Incorporating Global Information into Supervised Learning for Chinese Word Segmentation

Hai Zhao and Chunyu Kit
Department of Chinese, Translation and Linguistics,
City University of Hong Kong,
83 Tat Chee Avenue, Kowloon, Hong Kong, China
{haizhao,ctckit}@cityu.edu.hk

Abstract

This paper presents a novel approach to Chinese word segmentation (CWS) that attempts to utilize global information (GI) such as co-occurrence of sub-sequences and outputs of unsupervised segmentation in the whole text for further enhancement of the state-of-the-art performance of conditional random fields (CRF) learning. In the existing work of CWS, supervised and unsupervised learning seldom joined, and thus strengthened, with each other. Our attempt here is to integrate unsupervised learning into supervised learning for CWS. Our experimental results show that character-based CRF framework can effectively make use of global information for performance enhancement on top of the best existing results.

1 Introduction

Machine learning methods have shown their power in Chinese word segmentation (CWS) since the first International Chinese word segmentation Bakeoff in 2003 (Sproat and Emerson, 2003). The task of CWS is to segment an input sequence of characters into a sequence of words. Character-based tagging is a simple but effective formulation of the problem suitable for various competitive supervised machine learning models (Xue, 2003; Peng et al., 2004; Low et al., 2005; Tseng et al., 2005; Zhao et al., 2006a). However, all these models only make use of local information and neglect global information such as the occurrences of a sub-sequence in different sentences. This paper will explore how to integrate such information effectively into supervised learning for CWS.

Conventionally, unsupervised and supervised learning are considered two disjoint categories of main techniques for CWS. The latter relies on a pre-segmented corpus as training data, or at least a predefined lexicon; the former applies when such resources are unavailable (Brent, 1999; Ge et al., 1999; Peng and Schuurmans, 2001; Goldwater et al., 2006). Sophisticated technologies have been developed in both categories. A very interesting question we could not find a good answer in existing work and thus intend to explore in this paper is how the two can join each other effectively, in particular, how the former can be integrated into the latter for performance enhancement.

More specifically, we will integrate two kinds of global information features into our character-based tagging system for CWS and examine their effectiveness. One is whether a sub-sequence occurs in different sequences (namely, sentences), and the other is whether a sub-sequence is identified by unsupervised segmentation as a word. Our experiments show that both of them are effective in improving CRF model's performance on CWS as character-based tagging.

The remainder of the paper is organized as follows. The next section discusses how the two types of global information are extracted from the raw text input in question. Section 3 formulates our integration of such global information into CRF learning for CWS. Then, our experimental results will be presented in Section 4. Section 5 discusses related work on using global information in natural language processing tasks and on ensemble of supervised and unsupervised segmentation. Finally, we summarize our research contribution in Section 6.

2 Unsupervised Segmentation

In principle, unsupervised segmentation assumes no pre-segmented training data nor a pre-defined lexicon.

An unsupervised segmentation strategy has to follow some predefined criterion, e.g., mutual information, to recognize a sub-sequence as a word.

Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics, pages 66–74,

(Sproat and Shih, 1990) was an early comprehensive investigation in this direction using mutual information. Many successive works applied mutual information criterion with different ensemble methods (Chien, 1997; Sun et al., 1998; Zhang et al., 2000; Yamamoto and Church, 2001; SUN et al., 2004).

Kit proposed a compression-based unsupervised segmentation algorithm, named after description-length-gain (DLG) based segmentation (Kit and Wilks, 1999; Kit, 2000). This method was used for out-of-vocabulary words identification in (Kit and Liu, 2005).

(Feng et al., 2004) proposed a statistical criterion called *accessor variety* (AV) to measure how likely a sub-sequence is a word, and then to find the best segmentation pattern that maximizes a target function of accessor variety and the length of the sub-sequence as variants. (Jin and Tanaka-Ishii, 2006) proposed *branch entropy* as another criterion for unsupervised segmentation. Both criteria share a similar assumption as in the fundamental work by (Harris, 1970): If the uncertainty of successive tokens increases, then the location is at a border. We consider (Feng et al., 2004) and (Jin and Tanaka-Ishii, 2006) to be, respectively, the discrete and continuous formulation of a similar idea.

The idea behind AV or branch entropy criterion is that word boundary occurs at the point where the uncertainty of successive character increases. Either of these two criterions make use of the measurement of such an uncertainty.

In this paper, we adopt AV as our unsupervised segmentation criterion to find segmented unit candidates, for it handles low-frequent words well, as reported in (Feng et al., 2004). As a measure to evaluate how independent a sub-sequence is, and thus how likely it is a word, the accessor variety of a sub-sequence s of more than one character is defined as

$$AV(s) = min\{L_{av}(s), R_{av}(s)\} \quad (1)$$

where the left and right accessor variety $L_{av}(s)$ and $R_{av}(s)$ are, respectively, defined as the number of distinct predecessor and successor characters. Sub-sequences with an AV value below a given threshold are firstly discarded. The remaining sub-sequences are considered to be potentially meaningful words. In addition, Feng et al. also applied heuristic rules to remove sub-sequences that consist of a word and adhesive characters. In this study, we will drop all additional rules for the sake of simplification.

3 System Description of Supervised Segmentation

3.1 Chinese Word Segmentation via CRF Modeling

Chinese word segmentation (CWS) was first formulated as a character tagging problem in (Xue, 2003), via labeling each character's position in a word. For example, the segmentation,

'自然科学/的/研究/不断/深入',

'natural science / of / research / continuously / deepen',

receives the tag (label) sequence '*BMMESBEBEBE*' as segmentation result, where the four tags B, M, E and S stand, respectively, for the beginning, middle and ending positions in a word, and a single character as a word. A MaxEnt model was trained for such character tagging task in (Xue, 2003).

Conditional random field (CRF) (Lafferty et al., 2001) is a statistical sequence modeling framework that outperforms other popular models such as MaxEnt method. CRF was first applied to CWS in (Peng et al., 2004), treating CWS as a binary decision task for each Chinese character in the input: is it the beginning of a word?

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

$$P_\lambda(y|s) = \frac{1}{Z}\exp(\sum_{c\in C}\sum_k \lambda_k f_k(y_c, y_{c-1}, s, c)), \quad (2)$$

where $Y = \{y_c\}$ is the label sequence for the sentence, s is the sequence of unsegmented characters, Z is a normalization term, f_k is a feature function, λ_k is the corresponding feature weight, C is the tag set and c indexes into characters in the sequence being labeled. Our implementation of character-based tagging for CWS uses the CRF++ package by Taku Kudo [1].

3.2 Tag Set and Feature Templates of Baseline System

Existing work shows that a 6-tag set enables CRF learning to achieve a better segmentation performance than any other tag sets (Zhao et al., 2006b).

[1] http://chasen.org/taku/software/CRF++/

Table 1: Definition of 6-tag set

Word Length	1	2	3	4	5	6	7 or longer
Tag Sequence	S	BE	BB_2E	BB_2B_3E	BB_2B_3ME	BB_2B_3MME	$BB_2B_3M...ME$

Table 2: Feature templates for baseline system

Code	Type	Feature	Description
a	Unigram	$C_n, n = -1, 0, 1$	The previous (current, next) character
b	Bigram	$C_nC_{n+1}, n = -1, 0$	The previous (next) character and current character
		$C_{-1}C_1$	The previous character and next character
c	Punctuation, Date, Digital and Letter	$T_{-1}T_0T_1$	T_i is type of previous, current or next character

Thus, we opt for this tag set and corresponding n-gram feature templates as our baseline system.

The 6-tag set includes six tags: B, B_2, B_3, M, E, S. Table 1 illustrates how words with different lengths are tagged with this tag set. Feature templates are shown in Table 2.

We give an explanation to feature template (c). It defines five classes of character sets: numbers or characters whose meanings are numbers represent class 1, those characters whose meanings are date and time represent class 2, English letters represent class 3, punctuation labels represent class 4 and other characters represent class 5.

3.3 Global Feature Templates

The basic idea of using global information for CWS is to inform the supervised learner what subsequence is word candidate determined by global information. We do this by marking all characters in the sub-sequences that are identified by either of our criteria, in the same way as we label the pre-segmented training data with the 6-tag set for CRF training. Consider that almost all characters in the text can appear in different sentences and our tagging clique is exactly character, we exclude unigram from global information since character alone brings little useful global information.

As mentioned above, we currently apply two criteria to identify such sub-sequences as likely word candidates, one is whether they co-occur in different sentences and the other is whether they are recognized by AV-based segmentation. Henceforth, we refer to them as COS (Co-Occurrence Sub-sequence) and AVS (Accessor Variety based Sub-sequence), respectively, for brevity.

The next issue we need to determine is the length of such sub-sequences. As for COS, we only use one feature template to express it. Notice that COS could identify too many overlapped sub-sequences. For example, in the following two sentences,

a 菲律宾/总统/埃斯特拉达/宣布/说/，/军方/已经/救出了/11名/人质/。

Philippine / president / Estrada / declared /, /the army / has / rescued / 11 / hostages.

b 埃斯特拉达/二号/接受/一家/电视台/采访时/说/，/三军/参谋长/向/他/报告/人质/已经/获救/。

Estrada / on 2nd / accepted / a / TV channel/ interview / to say /, / the army / chief of staff / reported to / him / the hostages / have been / rescued.

All sub-sequences '埃斯特拉达'(Estrada, the ex-president of Philippine),'埃斯'(Es),'埃斯特'(Est),'埃斯特拉'(Estra), and '斯特拉'(strada) meet the COS criterion. However, only '埃斯特拉达' is the one intended. It is the Longest one. Accordingly, the sub-sequence of tags 'BB_2B_3ME', instead of 'BB_2EBE' or any other tag sub-sequences, is assigned to mark '埃斯特拉达' in both of sentences. Therefore, we will only tag the longest COS sub-sequence in each case for CRF training. To avoid inefficiency in matching and searching for COS sub-sequences with too low reusability elsewhere, we opt for an ad hoc length constraint, that is, only sub-sequences of 2 to 7 characters are considered. According to Zhao et al. (2006b), less than 1% words that are longer than seven-character exist in all kinds of corpora.

The case of AVS is different from COS, in that few sub-sequences identified by AVS overlap each other. Thus, we use different feature templates to represent sub-sequences with different lengths. But, still, for the sake of efficiency, we only consider six feature templates generated by AVS, namely, sub-sequences from bigram to seven-gram.

4 Experimental Results

The experiments are performed in all eight corpora from the second International Chinese Word Segmentation Bakeoff (Bakeoff-2005) and the third International Chinese Language Processing Bakeoff (Bakeoff-2006)[2] (Emerson, 2005; Levow, 2006). Corpus size information is list in Table 3.

Table 3: Corpus size of Bakeoff-2005 and 2006 in number of words

Bakeoff-2005	AS	CityU	MSRA	PKU
Training(M)	5.45	1.46	2.37	1.1
Test(K)	122	41	107	104
Bakeoff-2006	AS	CityU	CTB	MSRA
Training(M)	5.45	1.64	0.5	1.26
Test(K)	91	220	154	100

Word segmentation performance is measured by F-measure ($F = 2RP/(R+P)$), where the recall (R) and precision (P) are the proportions of the correctly segmented words to all words in, respectively, the gold-standard segmentation and a segmenter's output. All experimental results in Subsection 4.1 will be evaluated by F-measure.

COS and AVS features are accumulated from the training corpus and test corpus without any annotation. As for AVS feature, we mark all subsequences of length 2 to 7 for AV value larger than 1.

Our comparison with existing work will be conducted in closed test track of Bakeoff. The rule for the closed test is that no additional information beyond training corpus is allowed, while open test of Bakeoff is without such restrict.

We consider two kinds of baseline results, one achieved only with n-gram feature templates (a) and (b) defined in Table 2, and the other achieved with feature template (c) besides n-gram. [3]

Table 4: Comparisons of the best existing results and ours only with n-gram features in data of Bakeoff-2005

Participant (Site ID)	AS	CityU	MSRA	PKU
Tseng(14)	0.947	**0.943**	**0.964**	0.95
Asahara(15b)	**0.952**	0.941	0.958	0.941
Chen(27)	0.945	0.94	0.960	**0.95**
Best closed	0.952	0.943	0.964	0.95
Kit(33)	0.923	/	0.950	0.916
Zhou(13)	/ /	/ /	0.957 0.966*	0.946
Andrew	/	/	0.968	/
Zhang	0.951	0.951	0.971	0.951
Baseline	0.953	0.948	0.973	0.952
COS	**0.955**	0.954	**0.975**	0.952
AVS	0.954	**0.956**	**0.975**	**0.954**
COS+AVS	0.955	0.956	0.975	0.953
Err. Redu. (%)	4.2	22.8	30.6	6.0

4.1 Comparisons of Best Existing Results and Our Results

The comparison of our segmentation performance against the best in Bakeoff-2005 is presented in Table 4 and 5. All participants with at least a best performance in the closed test of Bakeoff-2005 are given in Table 4 (Asahara et al., 2005; Chen et al., 2005; Tseng et al., 2005). 'Best closed' in this Table means the *official* best results in closed test of Bakeoff-2005.

Some state-of-the-art results after Bakeoff-2005, (Andrew, 2006; ZHOU, 2005; Zhang et al., 2006), are given below the best closed results for further comparison. Note that (ZHOU, 2005) reported an unofficial result (with a star in Table 4) in MSRA corpus. The results of (Kit and Liu,

[2]Bakeoff was an international evaluation proceeding on Chinese word segmentation and named entity recognition held by SIGHAN, a special interest group for Chinese language processing within the association of computational linguistics (ACL). Three Bakeoffs have been held since 2003. Bakeoff-2006 is the latest contest. All corpora used in this study are accessible from the official Bakeoff-2005 and 2006 website, http://www.sighan.org/bakeoff2005 and http://www.sighan.org/bakeoff2006.

[3]As we see, feature template (c) or similar feature templates of character set do cause performance enhancement Low et al. (2005). Some researchers that participated closed test of Bakeoff-2005 and 2006 used such kind of feature templates ZHOU (2005); Tsai et al. (2006); Zhao et al. (2006a); Zhu et al. (2006), while others did not. Therefore, we differentiate between these two baseline results in order to make segmentation results fairly comparable.

2005) is demonstrated for further discussion. In Table 5, the best official results in open test of Bakeoff-2005 are also given.

Table 5: Comparison of the best existing results and ours with n-gram features and features of character set in data of Bakeoff-2005

	AS	CityU	MSRA	PKU
Best closed	0.952	0.943	0.964	0.95
Baseline	0.954	0.956	0.974	0.953
COS	0.958	0.961	**0.975**	0.953
AVS	**0.959**	**0.962**	**0.975**	**0.955**
COS+AVS	0.959	0.962	0.975	0.954
Err. Redu. (%)	14.6	33.3	30.6	8.0
Best open	0.956	0.962	0.972	0.969

A summary of the best results in the closed test of Bakeoff-2006 are presented in Table 6. All participants with at least a third best performance in the closed test of Bakeoff-2006 are given in this table (Carpenter, 2006; Tsai et al., 2006; Wang et al., 2006; Zhao et al., 2006a; Zhu et al., 2006).

Table 6: A summary of the best results in the closed test of Bakeoff-2006

Participant (Site ID)	AS	CityU	CTB	MSRA
Zhu(1)	0.944	0.968	0.927	0.956
Carpenter(9)	0.943	0.961	0.907	0.957
Tsai(15)	0.957	**0.972**	/	0.955
Zhao(20)	**0.958**	0.971	**0.933**	/
Zhang(26)	0.949	0.965	0.926	0.957
Wang(32)	0.953	0.970	0.930	**0.963**
Best closed	0.958	0.972	0.933	0.963

The comparison of our segmentation performance against the best in Bakeoff-2006 is presented in Table 7 and 8. The best official results in open test of Bakeoff-2006 are also given in Table 8.

The rates of error reduction from best closed results of Bakeoff to our method (COS+AVS) are also given in Table 4 through 8. We can observe that our system achieve outstanding performance improvement compared to the best existing results.

Table 7: Comparisons of the best existing results and ours only with n-gram features in data of Bakeoff-2006

	AS	CityU	CTB	MSRA
Best closed	0.958	0.972	0.933	0.963
Baseline	0.954	0.969	0.932	0.961
COS	0.955	0.971	0.938	0.962
AVS	**0.957**	**0.973**	**0.940**	**0.963**
COS+AVS	0.957	0.973	0.940	0.963
Err. Redu. (%)	n/a	3.6	10.4	0.0

Table 8: Comparison of the best existing results and ours with n-gram features and features of character set in data of Bakeoff-2006

	AS	CityU	CTB	MSRA
Best closed	0.958	0.972	0.933	0.963
Baseline	0.959	0.972	0.934	0.961
COS	**0.960**	0.974	0.938	0.963
AVS	**0.960**	**0.975**	**0.941**	**0.963**
COS+AVS	0.960	0.975	0.940	0.963
Err. Redu. (%)	4.8	10.7	10.4	0.0
Best open	0.961	0.977	0.944	0.979

4.2 Discussion

We see that our system demonstrates an significant performance improvement from baseline results and achieves further improvement on top of the state-of-the-art performance for the closed test. More interestingly, it also gives highly comparable results with the best of the open test in Bakeoff, where any extra resources were allowed.

We also find that it is not helpful too much when we attempt to use both COS and AVS features. This potentially suggests that COS and AVS are not independent in feature characteristics.

Since global information is adopted through word candidate information, some researchers may argue that a lexicon that is directly extracted from training corpus or any other linguistic resources can be more helpful for performance enhancement. However, this is not always truth. Existing work did show that a proper external lexicon can be useful for performance improvement (Low et al., 2005), while our empirical study show that lexicon extracted from training corpus will not improve the performance but cause performance loss. The key issue in lexicon usage for CWS is how

to find a good lexicon instead of using a lexicon. Our technique here just defined a method to find a good one, since we observed that it can always cause performance enhancement.

5 Related Work

Global information was shown to be useful in many NLP tasks, especially in named entity recognition (NER). An early work on global information extraction for NER was (Mikheev et al., 1998), via a hybrid system integrating hand-coded rules and machine learning methods. Another attempt at using global information in NER was (Borthwick, 1999), via an additional maximum entropy classifier that tries to correct mistakes by using reference resolution. Chieu and Ng (2002) reported a MaxEnt approach to NER using global features as well as local features, showing better results than Borthwick's reference resolution classifier.

Whole-sentence exponential language models was proposed in (Rosenfeld et al., 2001). Although intended for the modeling of whole sentences, this model can be directly applied to the modeling of whole corpus, and it is quite impressive for its modeling ability in the whole text and label interaction, and it then was adopted, through different formulizations or modifications, in many successive works (Finkel et al., 2005; Takamura et al., 2005; Nakagawa and Matsumoto, 2006).

Our approach of using global information is largely similar to that of (Chieu and Ng, 2002). However, we verify the effectiveness of unsupervised segmentation outputs to supervised segmentation of the training corpus with CRF learning for the first time. In addition, CWS is so primary processing task that it is not trivial to find those analogous features proposed for NER in (Chieu and Ng, 2002) before we consider unsupervised segmentation.

To our knowledge, ensemble of unsupervised and supervised segmentation is a brand-new research area for CWS, in which successful research work has not yet been reported so far. (Kit and Liu, 2005) used a simple divide-and-conquer strategy to integrate unsupervised and supervised segmentation. Their unsupervised segmentation method is DLG-based, while supervised segmentation used example-based learning. DLG was only applied to recognize new words among the sequences of mono-character items in the example-based segmentation output. However, this method was not particularly successful, although the potentials of OOV detection via DLG-based segmentation was illustrated.

The broadly used n-gram language model is not strictly concerned with the use of global information, although it does involve the extraction of global information from the entire training corpus.

(Gao et al., 2005) integrated a trigram model in their log-linear Model for CWS. However, their experimental results were produced in the sense of open test of Bakeoff-2003. Thus it is essentially incomparable between their results and other results for closed test, though our baseline system for closed test can achieve the results as the same level as their system for open test. [4].

(Wang et al., 2006) used a maximum entropy model incorporated with n-gram language model to perform segmentation. Their integration strategy is that the score of a decoding path will be modified by adding the bigram of words with a weight λ at the word boundaries. The modification of path score follows the following formula.

$$V[j,i] = ME[j,i] + min_{i-1}^{k=1}\{V[i-1,k] + \lambda \text{Bigram}(w_{k,i-1}, w_{i,j})\} \quad (3)$$

where $V[j,i]$ is the score of a local best path which ends at the j^{th} character, $ME[j,i]$ is the score output from the MaxEnt tagger, and the last word on the path is $w_{i,j} = c_i...c_j$, the weight parameter λ is optimized by the test set used in Bakeoff-2005. Note that ensemble parameters in Wang et al. (2006) were acquired from additional linguistic resource[5]. In contrast, our approach here start from an simple assumption, i.e., accessor variety, as a criteria of unsupervised segmentation in training corpus without annotation.

(Zhang et al., 2006) reported an ensemble learning method for CWS that integrated n-gram language model into a sub-word based tagging system. However, their n-gram language model was

[4]We do not use Bakeoff-2003 data sets for our experiments. In fact, our baseline system only with n-gram features could achieve F-scores 0.973,0.948,0.873 and 0.956 respectively for the four Bakeoff-2003 corpora, AS, CityU, CTB, and PKU, while the corresponding F-scores in Gao et al. (2005) (for open test) are 0.958, 0.954, 0.904 and 0.955, respectively.

[5]This violates the rule of closed test in the Bakeoff that nothing else than the training corpus was allowed for training. Consider the fact that MSRA2006 training corpus is a subset of MSRA2005 training corpus. This may explain why their result is so much better than that of second best.

trained with an annotated corpus instead of plain texts and intended to recognize known words. They adopted a weighting scheme as the basic strategy for integrating n-gram models. A confidence measure was calculated from joint probability of sub-word tagging procedure: If it is higher than an empirical threshold, then the sub-word tagging were applied; otherwise, stay with the dictionary-based n-gram model.

Comparison between these existing works and ours is given in Subsection 4.1, showing that our system achieves better results, in general.

In practice, CRF character-based tagging for CWS can be a heavy computation burden for many current hardware settings. This situation could become worse if more tags were introduced for the purpose of performance enhancement. Thus, a proper tradeoff between computational efficiency and performance is also an important issue with this learning framework. Our experience tells that the proposed method integrated with global information adds only some minor computational cost beyond the baseline, especially for COS feature generation.

6 Conclusion

In this paper, we have presented a novel approach to integrating global information such as outputs of unsupervised segmentation to supervised learning for Chinese word segmentation. We learnt no previous work on CWS that ever attempted to strengthen supervised learning with unsupervised learning outcomes. We provide evidence to show that character-based CRF modeling for CWS can make use of global information effectively, and accordingly achieve a performance better than the best records in the past, according to our experimental results with the latest Bakeoff data sets.

This ensemble strategy allows global features to be used in exactly the same way as local features. In this regard, our approach is straightforward, easy for implementation, and highly adaptable, besides its effectiveness and efficiency.

Acknowledgements

The research described in this paper was supported by the Research Grants Council of Hong Kong S.A.R., China, through the CERG grant 9040861 (CityU 1318/03H) and by City University of Hong Kong through the Strategic Research Grant 7002037. Dr. Hai Zhao was supported by a postdoctoral Research Fellowship in the Department of Chinese, Translation and Linguistics, City University of Hong Kong.

References

Galen Andrew. 2006. A hybrid markov/semi-markov conditional random field for sequence segmentation. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 465–472. Association for Computational Linguistics, Sydney, Australia.

Masayuki Asahara, Kenta Fukuoka, Ai Azuma, Chooi-Ling Goh, Yotaro Watanabe, Yuji Matsumoto, and Takashi Tsuzuki. 2005. Combination of machine learning methods for optimum Chinese word segmentation. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pages 134–137. Jeju Island, Korea.

Andrew Borthwick. 1999. *A Maximum Entropy Approach to Named Entity Recognition*. Ph.D. thesis, Computer Science Department, New York University.

Michael R. Brent. 1999. An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, 34:71–105.

Bob Carpenter. 2006. Character language models for Chinese word segmentation and named entity recognition. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 169–172. Association for Computational Linguistics, Sydney, Australia.

Aitao Chen, Yiping Zhou, Anne Zhang, and Gordon Sun. 2005. Unigram language model for Chinese word segmentation. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pages 138–141. Jeju Island, Korea.

Lee-Feng Chien. 1997. PAT-tree-based keyword extraction for Chinese information retrieval. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 50–58. Philadelphia.

Hai Leong Chieu and Hwee Tou Ng. 2002. Named entity recognition: A maximum entropy approach using global information. In *Proceedings of the 19th International Conference*

on Computational Linguistics (COLING 2002), page 190 ‑ 196. Taipei, Taiwan.

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.

Haodi Feng, Kang Chen, Xiaotie Deng, and Weimin Zheng. 2004. Accessor variety criteria for Chinese word extraction. *Computational Linguistics*, 30(1):75–93.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, page 363 ‑ 370. Association for Computational Linguistics, Ann Arbor, Michigan.

Jianfeng Gao, Mu Li, Andi Wu, and Chang-Ning Huang. 2005. Chinese word segmentation and named entity recognition: A pragmatic approach. *Computational Linguistics*, 31(4):531–574.

Xianping Ge, Wanda Pratt, and Padhraic Smyth. 1999. Discovering Chinese words from unsegmented text. In *SIGIR '99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 271–272. ACM, Berkeley, CA, USA.

Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. 2006. Contextual dependencies in unsupervised word segmentation. In *COLING/ACL 2006*, pages 673–670. Sidney, Australia.

Zellig Sabbetai Harris. 1970. Morpheme boundaries within words. In *Papers in Structural and Transformational Linguistics*, page 68 ‑ 77.

Zhihui Jin and Kumiko Tanaka-Ishii. 2006. Unsupervised segmentation of Chinese text by use of branching entropy. In *COLING/ACL 2006*, pages 428–435. Sidney, Australia.

Chunyu Kit. 2000. *Unsupervised Lexical Learning as Inductive Inference*. Ph.D. thesis, University of Sheffield.

Chunyu Kit and Xiaoyue Liu. 2005. An example-based Chinese word segmentation system for CWSB-2. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pages 146–149. Jeju Island, Korea.

Chunyu Kit and Yorick Wilks. 1999. Unsupervised learning of word boundary with description length gain. In M. Osborne and E. T. K. Sang, editors, *CoNLL-99*, pages 1–6. Bergen, Norway.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

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, Australia.

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.

Andrei Mikheev, Claire Grover, and Marc Moens. 1998. Description of the LTG system used for MUC-7. In Nancy A. Chinchor, editor, *Proceedings of the Seventh Message Understanding Conference(MUC-7)*. Fairfax, Virginia.

Tetsuji Nakagawa and Yuji Matsumoto. 2006. Guessing parts-of-speech of unknownwords using global information. In *COLING/ACL 2006*, pages 705–712. Sidney, Australia.

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.

Fuchun Peng and Dale Schuurmans. 2001. Self-supervised Chinese word segmentation. In *The 4th International Symposium on Intelligent Data Analysis*, pages 238–247. Lisbon, Portugal.

Ronald Rosenfeld, Stanley F. Chen, and Xiaojin Zhu. 2001. Whole-sentence exponential

language models: A vehicle for linguistic-statistical integration. *Computers Speech and Language*, 15(1):55 – 73.

Richard Sproat and Thomas Emerson. 2003. The first international Chinese word segmentation bakeoff. In *The Second SIGHAN Workshop on Chinese Language Processing*, pages 133–143. Sapporo, Japan.

Richard Sproat and Chilin Shih. 1990. A statistical method for finding word boundaries in Chinese text. *Computer Processing of Chinese and Oriental Languages*, 4(4):336–351.

Mao Song SUN, Ming XIAO, and Benjamin K. Tsou. 2004. Chinese word segmentation without using dictionary based on unsupervised learning strategy (in Chinese) (基于无指导学习策略的无词表条件下的汉语自动分词). *Chinese Journal of Computers*, 27(6):736–742.

Maosong Sun, Dayang Shen, and Benjamin K. Tsou. 1998. Chinese word segmentation without using lexicon and hand-crafted training data. In *COLING-ACL '98, 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, volume 2, pages 1265–1271. Montreal, Quebec, Canada.

Hiroya Takamura, Takashi Inui, and Manabu Okumura. 2005. Extracting semantic orientations of words using spin model. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, page 133 – 140. Association for Computational Linguistics, Ann Arbor, Michigan.

Richard Tzong-Han Tsai, Hsieh-Chuan Hung, Cheng-Lung Sung, Hong-Jie Dai, and Wen-Lian Hsu. 2006. On closed task of Chinese word segmentation: An improved CRF model coupled with character clustering and automatically generated template matching. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 108–117. Sydney, Australia.

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning. 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.

Xinhao Wang, Xiaojun Lin, Dianhai Yu, Hao Tian, and Xihong Wu. 2006. Chinese word segmentation with maximum entropy and N-gram language model. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 138–141. Sydney, Australia.

Nianwen Xue. 2003. Chinese word segmentation as character tagging. *Computational Linguistics and Chinese Language Processing*, 8(1):29–48.

Mikio Yamamoto and Kenneth W. Church. 2001. Using suffix arrays to compute term frequency and document frequency for all substrings in a corpus. *Computational Linguistics*, 27(1):1–30.

Jian Zhang, Jianfeng Gao, and Ming Zhou. 2000. Extraction of Chinese compound words – an experimental study on a very large corpus. In *Proceedings of the Second Chinese Language Processing Workshop*, pages 132–139. Hong Kong, China.

Ruiqiang Zhang, Genichiro Kikui, and Eiichiro Sumita. 2006. Subword-based tagging for confidence dependent Chinese word segmentation. In *COLING/ACL 2006*, pages 961–968. Sidney, Australia.

Hai Zhao, Chang-Ning Huang, and Mu Li. 2006a. An improved Chinese word segmentation system with conditional random field. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 162–165. Sydney, Australia.

Hai Zhao, Chang-Ning Huang, Mu Li, and Bao-Liang Lu. 2006b. Effective tag set selection in Chinese word segmentation via conditional random field modeling. In *Proceedings of PACLIC-20*, pages 87–94. Wuhan, China.

Guo Dong ZHOU. 2005. A chunking strategy towards unknown word detection in Chinese word segmentation. In Robert Dale, Kam-Fai Wong, Jian Su, and Oi Yee Kwong, editors, *IJCNLP 2005*, volume 3651 of *Lecture Notes in Computer Science*, pages 530–541. Jeju Island, Korea.

Mu-Hua Zhu, Yi-Lin Wang, Zhen-Xing Wang, Hui-Zhen Wang, and Jing-Bo Zhu. 2006. Designing special post-processing rules for SVM-based Chinese word segmentation. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 217–220. Sydney, Australia.