Effective Tag Set Selection in Chinese Word Segmentation via Conditional Random Field Modeling

Hai Zhao[1,2⋆], Chang-Ning Huang[2], Mu Li[2], and Bao-Liang Lu[1]

[1] Department of Computer Science and Engineering, Shanghai Jiao Tong University, 1954 Hua Shan Rd., Shanghai 200030, China
{zhaohai,blu}@cs.sjtu.edu.cn,
[2] Microsoft Research Asia,
49, Zhichun Road, Haidian District,
Beijing, China, 100080
cnhuang@msrchina.research.microsoft.com, muli@microsoft.com

Abstract. This paper is concerned with Chinese word segmentation, which is regarded as a character based tagging problem under conditional random field framework. It is different in our method that we consider both feature template selection and tag set selection, instead of feature template focused only method in existing work. Thus, there comes an empirical comparison study of performance among different tag sets in this paper. We show that there is a significant performance difference as different tag sets are selected. Based on the proposed method, our system gives the state-of-the-art performance.

1 Introduction

Chinese text is written without natural delimiters, so word segmentation is an essential first step in Chinese language processing. In this aspect, Chinese is quite different from English in which sentences of words delimited by white spaces. Though it seems very simple, Chinese word segmentation is not a trivial problem. Actually, it has been active area of research in computational linguistics for almost 20 years and has drawn more and more attention in the Chinese language processing community. To accomplish such a task, various technologies are developed [1].

To give a comprehensive comparison of Chinese segmentation on common test corpora, two International Chinese Word Segmentation Bakeoffs were held in 2003 and 2005, and there were 12 and 23 participants respectively [4], [5].

In all of proposed methods, character based tagging method [2] quickly rose in two Bakeoffs as a remarkable one with state-of-the-art performance. As reported in [5], the results of Bakeoff-2005 shows a general trend to a decrease in error rates from 3.9% to 2.8% compared to the results of Bakeoff-2003. Especially, two participants, Ng and Tseng, gave the best results in almost all test corpora [6], [7].

We continue to improve CRF-based tagging method of Chinese word segmentation on the track of Ng and Tseng in this study. It is different in our method that we consider

⋆ This work was finished while the first author visited Microsoft Research Asia

both feature template selection and tag set selection, instead of feature template focused only methods in previous work. That is, feature template selection was the main work if it was not the unique one before, while tag set is empirically specified aforehand.

There are two kinds of test schemes in Chinese word segmentation Bakeoff, open and closed test. In the open test participants are allowed to use training data and any other linguistic resource including other training corpora, proprietary dictionaries and so forth. In the closed test only training data are allowed to used for the particular corpus. No other data is allowed. In this study, we will limit our comparison in closed test because additional linguistical resource often varies from system to system.

The remainder of the paper is organized as follows. The next section is a simple introduction to conditional random field. Feature templates and tag sets are given in Section 3. In Section 4, our experimental results are demonstrated. We summarize our contribution in Section 5.

2 Conditional Random Field

Maximum entropy tagger was used in early character-based tagging for Chinese word segmentation [2], [3], while we choose linear-chain CRF as our learning model in this study. It can combine rich feature representation and probabilistic finite state model, too. In addition, it can avoid so-called 'label-bias' problem in some degree. Actually, such model was also proved to be very effective in many existing works [8].

Conditional random field (CRF) is a statistical sequence modeling framework first introduced into language processing in [9]. Work by Peng et al. first used this framework for Chinese word segmentation by treating it as a binary decision task, such that each Chinese character is labeled either as the beginning of a word or not.

The probability assigned to a label sequence for a particular sequence of characters by a CRF is given by the equation below:

$$p_\lambda(Y|W) = \frac{1}{Z(W)} \exp(\sum_{t \in T} \sum_{k} \lambda_k f_k(y_{t-1}, y_t, W, t)) \tag{1}$$

where $Y = \{y_i\}$ is the label sequence for the sentence, W is the sequence of unsegmented characters, $Z(W)$ is a normalization term, f_k is a feature function, and t indexes into characters in the label sequence.

3 Tag Sets and Feature Templates

Character based tagging method for Chinese segmentation, either based on maximum entropy or CRF, regards a segmentation procedure as tagging, which is described in detail in [10].

The probability model and corresponding feature function is defined over the set $H \times T$, where H is the set of possible contexts (or any predefined condition) and T is the set of possible tags. Generally, a feature function can be defined as follows,

$$f(h, t) = \begin{cases} 1, \text{ if } h = h_i \text{ is satisfied and } t = t_j \\ 0, \text{ otherwise,} \end{cases} \tag{2}$$

where $h_i \in H$ and $t_i \in T$.

For convenience, features are generally organized by some groups, which used to be called feature templates. For example, a bigram template C_1 stands for the next character occurring in the corpus after each character .

A feature template set we selected is shown in Table 1. We give an explanation to feature template $T_{-1}T_0T_1$. Here, T_{-1}, T_0 or T_1 stands for predefined class. There are four classes defined: numbers represent class 1, those characters whose meanings are dates represent class 2, English letters represent class 3, and other characters represent class 4. This feature template is improved from the corresponding one in [6]. Refer it for more details.

Table 1. Feature templates

Type	Feature	Function
Unigram	C_{-1}, C_0, C_1	The previous, current and next character
Bigram	$C_{-1}C_0, C_0C_1$	The previous (next) character and current character
	$C_{-1}C_1$	The previous character and next character
Punctuation	$Pu(C_0)$	Current character is a punctuation
Date, Digital and Letter	$T_{-1}T_0T_1$	Types of previous, current and next character

Table 2. Definitions of different tag sets

Tag set	Tags	Words in tagging
2-tag (Peng/Tseng)	B, E	B, BE, BEE, ...
4-tag (Xue/Ng)	B, M, E, S	S, BE, BME, BMME, ...
5-tag	B, B_2, M, E, S	$S, BE, BB_2E, BB_2ME, BB_2MME$, ...
6-tag	B, B_2, B_3, M, E, S	$S, BE, BB_2E, BB_2B_3E, BB_2B_3ME$, ...

As for tag set, there are two kinds of schemes are used to distinguish the character position in a word in the previous works. The detail can be referred to Table 2. Notice Xue and Ng use their 4-tag set in maximum entropy model. Two reported CRF methods, i.e., Peng and Tseng, used 2-tag set. Generally speaking, activated feature functions in practice like (2) are determined by both feature template and tag set. In the existing work, tag set is specified aforehand. Ng and Xue used a maximum entropy tagger with the same 4-tag set, while Peng and Tseng used CRF. CRF is a sequence learner, and word segmentation is often regarded a binary decision procedure. Thus they chose 2-

tag set. We will see tag set selection is as important as feature template set selection in the empirical comparison.

To effectively perform tagging for those longer words, we extend 4-tag set of Ng/Xue. The tag 'B_2' is added into 4-tag set to form 5-tag set, in which stands for the second character position in a Chinese word. Similarly, the tag 'B_3' is added into 5-tag set to form 6-tag set, in which stands for the third character position in a Chinese word. This extension is based on our observation of average weighted word length distribution in each corpus. We will give further discussions in Section 4.

4 Experiments

Eight corpora are available from Bakeoff-2003 and 2005. We use all of them to perform the evaluation. A summary of these corpora is shown in Table 3.

Table 3. Corpora statistics of Bakeoff-2003 and 2005

Provider	Corpus	Encoding	#Training words	#Test words	OOV rate
Academia Sinica	AS2003	Big5	5.8M	12K	0.022
	AS2005	Big5	5.45M	122K	0.043
Hong Kong City University	CityU2003	Big5	240K	35K	0.071
	CityU2005	Big5	1.46M	41K	0.074
U. Penn Chinese Treebank	CTB2003	GB	250K	40K	0.181
Beijing University	PKU2003	GB	1.1M	17K	0.069
	PKU2005	GB	1.1M	104K	0.058
Microsoft Research Asia	MSRA2005	GB	2.37M	107K	0.026

4.1 Experimental Results under Different Tag Sets

To observe the trends of performance under different feature template sets and different tag sets, we define another two feature template sets in addition to the feature template set defined in Table 3. Their definitions are shown in Table 4. The experimental results of CityU2003 and PKU2005 are shown in Table 5. From the experimental results, we can get a general trend of performance increasing from 2-tag to 6-tag set.

As we have shown above, TMPT-01 is the template set which gives the best performance when it combines with 6-tag set. TMPT-02 is a pure n-gram template set. It can be observed that TMPT-02 gives a substantial performance improvement when it works with 6-tag set. On the contrary, the best set, TMPT-01 will lose its top place when it works with 2-tag set.

Table 4. Four feature template sets

ID	Description
TMPT-01	Defined in Table 1.
TMPT-02	Add C_{-2}, C_2, and remove $Pu, T_{-1}T_0T_1$ in TMPT-01
TMPT-03	Add C_{-2} in TMPT-01
TMPT-04	Remove Pu and $T_{-1}T_0T_1$ in TMPT-01

Table 5. Experimental results of CityU2003 and PKU2005 under different feature template sets and tag sets

Feature Set	CityU2003				PKU2005			
	2-tag	4-tag	5-tag	6-tag	2-tag	4-tag	5-tag	6-tag
TMPT-01	0.9320	0.9472	0.9478	0.9483	0.9505	0.9522	0.9531	0.9536
TMPT-02	0.9302	0.9450	0.9461	0.9462	0.9461	0.9479	0.9499	0.9503
TMPT-03	0.9324	0.9471	0.9464	0.9467	0.9493	0.9499	0.9523	0.9526

There is another comparison between our system and Tseng's: we propose 8 groups of feature templates shown in Table 1, while there are 15 groups of selected feature templates in Tseng's system. However, with a selected appropriate tag set, our system gives its superior performance.

4.2 Comparisons of Best Existing Results and Our Results

The comparison between our results and best existing results are shown in Table 6[3]. There are two types of existing results. One is the best F score of Bakeoff-2003 and 2005 for each corpus under closed test. The other is the results of Tseng et al.. All of our results are performed under TMPT-01 (or TMPT-04[4]) and 6-tag set.

[3] The Third SIGHAN Chinese Language Processing Bakeoff has been held, the results will be presented at the 5th SIGHAN Workshop, to be held at ACL-COLING 2006 in Sydney, Australia, July 22-23, 2006. We also participate this Bakeoff, and our system with the techniques presented in this paper won four highest and two third highest F measures in six Chinese word segmentation tracks. Our results on Bakeoff-2006 appear in SIGHAN-2006 [11].

[4] Note that TMPT-04 is a feature template set only including n-gram ones. Researchers in CWS did not make an agree on what P and $T_{-1}T_0T_1$ are feature templates for closed test or not. Thus the results under TMPT-04 are demonstrated, too. We will see that our system gets state-of-the-art performance in either of feature template set.

Table 6. Comparisons of best existing results and our results in the corpora of Bakeoff-2003 and 2005

Participant	AS2003	CTB2003	CityU2003	PKU2003	AS2005	CityU2005	PKU2005	MSRA2005
Peng	0.956	0.849	0.928	0.941				
Tseng	0.970	0.863	0.947	0.953	0.947	0.943	0.950	0.964
Best of Bakeoff	0.961	0.881	0.940	0.951	0.952	0.943	0.95	0.964
Ours/TMPT-01	0.973	0.873	0.948	0.956	0.954	0.956	0.954	0.974
Ours/TMPT-04	0.973	0.872	0.947	0.956	0.953	0.948	0.952	0.974

4.3 Determine Effective Tag Sets

How to select an effective tag set for current segmentation task is an interesting problem. Since our task is to segment sequence into various length words, it is natural for us to analyze the word length distribution in a corpus.

Average weighted word length distribution of eight concerned training corpora are shown in Table 7. The length of a Chinese word is measured by the number of including characters. The numerical values in Table 7 are calculated through dividing the sum of lengths of words with specified length by the number of all words in the corpus. We may notice that average weighted word length of MSRA2005 and CTB2003 are the longest. As for CityU2005, though it is not the shortest one, the average length of all words which are longer than three-character is the shortest. This is what we are more interested in, since our tag set trends to extend the case with more tags.

Notice 6-tag set can label a five-character word without repeating its tags, that is, 'BB_2B_3ME', we may take average weighted word length of those words whose character lengths are larger than four as our experimental criteria to determine if 6-tag set should be taken or not. For example, we may calculate a value through dividing the sum of lengths of those words whose lengths are larger than four characters by the number of all words in corpus. The threshold is empirically set to 0.02 for our current task. If the obtained value is larger than the threshold, then we adopt 6-tag set, otherwise, we adopt a tag set with five or less tags. Comparison results are shown in Table 9. The experimental results show that CityU2005 corpus with shortest average weighted word length among eight corpora gets its higher performance at 5-tag set instead of 6-tag set though the difference is slight, while MSRA2005 and CTB2003 corpora with the longest word lengths win the most performance improvement from 5-tag to 6-tag set.

We have not discussed the tag set with more than six tags until now. The reason is still behind word length statistics. The distribution of words with different lengths in eight training corpora are shown in Table 8. We can see that the length of 99.89% words at least in all corpora are less than six characters. This also partially explains why 6-tag set works well in most cases. Actually, we tried tag sets with more than 6 tags, but we did not obtain obvious improved performance in almost all cases.

Table 7. Average weighted word length distribution of eight training corpora

Word Length	AS2003	AS2005	CTB2003	CityU2003	CityU2005	PKU2003	PKU2005	MSRA2005
Total	1.5458	1.5089	1.7016	1.6130	1.6275	1.6429	1.6455	**1.7101**
≥ 2	1.001	0.9378	1.2649	1.1190	1.1586	1.1708	1.1728	**1.2401**
≥ 3	0.2135	0.1804	0.3211	0.2648	*0.2479*	0.2692	0.2730	**0.3619**
≥ 4	0.0747	0.0730	0.1195	0.0887	*0.0688*	0.1208	0.1244	**0.2193**
≥ 5	0.0320	0.0334	0.0732	0.0252	*0.0150*	0.0390	0.0423	**0.1223**
≥ 6	0.0228	0.0241	0.0351	0.0133	*0.0072*	0.0105	0.0142	**0.0776**

Table 8. The distribution of words with different lengths in eight training corpora

Word Length	AS2003	AS2005	CTB2003	CityU2003	CityU2005	PKU2003	PKU2005	MSRA2005
≤ 5	0.9973	0.9974	0.9950	0.9981	0.9990	0.9985	0.9983	0.9899
6	0.0008	0.0007	0.0024	0.0010	0.0005	0.0007	0.0007	0.0037
≥ 7	0.0019	0.0019	0.0026	0.0009	0.0006	0.0008	0.0010	0.0063

Table 9. Relation between tag set and average weighted word length under feature template set TMPT-01

Participant	CTB2003	CityU2003	CityU2005	PKU2003	PKU2005	MSRA2005
6-tag	0.8727	0.9483	0.9563	0.9559	0.9536	0.9737
5-tag	0.8715	0.9478	0.9567	0.9554	0.9531	0.9724
Difference	**0.0012**	0.0005	*-0.0004*	0.0005	0.0005	**0.0013**
average weighted word length (Length≥ 5)	**0.0732**	0.0252	*0.0150*	0.0390	0.0423	**0.1223**

5 Conclusion

In this work, we have shown an appropriate tag set can give a substantial performance improvement of Chinese word segmentation for character based tagging method under CRF framework. Furthermore, we propose that average weighted word length distribution of the corpus can be taken as the criteria to choose tag set. Based on the proposed method, our system obtains state-of-the-art performance in all corpora of Bakeoff-2003 and 2005.

References

1. Jianfeng Gao, Mu Li, Andi Wu and Chang-Ning Huang. 2005. Chinese Word Segmentation and Named Entity Recognition: A Pragmatic Approach. *Computational Linguistics*, Vol. 31(4): 531-574.
2. Nianwen Xue. 2003. Chinese Word Segmentation as Character Tagging. *Computational Linguistics and Chinese Language Processing*, Vol. 8(1): 29-48.
3. Nianwen Xue and Libin Shen. 2003. Chinese Word Segmentation as LMR Tagging. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, in conjunction with ACL'03*, 176-179. Sapporo, Japan.
4. Richard Sproat and Thomas Emerson. 2003. The First International Chinese Word Segmentation Bakeoff. *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, 133-143. Sapporo, Japan.
5. Thomas Emerson. 2005. The Second International Chinese Word Segmentation Bakeoff. *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, 123-133. Jeju Island, Korea.
6. Jin Kiat Low, Hwee Tou Ng and Wenyuan Guo 2005. A Maximum Entropy Approach to Chinese Word Segmentation. *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, 161-164. Jeju Island, Korea.
7. Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, Christopher Manning. 2005. A Conditional Random Field Word Segmenter for Sighan Bakeoff 2005. *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, 168-171. Jeju Island, Korea.
8. Fuchun Peng, Fangfang Feng and Andrew McCallum. 2004. Chinese Segmentation and New Word Detection Using Conditional Random Fields. In *COLING 2004*, 562-568. Geneva, Switzerland, August 23-27, 2004.
9. John Lafferty, A. McCallum and F. Pereira. 2001. Conditional Random Field: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, 282-289. June 28-July 01, 2001.
10. Adwait Ratnaparkhi. 1996. A Maximum Entropy Part-of-speech Tagger. In *Proceedings of the Empirical Method in Natural Language Processing Conference*, 133-142. University of Pennsylvania.
11. Hai Zhao, Chang-Ning Huang and Mu Li. 2006. An Improved Chinese Word Segmentation System with Conditional Random Field. *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, 108-117. Sidney,Australia.