Chinese Word Segmentation with Conditional Support Vector Inspired Markov Models

Yu-Chieh Wu
[1]Dep. of Computer Science and Information Engineering; National Central University
[2]Finance Department
Ming Chuan University
Taipei, Taiwan
bcbb@db.csie.ncu.edu.tw

Jie-Chi Yang
Graduate Institute of Network Learning
National Central University
Taoyuan, Taiwan
yang@cl.ncu.edu.tw

Yue-Shi Lee
Dep. of Computer Science and Information Engineering
Ming Chuan University
Taoyuan, Taiwan
leeys@mcu.edu.tw

Abstract

In this paper, we present the proposed method of participating SIGHAN-2010 Chinese word segmentation bake-off. In this year, our focus aims to quick train and test the given data. Unlike the most structural learning algorithms, such as conditional random fields, we design an in-house development conditional support vector Markov model (CMM) framework. The method is very quick to train and also show better performance in accuracy than CRF. To give a fair comparison, we compare our method to CRF with three additional tasks, namely, CoNLL-2000 chunking, SIGHAN-3 Chinese word segmentation. The results were encourage and indicated that the proposed CMM produces better not only accuracy but also training time efficiency. The official results in SIGHAN-2010 also demonstrates that our method perform very well in traditional Chinese with fine-tuned features set.

1 Introduction

Since 2006 Chinese word segmentation bakeoff in SIGHAN-3 (Levow, 2006), this is the third time to join the competition (Wu et al., 2006, 2007). In this year, we join the SIGHAN bakeoff task in both traditional and simplified Chinese closed word segmentation. Unlike most western languages, there is no explicit space between words. The goal of word segmentation is to identify words given the sentence. This technique provides important features for downstream purposes. Examples include Chinese part-of-speech (POS) tagging (Wu et al., 2007), Chinese word dependency parsing (Wu et al., 2007, 2008).

With the rapid growth of structural learning algorithms, such as conditional random fields (CRFs) (Lafferty et al., 2001) and maximum-margin Markov models (M^3N) (Taskar et al., 2003) have received a great attention and become a prominent learning algorithm to many sequential labeling tasks. Examples include part-of-speech (POS) tagging (Shen et al., 2007) and syntactic phrase chunking (Suzuki et al., 2007). The Chinese word segmentation can also be treated as a character-based tagging task in (Xue and Converse, 2002). One feature of sequential labeling is that it aims at finding non-recursive chunk fragments in a given sentence. Among these approaches, CRF has been wildly used in recent SIGHAN bakeoff tasks (Jin and Chen, 2008; Levow, 2006).

Although these approaches do not suffer from so-called label-bias problems (Lafferty et al., 2001), one limitation is that they are inefficient to train with large-scale, especially large category data. On the other hand, non-structural learning approaches (e.g. maximum entropy models) which learn local predictors usually cost much better training time performance than structural learning algorithms. These methods condition on local context features and incorporate fix-length history information. Although higher order feature (longer history) maybe useful to some tasks, the exponential scaled inference time is also intractable in practice.

Support vector machines (SVMs) which is one of the state-of-the-art supervised learning algorithms have been widely employed as local classifiers to many sequential labeling tasks (Taku

and Matsumoto, 2001; Wu et al., 2006, 2008). Specially, the training time of linear kernel SVM with either L_1-norm (Joachims, 2006; Keerthi et al., 2008) or L_2-norm (Keerthi and DeCoste, 2005; Hsieh et al., 2008) can now be obtained in linear time. Even local classifier-based approaches have the drawbacks of label-bias problems, training nonstructural linear SVM is scalable to large-scale data. By means of so-called one-versus-all multiclass SVM training, it is also scalable to large-category data.

In this paper, we present our Chinese word segmentation based on the proposed conditional support vector Markov models for sequential labeling tasks, especially Chinese word segmentation. Unlike structural learning algorithms, our method can be simply trained without considering the entire structures and hence the training time scales linearly with the number of training examples. In this framework, to alleviate the ease of label-bias problems, the state transition probability is ignored. Instead, we merely utilize the property of label relationships between chunks (Wu et al., 2008). To demonstrate our method, we compare to several well-known structural learning algorithms, like CRF (Kudo et al., 2004), and SVM-HMM (Joachims et al., 2009) on two well-known data, namely, CoNLL-2000 syntactic chunking, SIGHAN-3 Chinese word segmentation tasks. By following this, we apply the model to the Chinese word segmentation tasks of SIGHAN-2010 this year. The empirical results showed that our method is not only fast but also achieving more superior accuracy than structural learning methods. In traditional Chinese, our method also achieves the state-of-the-art performance in accuracy with fined-tune features.

2 Conditional support vector Markov models

Traditional conditional Markov models (CMM) is to assign the tag sequence which maximizes the observation sequence.

$$P(s_1, s_2, ..., s_n \mid o_1, o_2, ..., o_n)$$

Where s_i is the tag of word i. For the first order left-to-right CMM, the chain rule decomposes the probabilistic function as:

$$P(s_1, s_2, ..., s_n \mid o_1, o_2, ..., o_n) = \prod_{i=1}^{n} P(s_i \mid s_{i-1}, o_i) \quad (1)$$

Therefore, we can employ a local classifier to predict $P(s_i \mid s_{i-1}, o_i)$ and the optimal tag sequence can be efficiently searched by using conventional Viterbi algorithm.

The graphic illustration of the K-th order left-to-right CMM is shown in Figure 1. The chain probability decompositions of the other K-th order CMM in Figure 1 are:

$$P(s,o) = \prod_{i=1}^{n} P(s_i \mid o_i) \quad (2)$$

$$P(s,o) = \prod_{i=2}^{n} P(s_i \mid o_i, s_{i-1}) \quad (3)$$

$$P(s,o) = \prod_{i=3}^{n} P(s_i \mid o_i, s_{i-1}, s_{i-2}) \quad (4)$$

$$P(s,o) = \prod_{i=3}^{n} P(s_i \mid o_i, s_{i-1}, \hat{s}_{i-1}) \quad (5)$$

Equations (2), (3), and (4) are merely standard zero, first and second order decompositions, while equation (5) is the proposed greedy second order CMM decomposition which will be discussed in next section.

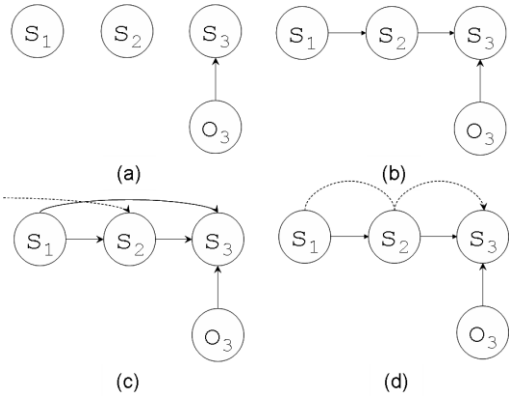

Figure 1: K-th order conditional Markov models: (a) the standard 0(zero) order CMM, (b) first order CMM, (c) second order CMM, and (d) the proposed second order CMM

The above decompositions merge the transition and emission probability with single function. McCallum et al. (2000) further combined the locally trained maximum entropy with the infered transition score. However, our conditional support vector Markov models make different chain probability. We replace the original transition probability with transition validity score, i.e.

$$P(s,o) = \prod_{i=2}^{n} \widetilde{P}(s_i \mid s_{i-1}) P(s_i \mid o_i) \quad (6)$$

$$P(s,o)=\prod_{i=3}^{n}\hat{P}(s_i \mid s_{i-1})P(s_i \mid o_i, s_{i-1}, \hat{s}_{i-1}) \quad (7)$$

The transition validity score is merely a Boolean flag which indicates the relationships between two neighbor labels. Equation (6) and (7) are zero-order and *our second order* chain probabilities. We will introduce the proposed inference algorithm and how to obtain the transition validity score automatically without concerning the change of chunk representation.

2.1 Tag transitions

In this paper, we do not explicitly adopt the state transitions for our CMM. Instead, a chunk-relation pair is used. Nevertheless, one important property to sequential chunk labeling is that there is only one phrase type in a chunk. For example, if the previous word is tagged as begin of noun phrase (B-NP), the current word must not be end of the other phrase (E-VP, E-PP, etc). Therefore, we only model relationships between chunk tags to generate valid phrase structure.

Wu et al. (2007, 2008) presented an automatic chunk pair relation construction algorithm which can handle so-called IOB1/IOB2/IOE1/IOE2 (Kudo and Matsumoto, 2001) chunk representation structures with either left-to-right or right-to-left directions. Here, we extend this idea and generalize to fit to more chunk tags. That is we can model the S-tag, B2, B3 tags with dividing the leading tags into two categories. For details can refer the literatures.

3 Empirical Results

Three large-scale and large-category dataset is used to evaluate the proposed method, namely, CoNLL-2000 syntactic chunking (Tjong Kim Sang and Buchholz, 2000), Chinese POS tagging, and three of SIGHAN-3 word segmentation tasks. Table 1 shows the statistics of those datasets.

Feature type	CoNLL-2000	SIGHAN-3
Unigram	w_{-2}~w_{+2}	w_{-2}~w_{+2}
Bigram	(w_{-2},w_{-1}),(w_{-1},w_0), (w_0,w_{+1}), (w_{+1},w_{+2}),(w_{+1},w_{-1})	(w_{-2},w_{-1}),(w_{-1},w_0), (w_0,w_{+1}), (w_{+1},w_{+2}),(w_{+1},w_{-1})
POS	p_{-2}~p_{+2}	
POS bigram	(p_{-2},p_{-1}),(p_{-1},p_0), (p_0,p_{+1}),(p_{+1},p_{+2}), (p_{+1},p_{-1})	
POS trigram	(p_{-2},p_{-1},p_0), (p_{-1},p_0,p_{+1}),(p_{-3},p_{-2},p_{-1}), (p_0,p_{+1},p_{+2}), (p_{+1},p_{+2},p_{+3})	
(Word+POS) bigram	(w_{-1},p_0),(w_{-2},p_{-1}) (w_0,p_{+1}), (w_{+1},p_{+2})	
Other features	2~4 suffix letters 2~4 prefix letters Orthographic feature (Wu et al., 2008)	AV feature of 2~6 grams (Zhou and Kit, 2007)

Figure 2: Feature templates used in experiments

CoNLL-2000 chunking task is a well-known and widely evaluated in many literatures (Suzuki et al., 2007; Ando and Zhang, 2005; Kudo and Matsumoto, 2001; Wu et al., 2008; Daumé III and Marcu, 2005). The training data was derived from Treebank WSJ section 15-18 while section 20 was used for testing. The goal is to find the non-recursive phrase structures in a sentence, such as noun phrase (NP), verb phrase (VP), etc. There are 11 phrase types in this dataset. We follow the previous best settings for SVMs (Kudo and Matsumoto, 2001; Wu et al., 2008). The IOE2 is used to represent the phrase structure and tagged the data with backward direction.

The training and testing data of the Chinese POS tagging is mainly derived from the Academic Sinica's balanced corpus (version 3.0). Seventy-five percent out of the data is used for training while the remaining 25% is used for testing. However, the task of the Chinese POS tagging is very different from classical English POS tagging in that there is no word boundary information in Chinese text. To achieve this, Ng and Low (2004) gave a successful study on Chinese POS tagging. Just as English phrase chunking, the IOB-tags can be used to represent the Chinese word and its part-of-speech tag. For example, the tag B-ADJ means the first character of a Chinese word which POS tag is ADJ (adjective). n this task, we simply use the IOB2 to represent the chunk structure. In this way, the tagger needs to recognize the chunk tag by considering 118 (59*2) categories at once.

As discussed in (Zhou and Kit, 2007), using more complex chunk representation bring better segmentation accuracy in several Chinese word segmentation benchmarks. It is very useful in particular to represent long Chinese word (in particular proper nouns). By following this line, we apply the six tags B, BI, I, IE, E, and S to represent the Chinese word. BI and IE are the *interior after begin* and *interior before end* of a chunk. B/I/E/S tags indicate the begin/interior/end/single of a chunk. Figure 2 lists the used feature set in both experiments.

3.1 Settings

We included the Liblinear with square loss (Hsieh et al., 2008) into our conditional Markov models as classification algorithms. In basic, the SVM was designed for binary classification problems. To port to multiclass problems, we adopted the well-known one-versus-all (OVA) method. One good property of OVA is that parameter estimation process can be trained indivi-

Table 2: SIGHAN-3 word segmentation results

SIGHAN-3	UPUC			MSRA			CityU		
Method	$F_{(\beta)}$	Training Time	Testing Time	$F_{(\beta)}$	Training Time	Testing Time	$F_{(\beta)}$	Training Time	Testing Time
Our method	93.86	0.06 hr	15.15 s	96.22	0.45 hr	15.41 s	97.26	0.26 hr	25.32 s
CRF	93.76	1.17 hr	23.48 s	96.11	3.63 hr	17.06 s	97.29	4.34 hr	31.29 s
SVM-HMM	Out-of-memory			Out-of-memory			Out-of-memory		
Best approach (Zhou and Kit, 2007)	94.28	N/A	N/A	96.34	N/A	N/A	97.43	N/A	N/A
Second best approach	93.30	N/A	N/A	96.30	N/A	N/A	97.20	N/A	N/A

Table 3: Official evaluation results of the traditional and simplified Chinese word segmentation tasks

Task	Literature					Computer				
	Recall	Precision	F1	OOV-RR	IV-RR	Recall	Precision	F1	OOV-RR	IV-RR
Traditional	0.942	0.942	0.942	0.788	0.958	0.948	0.957	0.952	0.666	0.977
Simplified	0.936	0.932	0.934	0.564	0.964	0.915	0.915	0.915	0.594	0.972
Task	Medicine					Finance				
	Recall	Precision	F1	OOV-RR	IV-RR	Recall	Precision	F1	OOV-RR	IV-RR
Traditional	0.953	0.957	0.955	0.798	0.966	0.964	0.962	0.963	0.812	0.975
Simplified	0.933	0.915	0.924	0.642	0.969	0.945	0.941	0.943	0.666	0.972

dually. This is in particularly useful to the tasks which involve training large number of features and categories (Wu et al., 2008). To obtain the probability output from SVM, we employ the sigmoid function with fixed parameter A=-2 and B=0 as (Platt, 1999).

3.2 Comparison to structural learning

The overall experimental results are summarized in Table 1. Column "All" denotes as the $F_{(\beta)}$ score of all chunk types, while "NP" is the $F_{(\beta)}$ score of the noun phrase only. The final two columns list the entire training and testing times.

As shown in Table 1, it is surprising that the proposed CMM outperforms the other structural learning methods, CRF and SVM-HMM. In terms of training time, our method shows substantial faster than CRF. However, in terms of testing time, our method is worse than CRF. The main reason is that we do not optimize the code and implementation. We trust this can be further improved.

Table 1: Syntactic chunking results of the proposed CMM and the selected structural learning methods.

Method	All	NP	Training Time	Testing Time
Our method	94.51	94.95	0.15 hr	13.72 s
CRF	93.67	93.93	0.88 hr	6.20 s
SVM-HMM	93.90	94.20	0.20 hr	13.60 s

Table 2 shows the experimental results of the SIGHAN-3 bake-off tasks. We ran and conducted the experiments with UPUC, MSRA, and CityU datasets. The final two rows in Table 5 list the top 1 and 2 scores of published papers.

Here, the SVM-HMM still suffer from the scalability problems. Similar to the findings in the Chinese POS tagging task, the zero-order CMM achieved the optimal accuracy among first-order, full second order and the proposed inference algorithms. The training time is still very efficient for most CMMs. In comparison to CRF, our method did clearly perform better accuracy (excepted for the CityU) and require much less training time. For example, for the CityU dataset, our 0-order CMM took less than 15 minutes to train, while the CRF takes 4.34 hours in training.

However, we observe that our CMM yielded better testing time speed than CRF in this task. We further exploit the trained SVM models and found that the produced weights were not as dense as CRF which produces many nonzero weights per category. In addition, we observed that our implementation worked very efficient in the small category tasks.

For the three datasets, our method produced very competitive results as previous best approach which also made use of CRF as classifiers.

Although we use the same techniques to derive global features (assessor variety (AV) feature with 2~6 grams) from both training and testing data, our CMMs and the conducted CRF could not perform as well as (Zhou and Kit, 2007). In our experiments, both CRF and CMMs received the same training set. Hence the CRF and our CMMs is comparable in this experiment.

3.3 Official Results in SIGHAN-2010

To apply CMM to SIGHAN-2010, we design the following strategy. First the classifier parameters,

feature set should be improved. To achieve this, 1/4 of the training data was used as development set, while the remaining 3/4 training data was used to train the classifier. Second, we combine multi-classifier to enhance the accuracy. The CRF and our CMM with basic feature set were trained to predict the initial labels of the testing data. Then the predicted labels were included as features to train the final-stage classifier. The final classifier is still our CMM. Third, the post-processing method (Low et al., 2005) is employed to enhance the unknown word segmentation.

Table 4 lists the empirical results of the development set. By validate with development data, we found that C=1.25 and use the E-BIES representation method (Wu et al., 2008) yields better accuracy than B-BIES (Zhou and Kit, 2007). Meanwhile, CRF seems to be suitable for B-BIES representation method.

The classifier parameters were fixed and then we try to search the optimal feature set via the incremental add-and-check method. That is, we use the initial feature set as basis and add one feature type from the pool and verify the goodness of the feature with the development data. Figure 3 figures out the used features of each pass.

In this year, the process was completely run-through for the traditional Chinese task. Unfortunately we have insufficient time to apply the same technique to Simplified Chinese task. Table 3 lists the official results in the SIGHAN 2010 Chinese word segmentation bake-off.

Table 4: Empirical results of the development set of single CRF and our CMM

Development dataset	Traditional Chinese		Simplified Chinese	
	B-BIES	E-BIES	B-BIES	E-BIES
Our method	97.40	97.42	97.34	97.37
CRF	97.07	97.10	97.07	96.96

4 Conclusion

In this paper, we investigate the issues of sequential chunk labeling and present the conditional support vector Markov models for this purpose. The experiments were conducted with two well-known datasets, includes CoNLL-2000 text chunking and SIGHAN-3 Chinese word segmentation. The experimental results showed that our method scales very well while achieving surprising good accuracy than structural learning methods. On the SIGHAN-3 task, the proposed method outperformed CRF, while substantially reduced the training time. We also apply such method to the SIGHAN-2010 traditional Chinese segmentation with fined tuned feature set. The result was also encouraged. Our approach obtains the best accuracy in this task. In terms of Simplified Chinese, we achieve mid-rank place due to the very limited time-constraint. In the future, we plan to completely adopt this method to the Simplified Chinese word segmentation with the elaborated feature selection metrics and the same post-processing method.

The full online demonstration of the proposed conditional support vector Markov models can be found at the web site[1].

Feature Name	Pass1: CRF/CMM	Pass2: CMM
Character	w_{-2}~w_{+2}	Feature set of Pass1
Character N-gram	$(w_{-2},w_{-1}),(w_{-1},w_0),(w_0,w_{+1}),(w_{+1},w_{+2}),(w_{+1},w_{-1})$	
Special Character flags (Low et al., 2005)	w_{-2}~w_{+2}	
Others	[2]AV feature and its 2-gram combinations	[2]AV feature and its 2-gram and 3-gram combinations
Future flags[1]	*N/A*	t_{+1}, t_{+2}, t_{+3},(t_0,t_{+2}),(t_{+1},t_{+2}),(w_0,t_{+1}),(w_0,t_{+2})

[1]Future flags: the predicted tags of previous classifier

Figure 3: Feature templates used in experiments

References

Rie K. Ando, and Tong Zhang. 2005. A high-performance semi-supervised learning method for text chunking, In *Proc. of ACL*, pp. 1-9.

Bernhard E. Boser, Isabelle M. Guyon, Vladimir N. Vapnik. 1992. A training algorithm for optimal margin classifiers, In *Proc. of COLT*, pp. 144-152.

Andrew McCallum, Dayne Freitag, and Fernando C. N. Pereira. 2000. Maximum entropy Markov models for information extraction and segmentation, In *Proc. of ICML*, pp. 591.598.

Hal Daumé III and Daniel Marcu. 2005. Learning as search optimization: approximate large margin methods for structured prediction, In *Proc. of ICML*, pp. 169-176.

Guangjin Jin and Xiao Chen. 2008. The Fourth International Chinese Language Proces sing Bakeoff: Chinese Word Segmentation Named Entity Recognition and Chinese POS Tagging. In *Proc. of the SIGHAN Workshop on Chinese Language Processing*, pp. 69-81.

[1] http://140.115.112.118/bcbb/Chunking.htm

Thorsten Joachims. 2006. Training linear SVMs in linear time, In *Proc. of KDD*, pp. 217-226.

Thorsten Joachims, Thomas Finley, and Chun-Nam Yu. 2009. Cutting-Plane Training of Structural SVMs, Machine Learning Journal, to appear.

Sathiya Keerthi and Dennis DeCoste. 2005. A modified finite Newton method for fast solution of large scale linear SVMs, JMLR, 6: 341-361.

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

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis, In *Proc. of EMNLP*, pp. 230-237.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data, In *Proc. of ICML*, pp. 282-289.

Gina-Anne Levow. 2006. The third international Chinese language processing bakeoff: Word segmentation and named entity recognition. In *Proc. of the SIGHAN Workshop on Chinese Language Processing*, pp. 108–117.

Jin Kiat Low, Hwee Tou Ng, and Wenyuan Guo. 2005. A maximum entropy approach to Chinese word segmentation. In *Proc. of the SIGHAN Workshop on Chinese Language Processing*, pp. 161-164.

Hwee Tou Ng and Jin Kiat Low. 2004. Chinese part-of-speech tagging. one-at-a-time or all-at-once? word-based or character-based? In *Proc. of EMNLP*, pp. 277-284.

John Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods, In Advances in Large Margin Classifiers.

Jun Suzuki, Akinori Fujino, and Hideki Isozaki. 2007. Semi-supervised structural output learning based on a hybrid generative and discriminative approach, In *Proc. of EMNLP-CoNLL*, pp. 791-800.

Jun Suzuki and Hideki Isozaki. 2008. Semi-Supervised Sequential Labeling and Segmentation using Giga-word Scale Unlabeled Data. In *Proc. of ACL*, pp. 665-673.

Ben Taskar, Carlos Guestrin, and Daphne Koller. 2003. Max-margin Markov networks, In *Proc. of NIPS*.

Eric F. Tjong Kim Sang, and Sabine Buchholz. 2000. Introduction to the CoNLL-2000 shared task: chunking. In *Proc. of CoNLL*, pp. 127-132.

Yu-Chieh Wu, Jie-Chi Yang, Yue-Shi Lee, and Show-Jane Yen. 2006. Efficient and robust phrase chunking using support vector machines, In *Asia Information Retrieval Symposium (AIRS)*, pp. 350-361.

Yu-Chieh Wu, Jie-Chi Yang, and Yue-Shi Lee. 2008. Description of the NCU Chinese Word Segmentation and Part-of-Speech Tagging for SIGHAN Bakeoff 2008, In *Proc. of the SIGHAN Workshop on Chinese Language Processing*, pp. 161-166, 2008.

Yu-Chieh Wu, Yue-Shi Lee, and Jie-Chi Yang. Robust and efficient multiclass SVM models for phrase pattern recognition, Pattern recognition, 41(9): 2874-2889, 2008.

Yu-Chieh Wu, Jie-Chi Yang, and Qian Xiang Lin. 2006. Description of the NCU Chinese word segmentation and named entity recognition System for SIGHAN Bakeoff 2006, In *Proc. of the SIGHAN Workshop on Chinese Language Processing*, pp. 209-212.

Yu-Chieh Wu, Yue-Shi Lee, and Jie-Chi Yang. 2008. Robust and efficient Chinese word dependency analysis with linear kernel support vector machines, In *Proc. of the COLING*, pp. 135-138.

Yu-Chieh Wu, Jie-Chi Yang, and Yue-Shi Lee. 2007. Multilingual deterministic dependency parsing framework using modified finite Newton method Support Vector Machines. In *Proc. of the EMNLP/CoNLL*, pp.1175-1181.

Tong Zhang, Fred Damerau, and David Johnson. 2002. Text chunking based on a generalization Winnow, JMLR, 2: 615-637.

Hai Zhao and Chunyu Kit. 2007. Incorporating global information into supervised learning for Chinese word segmentation, In *Proc. of PACLIC*, pp.66-74.