

# Chinese Word Segmentation Based on Large Margin Methods<sup>1</sup>

**Buzhou Tang, Xuan Wang, Xiaolong Wang**

Harbin Institute of Technology Shenzhen Graduate School

Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, 518055, China  
tangbuzhou@gmail.com wangxuan@insun.hit.edu.cn wangxl@insun.hit.edu.cn

---

## Abstract

*Chinese Word segmentation is the initial step for Chinese languages processing tasks, which transforms a Character string into a word sequence. Popular approaches to Chinese word segmentation treats Chinese Word Segmentation as a sequence labeling problem tagging each character in the sentence whether it is a single character word or the start, middle or end of a multi-character word. Lots of Discriminative models have been widely applied to this problem. In this paper, the large margin methods, which combine the advantages of two typical state-of-the-art methods, Support vector machines (SVMs) and Conditional Random Fields (CRFs), are presented for Chinese Word Segmentation by Character-based tagging. Their performances are also compared with the performance of SVMs and CRFs and other systems in the literature. Closed test on the PKU, MSR and HK CityU corpus of Sighan Bakeoff-2005 shows some performance improvements over SVMs, CRFs, and our system is competitive with the other best system in the literature.*

## Keywords:

*Chinese Word segmentation; Large Margin Methods; Support Vector Machines; Conditional Random Fields;*

---

---

<sup>1</sup> This research has been partially supported by the National Natural Science Foundation of China (No.60435020 and No.90612005), National 863 Program of China (No. 2007AA01Z194) and the Goal-oriented Lessons from the National 863 Program of China (No.2006AA01Z197).

## 1. Introduction

Words are basic semantic and syntactic units in natural language. Unlike English, there is no separator to mark word boundaries in Chinese and other Asian languages. Therefore, word segmentation is the initial step for most Asian languages processing tasks, which transforms a Character string into a word sequence.

In the case of Chinese, there are two main problems: segmentation ambiguities and unknown words (also called Out-Of-Vocabulary words, OOV) occurrences to affect the performance of Chinese Word Segmentation systems. Several models have been proposed in the literature, which can be classified into three categories. The first one is made up of simple pattern-based methods that make use of lexical features such as Forward Maximum Match (FMM) (Chen, 1999), Backward Maximum Match (BMM) (Liang, 1987) and Bi-Directed Maximum Match (BDMM) (Yaodong Chen, 2005). The second category is made up of rule-based methods that make use of syntactic relations between words. Fewest Words Match (FWM) (Xiaolong Wang, 1989) is a typical rule-based approach that makes the word number of the sequence fewest. Hockenmaier (Hockenmaier, 1998) used transformation-based error-driven learning. The third category is made up of statistical methods that extract the potential language features from training corpus. The typical language models are n-gram and Hidden Markov Model (HMM). Zhang (2003) used a hierarchical hidden Markov Model to incorporate lexical knowledge. Gao (2003) used a class-based model to refine words. All methods above are dictionary-based, which find the best word combination of a sentence according to rules and probabilities. After the seminal work of Xue (2003), which treats Chinese Word Segmentation as a sequence labeling problem tagging each character in the sentence whether it is a single character word or the start, middle or end of a multi-character word, word boundary tagging approach has been adopted currently. Character-based tagging and word-based tagging are two word boundary tagging approaches. Further, some complex language models such as Maximum Entropy (ME) (Xue, 2003) and Conditional Random Fields (CRFs) (Peng, 2004; Zhang, 2006) are also used for Chinese Word Segmentation. In addition, there are some other machine learning methods, such as Support Vector Machines (SVMs) (Goh, 2005) and Perceptron algorithm (Zhang, Y., 2007) introduced into this field. The experiments show that SVMs and CRFs are two typical state-of-the-art methods used for Chinese Word Segmentation. SVMs are non-linear discriminant classifiers for multi-classes problems, which can deal with high dimensional features with good generalization performance. On the other hand, CRFs are discriminative language models for sequence labeling, which can made full use

of correlations between neighboring labels of relevant words. Despite all this, they are both defective. SVMs ignore the correlations between neighboring labels, while CRFs are always locked into over-fitting problems. Recently, large margin methods related to SVMs, such as Max-Margin Markov Networks (M3Ns) and Hidden Markov Support Vectors (HMM-SVMs), are proposed to deal with sequence labeling problems, and successfully applied to several fields such as Chunking (Taskar, 2003) and Speech Recognition (Jiang, 2006) in natural language processing. This model combines the advantages of SVMs and CRFs.

In this paper, we introduce the large margin methods into Chinese Word Segmentation by Character-based tagging, and compare the performance of large margin methods with two typical state-of-the-art methods (SVMs and CRFs) and other systems in the literature. Closed test on the PKU, MSR and HK CityU corpus of Sighan Bakeoff-2005 (Thomas, 2005) shows some performance improvements over SVMs, CRFs and other systems. **The paper is organized as follows.**

## **2. Chinese Word Segmentation Based on Large Margin Methods**

Large Margin Methods are designed to deal with sequence labeling problems with large margin thinking of SVMs. In this section, we firstly introduced how to formalize the Chinese Word Segmentation as a sequence labeling problem, and then made a brief description of large margin methods. Finally, the details of Chinese Word Segmentation based on Large Margin Methods have been presented.

### **2.1 Character-based Chinese Word Segmentation**

Chinese Word Segmentation is essentially a decision-making process of two values which indicates whether we should split the sentence at the position of the current character. This word boundary tagging approach was first present by (Xue, et al., 2003) at 1st Sighan, and widely used in the latter word segmentation systems. There are two types of word boundary tagging approaches, which are character-based tagging and word-based tagging. The character-based tagging takes a Chinese Character as a token, while the word-based tagging takes a popular word that can not be separated as a token. In this paper, we focus on the character-based one. For word boundary tagging approaches, different tag sets have been proposed. The popular tag sets are 2-tag set {B, M}, 4-tag set {B, M, E, S} and 6-tag set {B, B1, B2, M, E, S}. The tag B indicates that the token is at the beginning of a word. In the 2-tag set, M indicates the other token except B. In the 4-tag set, S indicates that the token is

a single-character word, E indicates that the token is at the ending of a word and M indicates the other token. In the 6-tag set, B1 and B2 indicate the first and second tokens of the M in the 4-tags. Huihsin Tseng used two-tag set in a CRFs-based Chinese Word Segmentation system (Tseng, et al., 2005), Xue used a four-tag set in a ME-based system, and Hai Zhao (Zhao, et al., 2006) used a six-tag set in an improved CRFs-based system. Table 1 lists the different tag sets for Chinese Word Segmentation.

Table 1 Different tag sets for Chinese Word Segmentation

Tag set	Tags	Words in tagging
2-tag(Tseng)	B, M	B, BM, BMM, BMMM, ...
4-tag(Xue)	B, M, E, S	S, BE, BME, BMME, ...
5-tag(Zhao)	B, B1, M, E, S	S, BE, BB1E, BB1ME, BB1MME, ...
5-tag'	B, M, E1, E, S	S, BE, BE1E, BME1E, BMME1E, ...
6-tag(Zhao)	B, B1, B2, M, E, S	S, BE, BB1E, BB1B2E, BB1B2ME, BB1B2MME, ...
6-tag'	B, M, E2, E1, E, S	S, BE, BE1E, BE2E1E, BME2E1E, BMME2E1E, ...

In the Table1, the 5-tag' set and the 6-tag' set are derived from the 5-tag set and the 6-tag set. They emphasize the characters of a word in the end. Among them, 4-tag and 6-tag sets are the most common ones in literature.

After defining tag set, Chinese Word Segmentation can be formulated as a sequence labeling problem. Examples of Chinese Word Segmentation with different tag sets are shown in Table 2.

For tag set selection, Zhao (2006) presented an approach based on the average weighted word length distribution in the corpus. In our experiments, we did not completely adopt this method for tag set selection for memory limits. The detail description has been presented in section 3.

## 2.2 Chinese Word Segmentation based on Large Margin Methods

The task of sequence labeling problem is to find the best label sequence  $Y^* = y_1 y_2 \dots y_N$  for a given sequence of observations  $X = x_1 x_2 \dots x_N$ . Probabilistic graphical models are widely used for this problem. In probabilistic graphical models,  $Y^*$  is the most probable label sequence, i.e.  $Y^* = \arg \max_Y p(Y|X)$  in CRFs. While, in large margin methods,  $Y^*$  is the label sequence making highest score given by a linear discriminant function of features  $f_j$  with coefficients  $w_j$  such that:

$$h_w(x, y) = \sum_j w_j f_j(x, y) = w^T F(x, y) \quad (1)$$

Table 2 Examples of Chinese Word Segmentation with different tag sets for “希望(hope)/能(can)/听到(hear)/你 you/的(your)//大好消息(good news)/。” (we hope we can hear your good news.)

	2-tag	4-tag	5-tag	5-tag'	6-tag	6-tag'
希 望	B	B	B	B	B	B
	M	E	E	E	E	E
能	B	S	S	S	S	S
听 到	B	B	B	B	B	B
	M	E	E	E	E	E
你	B	S	S	S	S	S
的	B	S	S	S	S	S
大 好 消 息	B	B	B	B	B	B
	M	M	B1	E1	B1	E2
	M	M	M	M	B2	E1
	M	E	E	E	E	E
。	B	S	S	S	S	S

The large margin methods were proposed by Taskar (2003) and Joachims (2005) for labeling sequence data, which label a sequence using undirected markov chain as CRFs and maximize the classification boundary as SVMs. Figure 1 shows the chain structure of the large margin methods with single-node cliques<sup>2</sup> and pair wise cliques for sequence labeling. Differing from the conditional probability solution formulation of CRFs, large margin methods obtain the target label sequence  $Y^*$  by maximizing the distance between  $Y$  and classification margins like SVMs in Figure 2.

By the fundamental theorem of SVMs (Weston, et al, 1999), the distance over label sequence  $Y$  given  $X$  can be given by the linear discriminant function (1), and the coefficient  $w$  can be determined by maximizing the classification margins  $\gamma(x, y, y'; w) = [w^T F(x, y) - w^T F(x, y')]$ . To solve this optimization problem, Ben Taskar (year) used undirected markov chain to decompose the discriminant function and the classification margins respectively, which are given as follows:

<sup>2</sup> A clique is a fully connected subgraph.

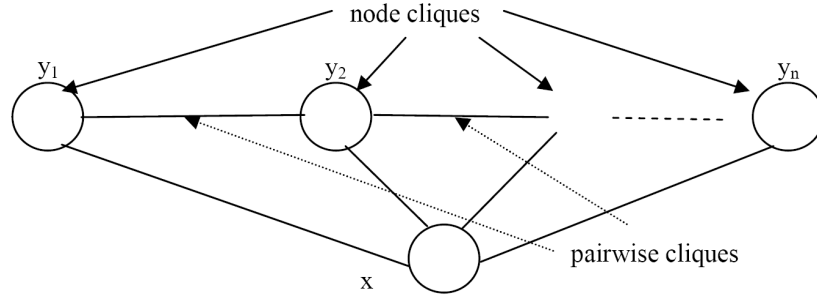


Figure 1 the chain structure of the large margin methods with single-node cliques and pairwise cliques.

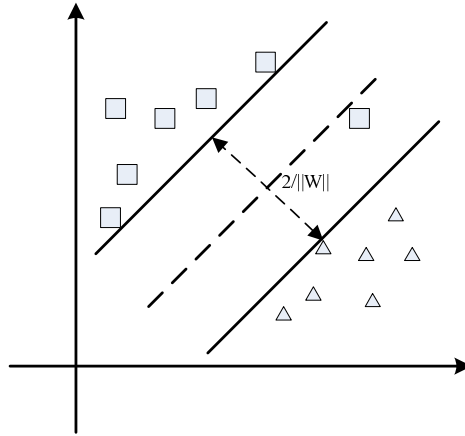


Figure 2 The margin of the SVM binary classifier

$$h_w(x, y) = w^T F(x, y) = \sum_j w_j^T F(x, y_j) = \sum_j w_j^T (f(x, y_j), f(x, y_{j-1}, y_j)) \quad (2)$$

$$\begin{aligned} \gamma(x, y, y'; w) &= [w^T F(x, y) - w^T F(x, y')] \\ &= \sum_j w_j^T ([f(x, y_j) - f(x, y'_j)], [f(x, y_{j-1}, y_j) - f(x, y'_{j-1}, y'_j)]) \quad (3) \\ &= w^T \Delta F(x, y, y') \end{aligned}$$

Similar to SVMs, non-negative slack variables are introduced into large margin methods for non-separable data i.e.  $\gamma(x, y, y'; w) \geq l(y, y') - \xi_x$  for  $\forall y' \neq y$ , where  $l(y, y')$  is the

loss function and  $\xi_x$  is the slack variant. In SVMs,  $l(y, y')$  is usually set to 0 or 1 indicating whether  $y'$  is the same as  $y$ , which is not excellent for sequence labeling problems. Therefore, it makes more sense to use a margin according to the structure of the sequence labels such as Hamming loss function, which computes per-label loss for each individual labels in  $y$ , given below.

$$l(y', y) = \sum_i^N I(y'_i \neq y_i) \quad (4)$$

where  $i$  is the  $i$ -th element of the label vector, and  $N$  is the length of the label vector. Hamming loss function has been proved more suitable for labeling sequence (Taskar, et al., 2003).

Taking all above given conditions into account, finding large margin for sequence labeling problem becomes the following quadratic optimization problem:

$$\begin{aligned} \min_{w, \epsilon} & \frac{1}{2} \|w\|^2 + \frac{C}{m} \sum_{x \in S} \xi_x \\ \text{s.t.} & w^T \Delta F(x, y, y') \geq l(y, y') - \xi_x, \\ & \forall y' \neq y, \xi_x \geq 0, (x, y) \in S \end{aligned} \quad (5)$$

where  $m$  is the number of samples.

For this problem, kernel trick can be applied as SVMs. Taskar, et al. presented an extended version of sequential minimal optimization (SMO) for parameters learning in (Taskar, et al., 2003). Collins, et al. presented an exponentiated gradient algorithm (EG) (Collins, et al., 2008) and T. Joachims presented a Cutting-Plane algorithm (Joachims, 2007) for training. For more details, readers can be referred to correlative references.

For Chinese word segmentation, the observation sequence  $X$  is a character sequence and the label sequence  $Y$  is composed of tags in the tag set.

### 3. Experiments and Results

#### 3.1 Data set

We used the data from Bakeoff-2005 shared task (Thomas Emerson, 2005) which can be obtained from <http://www.sighan.org/bakeoff2005/>. The data contains four corpora from

different sources: Peking University (PKU), Microsoft Research in Beijing (MSR), City University of Hong Kong (CITYU) and Academia Sinica (AS). All experiments performed on the computers with 64-bit 3.00GHz Intel(R) Pentium(R) D CPU and 2.0G memory. For the memory limits, only PKU, MSRA and CITYU were selected to perform the evaluation. A summary of these corpora is shown in Table 3.

Table 3 Partial Corpus of Sighan Bakeoff-2005

Corpus	Encoding	Training Size	Test Size	OOV rate
PKU	GB	1.1M	17K	0.058
MSR	GB	2.37M	107K	0.026
CITYU	Big5	1.46M	41k	0.074

### 3.2 Feature Templates

In Chinese Word Segmentation systems, features are usually extracted from feature template set which is another important factor affecting the performance of the systems. Two typical feature template sets we selected are shown in Table 4. The TMPT-6 proposed by Zhao (2006) is the best one for each corpus of Bakeoff-2005 under closed test, while TMPT-11 expands the window size of TMPT-6 to 5.

In our experiments, two types of features were used: node features and edge features. The node features are defined for each character-tag pair  $(x, y'_j)$  as:

$$f_j(x, y'_j) = \begin{cases} 1 & \text{if } TMPT(x) \text{ and } y'_j = t_j \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

Where  $TMPT(x)$  is the indicator function of the context related to  $x$  defined in Table 4 and  $t_j$  is a tag defined in Table 1.

The edge features are defined for each pair wise clique  $(y'_{j-1}, y'_j)$  as follows:

$$f_j(x, y'_{j-1}, y'_j) = \begin{cases} 1 & \text{if } y'_j = t_{j-1} \text{ and } y'_j = t_j \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

Where  $t_{j-1}$  and  $t_j$  are tags defined in Table 1. These features are the special features of CRFs and M3Ns, which contains rich information about correlations between neighboring labels.

Table 4 Two typical feature template sets for Chinese Word Segmentation

Template set	Type	Feature	Context description
TMPT-6	Unigram	$C_n, n=-1,0,1$	$C_0$ : the current token
	Bigram	$C_n C_{n+1}, n=-1,0,1$ $C-1C_1$	$C-1$ : the previous token $C-2$ : the token before the previous token
TMPT-11	Unigram	$C_n, n=-2,-1,0,1,2$	$C_1$ : the next token
	Bigram	$C_n C_{n+1}, n=-2,-1,0,1,2$ $C-1C_1$	$C_2$ : the token after the next token

### 3.3 Performance of Large Margin Methods versus other methods

In our experiments, Support Vector Machines (SVMs) and Conditional Random Fields (CRFs) were selected to compare with large margin methods, and Ben Taskar's Max-Margin Markov Networks (M3Ns) were selected as the representation of large margin methods. SVMs are discriminative classification based on large margin, while CRFs are discriminative models for sequence labeling. Both of them are state-of-the-art methods for Chinese Word Segmentation. We used LIBSVM (Lin, 2001) as the SVMs implementation, and CRF++<sup>3</sup> as the CRFs implementation. All other parameters in the toolkits were set by default. For M3Ns, we adopted linear kernel with Hamming loss function and first order. All other parameters were the same with SVMs. Table 5 shows the results of comparison using 4-tag shown in Table 2 and TMPT-6 shown in Table 4. M3Ns achieve the highest R-scores, P-scores and F-scores, but take the most time for training.

Moreover, we studied the affect of different combinations of tag sets and template sets on M3Ns. For tag sets, two typical tag sets (4-tag and 6-tag) and the new 6-tag set (6-tag') were selected for comparison. The experimental results are shown in Table 6. From the experimental results, we can get a general trend of increasing performance in the order 4-tag, 6-tag' and 6-tag for different tag sets, which is the same with the result of CRFs described in (Hai Zhao, et al., 2006). However, the TMPT-6 is not always the best one, which is different from the result of CRFs described in (Hai Zhao, et al., 2006). The best combinations are 6-tag plus TMPT-6 for PKU and CITYU, and 6-tag plus TMPT-11 for MSR, which are shown in font characters.

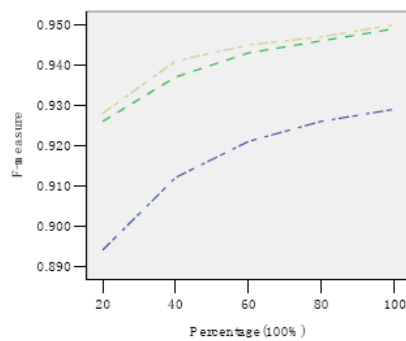
<sup>3</sup> <http://crfpp.sourceforge.net/>

Table 5 Results of comparison using 4-tag and TMPT-6.

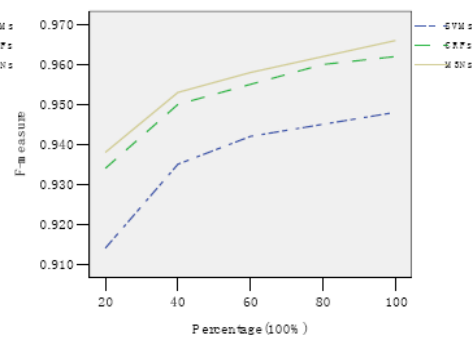
Corpus	Methods	R	P	F	#Parameters	Training time
PKU	M3Ns	<b>0.947</b>	<b>0.954</b>	<b>0.950</b>	415,269,6	42h56m
	CRFs	0.944	0.954	0.949	415,269,6	25h36m
	SVMs	0.925	0.932	0.929	415,268,0	10h52m
MSR	M3Ns	<b>0.965</b>	<b>0.966</b>	<b>0.966</b>	640,295,2	51h05m
	CRFs	0.962	0.962	0.962	640,295,2	40h57m
	SVMs	0.949	0.946	0.948	640,293,6	15h45m
CITYU	M3Ns	<b>0.943</b>	<b>0.948</b>	<b>0.946</b>	548,785,2	17h35m
	CRFs	0.938	0.941	0.939	548,785,2	15h28m
	SVMs	0.918	0.915	0.916	548,783,6	4h29m

Finally, in order to show the relationship between M3Ns and the data set size, we splited the training data into parts with different sizes: 20%, 40%, 60%, 80%, and compared M3Ns with SVMs and CRFs. Figure 3 shows the results of training data sets with different sizes using the tag set 4-tag shown in Table 2 and the feature template TMPT-6 shown in Table 4.

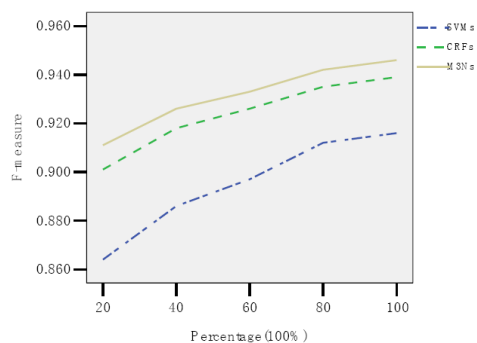
As we have shown above, M3Ns and CRFs are clearly superior to SVMs, and the F-measure increases by gradually increasing the size of corpus. Specially, M3Ns trained by 80% training corpus achieve the performance of CRFs trained by whole training corpus. Two factors lead to these results. On the one hand, M3Ns and CRFs take the correlations between neighboring labels into account, which is very important for word segmentation. On the other hand, the high generalization ability of the large margin methods makes M3Ns more robust than CRFs.



(a) The F-measure on PKU



(b) The F-measure on MSR



(c) The F-measure on CITYU

Figure 3 The results of training data sets with different sizes using 4-tag and TMPT-6.

Table 6 Comparisons of results under different tag sets and feature template sets

Tag set	Corpus	Template set	R	P	F	OOV	Roov	Riv
4-tag	PKU	TMPT-6	0.947	0.954	0.950	0.058	0.648	0.958
		TMPT-11	0.946	0.954	0.950	0.058	0.661	0.956
	MSR	TMPT-6	0.965	0.966	0.966	0.026	0.658	0.973
		TMPT-11	0.967	0.967	0.967	0.026	0.698	0.974
	CITYU	TMPT-6	0.943	0.948	0.946	0.074	0.685	0.964
		TMPT-11	0.942	0.947	0.945	0.074	0.702	0.961
6-tag	PKU	<b>TMPT-6</b>	<b>0.948</b>	<b>0.956</b>	<b>0.952</b>	<b>0.058</b>	<b>0.665</b>	<b>0.958</b>
		TMPT-11	0.946	0.954	0.950	0.058	0.661	0.956
	MSR	TMPT-6	0.970	0.970	0.970	0.026	0.707	0.977
		<b>TMPT-11</b>	<b>0.971</b>	<b>0.972</b>	<b>0.972</b>	<b>0.026</b>	<b>0.735</b>	<b>0.978</b>
	CITYU	<b>TMPT-6</b>	<b>0.945</b>	<b>0.947</b>	<b>0.946</b>	<b>0.074</b>	<b>0.689</b>	<b>0.965</b>
		TMPT-11	0.942	0.946	0.944	0.074	0.698	0.962
6-tag'	PKU	TMPT-6	0.948	0.955	0.952	0.058	0.658	0.958
		TMPT-11	0.947	0.955	0.951	0.058	0.672	0.957
	MSR	TMPT-6	0.970	0.967	0.969	0.026	0.668	0.978
		TMPT-11	0.971	0.971	0.971	0.026	0.730	0.978
	CITYU	TMPT-6	0.944	0.946	0.945	0.074	0.684	0.964
		TMPT-11	0.943	0.946	0.944	0.074	0.695	0.963

### 3.4 Comparisons with Related Works

In this section, we compared our results with the best existing results. The comparisons are

shown in Table 6. There are two types of existing results. One is the best F score of Bakeoff-2005 for each corpus under closed test. Second are the results of the character-based systems and word-based systems presented in literature. The character-based tagging systems are better in performance and the word-based tagging systems are lower in performance than the proposed system. All results of our experiments are selected from Table 7.

	PKU	MSR	CITYU
Best of Bakeoff2005	0.950	0.964	0.943
Tseng, 2005	0.950	0.964	0.943
<b>Ours</b>	<b>0.952</b>	<b>0.972</b>	0.946
Zhang <sup>4</sup> , 2006	0.945	0.964	0.946
Zhang <sup>5</sup> , 2006	0.951	0.971	0.951
Zhang, 2007	0.945	<b>0.972</b>	<b>0.951</b>

Table 7 Comparisons with Related Works

The Experimental results show that our system achieved the higher F measures than the best ones of Bakeoff-2005 for all corpora we selected and it is competitive with other word-based tagging systems. It is noteworthy that word-based tagging systems are better than character-based tagging systems under the same conditions (Ruiqiang Zhang, 2006; Zhao, 2007). This comparison with our system is not appropriate because it has been established that we can achieve appreciable improvement in the word-based tagging system.

#### 4. Conclusion and Discussion

In this paper, the large margin methods were presented for Chinese Word Segmentation by Character-based tagging and achieved good performance. However, it took too much time for training. Fortunately, the Cutting-Plane algorithm solved this problem. In future, we will focus the results of the large margin methods by Word-based tagging, and apply them to other fields.

---

<sup>4</sup> The pure sub-word CRF model.

<sup>5</sup> The confidence-based combination of the CRF and rule-based models.

## 5. References

- Chang, C.-C., Lin, C.-J., 2001, LIBSVM: a library for support vector machines, Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- Chen, S. F., and J. Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech and Language*, 13(4): 359-394.
- Chooi-Ling Goh, Masayuki Asahara, Yuji Matsumoto, 2005, Chinese Word Segmentation by Classification of Characters, *Computational Linguistics and Chinese Language Processing*, 10(3), pp. 381-396.
- Gao, J.-F., J. Goodman, M. Li, and K.-F. Lee. 2002. Toward a unified approach to statistical language modeling for Chinese. *ACM Trans, Asian Language Information Process*, 1(1): 3-33.
- Gao, J.-F., M. Li, and C.-N. Huang, 2003, Improved source-channel model for Chinese word segmentation, In the 41st Annual Meeting of the Association for Computational Linguistics, Sapporo, Japan, pp.272-279.
- Goh, Chooi-Ling and Asahara, Masayuki, Matsumoto, Yuji, 2005, Chinese Word Segmentation by Classification of Characters, In the Association for Computational Linguistics and Chinese Language Processing, Barcelona, Spain, pp.57-64.
- Hai Zhao, Chang-Ning Huang, Mu Li, Bao-Liang Lu, 2006, Effective Tag Set Selection in Chinese Word Segmentation via Conditional Random Field Modeling, The 20th Pacific Asia Conference on Language, Information and Computation (PACLIC-20), Wuhan, China, November 1-3, pp.87-94.
- Hai Zhao, Chunyu Kit, 2007, Effective Subsequence-based Tagging for Chinese Word Segmentation, (in Chinese) *Journal of Chinese Information Processing*, 21(5), pp.8-13.
- Hockenmaier, J., and C. Brew 1998. Error-driven Learning of Chinese word segmentation. In the 12th Pacific Conference on Language and Information. Singapore: 218-229.
- Hui Jiang, Xinwei Li, Chaojun Liu, 2006, Large margin hidden markov models for speech recognition, *IEEE transactions on audio, speech, and language processing*, 14, pp.1584-1595.
- 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*, Jeju Island, Korea, pp.168-171.
- I. Tsochantaridis, T. Joachims, T. Hofmann, Y. Altun, 2005, Large Margin methods for Structured and Interdependent Output Variables, *Journal of Machine Learning Research*, 6, pp.1453-1484.

- J. Weston, C. Watkins, 1999, Support vector machines for multi-class pattern recognition, Proceedings European Symposium on Artificial Neural Networks.
- Liang, N.-Y. 1987. Automatic word segmentation in written Chinese and an auto match word segmentation system-CDWS. (in Chinese) Journal of Chinese information processing, 1(2): 44-52.
- Michael Collins, Amir Globerson, Terry Koo, Xavier Carreras, PeterL, Bartlett, 2008, Exponentiated Gradient Algorithms for Conditional Random Fields and Max-Margin Markov Networks, Journal of Machine Learning Research, 9, pp. 1775-1822.
- Peng, F.-C., F.-F Feng, A. McCallum. 2004. Chinese segmentation and new word detection using conditional random fields, In Proceedings of the 20th International Conference on Computational Linguistics, Geneva, Switzerland, pp.562-568.
- Ruiqiang Zhang, Genichiro Kikui, Eiichiro Sumita, 2006, Subword-based tagging by Conditional Random Fields for Chinese word segmentation, In HLT/NAACL-2006, New York, pp.193-196.
- T. Joachims, T. Finley, Chun-Nam Yu, 2007, Cutting-Plane Training of Structural SVMs, under review.
- Taskar, B., Guestrin, C., & Koller, D, 2003, Max-margin markov networks. Advances in Neural Information Processing System 16.
- Thomas Emerson, 2005, The Second International Chinese Word Segmentation Bakeoff, Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing, pp.123-133, Jeju Island, Korea.
- Tseng H, Chang P, Andrew G, Jurafsky D, Manning C, 2005, A conditional random field word segmenter for SIGHAN Bakeoff 2005, Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing, pp.168-171.
- Xiaolong Wang, Kaizhu Wang, Zhongrong Li, Xiaohua Bai. 1989. Fewest Word Matching problem and its solution. (in Chinese) Chinese Science Bulletin, 34(13):1031-1032.
- Xue, N.-W., and L.-B. Shen. 2003. Chinese Word Segmentation as LMR Tagging. In the Second SIGHAN Workshop on Chinese Language Processing, Japan: 176-179.
- Yaodong Chen, Ting Wang and Huowang Chen. 2005. Using Directed Graph Based BDMM Algorithm for Chinese Word Segmentation. Computational Linguistics and Intelligent Text Processing : 214-217.
- Zhang, H.-P., Q. Liu, X.-Q. Cheng, H. Zhang, H.-K. Yu, 2003, Chinese Lexical Analysis Using Hierarchical Hidden Markov Model, In the Second SIGHAN workshop affiliated with 4th ACL, Sapporo Japan, pp.63-70.
- Zhang, Y. Clark, S, 2007, Chinese Segmentation with a Word-Based Perceptron Algorithm. Proceedings of the 45th Annual Meeting of the ACL. p.840-847.