A MMSM-based Hybrid Method for Chinese MicroBlog Word Segmentation

Xiao Sun*, Chengcheng Li, Chenyi Tang, Jiaqi Ye
AnHui Province Key Laboratory of Affective Computing and Advanced Intelligent Machine,
School of Computer and Information, Hefei University of Technology
Hefei, Anhui China, 230009.
School of Computer Science and Engineering, Dalian University of Technology
Dalian, Liaoning China, 116023.
sunx@hfut.edu.cn

Abstract

After years of researches, Chinese word segmentation has achieved quite high precisions for formal style text. However, the performance of segmentation is not so satisfying for MicroBlog corpora. In this paper we describe a scheme for Chinese word segmentation for, MicroBlog which integrates the character-based and word-based information in the directed graph generated by MMSM model. Word-level information is effective for analysis of known words, while character-level information is useful for analysis of unknown words. A multi-chain unequal states CRF model is proposed. The proposed multi-chain unequal states CRF has two state chains with unequal states which can recognize the POS tag simultaneously. The hybrid model was effective and adopted in real-world system.

1 Introduction

MicroBlog is an emerging application in the Web 2.0 era. On MicroBlog websites, users are able to post short messages less than a certain length, e.g., 140 English or Chinese characters, to communicate and share information with each other. After obtaining cleaned messages for a given user, we perform word segmentation for messages. In this paper, we use the system developed by Affective Computing and Natural Language Processing Group in Hefei University of Technology.

The system performs word segmentation and POS tagging simultaneously using a word lattice based re-ranking method proposed by Sun et al. [1]. Microblogs contain many out-of-vocabulary (OOV) words. To address the OOV problem, we also maintain a large up-to-date external vocabulary for word segmentation and POS tagging. To keep the vocabulary up-to-date, we import new words from two sources. The first is the Sogou New Word Dictionary which is updated weekly, and the second is the Sina Popular Word List, which is updated daily. The hybrid model for Chinese MicroBlog morphological analysis includes Chinese word segmentation, unknown word recognition and POS tagging. The foundation of the model is a directed segmentation graph based on the maximum matching and second-maximum matching (MMSM) model. Based on a known words system dictionary trained from the corpus, the MMSM model tries to build a directed graph with the candidate words and their parts-of-speech. In the directed graph, the character-level information and word-level information are combined, the HMM model is used to process the known words (words in system dictionary) using the word-level information; the proposed multi-chain unequal states CRF model is adopted to process the unknown words and their parts-of-speech using character-level information. Meanwhile, for the unknown word, which is the main difficulty in Chinese morphological analysis, both the word boundary and the parts-of-speech of the unknown words are unknown.

A multi-chain unequal states (MUS) CRF model is proposed here to process the unknown word segmentation and POS tagging. The proposed multi-chain CRF model has multi states chains for multi tasks. In our system, we adopted two states chains in which one states chain for the unknown words recognition and the other states chain for the unknown words POS tagging. The proposed MUS CRF model recognizes the unknown words from the sentence together with their POSs in one step, without using two separate linear-chain CRF models. The unknown words with their part-of-speech recognized by the multi-chain are added into the directed graph as candidates. With the directed segmentation

Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing, pages 99–105,
Tianjin, China, 20-21 DEC. 2012

graph and the proposed multi-chain CRF, the word-level information and character-level information are combined, Chinese word segmentation, unknown word recognition and POS tagging can be accomplished simultaneously.

2 The MMSM Directed Graph

The MMSM model acts as the basic framework in the hybrid model. The MMSM model (Huang and Sun, 2007) is a segmentation method that keeps the maximum and second-maximum segmentation result from a certain position in a sentence, and store the candidates of segmentation and POS tagging results in a directed graph, then some decoding algorithm is adopted to find the best path in the directed graph. With the MMSM model, all the possible segmentation paths and most lexical information like the POS information can be reserved for further use; little space cost is guaranteed by using the directed graph to store the segmentation paths; the context spaces are extended from single-dimension to multi-dimension; the MMSM model is also easy to be extended and add some new models in it.

The MMSM model is applied to build the original directed graph. Given a sentence, from a certain place if there are some candidates of segmentation words from the system dictionary, the MMSM model is applied to build the directed graph. Take the sentence “出生在聊城镇(Born in Liaocheng Town)” for example, the segmentation directed graph generated by the MMSM model is shown in figure 1. The labels after the words are POSs(parts-of-speech) defined in the PKU corpus.

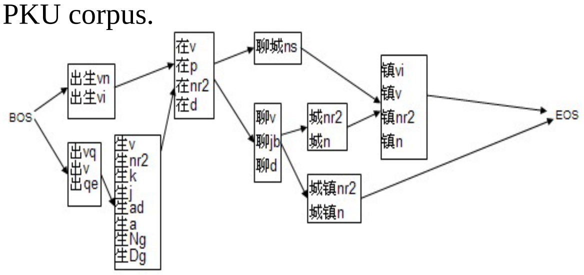

Figure 1. Segmentation directed graph by MMSM model

The word-based HMM model is trained and applied to assign cost for the nodes and edges in the directed graph by the MMSM model. The word-based HMM models were first used in English part-of-speech (POS) tagging (Charniak et al., 1993; Brants, 2000). This method identifies POS tags $T = t_1,\ldots, t_n$, given a sentence as a word sequence $W = w_1,\ldots,w_n$, where n is the number of words in the sentence. In Chinese language processing, the method is used with some modifications. Because each word in a sentence is not separated explicitly in Chinese, both segmentation of words and identification of the POS tags of the words must be done simultaneously. Given a sentence S, its most likely word sequence $\hat{\mathrm{W}}$ and POS sequence $\hat{\mathrm{T}}$ can be found as follows where W ranges over the possible segments of S ($w_1,\ldots,w_n = S$):

$$(\hat{\mathrm{W}},\hat{\mathrm{T}}) = \underset{\mathrm{W,T}}{\mathrm{argmax}}\, P(\mathrm{W,T} \mid \mathrm{S})$$

$$= \underset{\mathrm{W,T}}{\mathrm{argmax}} \frac{P(\mathrm{W,T,S})}{P(\mathrm{S})} = \underset{\mathrm{W,T}}{\mathrm{argmax}}\, P(\mathrm{W,T}) \qquad (1)$$

$$\approx \underset{\mathrm{W,T}}{\mathrm{argmax}} \prod_{\mathrm{i=1}}^{\mathrm{n}} P(\mathrm{w_i} \mid \mathrm{t_i}) P(\mathrm{t_i} \mid \mathrm{t_{i-1}})$$

$P(w_i|t_i)$ represents the cost of nodes, while $P(t_i|t_{i-1})$ represents the cost of edges in the directed graph. When building the directed graph, there could be some positions where exists no candidates of segmentation words and corresponding parts-of-speech. The MUS CRF model is applied from such positions to recognize the unknown words and their corresponding POS and then adds them to the directed graph.

3 Multi-chain Unequal States CRF Model

Conditional Random Fields (CRFs) (J. Lafferty et al, 2001) is considered as one of the best sequence labeling classifier. A sequence labeling problem can be viewed as following: given an observed sequence $\vec{\mathrm{x}}$, we hope to get a corresponding label sequence $\vec{\mathrm{y}}$ with maximum probability. All possible y_i in $\vec{\mathrm{y}}$ are assumed from a finite label set Y. For example, in a part-of-speech tagging problem, given a sentence $\vec{\mathrm{x}}$, the corresponding POS labels $\vec{\mathrm{y}}$ are hoped to be gotten. CRF is a kind of discriminative model, which aims to estimate the probability $p(\vec{\mathrm{y}} \mid \vec{\mathrm{x}})$ directly without estimating the marginal $p(\vec{\mathrm{x}})$. The Linear-chain CRF is,

$$P_\theta(\vec{y} \mid \vec{x}) = \frac{1}{Z_\theta} \prod_{t=1}^{T-1} \Phi_t(y_t, y_{t+1}, \vec{x}, t) \qquad (2)$$

Where

$$\Phi_t(y_t, y_{t+1}, \vec{x}, t) =$$
$$\exp(\sum_k \lambda_k f_k(y_t, y_{t-1}, \vec{x}, t))$$

The boundary and POS of the unknown word are both unknown. In order to solve the unknown word recognition and POS tagging, instead of adopting two separate linear-chain CRF models, a MUS CRF model is proposed in this paper. The multi-chain CRF includes one observe chain and two state chains. It is defined as follows:

Let $\vec{X}$ be an observed sequence, $\vec{Y}$ be a set of corresponding labels, and $\vec{W}$ be a set of higher-level labels. Then the distribution *p* is a multi-chain conditional random field if each state $\vec{x}_i$ in $\vec{X}$ corresponds to one state $\vec{y}_i$ in $\vec{Y}$ while each state $\vec{w}_i$ in $\vec{W}$ corresponds to several contiguous states in $\vec{X}$, the distribution is as follows:

$$\mathrm{P}_\theta(\vec{y} \mid \vec{x}) = \frac{1}{Z_\theta}\left(\prod_{t=1}^{T-1}\Phi_t(y_t, y_{t+1}, \vec{x}, t)\Psi_t(y_t, w_k, \vec{x}, t)\right) * \left(\prod_{k=1}^{K-1} T_k(w_k, w_{k+1}, k)\right) \quad (3)$$

Where

$$\Psi_t(y_t, w_k, \vec{x}, t) = \exp\left(\sum_i \lambda_i f_i(y_t, w_k, \vec{x}, t)\right)$$

$$T_k(w_k, w_{k+1}, k) = \exp\left(\sum_j \lambda_j f_j(w_k, w_{k+1}, k)\right)$$

$$\mathrm{Z}_\theta = \sum_{\vec{y}}\left\{\prod_{t=1}^{T-1}\Phi_t(y_t, y_{t+1}, \vec{x}, t)\Psi_t(y_t, w_k, \vec{x}, t)\right) * \left(\prod_{k=1}^{K-1} T_k(w_k, w_{k+1}, k)\right\}$$

In Chinese word segmentation and POS tagging, the $\vec{x}$ in the multi-chain CRF equation represents sequence of the Chinese characters, the x_i represents the *ith* character in the sentence. The $\vec{y}$ represents the positional tag sequence of $\vec{x}$, the y_i represents the positional tag of x_i. The $\vec{w}$ represents the POS tagging sequence of the sentence, the w_i represents the POS of the *ith* word in the sentence. Thus the MUS CRF can perform the Chinese word segmentation and POS tagging simultaneously without having to build two separate linear-chain CRF models. The feature functions *f* in equation (3) represents the features obtained from the contexts. The features templates will be discussed in the next subsection. The equations of MUS CRF can be easily derived from DCRF (Dynamic CRF) (Charles Sutton et al., 2006) and the parameter estimation for multi-chain CRF is almost the same as linear-chain CRF. The structure of the MUS CRF is shown in the following figure 2. The lines in the figure present the features between the nodes.

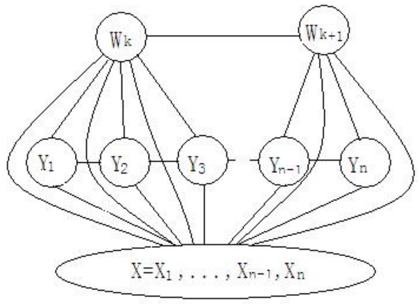

Figure 2, Multi-chain Unequal States CRF

The different between the DCRF and the proposed MUS CRF is that the top state chain in the MUS CRF does not have the same number of states as the bottom states chain. Just take the Chinese word segmentation and POS tagging for example. We should give each character in a sentence a corresponding label (Y_i) to mark its position in a word, a sequence of characters that form a word share a single POS label (W_k). The top state chain does not need so many states as the bottom state chain, so the complexity of computational cost drops down.

Given an input sentence, from the position that cannot be segmented, the multi-chain CRF is applied to recognize the unknown words and their related POSs. In our system, a 6-tag label set(Zhao,2006) is applied for Chinese word segmentation, which is shown in Table 1. Each character in the sentence is assigned a tag from the 6-tag label set to mark their position in a word.

Label	Position
B	The first position in a word
B_2	The second position in a word
B_3	The third position in a word
M	Other positions in a word with more than five characters except the last
E	The last position in a word
S	Single character word

Table 1. 6-tag label set for the Chinese word segmentation

The probability model and corresponding feature function is defined over the set H×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

$$f(h,t) = \begin{cases} 1 & if\ h = h_i \quad and\, t = t_i \\ 0 & else \end{cases} \quad (5)$$

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.

A feature template set for observe chain is shown in Table 2. C_i means the character at the *ith* poison. The C_iC_{i+1} means the combination of two characters C_i and C_{i+1}. The $C_{i-1}C_iC_{i+1}$ means the combination of three characters C_{i-1}, C_i, and C_{i+1}. In the table, $S(C_0)$ stands for predefined class of the character C_0. There are five classes predefined: numbers represent class 1, English letters represent class 2, punctuation represents class 3, Chinese characters represent class 4, and other characters represents class 5. We also import some outer lexical information like the outer dictionary to build the outer information template. The outer information template is derived from an outer lexical dictionary, which contains words and their lexical information selected from the internet and other formatted corpus. The words together with their POSs are stored in the dictionary. The maximum length of the word in the dictionary is five characters. The $T(C_0)$ represents the POS of the C_0 if C_0 exists as a word in the outer dictionary. The $L(C_0)$ represents the maximum length of word in the sentence around C_0 that exist in the outer dictionary. The $P(C_0)$ represent the position of the C_0 in the word exist in the outer dictionary.

Type	Label	Position
Unigram	1) C_{-2} 2) C_{-1} 3) C_0 4) C_1 5) C_2	The current character and characters around it.
Bigram	1) $C_{-2}C_{-1}$ 2) $C_{-1}C_0$ 3) C_0C_1 4) C_1C_2	The combination of two characters.
Trigram	1) $C_{-2}C_{-1}C_0$ 2) $C_{-1}C_0C_1$ 3) $C_0C_1C_2$	The combination of three characters
Style	1) $S(C_0)$	The predefined type of the current character
Outer Info.	1) $T(C_0)$ 2) $L(C_0)$ 3) $P(C_0)$	The information from outer dictionary.

Table 2. Feature templates

The proposed feature template is applied to train the MUS CRF model and recognize the unknown words together with their POSs. After the recognition, the unknown words are added into the directed graph. Take the “庄炎林担任庄希泉基金会主席(Yanlin Zhuang act as chairman of the Xiquan Zhuang Fund)” for example, The person name “庄炎林(Yanlin Zhuang)” and “庄希泉(Xiquan Zhuang)” do not exist in the system dictionary. The word-based MMSM model can not segment and POS tag them correctly. The MUS CRF is applied to recognize the unknown person name from the position where word-based model does not work. After the recognition, the two unknown person names are recognized together with their POSs(nr means person name) and added into the directed graph as shown in figure 3.

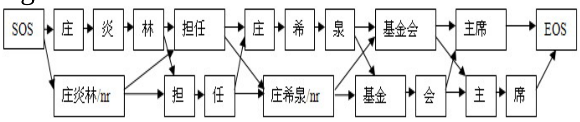

Figure 3. The directed graph after the unknown word recognition

4 Experimental and Results

We trained the hybrid model on the PKU2002 corpus, the PKU2002 corpus have 12 months corpus of Peoples’ Daily News of year 2002 that have been annotated. As the corpus are different from MicroBlog, so the final test result are not quite satisfying. The evaluation tools and standards for SIGHAN6 are adopted in the experiments. We present the results of our experiments in recall, precision and F-measure, which are defined in the equations below, as usual in such experiments.

$$recall = \frac{\#\text{ of correctly extracted words}}{\text{total }\#\text{ of words}}$$

$$precision = \frac{\#\text{ of correctly extracted words}}{\text{total }\#\text{ of recognized words}}$$

$$F - measure = \frac{2 \times recall \times precision}{recall + precision}$$

First the hybrid model was tested by using different size of training corpus with the same outer lexical dictionary (with the maximum length of word of five). The test corpus in our experiment is randomly selected 500KB raw corpus from the PKU corpus except the training corpus. The result is shown in table 3. The R in the table means recall; The P in the table means precision; The F in the table means F-measure. The R, P, F in the following tables has the same meaning. The IVR means recall of in-vocabulary words. The IVP means precision of in-vocabulary words. The IVF means F-measure of in-vocabulary words. The OOVR means recall of out-of-vocabulary words. The OOVP means precision of out-of-vocabulary words. The OOVF means F-measure of out-of-vocabulary words.

Train corpus	R	P	F
One month	0.9820	0.9853	0.9837
Two months	0.9829	0.9854	0.9841
Three months	0.9849	0.9879	0.9864
	IVR	IVP	IVF
One month	0.9838	0.9903	0.9870
Two months	0.9847	0.9894	0.9871
Three months	0.9859	0.9915	0.9887
	OOVR	OOVP	OOVF
One month	0.9456	0.8891	0.9165
Two months	0.9426	0.9027	0.9222
Three months	0.9574	0.8989	0.9272

Table 3. Chinese word segmentation result by using different size of training corpus.

In the experiments, as the size of training corpus increases, the training cost increases exponentially. It costs too much memories and time to train the model on four months corpus, so we only tested on one month, two months and three months corpus. We can see as the size of training corpus increases, the F-score of our model increases simultaneously.

We also tested the model using different outer dictionary. We adopted two different outer dictionaries, the maximum length of word in one dictionary is 4(DIC4), and the other is 5(DIC5). The first dictionary has about 100,000 words. The other has more than 300,000 words. The words in the dictionary are collected from the internet using our internet crawler. The training corpus in this experiment is the three months training corpus. The test corpus is randomly selected 500KB raw corpus. The result is shown in the following Table 4

Outer	R	P	F
DIC4	0.9784	0.9794	0.9789
DIC5	0.9849	0.9879	0.9864
	IVR	IVP	IVF
DIC4	0.9816	0.9859	0.9837
DIC5	0.9859	0.9915	0.9887
	OOVR	OOVP	OOVF
DIC4	0.8948	0.8227	0.8572
DIC5	0.9574	0.8989	0.9272

Table 4. Chinese word segmentation result by using different outer dictionary

The result of DIC5 is much better than the DIC4 because of the increasing of the maximum length of the word in the dictionary and the size of the dictionary.

We tested our POS tagging result using two training corpus. In the first experiment we trained one month corpus and in the second we trained two months corpus. The test corpus is randomly selected 500KB raw corpus. The result of POS tagging is in Table 5. The A in Table 5 means total accuracy of POS tagging. The IV-R means the POS tagging recall of in-vocabulary words. The OOV-R means the POS tagging recall of out-of-vocabulary words. The MT-R means POS tagging recall of multi-tag words.

Corpus	A	IV-R	OOV-R	MT-R
One month	0.9329	0.9518	0.6441	0.8972
Two months	0.9463	0.9711	0.6751	0.9064

Table 5. POS tagging result by using different size of training corpus.

We also deleted the outer dictionary for the multi-chain model and tested our model using the close test of SIGHAN6. We compared the Chinese word segmentation and POS tagging result with other participators' result (F-measure rank one in each corpus). We only adopted the close test of SIGHAN6 because we wanted to evaluate the model only. The Chinese word segmentation result is shown in Table 6

		R	P	F
CTB	Our	0.9620	0.9653	0.9636
	Rank1	0.9583	0.9596	0.9589
NCC	Our	0.9458	0.9329	0.9393
	Rank1	0.9402	0.9407	0.9405
SXU	Our	0.9658	0.9589	0.9623
	Rank1	0.9622	0.9625	0.9623

Table 6. Chinese word segmentation result of SIGHAN2007

We can see from the table that the hybrid model achieves competitive F-score and all the R-scores of the hybrid model are better than the rank one score in SIGHAN6. This is because the hybrid model combines the HMM model and CRF model together.

The POS tagging result on close test of SIGHAN6 is shown in Table 7

		A	IV-R	OOV-R	MT-R
CTB	Our	0.9456	0.9591	0.8032	0.9241
	Rank1	0.9428	0.9557	0.7522	0.9197
NCC	Our	0.9632	0.9801	0.7021	0.9340
	Rank1	0.9541	0.9738	0.5998	0.9195
PKU	Our	0.9503	0.9680	0.7102	0.9411
	Rank1	0.9411	0.9622	0.6057	0.9200

Table 7. POS tagging result of SIGHAN 2007

The hybrid model gets the highest score in Chinese POS tagging especially the OOV-R score in all corpuses. The MUS CRF in the hybrid model devotes a lot to this. The MUS CRF can recognize the POS of the unknown word and increase the performance of the whole model.

5 Conclusions

The MMSM model is adopted to combine the word-based HMM model and character-based

CRF model together. The word-based information is for known words segmentation and POS tagging while the character-based information is for the unknown words recognition and their POSs tagging. The MUS CRF is proposed to solve the unknown words recognition and their POS tagging synchronously. The adoption of the MUS CRF model decreases the computational cost of Dynamic CRF. Also it avoids using two separated linear-chain CRF models for the unknown word recognition and POS tagging. The hybrid model also decreases the computational cost without having to tagging all the characters in a sentence for Chinese word segmentation and POS tagging. Experimental results showed that the method achieves high accuracy compared to the state-of-the-art methods in both Chinese word segmentation and POS tagging.

The costs in the directed graph are encoded by the HMM model. We will adopted the CRF model to encode the cost in the directed graph, which will get rid of the limitations of hypothesis in the HMM model and combine more lexical information from the context in the directed graph to get higher precision.

Acknowledgments

The work is supported by the 863 National Advanced Technology Research Program of China (NO. 2012AA011103), and also supported by the Funding Project for AnHui Province Key Laboratory of Affective Computing and Advanced Intelligent Machine(1206c0805039), HeFei University of Technology. This project is also supported by the National Science Foundation for Post-doctoral Scientists of China (Grant No. 2012M511156) and China Postdoctoral Science Foundation(2012M511156).

References

Asahara, M. and Matsumoto, Y., 2003, Unknown Word Identification in Japanese Text Based on Morphological Analysis and Chunking, In *IPSJ SIG Notes Natural Language*, 2003-NL-154:47–54.

A. Berger, S. D. Pietra, and V. D. Pietra, A Maximum Entropy Approach to Natural Language Processing, *Computational Linguistics*, (22-1), March 1996.

Charles Sutton, Andrew McCallum and Khashayar Rohanimanesh. Dynamic Conditional Random Fields: Factorized Probabilistic Models for Labeling and Segmenting Sequence Data. *Journal of Machine Learning Research*. 2007:693-723

C. Sutton, A. McCallum, and K. Rohanimanesh, Dynamic Conditional Random Fields: Factorized Probabilistic Models for Labeling and Segmenting Sequence Data, *Journal of Machine Learning Research*, 2007(3):694-723.

E. Charniak, Hendrickson C., Jacobson N., and Perkowitz M. 1993. Equations for part of speech tagging. *In Proceedings of the Conference of the American Association for Artificial Intelligence*. 1993:784-789.

Gao J., Wu A., Li M., Huang C. N., Li H., Xia X., and Qin H. 2004. Adaptive Chinese word segmentation. *In Proceedings the 41st Annual Meeting of the Association for Computational Linguistics*, 2004:21-26.

Huang Degen and Sun Xiao. 2007. An Integrative Approach to Chinese Named Entity Recognition. *In Proceedings of Sixth International Conference on Advanced Language Processing and Web Information Technology*. 2007:171-176

H. T. Ng and J. K. Low. 2004. Chinese Part-Of-Speech Tagging: One-at-a-Time or All-at-Once? Word-Base or Character-Based? *In Proceedings of Conference on Empirical Methods in Natural Language Processing*. 2004.

J. Lafferty, A. McCallum, and F. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *International Conference on Machine Learning*, 2001.

K. P. Murphy, Y. Weiss, and M. I. Jordan. Loopy belief propagation for approximate inference: An empirical study. *In Conference on Uncertainty in Artificial Intelligence*, 1999:467-475.

Lafferty, J., McCallum, A., and Pereira, F. 2001. Conditional Random Field: Probabilistic models for segmenting and labeling sequence data. *In Proceedings of the International Conference on Machine Learning*. 2001:282-289

Lance A. Ramshaw and Mitchell P. Marcus. 1995. Text chunking using transformation-based learning. In Proceedings of the 3rd Workshop on Very Large Corpora, 1999:82-94.

Masayuki Asahara and Yuji Matsumoto. 2000. Extended Models and Tools for High-performance Parts-of-Speech Tagger. *In Proceedings of the 18th International Conference on Computational Linguistics*, 2000:21–27.

Masayuki Asahara, Chooi Ling Goh, Xiaojie Wang, and Yuji Matsumoto. 2003. Combining Segmenter and Chunker for Chinese Word Segmentation. *In Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing*, 2003:144–147.

Nianwen Xue. 2003. Chinese Word Segmentation as Character Tagging. *International Journal of Computational Linguistics and Chinese*, 8(1):29–48.

Peng, F., Feng, F., and McCallum, A. 2004. Chinese segmentation and new word detection using conditional random fields. *In Proceedings of the Computational Linguistics*, 2004:562-568.

Richard Sproat and Thomas Emerson. 2003. The First International ChineseWord Segmentation Bakeoff. *In Proceedings of the Second SIGHANWorkshop on Chinese Language Processing*, 2003:133–143.

S. M. Aji, G. B. Horn, and R. J. McEliece, On the convergence of iterative decoding on graphs with a single cycle. *In Proc. IEEE Int'l Symposium on Information Theory*, 1998.

Shi, W. 2005. Chinese Word Segmentation Based On Direct Maximum Entropy Model. *In Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing.*

Thorston Brants, 2000. TnT A Statistical Parts-of-Speech Tagger. *In Proceedings of Sixth Applied Natural Language Processing Conference.* 2000:224-231.

Tatsumi Yoshida, Kiyonori Ohtake, and Kazuhide Yamamoto. 2003. Performance Evaluation of Chinese Analyzers with Support Vector Machines. *Journal of Natural Language Processing*, 10(1):109–131.

Wu Y. C., Chang C. H. and Lee Y. S. 2006a. A general and multi-lingual phrase chunking model based on masking method. *Lecture Notes in Computer Science: Computational Linguistics and Intelligent Text Processing*, 3878: 144-155.

Wu Y. C., Fan T. K., Lee Y. S. and Yen S. J. 2006b. Extracting named entities using support vector machines, *Lecture Notes in Bioinformatics: Knowledge Discovery in Life Science Literature*, 3886: 91-103.

Wu Y. C., Lee Y. S., and Yang J. C. 2006c. The Exploration of Deterministic and Efficient Dependency Parsing. *In Proceedings of the 10th Conference on Natural Language Learning*.

Y. Shi, M. Wang, A Dual-layer CRFs Based Joint Decoding Method for Cascaded Segmentation and Labeling Tasks. *In International Joint Conferences on Artificial Intelligence,* 2007.

Zhao Hai, Huang Chang-Ning, and Li Mu, An Improved Chinese Word Segmentation System with Conditional Random Field, *In Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, 2006:162-165.