Field) [8]–[10] and the semi-CRF model [11], [12] model. The latent 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 the new words boundaries, which are combined with candidate POS tags and adopted to build the candidate entities for semi-CRF. In such a way, the scalability of the 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. The latent semi-CRF could detect the new words together with their POS tags synchronously. The contextual word-level information and character-level information for the new words could be fully used. In addition, the global information for new words called "Global Fragment Information" is proposed and adopted in the features set for model training and testing, which could obviously increases the precision of new word detection. The latent 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 computations cost of unnecessary POS tagging for incorrect candidate words are avoided.

The boundaries and the POS tags of the new words are both unknown. So given a sequence X that includes new words, we need not only segment the input sequence X (assigning BIO tags), but also assign POS tags to the segments with new words and assigning "O" to the segments without new words. There are too much candidate segments for the sequence and candidate POS tags for the segments. If we directly adopt the semi-CRF or LDCRF to detect the new words assign POS tags, the computation cost will be very large. For the semi-CRF model, all the candidate segments with all candidate POS tags have to be enumerated in model training and testing [13]. Furthermore, in the semi-CRF, a reasonable value of L (upper bound length of entities) has to be set for different tasks [13]. However, in the tasks of new word detection, the length of the new words in the sequence might be longer than the fixed L, thus the longer word cannot be detected correctly. We extended the semi-CRF and inserted the LDCRF to generate the candidate entities for the semi-CRF in model training and testing. In such a way, with the strength of the LDCRF model, we do not have to limit the L in the semi-CRF model. The LDCRF effectively learns the substructure of the input sequence X, and output the boundaries of the new words for the input sequence X. Furthermore, the advantage of LDCRF is that LDCRF can output N-best label sequences and their probabilities using efficient marginalization operations [9]. We use this characteristic and combine the Nbest outputs (candidate new words) from LDCRF with the possible tags to build the candidate entities for semi-CRF in model training and testing. We can adjust the number of the Nbest outputs from the LDCRF to control the computation cost and the precision of latent semi-CRF.

2. The Latent Semi-CRF Model

2.1 Definition of the Latent Semi-CRF Model

In order to extend the semi-CRF model, we follow the definitions in [13]. Let $x = \{x_0, x_1, \ldots, x_i \ldots\}$, $(0 \le i \le |x|)$ denotes a sequence of Chinese characters that includes new words to de detected. Let $y = \{y_0, y_1, \ldots, y_j, \ldots\}$, $(0 \le j \le |y|)$ denotes the output label sequence. Let $s = (s_1, \ldots, s_j, \ldots)$ denotes 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 $i = t_j$ and $i = u_j$, inclusive. In the tasks of new word detection, this means that all the characters in a new word share the same POS tag, each of the characters out of the new words has the tag "O". We assume that segments have a positive length bounded above by the pre-defined upper bound L ($1 \le t_j \le u_j \le |s|$, $u_j - t_j + 1 \le L$) 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 detection and POS tagging, a correct segmentation of the 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 the semi-CRFs defines a conditional probability of a state sequence y given an observation sequence x by:

$$p(y|x,\lambda) = \frac{\exp\left(\sum_j \sum_i \lambda_i f_i(y_{j-1}, y_j, x, t_j, u_j)\right)}{Z(x)} \quad (1)$$

where $f_i(y_{j-1}, y_j, x, t_j, u_j)$ is a feature function, s_j is the jth segment in s and $Z(x)$ is the normalization factor as defined for semi-CRF,

$$Z(x) = \exp\left(\sum_{s(x)} \sum_j \sum_i \lambda_i f_i(y_{j-1}, y_j, x, t_j, u_j)\right)$$

The $s(x)$ in normalization factor denotes all the candidate segments of x. From the Eq. (1), we can see that in order to get the original semi-CRF work for the new word detection and POS tagging, we have to enumerate all the candidate segments with different length for every x and enumerate all the POS tags for each candidate segment. These candidate entities made the inference of semi-CRF very expensive. So we adopted the LDCRF and used the Nbest output of the LDCRF to generate the candidate entities for the semi-CRF. We follow the definition of LDCRF described in Morency's work [8]. Let the input of the LDCRF be the sequence of Chinese characters. The LDCRF output the "BIO" boundary tags for the input sequence to mark the new words (or the characters out of the new words). Let the $Path_{NBEST} = \{path^1, \ldots, path^{NBEST}\}$ denote the Nbest output of the LDCRF for the input sequence x. The NBEST is a predefined const denoting the number of Nbest output paths. In the task of new word detection and POS tagging, the LDCRF is adopted to generate all the candidate new words first, and then the candidate POS tags are assigned to the candidate new words and the tag "O" are assigned to the single character out of the new words to build the candidate entities for the latent semi-CRF. Let the $s_{NBEST}(x)$ denotes all the candidate entities generated from the LDCRF for x. We replace the $s(x)$ in Eq. (1) with $s_{NBEST}(x)$ and get the latent semi-CRF:

$$p(y|x,\lambda) = \frac{1}{\exp\left(\sum_{s_{NBEST}(x)} \sum_j \sum_i \lambda_i f_i(y_{j-1}, y_j, x, t_j, u_j)\right)} \times \exp\left(\sum_j \sum_i \lambda_i f_i(y_{j-1}, y_j, x, t_j, u_j)\right) \quad (2)$$

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

2.2 Inference Algorithm for the Latent Semi-CRF

We revised the inference algorithm token from the semi-CRF [13] for latent semi-CRF. The inference algorithm for the latent 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:

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

where the 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:

$$y* = \arg\max_y \sum_{h:\forall h_i \in \mathrm{H}_{y_i}} P(h|x,\theta) \quad (4)$$

To estimate the label y_j^* of frame j, the marginal probabilities $P(h_j = a|x,\theta)$ are computed for all possible hidden states $a \in \mathrm{H}$. Then the marginal probabilities H_{y_j} and the label associated with the optimal set is chosen. In order to generate the candidate entities for the latent semi-CRF, first, we apply LDCRF to output Nbest label sequence for the new words boundaries and their probabilities. The candidate entities for the semi-CRF are the segments with labels, but the Nbest results from the LDCRF only include the information of all the candidate segments for the new words without proper labels (POS tags). So we assign the possible POS tags to the candidate new words to set up the candidate entities for the latent 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 the Nbest output of LDCRF is $S_{NBEST}(x)$, which includes all the candidate new words boundaries sequences together with the proper tags (POS tags or "O").

The inference algorithm for latent semi-CRF is to get the result of the $\arg\max_{s \in S_{NBEST}(x)} P(s|x,\lambda)$ We use $F(x,s)$ to denote $\sum_j \sum_i f_i(y_{j-1}, y_j, x, t_j, u_j)$, we use $\vec{\lambda}$ to denote the weight vector for $F(x,s)$ and use $\vec{f}(y_{j-1}, y_j, x, t_j, u_j)$ to denote $\sum_i f_i(y_{j-1}, y_j, x, t_j, u_j)$, so that the former $\arg\max_{s \in S_{NBEST}(x)} P(s|x,\lambda)$ can be rewritten as

$$\arg\max_{s \in S_{NBEST}(x)} \vec{\lambda} \cdot F(x,s) = \arg\max_{s \in S_{NBEST}(x)} \vec{\lambda} \cdot \sum_j \vec{f}(y_{j-1}, y_j, x, t_j, u_j) \quad (5)$$

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

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

where the *len* denotes the $|s^i_{end(i)}|$, which is 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 Latent Semi-CRF

We revised the learning algorithm token from semi-CRF [13] for parameter estimation in latent semi-CRF. First, we train the LDCRF model in order to generate the candidate entities for training latent semi-CRF. For the LDCRF model we use the following objective function to learn the 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 the conditional log-likelihood of the training data. The second term is the log of a Gaussian prior with variance σ^2 [8]. We here adopted the Limited-memory BFGS Method (L-BFGS) to estimate the parameter. The L-BFGS algorithm is currently the most effective optimization method for CRF parameter estimation [15]. As the parameter θ of the LDCRF is already known, so it is possible for the LDCRF to generate the Nbest results (boundaries for all the candidate new words) from the input sequence to build the candidate entities for training latent semi-CRF. We generate the candidate entities for training latent semi-CRF in the same way as we did in inference algorithm for latent semi-CRF: adding possible tags to the segments in the Nbest results from the LDCRF. The candidate entities set generated from the Nbest output of LDCRF is $S_{NBEST}(x)$, which includes all the candidate new words together with their POS tags. For the original semi-CRF, over a given training set $T = \{(x_l, s_l)\}_{l=1}^N$, we express the log-likelihood over the training sequences as:

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

In the latent semi-CRF, for a given input sequence x_l, we use the $S_{NBEST}(x_l)$ to replace the set of all candidate entities, which means the candidate entities generated from the Nbest result of x_l. For convenience, we use the $S^*(x_l)$ to denotes $S_{NBEST}(x_l)$. So the gradient of $L(\lambda)$ is the following:

$$\begin{aligned} \Delta L(\lambda) &= \sum_l F(x_l, s_l) - \frac{\sum_{s^*(x_l)} F(s^*(x_l), x_l) e^{\vec{\lambda} \cdot F(s^*(x_l), x_l)}}{Z_\lambda(x_l)} \\ &= \sum_l F(x_l, s_l) - E_{P(s^*(x_l)|\vec{\lambda})} F(x_l, s^*(x_l)) \end{aligned} \tag{9}$$

The $E_{P(s^*(x_l)|\vec{\lambda})} F(x_l, s^*(x_l)$ is the expected value of the features under the current weight vector (parameter). From the Eq. (9) we can see that we only have to consider the candidate entities generated by the Nbest result of the LDCRF, which obviously reduce the computation cost for the latent semi-CRF.

2.4 New-Words-Generating Framework

The new words are words that do not exist in system dictionary, so it is difficult to regenerate the new words for training the 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 the new words under such frameworks may not confirm to the new words existing in real text. In real text, sometimes the new words appear more times than the known words do in specific texts. We here proposed a framework for new words training and testing. The generation of the new words in the real text has two factors, the first is a system basic dictionary 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 dictionary. We analyzed the PKU corpus [14], which includes the all the news text of People's Daily in year 2000 classified and separated by the months to build basic dictionary of proper size for the new words generation. Several consequent months of PKU corpus are used to build the basic dictionary, and then the following one month corpus is used as the corpus with new words, which are also called new words corpus. The characteristic of new words generates under such framework is in accordance with the new words in real text, so the proposed new-word-generating framework has the expansibility to adopt sundry new words.

In order to generate the proper size of basic dictionary, we count the number of the words in all the possible consequent of corpus and the new words in the following one month. It can be seen from the survey that after half year, the percentage of the new words is becoming steady. The POSs distribution of the new words in June and July is shown in Table 1 (Only top 10 POS tags are listed). The POS of the new words is directly token from the PKU corpus and the definition of the POS can be referred to the paper of Yu et al. [14]. The distribution of the POS has some statistical laws, such as the most POS tags of the new words are noun (n), the top 10 of the POS tags in Table 1 are almost the same. The distribution and the characteristic of the new words under the proposed framework are in accordance

Table 1 The distribution of the POSs of new words.

June			July		
POS	Count	Percentage	POS	Count	Percentage
n	4000	0.417188	n	4298	0.460961
m	1619	0.168857	m	1726	0.185114
nz	1111	0.115874	nz	1106	0.118619
ns	521	0.054339	j	519	0.055663
j	451	0.047038	v	361	0.038717
v	441	0.045995	l	210	0.022523
nr	353	0.036817	i	184	0.019734
l	210	0.021902	nr	178	0.019091
i	194	0.020234	ns	113	0.012119
t	139	0.014497	t	107	0.011476

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

2.5 Global Fragment Features

For the new words, the global information in the context is important for the detection. Certain new words appear considerable times in certain context. Although they can not be segmented correctly, the fragments generated by the 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 the segmentation by the HMM 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 detection. For example, the "正龙拍虎" appears 10 times in a certain Chinese news text, which are all segmented into "正/龙/拍/虎" after the segmentation. We could find the "正/龙/拍/虎" are fragments, which have the following global features: The fragments appear 10 times in the context and the length of the fragments is 4 Chinese characters. These global features could be used as features for LDCRF and latent semi-CRF, which are listed in Table 2. In Table 2, the position means the start position of the character or word in the fragment.

3. Features and Templates for Latent Semi-CRF Model

First, we need to define the features and templates for the training and testing LDCRF model. The PKU corpus of July is divided into 80% for training and 20% for testing. Take the sentence "多来米/nz 中文/nz 网/n" (DuoLaiMi Chinese Net) in the training corpus for example, the "多来米 (DuoLaiMi)" is a new words according to the basic dictionary.

Table 2 Global fragment features for new word detection and POS tagging.

Global fragment features 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
Global fragment features for latent 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

Table 3 The 5-tag label set for word boundaries.

Labels	Description
B	The first character in a word
I	Middle character in a word with more than two characters.
E	The last character word in new word
S	Single character new word
O	Other character in known words

Table 4 The template of the 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 character
Trigram	$C_{-1}C_0C_1$	The combination of three character
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 the C_0, $C_{-1}C_0$, C_0C_1 exists in words of the basic dictionary or not.
UW Dic	$N(C_0)$, $N(C_{-1}C_0)$, $N(C_0C_1)$	Whether the C_0, $C_{-1}C_0$, C_0C_1 exists in words of the UW dictionary or not.

To train the LDCRF model, we used the 5-tag labels set described in Table 3 as the boundary labels. For example, the "多" is labeled with "B", which means it is the first character in the new word. The "来" is labeled with "I", which means it is the middle character in the new word. The "米" labeled with "E", which means it is the last 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 4.

The GFF in the table means "Global fragment features", the UW Dic denotes "new words dictionary". The predefined $S(C_0)$ for the characters are five classes: Class 1 represents numbers; Class 2 represents English letters; Class 3 represents punctuation; Class 4 represents Chinese characters; Class 5 represents other characters. The basic dictionary is built by the corpus from January to June. The new

Table 5 Templates for the latent semi-CRF.

Type	Feature	Description
Unigram	W_0, $G(W_0)$	Unigram features for the current segment. The $G(W_0)$ is global fragment features.
	$N(W_0(C_n))$ $(1 \le n \le \|W_0\|)$ $N(W_0(C_0C_1))$ $N(W_0(C_0C_1C_2))$	Whether character sequence C_n, C_0C_1, $C_0C_1C_2$ in the segment exist in some word of the new words dictionary or not. If so, this template outputs words and POS tags from new words dictionary.
Bigram	W_{-1}/W_0, W_0/W_1	The combination of two words in two segments.
Trigram	$W_{-2}W_{-1}W_0$ $W_{-1}W_0W_1$ $W_0W_1W_2$	The combination of three words in three segments.
Length	$L(W_0)$	The length of the word in current segment

dictionary includes all the new words in the training corpus (we also collected some new words from the internet). The HMM segmenter is built from the basic dictionary using forward maximum matching (FMM) method. The LDCRF is character-based, but we also imported the information from the outer dictionary as features from "Basic Dic" and "UW Dic" template. The $B(C_{-1}C_0)$ means whether the character sequence $C_{-1}C_0$ exists in some word of the basic dictionary. The new words dictionary is very useful for detecting the new words like the person name, because the set of the last name for a person is limited. The "UW Dic" template is built according to the new words dictionary.

In a latent semi-CRF learner, features no longer apply to individual words, but instead are 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 [13]. 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. The templates and features for the latent semi-CRF are listed in the following Table 5.

4. Experiments and Results

In order to get the proper hidden states for the LDCRF and Nbest variable for latent semi-CRF in training and testing, we first performed the cross-validation by using the corpus of July and test the overall F-score of the new word detection and POS tagging. We finally set the number of hidden states to 4 and the Nbest to 30 for LDCRF and latent semi-CRF, which is according with the characteristics of the new words. 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 the 20% test corpus of July. We first evaluated the recall of the new word detection by their POS tags. The results are shown in Table 6. In Table 6, the "Count" column is the same as in Table 1, the "Detected" column means the number of the new words detected with the corresponding POS tags, and the "Correct" column

Table 6 Distribution of detected new words by their POS tags.

POS	Count	Detected	Correct	R(%)	P(%)
n	4298	4498	3664	85.25	81.46
m	1726	1778	1659	96.12	93.31
nz	1106	1124	1002	90.60	89.15
j	519	345	344	66.28	99.71
v	361	280	215	59.56	76.79
l	210	160	110	52.38	68.75
i	184	114	95	51.63	83.33
nr	178	140	102	57.30	72.86
ns	113	120	107	94.69	89.17
t	107	112	100	93.45	89.29
Others	549	530	485	88.34	91.51
All	9351	9201	7883	84.30	85.68

Table 7 Comparison with other models.

	R(%)	P(%)	F(%)	T_{train}	T_{test}
HMM+Rules	75.43	70.61	72.94	1	1
SVM+ME	71.32	89.11	79.23	4.12	2.38
CRF+ME	80.59	79.61	80.1	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
Latent Semi-CRF	86.55	87.98	87.26	3.89	1.48

means the correct new words with correct POS tags. We get quite satisfactory precision by using the proposed method. As there is no single standard definition of 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 can not be detected and given the POS correctly. The POS tagging is not quite satisfied because we only apply the latent semi-CRF model to guess the POS tags for new words, the information of the POS of known word could not be used. We will apply the latent semi-CRF to detect the known words together with new words so that the lexical information of the known words can be fully used. The precision of the new word detection can be further increased.

In order to compare the latent semi-CRF model with other models, we built five other Chinese new word detection models, which are listed in Table 7. We first adopted the HMM model with some rules to build the new word detection and POS tagging tools, based on the methods proposed by [16] and [17], which can be treated as the baseline. The "SVM+ME" model is based on the [18]. The "CRF+ME" model is partly based on the [19], and the POS tagging is still based on the method proposed by [18]. 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 detection and POS tagging. We adopted cross validation to get the optimum parameters for all these models. In Table 7, in order to compare the computational cost between models, we calculated the training (T_{train}) and testing (T_{test})

Table 8 Effectiveness of global fragment features.

	R(%)	P(%)	F(%)
No GFF	85.83	86.62	86.22
With GFF	86.55	87.98	87.26

corpus of each method. In order to make the comparison clear, in the training time column (T_{train}), we set the time of model "HMM+Rules" to 1. In the testing time column (T_{test}), we also set the time of model "HMM+Rules" to 1. In order to test the effectiveness of our proposed "Global Fragment Features", we deleted the GFF from latent semi-CRF model (No GFF) and compare the result with the model with Global Fragment Features (With GFF). The results are listed in Table 8. Although our method tackled especially the overlapping cases, the results turned out to be not so satisfactory. There are 3257 and 913 overlapping cases in the training and testing data, respectively. Out of the 913 cases in the testing data, only 56 cases have been detected. In the phrase "酒台/n 上/s 铺/v 着/u 新/a 台/q 布/n" (the bar is covered with new table cloth), the new word "酒台" (bar) has been detected (but the new word "台布" (table cloth) has not been detected). Unfortunately, there are still many cases which could not be detected. For example, in "到/v 一/m 年终/t 了/y" (When a year ended), the new word "终了" (ended) could not be detected, and in "不断/d 引/v 动人/a 们/k 的/u" (continuously attract the people's ...), the new word "引动" (attract) could not be detected as well. We still need to find an alternate approach to solve this problem.

There are some incorrect cases when two (or more) consecutive new words exist. For example, "开怀狂饮" (drink wildly and happily), should be two new words, "开怀/d" (happily) and "狂饮/v" (drink wildly), but this model combined them to produce only one new word. Other examples of consecutive new words are "二三流货色" (second third class items, 二三流/d 货色/n), "迎宾送客" (to welcome and see off customers, 迎宾/vn 送客/vn) and "京腔京韵" (Peking slang, 京腔/n 京韵/n).

There exist also some inconsistencies in the pattern of words. For example, "甲等/n 奖/n" (first prize) is considered as two words but "一等奖/n" (first prize) is one word, and "办学/vn 史/Ng" (the history of building school) as two words but "建设史/n" (the history of construction) and "发展史/n" (the history of development) as one word, even they have the same suffixes. Our model combined "甲等奖" and "办学史" as one word, but were considered as errors since the original segmentation were separate words. Normally if a word is a frequently used word, then it will be considered as one word, or else, it will be separated as two words. Furthermore, if the word before the suffix is a monosyllabic word, then it will be combined with the suffix, or else, it will become two words. These are some of the special rules that have been defined by the Peking University Corpus which perhaps can only be corrected by defining some rules according to their standard.

5. Conclusions and Future Work

In order to detect the Chinese new words with their POSs in the real text or on the internet, we proposed a latent semi-CRF model, which combines the strength of the LDCRF and the semi-CRF model. By importing the LDCRF model in the latent semi-CRF model, we do not have to enumerate all the candidate entities as the semi-CRF do. The proposed latent semi-CRF adopts the candidate entities generated from the Nbest results of the LDCRF, which obviously decreases the computation 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 to build the basic dictionary and training/testing corpus for the latent semi-CRF model. Under such framework the new words for training are in accordance with the characteristic of the Chinese new words in real text, hence the framework 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 the new words is a very useful feature for Chinese new word detection and POS tagging, which was adopted for training and testing the latent semi-CRF model. There exist some phenomena which cannot be analyzed only with bigram features in new word detection. To improve accuracy, trigram or more general n-gram features would be useful. Latent semi-CRF has capability of handling such features. We also need a practical features selection which effectively trades between accuracy and efficiency. We will apply the proposed latent semi-CRF model in Chinese word segmentation and POS tagging, in such a way that more context information such as the known words together with their POS tags for the new word will be imported and the precision of the new word detection and POS tagging can be further increased.

References

[1] C. Goh, M. Asahara, and Y. Matsumoto, "Chinese unknown word identification using character-based tagging and chunking," Proc. 41st Annual Meeting on Association for Computational Linguistics, pp.197–200, 2003.

[2] J. Nie, M. Hannan, and W. Jin, "Unknown word detection and segmentation of Chinese using statistical and heuristic knowledge," Communications of COLIPS, vol.5, no.1, pp.47–57, 1995.

[3] C. Chen, M. Bai, and K. Chen, "Category guessing for chinese unknown words," Proc. Natural Language Processing Pacific Rim Symposium, pp.35–40, 1997.

[4] R. Sproat, C. Shih, W. Gale, and N. Chang, "A stochastic finite-state word-segmentation algorithm for Chinese," Computational Linguistics, vol.22, no.2, pp.377–404, 1996.

[5] K. Chen and M. Bai, "Unknown word detection for Chinese by a corpus-based learning method," Computational Linguistics, vol.3, no.1, pp.27–44, 1998.

[6] G. Fu and K. Luke, "Chinese unknown word identification using class-based LM," Lect. Notes Comput. Sci. (IJCNLP 2004),

vol.3248, pp.704–713, 2005.
[7] G. Fu and K. Luke, "An integrated approach for Chinese word segmentation," Proc. PACLIC, vol.17, pp.80–87, 2003.
[8] L. Morency, A. Quattoni, and T. Darrell, "Latent-dynamic discriminative models for continuous gesture recognition," Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2007.
[9] X. Sun, L. Morency, D. Okanohara, and J. Tsujii, "Modeling latent-dynamic in shallow parsing: A latent conditional model with improved inference," Proc. 22nd International Conference on Computational Linguistics, pp.841–848, 2008.
[10] X. Sun and J. Tsujii, "Sequential labeling with latent variables: An exact inference algorithm and its efficient approximation," Proc. 12th Conference of the European Chapter of the ACL, pp.772–780, 2009.
[11] C. Sutton, A. McCallum, and K. Rohanimanesh, "Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data," J. Machine Learning Research, vol.8, pp.693–723, MIT Press Cambridge, MA, USA, 2007.
[12] D. Okanohara, Y. Miyao, Y. Tsuruoka, and J. Tsujii, "Improving the scalability of semi-markov conditional random fields for named entity recognition," Ratio, vol.1, no.21646, pp.42–19, 2006.
[13] S. Sarawagi and W. Cohen, "Semi-markov conditional random fields for information extraction," Advances in Neural Information Processing Systems, vol.17, pp.1185–1192, Citeseer, 2005.
[14] S. Yu, H. Duan, X. Zhu, B. Swen, and B. Chang, "Specification for corpus processing at Peking University: Word segmentation, POS tagging and phonetic notation," J. Chinese Language and Computing, vol.13, pp.121–158, 2003.
[15] D. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," Mathematical Programming, vol.45, pp.503–528, Springer, 1989.
[16] A. Chen, "Chinese word segmentation using minimal linguistic knowledge," Proc. Second SIGHAN Workshop on Chinese Language Processing, pp.148–151, 2003.
[17] G. Zhou, "A chunking strategy towards unknown word detection in Chinese word segmentation," Lect. Notes Comput. Sci., vol.3651, pp.530–541, Springer, 2005.
[18] G. Goh, M. Asahara, and Y. Matsumoto, "Machine learning-based methods to Chinese unknown word detection and POS tag guessing," J. Chinese Language and Computing, vol.16, pp.185–206, 2006.
[19] F. Peng, F. Feng, and A. McCallum, "Chinese segmentation and new word detection using conditional random fields," COLING '04: Proc. 20th international conference on Computational Linguistics, Association for Computational Linguistics, pp.562–569, 2004.

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 the Dalian Nationality University. He got his double-doctor's degree in Dalian University of Technology of China and the University of Tokushima of Japan. His research interests include Natural Language Processing, Machine Translation, Chinese Lexical Analysis.

Degen Huang was born in 1965. He is the Ph.D. and professor in the Dalian University of Technology. His main research interests include Natural Language Processing and Machine Translation. He is now working at the Department of Computer Science and Engineering, Dalian University of Technology, No.2 LingGong Road, Ganjingzi District, Dalian, P.R.China.

Fuji 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, Bejing, China.He also 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.