Subword-based Tagging by Conditional Random Fields for 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 two approaches to improve Chinese word segmentation: a subword-based tagging and a confidence measure approach. We found the former achieved better performance than the existing character-based tagging, and the latter improved segmentation further by combining the former with a dictionary-based segmentation. In addition, the latter can be used to balance out-of-vocabulary rates and in-vocabulary rates. By these techniques we achieved higher F-scores in CITYU, PKU and MSR corpora than the best results from Sighan Bakeoff 2005.

1 Introduction

The character-based "IOB" tagging approach has been widely used in Chinese word segmentation recently (Xue and Shen, 2003; Peng and McCallum, 2004; Tseng et al., 2005). Under the scheme, each character of a word is labeled as 'B' if it is the first character of a multiple-character word, or 'O' if the character functions as an independent word, or 'I' otherwise." For example, "全(whole) 北京市(Beijing city)" is labeled as "全(whole)/O 北(north)/B 京(capital)/I 市(city)/I".

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 into a character-based one. Taking the same example mentioned above, "全(whole) 北京市(Beijing city)" is labeled as "全(whole)/O 北京(Beijing)/B 市(city)/I" in the subword-based tagging, where "北京(Beijing)/B" is labeled as one unit. We will give a detailed description of this approach in Section 2.

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

In addition, we found a clear weakness with the IOB tagging approach: It yields a very low in-vocabulary (IV) 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 conditional random fields (CRF) for the IOB tagging, yielded very high R-oovs 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 the weakness. By this approach we can change R-oovs and R-ivs and find an optimal tradeoff. This approach will be described in Section 2.2.

In the followings, we illustrate our word segmentation process in Section 2, where the subword-based tagging is implemented by the CRFs method. Section 3 presents our experimental results. Section 4 describes current state-of-the-art methods for Chinese word segmentation, with which our results were compared. Section 5 provides the concluding remarks.

2 Our Chinese word segmentation process

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 subword-based tagging by the CRF for recognizing OOVs, and a confidence-dependent word segmentation used for merging the results of both the dictionary-based and the IOB tagging. An example exhibiting each step's results is also given in the figure.

Since the dictionary-based approach is a well-known method, we skip its technical descriptions. However, keep in mind that the dictionary-based approach can produce a higher R-iv rate. We will use this advantage in the confidence measure approach.

2.1 Subword-based IOB tagging using CRFs

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

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 193–196,
New York, June 2006.

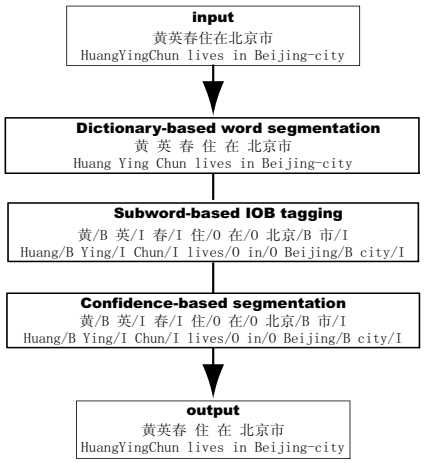

Figure 1: Outline of word segmentation process

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 belonging to the subset, and assigned IOB tags to them. For a character-based IOB tagger, there is only one possibility of re-segmentation. However, there are multiple choices for a 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. Of course, backward maximal match (BMM) or other approaches are also applicable. We did not conduct comparative experiments because trivial differences of these approaches may not result in significant consequences to the subword-based approach.

In the third step, we used the CRFs approach to train the IOB tagger (Lafferty et al., 2001) on the training data. We downloaded and used the package "CRF++" from the site "http://www.chasen.org/taku/software." According to the CRFs, the probability of an IOB tag sequence, $T = t_0 t_1 \cdots t_M$, given the word sequence, $W = w_0 w_1 \cdots w_M$, is defined by

$$p(T|W) = \exp\left(\sum_{i=1}^{M}\left(\sum_k \lambda_k f_k(t_{i-1}, t_i, W) + \sum_k \mu_k g_k(t_i, W)\right)\right)/Z, \quad (1)$$
$$Z = \sum_{T=t_0 t_1 \cdots t_M} p(T|W)$$

where we call $f_k(t_{i-1}, t_i, W)$ bigram feature functions because the features trigger the previous observation t_{i-1} and current observation t_i simultaneously; $g_k(t_i, W)$, the unigram feature functions because they trigger only current observation t_i. λ_k and μ_k are the model parameters corresponding to feature functions f_k and g_k respectively.

The model parameters were trained by maximizing the log-likelihood of the training data using L-BFGS gradient descent optimization method. In order to overcome overfitting, a gaussian prior was imposed in the training.

The types of unigram features used in our experiments included the following types:

$$w_0, w_{-1}, w_1, w_{-2}, w_2, w_0 w_{-1}, w_0 w_1, w_{-1} w_1, w_{-2} w_{-1}, w_2 w_0$$

where w stands for word. The subscripts are position indicators. 0 means the current word; $-1, -2$, the first or second word to the left; $1, 2$, the first or second word to the right.

For the bigram features, we only used the previous and the current observations, $t_{-1} t_0$.

As to feature selection, we simply used absolute counts for each feature in the training data. We defined a cutoff value for each feature type and selected the features with occurrence counts over the cutoff.

A forward-backward algorithm was used in the training and viterbi algorithm was used in the decoding.

2.2 Confidence-dependent word segmentation

Before moving to this step in Figure 1, 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 results with higher R-ivs but lower R-oovs 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:

$$CM(t_{iob}|w) = \alpha CM_{iob}(t_{iob}|w) + (1 - \alpha)\delta(t_w, t_{iob})_{ng} \quad (2)$$

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, is defined as

$$CM_{iob}(t|w_i) = \frac{\sum_{T=t_0 t_1 \cdots t_M, t_i=t} P(T|W, w_i)}{\sum_{T=t_0 t_1 \cdots t_M} P(T|W)}$$

where the numerator is a sum of all the observation sequences with word w_i labeled as t.

$\delta(t_w, t_{iob})_{ng}$ denotes the contribution of the dictionary-based segmentation. It 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. 2, α is a weighting between the IOB tagging and the dictionary-based word segmentation. We found the value 0.7 for α, empirically.

By Eq. 2 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 a real application, a satisfactory tradeoff between R-ivs and R-oovs could find through tuning the confidence threshold. In Section 3.2 we will present the experimental segmentation results of the confidence measure approach.

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 (AS), City University of Hong Kong (CITYU), Peking University (PKU) and Microsoft Research in Beijing (MSR). Since this work was to evaluate the proposed subword-based IOB tagging, we carried out the closed test only. Five metrics were used to evaluate segmentation results: recall(R), precision(P), F-score(F), OOV rate(R-oov) and IV rate(R-iv). For detailed info. of the corpora and these scores, refer to (Emerson, 2005).

For the dictionary-based approach, we extracted a word list from the training data as the vocabulary. Trigram LMs were generated using the SRI LM toolkit for disambiguation. Table 1 shows the performance of the dictionary-based segmentation. Since there were some single-character words present in the test data but not in the training data, the R-oov rates were not zero in this experiment. In fact, there were no OOV recognition. Hence, this approach produced lower F-scores. However, the R-ivs were very high.

3.1 Effects of the Character-based and the subword-based tagger

The main difference between the character-based and the word-based is the contents of the lexicon subset used for re-segmentation. For the character-based tagging, we used all the Chinese characters. For the subword-based tagging, we added another 2000 most frequent multiple-character words to the lexicons for tagging. The segmentation results of the dictionary-based were re-segmented using the FMM, and then labeled with "IOB" tags by the CRFs. The segmentation results using CRF tagging are shown in Table 2, where the upper numbers of each slot were produced by the character-based approach while the lower numbers were of the subword-based. We found that the proposed subword-based approaches were effective in CITYU and MSR corpora, raising the F-scores from 0.941 to 0.946 for CITYU corpus, 0.959 to 0.964 for MSR corpus. There were no F-score changes for AS and PKU corpora, but the recall rates were improved. Comparing Table 1 and 2, we found the CRF-modeled IOB tagging yielded better segmentation than the dictionary-based approach. However, the R-iv rates were getting worse in return for higher R-oov rates. We will tackle this problem by the confidence measure approach.

	R	P	F	R-oov	R-iv
AS	0.941	0.881	0.910	0.038	0.982
CITYU	0.928	0.851	0.888	0.164	0.989
PKU	0.948	0.912	0.930	0.408	0.981
MSR	0.968	0.927	0.947	0.048	0.993

Table 1: Our segmentation results by the dictionary-based approach for the closed test of Bakeoff 2005, very low R-oov rates due to no OOV recognition applied.

	R	P	F	R-oov	R-iv
AS	0.951	0.942	0.947	0.678	0.964
	0.953	0.940	0.947	0.647	0.967
CITYU	0.939	0.943	0.941	0.700	0.958
	0.950	0.942	0.946	0.736	0.967
PKU	0.940	0.950	0.945	0.783	0.949
	0.943	0.946	0.945	0.754	0.955
MSR	0.957	0.960	0.959	0.710	0.964
	0.965	0.963	0.964	0.716	0.972

Table 2: Segmentation results by a pure subword-based IOB tagging. The upper numbers are of the character-based and the lower ones, the subword-based.

3.2 Effect of the confidence measure

In section 2.2, we proposed a confidence measure approach to re-evaluate the results of IOB tagging by combinations of the results of the dictionary-based segmentation. The effect of the confidence measure is shown in Table 3, where we used $\alpha = 0.7$ and confidence threshold $t = 0.8$. In each slot, the numbers on the top were of the character-based approach while the numbers on the bottom were the subword-based. We found the results in Table 3 were better than those in Table 2 and Table 1, which prove that using confidence measure approach achieved the best performance over the dictionary-based segmentation and the IOB tagging approach. The act of confidence measure made a tradeoff between R-ivs and R-oovs, yielding higher R-oovs than Table 1 and higher R-

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

Table 3: Effects of combination using the confidence measure. The upper numbers and the lower numbers are of the character-based and the subword-based, respectively

	AS	CITYU	MSR	PKU
Bakeoff-best	0.952	0.943	0.964	0.950
Ours	0.951	0.951	0.971	0.951

Table 4: Comparison our results with the best ones from Sighan Bakeoff 2005 in terms of F-score

ivs than Table 2.

Even with the use of confidence measure, the word-based IOB tagging still outperformed the character-based IOB tagging. It proves the proposed word-based IOB tagging was very effective.

4 Discussion and Related works

The IOB tagging approach adopted in this work is not a new idea. It was first used in Chinese word segmentation by (Xue and Shen, 2003), where maximum entropy methods were used. Later, this approach was implemented by the CRF-based method (Peng and McCallum, 2004), which was proved to achieve better results than the maximum entropy approach because it can solve the label bias problem (Lafferty et al., 2001).

Our main contribution is to extend the IOB tagging approach from being a character-based to a subword-based. We proved the new approach enhanced the word segmentation significantly. Our results are listed together with the best results from Bakeoff 2005 in Table 4 in terms of F-scores. We achieved the highest F-scores in CITYU, PKU and MSR corpora. We think our proposed subword-based tagging played an important role for the good results. Since it was a closed test, some information such as Arabic and Chinese number and alphabetical letters cannot be used. We could yield a better results than those shown in Table 4 using such information. For example, inconsistent errors of foreign names can be fixed if alphabetical characters are known. For AS corpus, "Adam Smith" are two words in the training but become a one-word in the test, "AdamSmith". Our approaches produced wrong segmentations for labeling inconsistency.

Another advantage of the word-based IOB tagging over the character-based is its speed. The subword-based approach is faster because fewer words than characters were labeled. We found a speed up both in training and test.

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 delicately. 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.

5 Conclusions

In this work, we proposed a subword-based IOB tagging method for Chinese word segmentation. Using the CRFs approaches, we prove that it outperformed the character-based method using the CRF approaches. We also successfully employed the confidence measure to make a confidence-dependent word segmentation. This approach is effective for performing desired segmentation based on users' requirements to R-oov and R-iv.

Acknowledgements

The authors appreciate the reviewers' effort and good advice for improving the paper.

References

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

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.

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.

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*.