Which Performs Better on In-Vocabulary Word Segmentation: Based on Word or Character?

Zhenxing Wang[1,2], Changning Huang[2] and Jingbo Zhu[1]
1 Institute of Computer Software and Theory, Northeastern University,
Shenyang, China, 110004
2 Microsoft Research Asia, 49, Zhichun Road,
Haidian District, Beijing, China, 100080
zxwang@ics.neu.edu.cn
v-cnh@microsoft.com
zhujingbo@mail.neu.edu.cn

Abstract*

Since the first Chinese Word Segmentation (CWS) Bakeoff on 2003, CWS has experienced a prominent flourish because Bakeoff provides a platform for the participants, which helps them recognize the merits and drawbacks of their segmenters. However, the evaluation metric of bakeoff is not sufficient enough to measure the performance thoroughly, sometimes even misleading. One typical example caused by this insufficiency is that there is a popular belief existing in the research field that segmentation based on word can yield a better result than character-based tagging (CT) on in-vocabulary (IV) word segmentation even within closed tests of Bakeoff. Many efforts were paid to balance the performance on IV and out-of-vocabulary (OOV) words by combining these two methods according to this belief. In this paper, we provide a more detailed evaluation metric of IV and OOV words than Bakeoff to analyze CT method and combination method, which is a typical way to seek such balance. Our evaluation metric shows that CT outperforms dictionary-based (or so called word-based in general) segmentation on both IV and OOV words within Bakeoff closed tests. Furthermore, our analysis shows that using confidence measure to combine the two segmentation results should be under certain limitation.

1 Introduction

Chinese Word Segmentation (CWS) has been witnessed a prominent progress in the last three Bakeoffs (Sproat and Emerson, 2003), (Emerson, 2005), (Levow, 2006). One of the reasons for this progress is that Bakeoff provides standard corpora and objective metric, which makes the result of each system comparable. Through those evaluations researchers can recognize the advantage and disadvantage of their methods and improve their systems accordingly. However, in the evaluation metric of Bakeoff, only the overall F measure, precision, recall, IV (in-vocabulary) recall and OOV (out-of-vocabulary) recall are included and such a metric is not sufficient to give a completely measure on the performance, especially when the performance on IV and OOV word segmentation need to be evaluated. An important issue is that segmentation based on which, word or character, can yield the better performance on IV words. We give a detailed explanation about this issue as following.

Since CWS was firstly treated as a character-based tagging task (we call it "CT" for short hereafter) in (Xue and Converse, 2002), this method has been widely accepted and further developed by researchers (Peng et al., 2004), (Tseng et al., 2005), (Low et al., 2005), (Zhao et al., 2006). Relatively to dictionary-based

* The work is done when the first author is working in MSRA as an intern.

segmentation (we call it "DS" for short hereafter), CT method can achieve a higher accuracy on OOV word recognition and a better performance of segmentation in whole. Thus, CT has drawn more and more attention and became the dominant method in the Bakeoff 2005 and 2006.

Although CT has shown its merits in word segmentation task, some researchers still hold the belief that on IV words DS can perform better than CT even in the restriction of Bakeoff closed test. Consequently, many strategies are proposed to balance the IV and OOV performance (Goh et al., 2005), (Zhang et al., 2006a). Among these strategies, the confidence measure used to combine the results of CT and DS is a straight-forward one, which is introduced in (Zhang et al., 2006a). The basic assumption of such combination is that DS method performs better on IV words and Zhang derives this belief from the fact that DS achieves higher IV recall rate as Table 1 shows. In which AS, CityU, MSRA and PKU are four corpora used in Bakeoff 2005 (also see Table 2 for detail). We provide a more detailed evaluation metric to analyze these two methods, including precision and F measure of IV and OOV respectively and our experiments show that CT outperforms DS on both IV and OOV words within Bakeoff closed test. The precision and F measure are existing metrics and the definitions of them are clear. Here we just employ them to evaluate segmentation results. Furthermore, our error analysis on the results of combination reveals that confidence measure in (Zhang et al., 2006a) has a representation flaw and we propose an EIV tag method to revise it. Finally, we give an empirical comparison between existing pure CT method and combination, which shows that pure CT method can produce state-of-the-art results on both IV word and overall segmentation.

Corpus	R_{IV}		R_{OOV}	
	DS	CT	DS	CT
AS	0.982	0.967	0.038	0.647
CityU	0.989	0.967	0.164	0.736
MSRA	0.993	0.972	0.048	0.716
PKU	0.981	0.955	0.408	0.754

Table 1 IV and OOV recall in (Zhang et al., 2006a)

The rest of this paper is organized as follows. In Section 2, we give a brief introduction to Zhang's DS method and subword-based tagging, which is a special CT method. And by comparing the results of this special CT method and DS according our detailed metric, we show that CT performs better on both IV and OOV. We review in Section 3 how confidence measure works and indicate its representation flaw. Furthermore, an "EIV" tag method is proposed to revise the confidence measure. In Section 4, the experimental results of existing pure CT method are demonstrated to compare with combination result, based on which we discuss the related work. In Section 5, we conclude the contributions of this paper and discuss the future work.

2 Comparison between DS and CT Based on Detailed Metric

We proposed a detailed evaluation metric for IV and OOV word identification in this section and experiments based on the new metric show that CT outperforms DS not only on OOV words but also on IV words with F-measure of IV. All the experiments in this paper conform to the constraints of closed test in Bakeoff 2005 (Emerson, 2005). It means that any resource beyond the training corpus is excluded. We first review how DS and CT work and then present our evaluation metric and experiment results. There is one thing should be emphasized, by comparing DS and CT result we just want to verify that our new metric can show the performance on IV words more objectively. Since either DS or CT implementation has specific setting here we should not extend the comparison result to a general sense between those generative models and discriminative models.

2.1 Dictionary-based segmentation

For the dictionary-based word segmentation, we collect a dictionary from training corpus first. Instead of Maximum Match, trigram language model[2] trained on training corpus is employed for disambiguation. During the disambiguation procedure, a beam search decoder is used to seek the most possible segmentation. Since the setting in our paper is consistent with the closed test of

[2] Language model used in this paper is SLRIM from http://www.speech.sri.com/projects/srilm/

Bakeoff, we can only use the information we learn from training corpus though other open resources may be helpful to improve the performance further. For detail, the decoder reads characters from the input sentence one at a time, and generates candidate segmentations incrementally. At each stage, the next incoming character is combined with an existing candidate in two different ways to generate new candidates: it is either appended to the last word in the candidate, or taken as the start of a new word. This method guarantees exhaustive generation of possible segmentations for any input sentence. However, the exponential time and space of the length of the input sentence are needed for such a search and it is always intractable in practice. Thus, we use the trigram language model to select top B (B is a constant predefined before search and in our experiment 3 is used) best candidates with highest probability at each stage so that the search algorithm can work in practice. Finally, when the whole sentence has been read, the best candidate with the highest probability will be selected as the segmentation result. Here, the term "dictionary-based" is exactly the method implemented in (Zhang et al., 2006a), it does not mean the generative language model in general.

2.2 Character-based tagging

Under CT scheme, each character in one sentence is labeled as 'B' if it is the beginning of a word, 'O' tag means the current character is a single-character word, other character is labeled as 'I'. For example, "全中国 (whole China)" is labeled as "全 (whole)/O 中 (central)/B 国 (country)/I".

In (Zhang et al., 2006a), the above CT method is developed as subword-based tagging. First, the most frequent multi-character words and all single characters in training corpus are collected as subwords. During the subword-based tagging, a subword is viewed as an unit instead of several separate characters and given only one tag. For example, in subword-based tagging, "全中国 (whole China)" is labeled as " 全 (whole)/O 中国 (China)/O", if the word "中国 (China)" is collected as a subword. As the preprocessing, both training and test corpora are segmented by maximum match with subword set as dictionary. After this preprocessing, every sentence in both training and test corpora becomes subword sequence. Finally, the tagger is trained by CRFs approach[3] on the training data. Although word information is integrated into this method, it still works in the scheme of "IOB" tagging. Thus, we still call subword-based tagging as a special CT method and in the reminder of this paper "CT" means subword-based tagging in Zhang's paper and "Pure CT" means CT without subword.

2.3 A detailed evaluation metric

In this paper, data provided by Bakeoff 2005 is used in our experiments in order to compare with the published results in (Zhang et al., 2006a). The statistics of the corpora for Bakeoff 2005 are listed in Table 2 (Emerson, 2005).

Corpus	Encoding	#Training words	#Test words	OOV rate
AS	Big5	5.45M	122K	0.043
CityU	Big5	1.46M	41K	0.074
MSRA	GB	2.37M	107K	0.026
PKU	GB	1.1M	104K	0.058

Table 2 Corpora statistics of Bakeoff 2005

Evaluation standard is also provided by Bakeoff, including overall precision, recall, F measure, IV recall and OOV recall (Sproat and Emerson, 2003), (Emerson, 2005). However, some important metrics, such as F measure and precision of both IV and OOV words are omitted, which are necessary when the performance of IV or OOV word identification need to be judged. Thus, in order to judge the results of each experiment, a more detailed evaluation with precision and F measure of both IV and OOV words included is used. To calculate the IV and OOV precision and recall, we firstly divide words of the segmenter's output and gold data into IV word and OOV word sets respectively with the dictionary collected from the training corpus. Then, for IV and OOV word sets respectively, the IV (or OOV) recall is the proportion of the correctly segmented IV (or OOV) word tokens to all IV (or OOV) word tokens in the gold data, and IV (or OOV) precision is the proportion of the correctly segmented IV

[3] CRF tagger in this paper is implemented by CRF++ downloaded from http://crfpp.sourceforge.net/

(or OOV) word tokens to all IV (or OOV) word tokens in the segmenter's output. One thing have to be emphasized is that the single character in test corpus will be defined as OOV if it does not appear in training corpus. We will see later in this section, by this evaluation, some facts covered by the bakeoff evaluation can be illustrated by our new evaluation metric.

Here, we repeat two experiments described in (Zhang et al., 2006a), namely dictionary-based approach and subword-based tagging. For CT method, top 2000 most frequent multi-character words and all single characters in training corpus are selected as subwords and the feature templates used for CRF model is listed in Table 3. We present all the segmentation results in Table 6 to see the strength and weakness of each method conveniently.

Based on IV and OOV recall as we show in Table 1, Zhang argues that the DS performs better on IV word identification while CT performs better on OOV words. But we can see from the results in Table 6 (the lines about DS and CT), the IV precision of DS approach is much lower than that of CT on all the four corpora, which also causes a lower F measure of IV. The reason for low IV precision of DS is that many OOV words are segmented into two IV words by DS. For example, OOV word “歌唱班(choral)” is segmented into“歌唱(sing) 班(class)” by DS. These wrongly identified IV words increase the number of all IV words in the segmenter's output and cause the low IV precision of the DS result. Since the F measure of IV is a more reasonable metric of performance of IV than IV recall only, Table 6 shows that CT method outperforms the DS on IV word segmentation over all four corpora. The comparison also shows that CT outperforms the DS on OOV and overall segmentation as well.

Type	Feature	Function
Unigram	C_{-2}, C_{-1}, C_0, C_1, C_2	Previous two, current and next two subword
Bigram	$C_{-2}\ C_{-1}$, $C_{-1}\ C_0$, $C_0\ C_1$, $C_1\ C_2$	Two adjacent subwords
Jump	$C_{-1}\ C_1$	Previous character and next subwords

Table 3 Feature templates used for CRF in our experiments

3 Balance between IV and OOV Performance

There are other strategies such as (Goh et al., 2005) trying to seek balance between IV and OOV performance. In (Goh et al, 2005), information in a dictionary is used in a statistical model. In this way, the dictionary-based approach and the statistical model are combined. We choose the confidence measure to study because it is straight-forward. We show in this section that there is a representation flaw in the formula of confidence measure in (Zhang et al., 2006a). And we propose an “EIV” tag method to solve this problem. Our experiments show that confidence measure with EIV tag outperforms CT and DS alone.

3.1 Confidence measure

Confidence Measure (CM) means to seek an optimal tradeoff between performance on IV and OOV words. The basic idea of CM comes from the belief that CT performs better on OOV words while DS performs better on IV words. When both results of CT and DS are available, the CM can be calculated according to the following formula in (Zhang et al., 2006a):

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

Here, w is a subword, t_{iob} is “IOB” tag given by CT and t_w is “IOB” tag generated by DS. In the first term of the right hand side of the formula, $CM_{iob}(t_{iob} \mid w)$ is the marginal probability of t_{iob} (we call this marginal probability “MP” for short). And in the second term, $\delta(t_w, t_{iob})_{ng}$ is a Kronecker delta function, returning 1 if and only if t_w and t_{iob} are identical, else returning 0. But if $\delta(t_w, t_{iob})_{ng} = 1$, there is no requirement of replacement at all. While if $\delta(t_w, t_{iob})_{ng} = 0$, when $t_w \neq t_{iob}$, CM depends on the first term of its right hand side only and α is unnecessary to be set as a weight. Finally, α in the formula is a weight to seek balance between CT tag and DS tag. Another parameter here is a threshold t for the CM. If CM is less than t, t_w replaces t_{iob} as the final tag, otherwise t_{iob} will be remained as the final tag. However, two parameters in the CM, namely α and t, are unnecessary, because when MP is greater than or equal to t/α, t_{iob} will be kept, otherwise it will be replaced with t_w. Thus, the CM ultimately is the marginal probability of the "IOB" tag (MP). In the experiment of this paper, MP is used as CM because it is equivalent to Zhang's CM but more convenient to express.

3.2 Experiments and error analysis about combination

We repeat the experiments about CM in Zhang's paper (Zhang et al., 2006a) and show that there is a representation flaw in the CM formula. Furthermore, we propose an EIV tag method to make CM yield a better result.

In this paper, α = 0.8 and t = 0.7 (Parameters in two papers, Zhang et al. 2006a and Zhang et al. 2006b, are different. And our parameters are consistent with Zhang et al. 2006b which is confirmed by Dr Zhang through email) are used in CM, namely MP= 0.875 is the threshold. Here, in Table 4, we provide some statistics on the results of CT when MP is less than 0.875. From Table 4 we can see that even with MP less than 0.875, most of the subwords are still tagged correctly by CT and should not be revised by DS result. Besides, lots of the subwords with low MP contained by OOV words in test data, especially for the corpus whose OOV rate is high (i.e. on CityU corpus more than one third subwords with low MP belong to OOV word) and performance on OOV recognition is the advantage of CT rather than that of DS approach. Thus when combining the results of the two methods, it is the t_{iob} should be maintained if the subword is contained by an OOV word. Therefore, the CM formula seems somewhat unreasonable.

The error analysis about how many original errors are eliminated and how many new errors are introduced by CM is provided in Table 5 (the columns about CM). Table 5 illustrates that, after combining the two results, most original errors on IV words are corrected because DS can achieve higher IV recall as described in Zhang's paper. But on OOV part, more new errors are introduced by CM and these new errors decrease the precision of the IV words. For example, the OOV words "警卫队员 (guard member)" and " 设计费 (design fee)" is recognized correctly by CT but with low CM. In the combining procedure, these words are wrongly split as IV errors: "警卫 (guard) 队员 (member)" and "设计 (design) 费 (fee)". Thus, for two corpora (i.e. CityU and AS), F measure of IV and overall F measure decreases since there are more new errors introduced than original ones eliminated and only on the other two corpora (MSRA and PKU), overall F measure of combination method is higher than CT alone, which is shown in Table 6 by the lines about combination.

3.3 EIV tag method

Since combining the two results by CM may produce an even worse performance in some case, it is worthy to study how to use this CM to get an enhanced result. Intuitively, if we can change only the CT tags of the subwords which contained in IV word while keep the CT tags of those contained in OOV words unchanged, we will improve the final result according to our error analysis in Table 5. Unfortunately, only from the test data, we can get the information whether a subword contained in an IV word, just as what we do to get Table 4. However, we can get an approximate estimation from DS result. When using subwords to re-segment DS result[4], all the fractions re-segmented out of multiple-character words, including both multiple-character words and single characters, will be given an "EIV" tag, which means that the current multiple-character word or single character is contained in an IV word with high probability. For example, "人力资源 (human resource)" in DS result is a whole word. However, only "资源 (resource)" belongs to the subword set, so during the re-segmentation "人力资源 (human resource)" will be re-segmented as "人 (people) 力 (force) 资源 (resource)". All these three fractions will be labeled with an "EIV" tag respectively. It is reasonable because all the multiple-character words in the DS result can match an IV word. After this procedure, when combining

[4] For the detail, please refer to (Zhang et al., 2006a).

Corpus	AS	CityU	MSRA	PKU
# subword tokens belong to IV	10010	4404	9552	9619
# subword tokens belong to IV and tagged correctly by CT	7452	3434	7452	7213
# subword tokens belong to IV and tagged wrongly by CT	2558	970	2100	2406
# subword tokens belong to OOV	5924	2524	2685	3580
# subword tokens belong to OOV and tagged correctly by CT	3177	1656	1725	2208
# subword tokens belong to OOV and tagged wrongly by CT	2747	868	960	1372

Table 4 Results of CT when MP is less than 0.875

Corpus	AS		CityU		MSRA		PKU	
Method	CM	EIV	CM	EIV	CM	EIV	CM	EIV
#original errors eliminated on IV	1905	1003	904	469	1959	1077	1923	1187
#original errors eliminated on OOV	755	75	155	80	104	30	230	76
#original errors eliminated totally	2660	1078	1059	549	2063	1107	2153	1263
#new errors introduced on IV	441	185	80	50	148	68	211	118
#new errors introduced on OOV	2487	77	1320	103	1517	57	1681	58
# new errors introduced totally	2928	262	1400	153	1665	125	1892	176

Table 5 Error analysis of confidence measure with and without EIV tag

the two results, only the CT tag with EIV tags and low MP will be replaced by DS tag, otherwise the original CT tag will be maintained. Under this condition the errors introduced by OOV will not happen and enhanced results are listed in Table 6 lines about EIV. We can see that on all four corpora the overall F measure of EIV result is higher than that of CT alone, which show that our EIV method works well. Now, let's check what changes happened in the number of error tags after EIV condition added into the CM. We can see from the Table 5 columns about EIV, there are more errors eliminated than the new errors introduced after EIV condition added into CM and most CT tags of subwords contained in OOV words maintained unchanged as we supposed. And then, our results (in Table 6 lines about EIV) are comparable with that in Zhang's paper. Thus, there may be some similar strategies in Zhang's CM too but not presented in Zhang's paper.

4 Discussion and Related Works

Although the method such as confidence measure can be helpful at some circumstance, our experiment shows that pure character-based tagging (pure CT) can work well with reasonable features and tag set. In (Zhao et al., 2006), an enhanced CRF tag set is proposed to distinguish different positions in the multi-character words when the word length is less than 6. In this method, feature templates are almost the same as shown in Table 3 with a 3-character window and a 6-tag set {B, B2, B3, M, E, O} is used. Here, tag B and E stand for the first and the last position in a multi-character word, respectively. S stands up a single-character word. B2 and B3 stand for the second and the third position in a multi-character word, whose length is larger than two-character or three-character. M stands for the fourth or more rear position in a multi-character word, whose length is larger than four-character.

In Table 6, the lines about "pure CT" provide the results generated by pure CT with 6-tag set. We can see from the Table 6 this pure CT approach achieves the state-of-the-art results on all the corpora. On three of the four corpora (AS, MSRA and PKU) this pure CT method gets the best result. Even on IV word, this pure CT approach outperforms Zhang's CT method and produces comparable results with combination with EIV tags, which shows that pure CT method can perform well on IV words too. Moreover, this character-based tagging approach is more clear and simple than the confidence measure method.

Although character-based tagging became mainstream approach in the last two Bakeoffs, it does not mean that word information is valueless in Chinese word segmentation. A word-based perceptron algorithm is proposed recently (Zhang and Clark, 2007), which views Chinese word segmentation task from a new angle instead of character-based tagging and gets comparable results with the best results of Bakeoff.

Corpus	Method	R	P	F	R_{IV}	P_{IV}	F_{IV}	R_{OOV}	P_{OOV}	F_{OOV}
AS	DS	0.943	0.881	0.911	0.984	0.892	0.935	0.044	0.217	0.076
	CT	0.954	0.938	0.946	0.967	0.960	0.964	0.666	0.606	0.635
	Combination	0.958	0.929	0.943	0.980	0.945	0.962	0.487	0.593	0.535
	EIV tag	0.960	0.942	0.951	0.973	0.962	0.968	0.667	0.624	0.645
	Pure CT	0.958	0.947	0.953	0.971	0.963	0.967	0.682	0.618	0.648
CityU	DS	0.928	0.848	0.886	0.989	0.865	0.923	0.162	0.353	0.223
	CT	0.947	0.940	0.944	0.963	0.964	0.964	0.739	0.717	0.728
	Combination	0.954	0.922	0.938	0.984	0.938	0.961	0.581	0.693	0.632
	EIV tag	0.953	0.949	0.951	0.970	0.968	0.969	0.744	0.750	0.747
	Pure CT	0.947	0.948	0.948	0.967	0.973	0.970	0.692	0.660	0.676
MSRA	DS	0.969	0.927	0.947	0.994	0.930	0.961	0.036	0.358	0.066
	CT	0.963	0.964	0.963	0.970	0.979	0.975	0.698	0.662	0.680
	Combination	0.977	0.961	0.969	0.990	0.970	0.980	0.511	0.653	0.574
	EIV tag	0.972	0.970	0.971	0.980	0.982	0.981	0.696	0.679	0.688
	Pure CT	0.972	0.975	0.973	0.978	0.986	0.982	0.750	0.632	0.686
PKU	DS	0.948	0.911	0.929	0.981	0.920	0.950	0.403	0.711	0.515
	CT	0.944	0.945	0.945	0.955	0.966	0.961	0.763	0.727	0.745
	Combination	0.955	0.942	0.949	0.973	0.953	0.963	0.664	0.782	0.718
	EIV tag	0.950	0.952	0.951	0.961	0.970	0.966	0.768	0.753	0.760
	Pure CT	0.946	0.957	0.951	0.956	0.973	0.964	0.672	0.580	0.623

Table 6 Results of different approach used in our experiments (White background lines are the results we repeat Zhang's methods and they have some trivial difference with Table 1.)

Therefore, the most important thing worth to pay attention in future study is how to integrate linguistic information into the statistical model effectively, no matter character or word information.

5 Conclusions and Future Work

In this paper, we first provided a detailed evaluation metric, which provides the necessary information to judge the performance of each method on IV and OOV word identification. Second, by this evaluation metric, we show that character-based tagging outperforms dictionary-based segmentation not only on OOV words but also on IV words within Bakeoff closed tests. Furthermore, our experiments show that confidence measure in Zhang's paper has a representation flaw and we propose an EIV tag method to revise the combination. Finally, our experiments show that pure character-based approach also can achieve good IV word and overall performance. Perhaps, there are two reasons that existing combination results don't outperform the pure CT. One is that most information contained in statistic language model is already captured by the CT feature templates in CRF framework. The other is that confidence measure may not be the effective way to combine the DS and CT results.

In the future work, our research will focus on how to integrate word information into CRF features rather than using it to modify the results of CRF tagging. In this way, we can capture the word information meanwhile avoid destroying the optimal output of CRF tagging.

Acknowledgement

The authors appreciate Dr. Hai Zhao in City University of Hong Kong and Dr. Ruiqiang Zhang in Spoken Language Communications Lab, ATR, Japan providing a lot of help for this paper. Thank those reviewers who gave the valuable comment to improve this paper.

References

Thomas Emerson. 2005. The Second International Chinese Word Segmentation Bakeoff. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pages 123-133, Jeju Island, Korea:

Chooi-Ling Goh, Masayuku Asahara and Yuji Matsumoto. 2005. Chinese Word Segmentatin by Classification of Characters. *Computational Linguistics and*

Chinese Language Processing, Vol. 10(3): pages 381-396.

Gina-Anne Levow. 2006. The Third International Chinese Language Processing Bakeoff: Word Segmentation and Named Entity Recognition. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing* , pages 108-117, Sydney: July.

Jin Kiat Low, Hwee Tou Ng, and Wenyuan Guo. 2005. A Maximum Entropy Approach to Chinese Word Segmentation. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pages 161-164, Jeju Island, Korea.

Fuchun Peng, Fangfang Feng, and Andrew McCallum. 2004. Chinese segmentation and new word detection using conditional random fields. In COLING 2004, pages 562–568. Geneva, Switzerland.

Richard Sproat and Thomas Emerson. 2003. The First International Chinese Word Segmentation Bakeoff. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, pages 133-143, Sapporo, Japan: July 11-12,

Huihsin Tseng, Pichuan Chang et al. 2005. A Conditional Random Field Word Segmenter for SIGHAN Bakeoff 2005. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pages 168-171, Jeju Island, Korea.

Neinwen Xue and Susan P. Converse. 2002. Combining Classifiers for Chinese Word Segmentation. In *Proceedings of the First SIGHAN Workshop on Chinese Language Processing*, pages 63-70, Taipei, Taiwan.

Ruiqiang Zhang, Genichiro Kikui and Eiichiro Sumita. 2006a. Subword-based Tagging by Conditional Random Fields for Chinese Word Segmentation. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion volume*, pages 193-196. New York, USA.

Ruiqiang Zhang, Genichiro Kikui and Eiichiro Sumita. 2006b. Subword-based Tagging for Confidence-dependent Chinese Word Segmentaion. In *Proceedings of the COLING/ACL, Main Conference Poster Sessions*, pages 961-968. Sydney, Australia.

Yue Zhang and Stephen Clark. 2007. Chinese Segmentation with a Word-Based Perceptron Algorithm. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 840-847. Prague, Czech Republic.

Hai Zhao, Changning Huang et al. 2006. Effective Tag Set Selection in Chinese Word Segmentation via Conditional Random Field Modeling. In *Proceedings of PACLIC-20*. pages 87-94. Wuhan, China, Novemeber.