Exploiting Unlabeled Text with Different Unsupervised Segmentation Criteria 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 improve Chinese word segmentation (CWS) that attempts to utilize unlabeled data such as training and test data without annotation for further enhancement of the state-of-the-art performance of supervised learning. The lexical information plays the role of information transformation from unlabeled text to supervised learning model. Four types of unsupervised segmentation criteria are used for word candidate extraction and the corresponding word likelihood computation. The information output by unsupervised segmentation criteria as features therefore is integrated into supervised learning model to strengthen the learning for the matching subsequence. The effectiveness of the proposed method is verified in data sets from the latest international CWS evaluation. Our experimental results show that character-based conditional random fields framework can effectively make use of such information from unlabeled data for performance enhancement on top of the best existing results.

1 Introduction

The task of Chinese word segmentation (CWS) is to segment an input sequence of characters into a sequence of words. It is also a preprocessing task shared by many Asian languages without overt word delimiters. CWS was first formulated as a character tagging problem in [1], via labeling each character's position in a word. For example, the segmentation for following sentences,

他 / 来自 / 墨西哥 。
(he / comes from / Mexico.),

receives the tag (label) sequence $SBEBME$ as segmentation result, where the four tags B, M and E stand for the beginning, middle and ending positions in a word, and S for a single character as a word. A Maximum Entropy (MaxEnt) model was trained for such a tagging task in [1]. Many supervised learning methods have been successfully applied to CWS since the First International Chinese Word Segmentation Bakeoff in 2003 [2]. Among them, the character tagging is a particularly simple but effective formulation of the problem suitable for various competitive supervised machine learning models such as MaxEnt, conditional random fields (CRFs), and support vector machines. [1, 3–8].

Advances in Natural Language Processing and Applications
Research in Computing Science 33, 2008, pp. 93-104

Received 07/10/07
Accepted 07/12/07
Final Version 17/01/08

However, few existing methods make use of non-local information of any given sequences as a source of knowledge. In this paper, we will explore a new approach to integrating such useful information from the unlabeled text into the supervised learning for CWS. It attempts to utilize lexicon information derived by various word likelihood criteria, which were intended for unsupervised word segmentation techniques. We know that an unsupervised segmentation strategy has to follow some predefined criterion about how likely a target substring, as a word candidate, is to be a true word. It is important to examine how such information which usually appears as a goodness score for a word candidate can be exploited to facilitate a supervised learning mode for CWS. In this study, we will examine four kinds of such criteria, frequency of substring after reduction, description length gain, accessor variety, and boundary entropy. All of them will be represented as features for integration into our character tagging system for CWS, and their effectiveness will be evaluated using the large-scale data sets for the previous Bakeoff.

The remainder of the paper is organized as follows. The next section describes the baseline supervised learning with the CRFs model for CWS. Section 3 discusses four criteria for unsupervised word extraction and formulates our approach to integrating them to the CRFs learning for CWS. Then, our experimental results are presented in Section 4. Section 5 discusses related work and the possibilities of semi-supervised learning with CRFs. Finally, we summarize our research achievements to conclude the paper in Section 6.

2 Supervised Learning for Word Segmentation

CRFs [9] is a statistical sequence modeling framework that is reported to outperform other popular learning models including MaxEnt method in a number of natural language processing (NLP) applications[10]. CRFs is first applied to CWS in [3], treating CWS as a binary decision task to determine whether a Chinese character in the input is the beginning of a word.

The probability assigned to a label sequence for an unsegmented sequence of characters by a CRFs is given by the equation below:

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

where y 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 and λ_k is the respective weight, C is the label(or tag) set, and c indexes into characters in the sequence to be labeled. For CRFs learning, we use the CRF++ package with necessary modification for training speedup [1].

It is shown in our previous work that the CRFs learning achieves a better segmentation performance with a 6-tag set than any other tag set [11]. Thus, we opt for using this tag set and its six n-gram feature templates as the baseline for our evaluation. The six tags are B, B_2, B_3, M, E and S. Accordingly, we have the tag sequences S,

[1] http://crfpp.sourceforge.net/

BE, BB_2E, BB_2B_3E, BB_2B_3ME and $BB_2B_3M\cdots ME$ for characters in a word of length 1, 2, 3, $\cdots$, and 6 (and above), respectively. The six n-gram feature templates are C_{-1}, C_0, C_1, $C_{-1}C_0$, C_0C_1 and $C_{-1}C_1$, where 0, -1 and 1 stand for the positions of the current, previous and next characters, respectively.

3 Unsupervised Segmentation Criteria

In general, unsupervised segmentation assumes no pre-segmented data for training and no pre-defined lexicon. It has to follow some predefined criterion to identify word candidates and assign to them a goodness score to indicate their word likelihood. In this study, we explore the effectiveness of utilizing four existing unsupervised segmentation criteria to facilitate supervised learning for CWS. Each of them is applied to compute a goodness score $g(s)$ for an n-gram substring s in the input text. In principle, the higher the goodness score for a substring, the more likely it is to be a true word. We consider all available substrings in the input as possible word candidates for each criterion.

3.1 Frequency of Substring after Reduction

Frequency is not a reliable estimator for how likely a substring is to be a word, although we feel like that a more frequent substring seems to have a better chance to be a true word. Statistical substring reduction [12] is perhaps a workable idea to turn the frequency into a good word-hood criterion. Its underlying assumption is that if two overlapping substrings have the same frequency, then the shorter one can be discarded as a word candidate. To integrate such frequency information after substring reduction (FSR) into our CRFs learning, we define a goodness score as follows,

$$g_{FSR}(s) = \log(p(s)), \tag{1}$$

where $p(s)$ is the frequency of s. That is, we take the logarithm value of the frequency as the goodness score for s. Such a score is the number of bits needed to encode s in a sense of information theory, if the base for the logarithm is 2.

3.2 Description Length Gain

It is proposed in [13], as a goodness measure for a compression-based method for unsupervised word segmentation. The DLG from extracting all occurrences of a substring $s = x_i x_{i+1} ... x_j$ (also denoted as $x_{i..j}$) as a candidate word from a corpus $X = x_1 x_2 ... x_n$ (with a vocabulary V, here, a character list) is defined as

$$DLG(x_{i..j}) = L(X) - L(X[r \rightarrow x_{i..j}] \oplus x_{i..j})$$

where $X[r \rightarrow x_{i..j}]$ represents the resultant corpus from replacing all instances of $x_{i..j}$ with a trace symbol r throughout X and $\oplus$ denotes string concatenation. $L(\cdot)$ is the empirical description length of a corpus in bits that can be estimated as below, following classic information theory [14, 15].

$$L(X) \doteq -|X| \sum_{x \in V} p(x) \log_2 p(x)$$

where $|\cdot|$ denotes the length of a string in number of characters. To effectively integrate DLG into our CRFs model, we define $g_{DLG}(s) = \log(DLG(s))$.

3.3 Accessor Variety

This criterion is formulated in [16]. It has a nice performance in extraction of low-frequent words as reported in [16]. As a measure to evaluate how independent a subsequence is and hence how likely it is a true word, the accessor variety of a substring s is defined as

$$AV(s) = \min\{L_{av}(s), R_{av}(s)\} \tag{2}$$

where the left and right accessor variety $L_{av}(s)$ and $R_{av}(s)$ are defined, respectively, as the number of the distinct predecessor and successor characters. For the similar reason as in Section 3.1, the goodness score $g_{AV}(s)$ for s is set to the logarithm value of AV, $\log(AV(s))$.

3.4 Boundary Entropy

The branching entropy or boundary entropy (BE) is formulated as a criterion for unsupervised segmentation in a number of previous works [17–20]. The local entropy for a given substring $s = x_{i..j}$,

$$h(x_{i..j}) = -\sum_{x \in V} p(x|x_{i..j}) \log p(x|x_{i..j}), \tag{3}$$

indicates the average uncertainty next to $x_{i..j}$. Two scores, namely, $h_L(x_{i..j})$ and $h_R(x_{i..j})$, can be defined for the two directions to extend $x_{i..j}$. Also, let $h_{min} = \min\{h_R, h_L\}$ in a similar manner for AV in equation (2), and then we define $g_{BE}(s) = \log(h_{min})$.

The two criteria AV and BE share a similar assumption as in the pioneering work [21]: If the uncertainty of successive tokens increases, then the location is likely to be at a boundary. In this sense, they are various formulation for a similar idea.

3.5 Feature Templates to Incorporate Unsupervised Segmentation Criteria

The basic idea of exploiting information derived by different unsupervised segmentation criteria is to inform a supervised learner of how likely a substring is to be a true word according to a particular criterion.

To make best of such information, suitable feature templates need to be used to represent word candidates with different lengths. According to [11], less than 1% words are longer than 6-character in segmented corpora of Chinese. Thus, we consider only n-gram of no more than five-character long for feature generation in this work.

We use CRFs as an ensemble model to integrate these features. For each unsupervised segmentation criterion, we consider two types of features. One is concerned with word matching for an n-gram s, which is formulated as a feature function,

$$f_n(s) = \begin{cases} 1, & \text{if } s \in L \\ 0, & \text{otherwise,} \end{cases} \tag{4}$$

to indicate whether s belongs to the word candidate list L. Heuristic rules are applied in [16] to remove substrings that consist of a word and adhesive characters for AV criterion. In this study, we do not use any heuristic rules. For each criterion, we only set a default threshold, namely, 0, to get the corresponding list: $g_{FSR}(s) > 0$, $g_{AV}(s) > 0$, $h_{min} > 0$, and $DLG(s) > 0$. The other is concerned with word likelihood information. A feature template for an n-gram string s with a score $g(s)$ is formulated as,

$$f_n(s, g(s)) = \begin{cases} t, & \text{if } t \leq g(w) < t+1 \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

where t is an integer to discretize the word likelihood score. For an overlap character of several word candidates, we choose the one with the greatest goodness score to activate the above feature functions for that character. This makes the feature representation robust enough to cope with many infrequent candidates. In this way, feature values will not be sensitive to the threshold about word candidate list generation. Feature function (5) can be actually generated from all possible substrings occurring in the given text. Note that all t in (5) are not parameters but as feature values in the system. Our system is basically parameter-free.

4 Evaluation

Our approach is evaluated in all four corpora from the Third International Chinese Language Processing Bakeoff (Bakeoff-3) [2] [22]. , where corpus size information can be found in Table 1. Word segmentation performance is measured by F-measure, $F = 2RP/(R+P)$, where the recall R and precision P are respectively the proportions of the correctly segmented words to all words in the gold-standard segmentation and a segmenter's output [3]. The recall of out-of-vocabulary words (OOVs), R_{OOV}, is also given to measure the effectiveness of OOV identification.

Table 1. Corpus size of Bakeoff-3 in number of words

Corpus	AS[a]	CityU[b]	CTB[c]	MSRA[d]
Training (M)	5.45	1.64	0.5	1.26
Test (K)	91	220	154	100

[a] Academia Sinica Corpus.
[b] City University of Hong Kong Corpus.
[c] Corpus by University of Pennsylvania and University of Colorado
[d] Microsoft Research Asia Corpus.

[2] http://www.sighan.org/bakeoff2006.
[3] A standard scoring tool is available at http://www.sighan.org/bakeoff2003/score.

4.1 Performance Comparison with Different Criteria

We take the system described in Section 2 as baseline for comparison. Features generated by unsupervised segmentation criteria according to formulae (4) and (5) are derived from the unannotated training and test text. They are integrated the baseline system for evaluation. Our evaluation results are in Table 2 and 3[4].

Table 2. Performance comparison: features derived by different criteria

Criterion	F-score				R_{OOV}			
	AS	CityU	CTB	MSRA	AS	CityU	CTB	MSRA
Baseline	0.9539	0.9691	0.9321	0.9609	0.6699	0.7815	0.7095	**0.6658**
$AV_{(1)}$[a]	0.9566	0.9721	0.9373	0.9630	0.6847	0.7997	0.7326	0.6584
$AV_{(2)}$[b]	0.9573	0.9740	**0.9428**	**0.9634**	0.6842	0.8002	**0.7581**	0.6523
$BE_{(1)}$	0.9566	0.9721	0.9373	0.9630	0.6847	0.7997	0.7326	0.6584
$BE_{(2)}$	**0.9584**	**0.9743**	0.9421	0.9633	**0.6930**	**0.8029**	0.7569	0.6493
$FSR_{(1)}$	0.9565	0.9715	0.9367	0.9621	0.6782	0.7931	0.7299	0.6628
$FSR_{(2)}$	0.9575	0.9735	0.9415	0.9630	0.6775	0.7994	0.7557	0.6420
$DLG_{(1)}$	0.9554	0.9708	0.9395	0.9616	0.6738	0.7883	0.7459	0.6514
$DLG_{(2)}$	0.9560	0.9718	0.9401	0.9617	0.6793	0.7970	0.7528	0.6531

[a] (1): using feature formula (4) for all criteria
[b] (2): using feature formula (5) for all criteria

From Table 2, we can see that every unsupervised segmentation criteria do lead to performance improvement upone the baseline. It is also shown that all criteria give further performance improvement upon the cases without word likelihood information included in features, though the improvement is slight in many cases. Among those criteria, it can be observed that AV and BE give the most competitive performance while DLG the least. Such difference is due to supervised learning model and how unsupervised segmentation information is integrated. Although our results show AV and BE are the best criteria to improve supervised learning for CWS, this does not necessarily mean that they are better unsupervised segmentation criteria than DLG for unsupervised segmentation when working alone. Actually, DLG gives better results for unsupervised segmentation as reported in [24]. It is also worth noting that AV and BE results in all evaluation corpora just do as they for unsupervised segmentation tasks [24], revealing the intrinsic similarity of these two criteria.

Table 3 integrating all features from two or more criteria does not necessarily lead to any further improvement. This suggests that though each criterion does give its useful insight to word boundary in Chinese text, their characteristics overlap very much.

[4] The results here are slightly different from those in [23] for the same experimental settings, the only cause to the difference, as we can see, it is the different CRFs implementation this time.

Table 3. Performance improvement with features derived from various unsupervised criterion combinations using feature formula (5)

Criteria				F-score				R_{OOV}			
AV	BE	FSR	DLG	AS	CityU	CTB	MSRA	AS	CityU	CTB	MSRA
				0.9539	0.9691	0.9321	0.9609	0.6699	0.7815	0.7095	**0.6658**
+				0.9573	0.9740	**0.9428**	**0.9634**	0.6842	0.8002	**0.7581**	0.6523
	+			**0.9584**	**0.9743**	0.9421	0.9633	**0.6930**	**0.8029**	0.7569	0.6493
+	+			0.9570	0.9726	0.9421	**0.9635**	0.6798	0.7990	0.7550	0.6614
+		+		0.9574	**0.9739**	0.9425	0.9633	0.6860	0.7975	0.7596	**0.6637**
+			+	0.9569	0.9733	0.9425	0.9629	0.6829	0.7995	0.7587	0.6417
	+	+		**0.9575**	0.9725	0.9423	0.9631	0.6842	0.7996	0.7581	0.6464
	+		+	0.9570	0.9734	**0.9428**	0.9627	0.6860	0.7979	**0.7638**	0.6452
		+	+	0.9573	0.9732	0.9416	0.9632	0.6769	0.8017	0.7541	0.6514
+	+	+	+	**0.9575**	0.9729	0.9413	0.9630	**0.6878**	**0.8039**	0.7535	0.6361

4.2 Comparison against Existing Results

We compare our performance with those best ones 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 a constraint.

A summary of the best results in the closed test of Bakeoff-3 are presented in Table 4 for a comparison with ours. Our results are obtained by integrating BE features into the baseline system. All six participants with at least a third best performance in the closed test of Bakeoff-3 are given in this table [25, 7, 26, 27, 6, 8].

Table 4. Comparisons of the best existing results and ours in data of Bakeoff-3 (F-scores)

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 of Bakeoff-3		0.958	0.972	0.933	0.963
Ours		0.958	0.974	0.942	0.963
Error Reduction (%)		-	7.1	13.4	-

From Table 4, we see that our system demonstrates a significant improvement upon the baseline and achieves a better performance on top of the state-of-the-art as in

Bakeoff-3. Especially, our results are achieved only with n-gram information alone, while some official results of Bakeoff-3 were involved in features or techniques that are only allowed in open test [26, 6, 8][5].

To check if those results with slight difference are statistical significant, we perform some statistical significance tests in the results of closed test. Following the previous work [2] and assuming the binomial distribution is appropriate for our experiments, we may compute 95% confidence interval as $\pm 2\sqrt{p'(1-p')/n}$ according to the Central Limit Theorem for Bernoulli trials [28], where n is the number of trials (words). We suppose that the recall represents the probability of correct word identification, and the precision represents the probability that a character string that has been identified as a word is really a word. Thus two types of intervals, C_r and C_p, can be computed, respectively, as p' is set to r and p. One can determine if two results are significantly different at a 95% confidence level by checking whether their confidence intervals overlap. The values of C_r and C_p for the best existing results and ours are in Table 5, where the data of each row with head 'bakeoff-3' are from [22].

Table 5. Statistical significance: comparisons of the best closed results of Bakeoff-3 and ours

Corpus	#word	Best	R	C_r	P	C_p	F
AS	91K	Bakeoff-3	0.961	$\pm$0.001280	0.955	$\pm$0.001371	0.958
		Ours	0.964	$\pm$0.001235	0.953	$\pm$0.001403	0.958
CityU	220K	Bakeoff-3	0.973	$\pm$0.000691	0.972	$\pm$0.000703	0.972
		Ours	0.974	$\pm$0.000679	0.974	$\pm$0.000679	0.974
CTB	154K	Bakeoff-3	0.940	$\pm$0.001207	0.926	$\pm$0.001330	0.933
		Ours	0.947	$\pm$0.001142	0.937	$\pm$0.001238	0.943
MSRA	100K	Bakeoff-3	0.964	$\pm$0.001176	0.961	$\pm$0.001222	0.963
		Ours	0.960	$\pm$0.001239	0.967	$\pm$0.001130	0.963

4.3 Early Results in Open Test

Until now, we only consider using the plain text from training and test. Some external segmented corpora are used to strengthen the current segmentation task in [4] and [6]. However, it is well known that segmented corpus is not always easily obtained. Thus it will be meaningful to extend our approach to external unlabeled text, which can be easily obtained as any requirement. Here we report some early segmentation results on using such external resources.

The unlabeled text that we adopt is that of People's Daily[6] from 1993 to 1997 of about 100M characters. The evaluation corpora are CTB and MSRA of Bakeoff-3 that

[5] Although manifestly prohibited by the closed test rules of Bakeoffs, character type features are used in [6] and [8], and a key parameter is estimated by using an external segmented corpus in [26].

[6] This is the most popular official newspaper in mainland of China.

are also in GB encode. AV is selected as the criterion for the goodness score computation. The results are given in Table 6.

Table 6. Performances using external unlabeled data or lexicon

Corpus	Metric	Text[a]	Dict[b]	Text+Dict	Best open
CTB	F-score	0.9401	0.9412	0.9443	0.944
	R_{OOV}	0.7382	0.7412	0.7565	0.768
MSRA	F-score	0.9674	0.9681	0.9716	0.979
	R_{OOV}	0.6905	0.6905	0.7140	0.839

[a] Using AV features from People's Daily text.
[b] Using features from external dictionary.

Note that the way that we extract useful information from unlabeled data is to make use of a word candidate list. It is also a natural way to integrate an external dictionary. The results of using the same online dictionary from Peking University and feature representation as [4] are also given in Table 6. There are about 108,000 words of length one to four characters in this dictionary[7].

We obtain two competitive results compared to the best existing results only by using unlabeled data and an external lexicon, while two best official results in Bakeoff-3 were obtained through large scale external segmented corpora, lexicons and named entity information [29,30]. This shows that our approach is also effective to exploit external unlabeled data.

5 Discussion and Related Work

In this study, we explore a combination of fully supervised and unsupervised learning for Chinese word segmentation. It is not sure at the first glance whether unsupervised segmentation in the same supervised data can help supervised learning. However, if unsupervised technique can extract global information of the whole text instead from local context inside a sequence, then we can expect the effectiveness, since each type of unsupervised segmentation criterion makes global statistics through the whole text.

When we are applying unlabeled data to improve supervised learning, semi-supervised method is actually introduced into this field. Since Researchers developed techniques for structural semi-supervised learning scheme for large scale unlabeled data in linguistics. As a sequence labeling tool, CRFs with revision for semi-supervised learning has been developed recently.

Semi-supervised CRFs based on a minimum entropy regularizer was proposed in [31]. Its parameters are estimated to maximize the likelihood of labeled data and the negative conditional entropy of the unlabeled data.

[7] It is available from http://ccl.pku.edu.cn/doubtfire/Course/Chinese%20Information %20Processing/Source_Code/Chapter_8/Lexicon_full_2000.zip

In [32], a semi-supervised learning approach was proposed based on a hybrid generative and discriminative approach. Defining the objective function of a hybrid model in log-linear form, discriminative structured predictor (i.e., CRFs) and generative model(s) that incorporate unlabeled data are integrated. Then, the generative model attached unlabeled data is used to increase the sum of the discriminant functions during the parameter estimation.

The idea in our work is close to that of [32]. However, considering that supervised learning for CWS is often a large scale task in computation and lexical information is traditionally used for information transformation. Unsupervised word extraction methods are directly adopted to output lexical information for discriminant model. Our method has been shown efficient and effective in this way.

6 Conclusion

In this paper, we have presented an ensemble learning approach to take advantage of unlabeled data for Chinese word segmentation.

The lexical information plays the central role in information transformation from unlabeled data to supervised learning model. Four types of unsupervised segmentation methods are considered and formulated as word candidate extraction and the respective goodness score computation. Such information about outputs of unsupervised word extraction is integrated as features into CRFs learning model. The effectiveness of different unsupervised criteria for word extraction is studied. We provide evidence to show that character-based CRFs modeling for CWS can take advantage of unlabeled data, especially, the unlabeled text of training corpus and test corpus, 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.

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.

References

1. Xue, N.: Chinese word segmentation as character tagging. Computational Linguistics and Chinese Language Processing **8** (2003) 29–48
2. Sproat, R., Emerson, T.: The first international Chinese word segmentation bakeoff. In: The Second SIGHAN Workshop on Chinese Language Processing, Sapporo, Japan (2003) 133–143
3. Peng, F., Feng, F., McCallum, A.: Chinese segmentation and new word detection using conditional random fields. In: COLING 2004, Geneva, Switzerland (2004) 562–568
4. Low, J.K., Ng, H.T., Guo, W.: A maximum entropy approach to Chinese word segmentation. In: Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing, Jeju Island, Korea (2005) 161–164

5. Tseng, H., Chang, P., Andrew, G., Jurafsky, D., Manning, C.: A conditional random field word segmenter for SIGHAN bakeoff 2005. In: Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing, Jeju Island, Korea (2005) 168–171
6. Zhao, H., Huang, C.N., Li, M.: An improved Chinese word segmentation system with conditional random field. In: Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, Sydney, Australia (2006) 162–165
7. Tsai, R.T.H., Hung, H.C., Sung, C.L., Dai, H.J., Hsu, W.L.: 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, Sydney, Australia (2006) 108–117
8. Zhu, M.H., Wang, Y.L., Wang, Z.X., Wang, H.Z., Zhu, J.B.: Designing special post-processing rules for SVM-based Chinese word segmentation. In: Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, Sydney, Australia (2006) 217–220
9. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc. (2001) 282–289
10. Rosenfeld, B., Feldman, R., Fresko, M.: A systematic cross-comparison of sequence classifiers. In: SDM 2006, Bethesda, Maryland (2006) 563–567
11. Zhao, H., Huang, C.N., Li, M., Lu, B.L.: Effective tag set selection in Chinese word segmentation via conditional random field modeling. In: Proceedings of the 20th Asian Pacific Conference on Language, Information and Computation, Wuhan, China (2006) 87–94
12. Lü, X., Zhang, L., Hu, J.: Statistical substring reduction in linear time. In et al., K.Y.S., ed.: Proceeding of the 1st International Joint Conference on Natural Language Processing (IJCNLP-2004). Volume 3248 of Lecture Notes in Computer Science., Sanya City, Hainan Island, China, Springer (2004) 320–327
13. Kit, C., Wilks, Y.: Unsupervised learning of word boundary with description length gain. In Osborne, M., Sang, E.T.K., eds.: CoNLL-99, Bergen, Norway (1999) 1–6
14. Shannon, C.E.: A mathematical theory of communication. The Bell System Technical Journal **27** (1948) 379–423, 623–656
15. Cover, T.M., Thomas, J.A.: Elements of Information Theory. John Wiley and Sons, Inc., New York (1991)
16. Feng, H., Chen, K., Deng, X., Zheng, W.: Accessor variety criteria for Chinese word extraction. Computational Linguistics **30** (2004) 75–93
17. Tung, C.H., Lee, H.J.: Identification of unknown words from corpus. Computational Proceedings of Chinese and Oriental Languages **8** (1994) 131–145
18. Chang, J.S., Su, K.Y.: An unsupervised iterative method for Chinese new lexicon extraction. Computational Linguistics and Chinese Language Processing **2** (1997) 97–148
19. Huang, J.H., Powers, D.: Chinese word segmentation based on contextual entropy. In Ji, D.H., Lua, K.T., eds.: Proceedings of the 17th Asian Pacific Conference on Language, Information and Computation, Sentosa, Singapore, COLIPS Publication (2003) 152–158
20. Jin, Z., Tanaka-Ishii, K.: Unsupervised segmentation of Chinese text by use of branching entropy. In: COLING/ACL 2006, Sidney, Australia (2006) 428–435
21. Harris, Z.S.: Morpheme boundaries within words. In: Papers in Structural and Transformational Linguistics. (1970) 68 - 77
22. Levow, G.A.: The third international Chinese language processing bakeoff: Word segmentation and named entity recognition. In: Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, Sydney, Australia (2006) 108–117
23. Zhao, H., Kit, C.: Unsupervised segmentation helps supervised learning of character tagging for word segmentation and named entity recognition. In: The Sixth SIGHAN Workshop on Chinese Language Processing, Hyderabad, India (2008)

24. Zhao, H., Kit, C.: An empirical comparison of goodness measures for unsupervised Chinese word segmentation with a unified framework. In: The Third International Joint Conference on Natural Language Processing (IJCNLP-2008), Hyderabad, India (2008)
25. Carpenter, B.: Character language models for Chinese word segmentation and named entity recognition. In: Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, Sydney, Australia, Association for Computational Linguistics (2006) 169–172
26. Wang, X., Lin, X., Yu, D., Tian, H., Wu, X.: Chinese word segmentation with maximum entropy and N-gram language model. In: Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, Sydney, Australia (2006) 138–141
27. Zhang, M., Zhou, G.D., Yang, L.P., Ji, D.H.: Chinese word segmentation and named entity recognition based on a context-dependent mutual information independence model. In: Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, Sydney, Australia (2006) 154–157
28. Grinstead, C., Snell, J.L.: Introduction to Probability. American Mathematical Society, Providence, RI (1997)
29. Jacobs, A.J., Wong, Y.W.: Maximum entropy word segmentation of Chinese text. In: Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, Sydney, Australia (2006) 108–117
30. Liu, W., Li, H., Dong, Y., He, N., Luo, H., Wang, H.: France Telecom R&D Beijing word segmenter for SIGHAN bakeoff 2006. In: Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, Sydney, Australia (2006) 108–117
31. Jiao, F., Wang, S., Lee, C.H., Greiner, R., Schuurmans, D.: Semi-supervised conditional random fields for improved sequence segmentation and labeling. In: Proc. of COLING/ACL-2006, Sydney, Australia (2006) 209 - 216
32. Suzuki, J., Fujino, A., Isozaki, H.: Semi-supervised structured output learning based on a hybrid generative and discriminative approach. In: Proceedings of the 45rd Annual Meeting of the Association for Computational Linguistics (ACL'07), Prague, Czech, Association for Computational Linguistics (2007) 791