# Subword-based Tagging for Confidence-dependent Chinese Word Segmentation

**Ruiqiang Zhang**[1,2] and **Genichiro Kikui**[*] and **Eiichiro Sumita**[1,2]
[1]National Institute of Information and Communications Technology
[2]ATR Spoken Language Communication Research Laboratories
2-2-2 Hikaridai, Seiika-cho, Soraku-gun, Kyoto, 619-0288, Japan
{ruiqiang.zhang,eiichiro.sumita}@atr.jp

## Abstract

We proposed a subword-based tagging for Chinese word segmentation to improve the existing character-based tagging. The subword-based tagging was implemented using the maximum entropy (MaxEnt) and the conditional random fields (CRF) methods. We found that the proposed subword-based tagging outperformed the character-based tagging in all comparative experiments. In addition, we proposed a confidence measure approach to combine the results of a dictionary-based and a subword-tagging-based segmentation. This approach can produce an ideal tradeoff between the in-vocaulary rate and out-of-vocabulary rate. Our techniques were evaluated using the test data from Sighan Bakeoff 2005. We achieved higher F-scores than the best results in three of the four corpora: PKU(0.951), CITYU(0.950) and MSR(0.971).

## 1 Introduction

Many approaches have been proposed in Chinese word segmentation in the past decades. Segmentation performance has been improved significantly, from the earliest maximal match (dictionary-based) approaches to HMM-based (Zhang et al., 2003) approaches and recent state-of-the-art machine learning approaches such as maximum entropy (MaxEnt) (Xue and Shen, 2003), support vector machine (SVM) (Kudo and Matsumoto, 2001), conditional random fields (CRF) (Peng and McCallum, 2004), and minimum error rate training (Gao et al., 2004). By analyzing the top results in the first and second Bakeoffs, (Sproat and Emerson, 2003) and (Emerson, 2005), we found the top results were produced by direct or indirect use of so-called "IOB" tagging, which converts the problem of word segmentation into one of character tagging so that part-of-speech tagging approaches can be used for word segmentation. This approach was also called "LMR" (Xue and Shen, 2003) or "BIES" (Asahara et al., 2005) tagging. Under the scheme, each character of a word is labeled as "B" if it is the first character of a multiple-character word, or "I" otherwise, and "O" if the character functioned as an independent word. For example, "全(whole) 北京市(Beijing city)" is labeled as "全/O 北/B 京/I 市/I". Thus, the training data in word sequences are turned into IOB-labeled data in character sequences, which are then used as the training data for tagging. For new test data, word boundaries are determined based on the results of tagging.

While the IOB tagging approach has been widely used in Chinese word segmentation, we found that so far all the existing implementations were using character-based IOB tagging. In this work we propose a subword-based IOB tagging, which assigns tags to a pre-defined lexicon subset consisting of the most frequent multiple-character words in addition to single Chinese characters. If only Chinese characters are used, the subword-based IOB tagging is downgraded to a character-based one. Taking the same example mentioned above, "全北京市" is la-

[*] Now the second author is affiliated with NTT.



*Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 961–968,
Sydney, July 2006. 

beled as "全/O 北京/B 市/I" in the subword-based tagging, where "北京/B" is labeled as one unit. We will give a detailed description of this approach in Section 2.

There exists a clear weakness with the IOB tagging approach: It yields a very low in-vocabulary rate (R-iv) in return for a higher out-of-vocabulary (OOV) rate (R-oov). In the results of the closed test in Bakeoff 2005 (Emerson, 2005), the work of (Tseng et al., 2005), using CRFs for the IOB tagging, yielded a very high R-oov in all of the four corpora used, but the R-iv rates were lower. While OOV recognition is very important in word segmentation, a higher IV rate is also desired. In this work we propose a confidence measure approach to lessen this weakness. By this approach we can change the R-oov and R-iv and find an optimal tradeoff. This approach will be described in Section 2.3.

In addition, we illustrate our word segmentation process in Section 2, where the subword-based tagging is described by the MaxEnt method. Section 3 presents our experimental results. The effects using the MaxEnts and CRFs are shown in this section. Section 4 describes current state-of-the-art methods with Chinese word segmentation, with which our results were compared. Section 5 provides the concluding remarks and outlines future goals.

## 2 Chinese word segmentation framework

Our word segmentation process is illustrated in Fig. 1. It is composed of three parts: a dictionary-based N-gram word segmentation for segmenting IV words, a maximum entropy subword-based tagger for recognizing OOVs, and a confidence-dependent word disambiguation used for merging the results of both the dictionary-based and the IOB-tagging-based. An example exhibiting each step's results is also given in the figure.

### 2.1 Dictionary-based N-gram word segmentation

This approach can achieve a very high R-iv, but no OOV detection. We combined with it the N-gram language model (LM) to solve segmentation ambiguities. For a given Chinese character sequence, $C = c_0 c_1 c_2 \dots c_N$, the problem of word segmentation can be formalized as finding a word sequence,

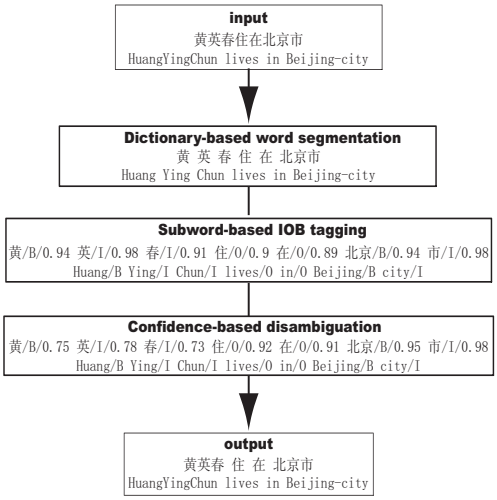


Figure 1: Outline of word segmentation process

$W = w_{t_0} w_{t_1} w_{t_2} \dots w_{t_M}$, which satisfies

$$w_{t_0} = c_0 \dots c_{t_0}, \quad w_{t_1} = c_{t_0+1} \dots c_{t_1}$$
$$w_{t_i} = c_{t_{i-1}+1} \dots c_{t_i}, \quad w_{t_M} = c_{t_{M-1}+1} \dots c_{t_M}$$
$$t_i > t_{i-1}, \quad 0 \le t_i \le N, \quad 0 \le i \le M$$

such that

$$\begin{aligned} W &= \arg\max_W P(W|C) = \arg\max_W P(W)P(C|W) \\ &= \arg\max_W P(w_{t_0} w_{t_1} \dots w_{t_M}) \delta(c_0 \dots c_{t_0}, w_{t_0}) \\ &\quad \delta(c_{t_0+1} \dots c_{t_1}, w_{t_1}) \dots \delta(c_{t_{M-1}+1} \dots c_M, w_{t_M}) \end{aligned} \tag{1}$$

We applied Bayes' law in the above derivation. Because the word sequence must keep consistent with the character sequence, $P(C|W)$ is expanded to be a multiplication of a Kronecker delta function series, $\delta(u, v)$, equal to 1 if both arguments are the same and 0 otherwise. $P(w_{t_0} w_{t_1} \dots w_{t_M})$ is a language model that can be expanded by the chain rule. If trigram LMs are used, we have

$$P(w_0)P(w_1|w_0)P(w_2|w_0 w_1) \cdots P(w_M|w_{M-2} w_{M-1})$$

where $w_i$ is a shorthand for $w_{t_i}$ .

Equation 1 indicates the process of dictionary-based word segmentation. We looked up the lexicon to find all the IVs, and evaluated the word sequences by the LMs. We used a beam search (Jelinek, 1998) instead of a viterbi search to decode the best word

sequence because we found that a beam search can speed up the decoding. N-gram LMs were used to score all the hypotheses, of which the one with the highest LM scores is the final output. The experimental results are presented in Section 3.1, where we show the comparative results as we changed the order of LMs.

### 2.2 Subword-based IOB tagging

There are several steps to train a subword-based IOB tagger. First, we extracted a word list from the training data sorted in decreasing order by their counts in the training data. We chose all the single characters and the top multi-character words as a lexicon subset for the IOB tagging. If the subset consists of Chinese characters only, it is a character-based IOB tagger. We regard the words in the subset as the subwords for the IOB tagging.

Second, we re-segmented the words in the training data into subwords of the subset, and assigned IOB tags to them. For the character-based IOB tagger, there is only one possibility for re-segmentation. However, there are multiple choices for the subword-based IOB tagger. For example, "北 京 市(Beijing-city)" can be segmented as "北 京 市(Beijing-city)/O," or "北 京(Beijing)/B 市(city)/I," or "北(north)/B 京(capital)/I 市(city)/I." In this work we used forward maximal match (FMM) for disambiguation. Because we carried out FMMs on each words in the manually segmented training data, the accuracy of FMM was much higher than applying it on whole sentences. Of course, backward maximal match (BMM) or other approaches are also applicable. We did not conduct comparative experiments due to trivial differences in the results of these approaches.

In the third step, we used the maximum entropy (MaxEnt) approach (the results of CRF are given in Section 3.4) to train the IOB tagger (Xue and Shen, 2003). The mathematical expression for the MaxEnt model is

$$P(t|h) = \exp\left(\sum_i \lambda_i f_i(h,t)\right)/Z, \quad Z = \sum_t P(t|h) \quad (2)$$

where $t$ is a tag, "I,O,B," of the current word; $h$, the context surrounding the current word, including word and tag sequences; $f_i$, a binary feature equal to 1 if the $i$-th defined feature is activated and 0 otherwise; $Z$, a normalization coefficient; and $\lambda_i$, the weight of the $i$-th feature.

Many kinds of features can be defined for improving the tagging accuracy. However, to conform to the constraints of closed test in Bakeoff 2005, some features, such as syntactic information and character encodings for numbers and alphabetical characters, are not allowed. Therefore, we used the features available only from the provided training corpus.

- Contextual information:
  $w_0, t_{-1}, w_0t_{-1}, w_0t_{-1}w_1, t_{-1}w_1, t_{-1}t_{-2}, w_0t_{-1}t_{-2}$,
  $w_0w_1, w_0w_1w_2, w_{-1}, w_0w_{-1}, w_0w_{-1}w_1$,
  $w_{-1}w_1, w_{-1}w_{-2}, w_0w_{-1}w_{-2}, w_1, w_1w_2$
  where $w$ stands for word and $t$, for IOB tag. The subscripts are position indicators, where 0 means the current word/tag; $-1, -2$, the first or second word/tag to the left; $1, 2$, the first or second word/tag to the right.

- Prefixes and suffixes. These are very useful features. Using the same approach as in (Tseng et al., 2005), we extracted the most frequent words tagged with "B", indicating a prefix, and the last words tagged with "I", denoting a suffix. Features containing prefixes and suffixes were used in the following combinations with other features, where $p$ stands for prefix; $s$, suffix; $p_0$ means the current word is a prefix and $s_1$ denotes that the right first word is a suffix, and so on.
  $p_0, w_0p_{-1}, w_0p_1, s_0, w_0s_{-1}, w_0s_1$,
  $p_0w_{-1}, p_0w_1, s_0w_{-1}, s_0w_{-2}$

- Word length. This is defined as the number of characters in a word. The length of a Chinese word has discriminative roles for word composition. For example, single-character words are more apt to form new words than are multiple-character words. Features using word length are listed below, where $l_0$ means the word length of the current word. Others can be inferred similarly.
  $l_0, w_0l_{-1}, w_0l_1, w_0l_{-1}l_1, l_0l_{-1}, l_0l_1$

As to feature selection, we simply adopted the absolute count for each feature in the training data as

the metric, and defined a cutoff value for each feature type.

We used IIS to train the maximum entropy model. For details, refer to (Lafferty et al., 2001).

The tagging algorithm is based on the beam-search method (Jelinek, 1998). After the IOB tagging, each word is tagged with a B/I/O tag. The word segmentation is obtained immediately. The experimental effect of the word-based tagger and its comparison with the character-based tagger are made in section 3.2.

### 2.3 Confidence-dependent word segmentation

In the last two steps we produced two segmentation results: the one by the dictionary-based approach and the one by the IOB tagging. However, neither was perfect. The dictionary-based segmentation produced a result with a higher R-iv but lower R-oov while the IOB tagging yielded the contrary results. In this section we introduce a confidence measure approach to combine the two results. We define a confidence measure, $CM(t_{iob}|w)$, to measure the confidence of the results produced by the IOB tagging by using the results from the dictionary-based segmentation. The confidence measure comes from two sources: IOB tagging and dictionary-based word segmentation. Its calculation is defined as:

$$CM(t_{iob}|w) = \alpha CM_{iob}(t_{iob}|w) + (1-\alpha)\delta(t_w, t_{iob})_{ng} \quad (3)$$

where $t_{iob}$ is the word $w$'s IOB tag assigned by the IOB tagging; $t_w$, a prior IOB tag determined by the results of the dictionary-based segmentation. After the dictionary-based word segmentation, the words are re-segmented into subwords by FMM before being fed to IOB tagging. Each subword is given a prior IOB tag, $t_w$. $CM_{iob}(t|w)$, a confidence probability derived in the process of IOB tagging, which is defined as

$$CM_{iob}(t|w) = \frac{\sum_{h_i} P(t|w, h_i)}{\sum_t \sum_{h_i} P(t|w, h_i)}$$

where $h_i$ is a hypothesis in the beam search. $\delta(t_w, t_{iob})_{ng}$ denotes the contribution of the dictionary-based segmentation.

$\delta(t_w, t_{iob})_{ng}$ is a Kronecker delta function defined as

$$\delta(t_w, t_{iob})_{ng} = \begin{cases} 1 & \texttt{if } t_w = t_{iob} \\ 0 & \texttt{otherwise} \end{cases}$$

In Eq. 3, $\alpha$ is a weighting between the IOB tagging and the dictionary-based word segmentation. We found an empirical value 0.8 for $\alpha$.

By Eq. 3 the results of IOB tagging were re-evaluated. A confidence measure threshold, $t$, was defined for making a decision based on the value. If the value was lower than $t$, the IOB tag was rejected and the dictionary-based segmentation was used; otherwise, the IOB tagging segmentation was used. A new OOV was thus created. For the two extreme cases, $t = 0$ is the case of the IOB tagging while $t = 1$ is that of the dictionary-based approach. In Section 3.3 we will present the experimental segmentation results of the confidence measure approach. In a real application, we can actually change the confidence threshold to obtain a satisfactory balance between R-iv and R-oov.

An example is shown in Figure 1. In the stage of IOB tagging, a confidence is attached to each word. In the stage of confidence-based, a new confidence was made after merging with dictionary-based results where all single-character words are labeled as "O" by default except "Beijing-city" labeled as "Beijing/B" and "city/I".

## 3 Experiments

We used the data provided by Sighan Bakeoff 2005 to test our approaches described in the previous sections. The data contain four corpora from different sources: Academia sinica, City University of Hong Kong, Peking University and Microsoft Research (Beijing). The statistics concerning the corpora is listed in Table 3. The corpora provided both unicode coding and Big5/GB coding. We used the Big5 and CP936 encodings. Since the main purpose of this work is to evaluate the proposed subword-based IOB tagging, we carried out the closed test only. Five metrics were used to evaluate the segmentation results: recall (R), precision (P), F-score (F), OOV rate (R-oov) and IV rate (R-iv). For a detailed explanation of these metrics, refer to (Sproat and Emerson, 2003).

| Corpus | Abbrev. | Encodings | Training size (words) | Test size (words) |
|---|---|---|---|---|
| Academia Sinica | AS | Big5/Unicode | 5.45M | 122K |
| Beijing University | PKU | CP936/Unicode | 1.1M | 104K |
| City University of Hong Kong | CITYU | Big5/Unicode | 1.46M | 41K |
| Microsoft Research (Beijing) | MSR | CP936/Unicode | 2.37M | 107K |

Table 1: Corpus statistics in Sighan Bakeoff 2005

### 3.1 Effects of N-gram LMs

We obtained a word list from the training data as the vocabulary for dictionary-based segmentation. N-gram LMs were generated using the SRI LM toolkit. Table 2 shows the performance of N-gram segmentation by changing the order of N-grams.

We found that bigram LMs can improve segmentation over unigram, though we observed no effect from the trigram LMs. For the PKU corpus, there was a relatively strong improvement due to using bigrams rather than unigrams, posssibly because the PKU corpus' training size was smaller than the others. For a sufficiently large training corpus, the unigram LMs may be enough for segmentation. This experiment revealed that language models above bigrams do not improve word segmentation. Since there were some single-character words present in test data but not in the training data, the R-oov rates were not zero in this experiment. In fact, we did not use any OOV detection for the dictionary-based approach.

### 3.2 Comparisons of Character-based and Subword-based tagger

In Section 2.2 we described the character-based and subword-based IOB tagging methods. The main difference between the two is the lexicon subset used for re-segmentation. For the subword-based IOB tagging, we need to add some multiple-character words into the lexicon subset. Since it is hard to decide the optimal number of words to add, we test three different lexicon sizes, as shown in Table 3. The first one, s1, consisting of all the characters, is a character-based approach. The second, s2, added 2,500 top words from the training data to the lexicon of s1. The third, s3, added another 2,500 top words to the lexicon of s2. All the words were among the most frequent in the training corpora. After choosing the subwords, the training data were re-segmented using the subwords by FMM. The final lexicons were collected again, consisting of single-character words and multiple-character words. Table 3 shows the sizes of the final lexicons. Therefore, the minus of the lexicon size of s2 to s1 are not 2,500, exactly.

| | AS | CITYU | MSR | PKU |
|---|---|---|---|---|
| s1 | 6,087 | 4,916 | 5,150 | 4,685 |
| s2 | 8,332 | 7,338 | 7,464 | 7,014 |
| s3 | 10,876 | 9,996 | 9,990 | 9,053 |

Table 3: Three different vocabulary sizes used in subword-based tagging. s1 contains all the characters. s2 and s3 contains some common words.

The segmentation results of using three lexicons are shown in Table 4. The numbers are separated by a "/" in the sequence of "s1/s2/s3." We found although the subword-based approach outperformed the character-based one significantly, there was no obvious difference between the two subword-based approaches, s2 and s3, adding respective 2,500 and 5,000 subwords to s1. The experiments show that we cannot find an optimal lexicon size from 2,500 to 5,000. However, there might be an optimal point less than 2,500. We did not take much effort to find the optimal point, and regarded 2,500 as an acceptable size for practical usages.

The F-scores of IOB tagging shown in Table 4 are better than that of N-gram word segmentation in Table 2, which proves that the IOB tagging is effective in recognizing OOV. However, we found there was a large decrease in the R-ivs, which shows the weakness of the IOB tagging approach. We use the confidence measure approach to deal with this problem in next section.

### 3.3 Effects of the confidence measure

Up to now we had two segmentation results by using the dictionary-based word segmentation and the IOB tagging. In Section 2.3, we proposed a confidence measure approach to re-evaluate the results of IOB tagging by combining the two results. The effects of

| | R | P | F | R-oov | R-iv |
|---|---|---|---|---|---|
| AS | 0.934/0.942/0.941 | 0.884/0.881/0.881 | 0.909/0.910/0.910 | 0.041/0.040/0.038 | 0.975/0.983/0.982 |
| CITYU | 0.924/0.929/0.928 | 0.851/0.851/0.851 | 0.886/0.888/0.888 | 0.162/0.162/0.164 | 0.984/0.990/0.989 |
| PKU | 0.938/0.949/0.948 | 0.909/0.912/0.912 | 0.924/0.930/0.930 | 0.407/0.403/0.408 | 0.971/0.982/0.981 |
| MSR | 0.965/0.969/0.968 | 0.927/0.927/0.927 | 0.946/0.947/0.947 | 0.036/0.036/0.048 | 0.991/0.994/0.993 |

Table 2: Segmentation results of dictionary-based segmentation in closed test of Bakeoff 2005. A "/" separates the results of unigram, bigram and trigram.

| | R | P | F | R-oov | R-iv |
|---|---|---|---|---|---|
| AS | 0.922/0.942/0.943 | 0.914/0.930/0.930 | 0.918/0.936/0.937 | 0.641/0.628/0.609 | 0.935/0.956/0.959 |
| CITYU | 0.906/0.933/0.934 | 0.905/0.929/0.927 | 0.906/0.931/0.930 | 0.668/0.671/0.671 | 0.925/0.954/0.955 |
| PKU | 0.913/0.934/0.936 | 0.922/0.938/0.940 | 0.918/0.936/0.938 | 0.744/0.724/0.713 | 0.924/0.946/0.949 |
| MSR | 0.929/0.953/0.953 | 0.934/0.955/0.952 | 0.932/0.954/0.952 | 0.656/0.684/0.665 | 0.936/0.961/0.961 |

Table 4: Segmentation results by the pure subword-based IOB tagging. The separator "/" divides the results by three lexicon sizes as illustrated in Table 3. The first is character-based (s1), while the other two are subword-based with different lexicons (s2/s3).

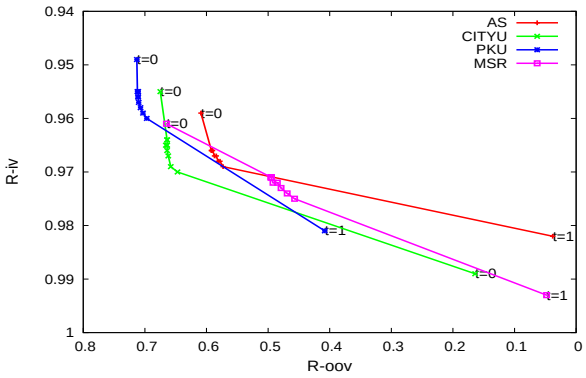


Figure 2: R-iv and R-oov varing as the confidence threshold, $t$.

the confidence measure are shown in Table 5, where we used $\alpha = 0.8$ and confidence threshold $t = 0.7$. These are empirical numbers. We obtained the optimal values by multiple trials on held-out data. The numbers in the slots of Table 5 are divided by a separator "/" and displayed as the sequence "s1/s2/s3", just as Table 4. We found that the results in Table 5 were better than those in Table 4 and Table 2, which proved that using the confidence measure approach yielded the best performance over the N-gram segmentation and the IOB tagging approaches.

Even with the use of the confidence measure, the subword-based IOB tagging still outperformed the character-based IOB tagging, proving that the proposed subword-based IOB tagging was very effective. Though the improvement under the confidence measure was decreasing, it was still significant.

We can change the R-oov and R-iv by changing the confidence threshold. The effect of R-oov and R-iv's varing as the threshold is shown in Fig. 2, where R-oovs and R-ivs are moving in different directions. When the confidence threshold $t = 0$, the case for the IOB tagging, R-oovs are maximal. When $t = 1$, representing the dictionary-based segmentation, R-oovs are the minimal. The R-oovs and R-ivs varied largely at the start and end point but little around the middle section.

### 3.4 Subword-based tagging by CRFs

Our proposed approaches were presented and evaluated using the MaxEnt method in the previous sections. When we turned to CRF-based tagging, we found a same effect as the MaxEnt method. Our subword-based tagging by CRFs was implemented by the package "CRF++" from the site "http://www.chasen.org/~taku/software."

We repeated the previous sections' experiments using the CRF approach except that we did one of the two subword-based tagging, the lexicon size s3. The same values of the confidence measure threshold and $\alpha$ were used. The results are shown in Table 6.

We found that the results using the CRFs were much better than those of the MaxEnts. However, the emphasis here was not to compare CRFs and MaxEnts but the effect of subword-based IOB tagging. In Table 6, the results before "/" are the character-based IOB tagging and after "/", the subword-based. It was clear that the subword-based approaches yielded better results than the character-based approach though the improvement was not as higher as that of the MaxEnt approaches. There was

| | R | P | F | R-oov | R-iv |
|---|---|---|---|---|---|
| AS | 0.938/0.950/0.953 | 0.945/0.946/0.951 | 0.941/0.948/0.948 | 0.674/0.641/0.606 | 0.950/0.964/0.969 |
| CITYU | 0.932/0.949/0.946 | 0.944/0.933/0.944 | 0.938/0.941/0.945 | 0.705/0.597/0.667 | 0.950/0.977/0.968 |
| PKU | 0.941/0.948/0.949 | 0.945/0.947/0.947 | 0.943/0.948/0.948 | 0.672/0.662/0.660 | 0.958/0.966/0.966 |
| MSR | 0.944/0.959/0.961 | 0.959/0.964/0.963 | 0.951/0.961/0.962 | 0.671/0.674/0.631 | 0.951/0.967/0.970 |

Table 5: Effects of combination using the confidence measure. Here we used $\alpha = 0.8$ and confidence threshold $t = 0.7$. The separator "/" divides the results of s1, s2, and s3.

no change on F-score for AS corpus, but a better recall rate was found. Our results are better than the best one of Bakeoff 2005 in PKU, CITYU and MSR corpora.

Detailed descriptions about subword tagging by CRF can be found in our paper (Zhang et al., 2006).

## 4 Discussion and Related works

The IOB tagging approach adopted in this work is not a new idea. It was first implemented in Chinese word segmentation by (Xue and Shen, 2003) using the maximum entropy methods. Later, (Peng and McCallum, 2004) implemented the idea using the CRF-based approach, which yielded better results than the maximum entropy approach because it could solve the label bias problem (Lafferty et al., 2001). However, as we mentioned before, this approach does not take advantage of the prior knowledge of in-vocabulary words; It produced a higher R-oov but a lower R-iv. This problem has been observed by some participants in the Bakeoff 2005 (Asahara et al., 2005), where they applied the IOB tagging to recognize OOVs, and added the OOVs to the lexicon used in the HMM-based or CRF-based approaches. (Nakagawa, 2004) used hybrid HMM models to integrate word level and character level information seamlessly. We used confidence measure to determine a better balance between R-oov and R-iv. The idea of using the confidence measure has appeared in (Peng and McCallum, 2004), where it was used to recognize the OOVs. In this work we used it more than that. By way of the confidence measure we combined results from the dictionary-based and the IOB-tagging-based and as a result, we could achieve the optimal performance.

Our main contribution is to extend the IOB tagging approach from being a character-based to a subword-based one. We proved that the new approach enhanced the word segmentation significantly in all the experiments, MaxEnts, CRFs and using confidence measure. We tested our approach using the standard Sighan Bakeoff 2005 data set in the closed test. In Table 7 we align our results with some top runners' in the Bakeoff 2005.

Our results were compared with the best performers' results in the Bakeoff 2005. Two participants' results were chosen as bases: No.15-b, ranked the first in the AS corpus, and No.14, the best performer in CITYU, MSR and PKU. . The No.14 used CRF-modeled IOB tagging while No.15-b used MaxEnt-modeled IOB tagging. Our results produced by the MaxEnt are denoted as "ours(ME)" while "ours(CRF)" for the CRF approaches. We achieved the highest F-scores in three corpora except the AS corpus. We think the proposed subword-based approach played the important role for the achieved good results.

*A second advantage of the subword-based IOB tagging over the character-based is its speed.* The subword-based approach is faster because fewer words than characters needed to be labeled. We observed a speed increase in both training and testing. In the training stage, the subword approach was almost two times faster than the character-based.

## 5 Conclusions

In this work, we proposed a subword-based IOB tagging method for Chinese word segmentation. The approach outperformed the character-based method using both the MaxEnt and CRF approaches. We also successfully employed the confidence measure to make a confidence-dependent word segmentation. By setting the confidence threshold, R-oov and R-iv can be changed accordingly. This approach is effective for performing desired segmentation based on users' requirements to R-oov and R-iv.

| | R | P | F | R-oov | R-iv |
|---|---|---|---|---|---|
| AS | 0.953/0.956 | 0.944/0.947 | 0.948/0.951 | 0.607/0.649 | 0.969/0.969 |
| CITYU | 0.943/0.952 | 0.948/0.949 | 0.946/0.951 | 0.682/0.741 | 0.964/0.969 |
| PKU | 0.942/0.947 | 0.957/0.955 | 0.949/0.951 | 0.775/0.748 | 0.952/0.959 |
| MSR | 0.960/0.972 | 0.966/0.969 | 0.963/0.971 | 0.674/0.712 | 0.967/0.976 |

Table 6: Effects of using CRF. The separator "/" divides the results of s1, and s3.

| Participants | R | P | F | R-oov | R-iv |
|---|---|---|---|---|---|
| Hong Kong City University | | | | | |
| ours(CRF) | 0.952 | 0.949 | 0.951 | 0.741 | 0.969 |
| ours(ME) | 0.946 | 0.944 | 0.945 | 0.667 | 0.968 |
| 14 | 0.941 | 0.946 | 0.943 | 0.698 | 0.961 |
| 15-b | 0.937 | 0.946 | 0.941 | 0.736 | 0.953 |
| Academia Sinica | | | | | |
| 15-b | 0.952 | 0.951 | 0.952 | 0.696 | 0.963 |
| ours(CRF) | 0.956 | 0.947 | 0.951 | 0.649 | 0.969 |
| ours(ME) | 0.953 | 0.943 | 0.948 | 0.608 | 0.969 |
| 14 | 0.95 | 0.943 | 0.947 | 0.718 | 0.960 |
| Microsoft Research | | | | | |
| ours(CRF) | 0.972 | 0.969 | 0.971 | 0.712 | 0.976 |
| 14 | 0.962 | 0.966 | 0.964 | 0.717 | 0.968 |
| ours(ME) | 0.961 | 0.963 | 0.962 | 0.631 | 0.970 |
| 15-b | 0.952 | 0.964 | 0.958 | 0.718 | 0.958 |
| Peking University | | | | | |
| ours(CRF) | 0.947 | 0.955 | 0.951 | 0.748 | 0.959 |
| 14 | 0.946 | 0.954 | 0.950 | 0.787 | 0.956 |
| ours(ME) | 0.949 | 0.947 | 0.948 | 0.660 | 0.966 |
| 15-b | 0.93 | 0.951 | 0.941 | 0.76 | 0.941 |

Table 7: List of results in Sighan Bakeoff 2005

## Acknowledgements

The authors thank the reviewers for the comments and advice on the paper. Some related software for this work will be released very soon.

## References

Masayuki Asahara, Kenta Fukuoka, Ai Azuma, Chooi-Ling Goh, Yotaro Watanabe, Yuji Matsumoto, and Takashi Tsuzuki. 2005. Combination of machine learning methods for optimum chinese word segmentation. In *Forth SIGHAN Workshop on Chinese Language Processing, Proceedings of the Workshop*, pages 134–137, Jeju, Korea.

Thomas Emerson. 2005. The second international chinese word segmentation bakeoff. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, Jeju, Korea.

Jianfeng Gao, Andi Wu, Mu Li, Chang-Ning Huang, Hongqiao Li, Xinsong Xia, and Haowei Qin. 2004. Adaptive chinese word segmentation. In *ACL-2004*, Barcelona, July.

Frederick Jelinek. 1998. *Statistical methods for speech recognition*. the MIT Press.

Taku Kudo and Yuji Matsumoto. 2001. Chunking with support vector machine. In *Proc. of NAACL-2001*, pages 192–199.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proc. of ICML-2001*, pages 591–598.

Tetsuji Nakagawa. 2004. Chinese and japanese word segmentation using word-level and character-level information. In *Proceedings of Coling 2004*, pages 466–472, Geneva, August.

Fuchun Peng and Andrew McCallum. 2004. Chinese segmentation and new word detection using conditional random fields. In *Proc. of Coling-2004*, pages 562–568, Geneva, Switzerland.

Richard Sproat and Tom Emerson. 2003. The first international chinese word segmentation bakeoff. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, Sapporo, Japan, July.

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning. 2005. A conditional random field word segmenter for Sighan bakeoff 2005. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, Jeju, Korea.

Nianwen Xue and Libin Shen. 2003. Chinese word segmentation as LMR tagging. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*.

Huaping Zhang, HongKui Yu, Deyi xiong, and Qun Liu. 2003. HHMM-based Chinese lexical analyzer ICTCLAS. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, pages 184–187.

Ruiqiang Zhang, Genichiro Kikui, and Eiichiro Sumita. 2006. Subword-based tagging by conditional random fields for chinese word segmentation. In *Proc. of HLT-NAACL*.