Computational Linguistics and Chinese Language Processing
Vol. 17, No. 3, September 2012, pp. 45-86

Enhancement of Feature Engineering for Conditional Random Field Learning in Chinese Word Segmentation Using Unlabeled Data

Mike Tian-Jian Jiang[*+], Cheng-Wei Shih[†+], Ting-Hao Yang[+],

Chan-Hung Kuo[+], Richard Tzong-Han Tsai[‡] and Wen-Lian Hsu[*†+]

Abstract

This work proposes a unified view of several features based on frequent strings extracted from unlabeled data that improve the conditional random fields (CRF) model for Chinese word segmentation (CWS). These features include character-based *n*-gram (CNG), accessor variety based string (AVS) and its variation of left-right co-existed feature (LRAVS), term-contributed frequency (TCF), and term-contributed boundary (TCB) with a specific manner of boundary overlapping. For the experiments, the baseline is the *6-tag*, a state-of-the-art labeling scheme of CRF-based CWS, and the data set is acquired from the 2005 CWS Bakeoff of Special Interest Group on Chinese Language Processing (SIGHAN) of the Association for Computational Linguistics (ACL) and SIGHAN CWS Bakeoff 2010. The experimental results show that all of these features improve the performance of the baseline system in terms of *recall*, *precision*, and their harmonic average as F_1 *measure score*, on both accuracy (F) and out-of-vocabulary recognition (F_{OOV}). In particular, this work presents compound features involving LRAVS/AVS and TCF/TCB that are competitive with other types of features for CRF-based CWS in terms of F and F_{OOV}, respectively.

Keywords: Conditional Random Fields, Word Segmentation, Accessor Variety, Term-contributed Frequency, Term-contributed Boundary.

[*] Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan.
[†] Institute of Information System and Application, National Tsing Hua University, Hsinchu, Taiwan.
[‡] Department of Computer Science & Engineering, Yuan Ze University, Taoyuan, Taiwan.
E-mail: thtsai@saturn.yzu.edu.tw
[+] Institute of Information Science, Academia Sinica, Taipei, Taiwan.
E-mail: {tmjiang, dapi, tinghaoyang, laybow, hsu}@iis.sinica.edu.tw

1. Introduction

Background

Many intelligent text processing tasks, such as information retrieval, text-to-speech, and machine translation assume the ready availability of a tokenization into words, which is relatively straightforward in languages with word delimiters (*e.g.*, space) but is a little difficult for Asian languages, such as Chinese and Japanese.

Chinese word segmentation (CWS) has been an active area of research in computational linguistics for two decades. SIGHAN, the Special Interest Group for Chinese Language Processing of the Association for Computational Linguistics, has conducted five word segmentation bakeoffs (Emerson, 2005; Jin & Chen, 2007; Levow, 2006; Sproat & Emerson, 2003; Zhao & Liu, 2010). After years of intensive research, CWS has achieved high accuracy, but the issue of out-of-vocabulary (OOV) word recognition remains.

The State of the Art of CWS

Traditional approaches for CWS adopt a dictionary and rules to segment unlabeled texts, such as the work of Ma and Chen (2003). In recent years, there has been a potent trend of using statistical machine learning models, especially the conditional random fields (CRF) (Lafferty *et al.*, 2001), which displays moderate performance for the sequential labeling problem and achieves competitive results with character-position based methods(Zhao *et al.*, 2010).

Unsupervised Feature Selection for CWS

In this work, unsupervised feature selection for CWS is based on frequent strings that are extracted automatically from unlabeled corpora. For convenience, these features are referred to as *unsupervised features* in the rest of this paper. Unsupervised features are suitable for closed training evaluation where external resources or extra information is not allowed, especially for cross-domain tasks, such as SIGHAN CWS bakeoff 2010(Zhao & Liu, 2010). Without proper knowledge, the closed training evaluation of word segmentation can be difficult with OOV words, where frequent strings collected from the test data may help. For incorporating unsupervised features into character-position based CRF for CWS, Zhao and Kit (2007) tried strings based on *accessor variety* (AV), which was developed by Feng *et al.* (2004), and based on *co-occurrence strings* (COS). Jiang *et al.* (2010) applied a feature similar to COS, called *term-contributed boundary* (TCB).

According to Zhao and Kit (2007), AV-based string (AVS) is one of the most effective unsupervised features for CWS by character-position based CRF. One motivation here is to seek deeper understanding of AVS's success. This work suspects that, since AVS is designed to keep overlapping substrings via the outer structure of a string while COS/TCB is usually selected via the inner structure of a string with its longest-first (*i.e.*, non-overlapping) nature before integration into CRF, combining overlapping and outer information with non-overlapping and inner information may enhance CRF-based CWS. Hence, a series of experiments is conducted to examine this hypothesis.

The remainder of the article is organized as follows. Section 2 briefly introduces CRF. Common unsupervised features based on the concept of frequent strings are explained in Section 3. Section 4 discusses related works. Section 5 describes the design of the labeling scheme and feature templates, along with a framework that is able to encode those overlapping features in a unified way. Details about the experiment are reported in Section 6. Finally, the conclusion is presented in Section 7.

2. Conditional Random Fields

Conditional random fields (CRF) are undirected graphical models trained to maximize a conditional probability of random variables X and Y, and the concept is well established for the sequential labeling problem (Lafferty *et al.*, 2001). Given an input sequence (or observation sequence) $X = x_1 ... x_T$ and a label sequence $Y = y_1 ... y_T$, a conditional probability of linear-chain CRF with parameters $\Lambda = \lambda_1 ... \lambda_n$ can be defined as:

$$P_\lambda(Y \mid X) = \frac{1}{Z_X} \exp\left(\sum_{t=1}^{T} \sum_k \lambda_k f_k(y_{t-1}, y_t, X, t) \right). \tag{1}$$

where Z_X is the normalization constant that makes probability of all label sequences sum to one; $f_k(y_{t-1}, y_t, X, t)$ is a feature function which is often binary valued, but can be real valued; and λ_k is a learned weight associated with feature f_k.

The feature functions can measure any aspect of state transition $y_{t-1} \rightarrow y_t$, and the entire observation sequence *X* is centered at the current position *t*.

Given the model defined in (1), the most probable labeling sequence for an input sequence *X* is as follows:

$$y^* = \underset{Y}{\operatorname{argmax}}\, P_\Lambda(Y \mid X) \quad . \tag{2}$$

Equation (2) can be efficiently calculated by dynamic programming using the Viterbi algorithm. More details about the concepts of CRF and learning parameters could be found in Wallach (2004). For sequential labeling tasks, like CWS, a linear-chain CRF is currently one of the most popular choices.

3. Unified View via Frequent String

3.1 Character-based *N-gram*

The word boundary and the word frequency are the standard notions of frequency in corpus-based natural language processing. Word-based *n*-gram is an intuitive and effective solution of language modeling. For languages without explicit word boundaries, such as Chinese, character-based *n*-gram (CNG) is usually insufficient. For example, consider some sample texts in Chinese:

- "自然科學的重要性" (the importance of natural science), and
- "自然科學的研究是唯一的途徑" (natural science research is the only way),

where many character-based *n*-grams can be extracted, but some of them are out of context, such as "然科" (so; discipline) and "學的" (study; of), even when they are relatively frequent. For the purpose of interpreting overlapping behavior of frequent strings, however, character-based *n*-grams could still be useful for baseline analysis and implementation.

3.2 Reduced *N*-gram

The lack of correct information about the actual boundary and frequency of a multi-character/word expression's occurrence has been researched in different languages. The distortion of phrase boundaries and frequencies was first observed in the Vodis Corpus, where the word-based bigram "RAIL ENQUIRIES" and word-based trigram "BRITISH RAIL ENQUIRIES" were estimated and reported by O'Boyle (1993) and Ha *et al.* (2005). Both of them occur 73 times, which is a large number for such a small corpus. "ENQUIRIES" follows "RAIL" with a very high probability when "BRITISH" precedes it. When "RAIL" is preceded by words other than "BRITISH," however, "ENQUIRIES" does not occur, but words like "TICKET" or "JOURNEY" may. Thus, the bigram "RAIL ENQUIRIES" gives a misleading probability that "RAIL" is followed by "ENQUIRIES" irrespective of what precedes it.

A common solution to this problem is that, if some *n*-grams consist of others, then the frequencies of the shorter ones have to be discounted with the frequencies of the longer ones. For Chinese, Lin & Yu (2011) reported a similar problem and its corresponding solution in the sense of *reduced n-gram* of Chinese characters. By excluding *n*-grams with their numbers of appearance that fully depend on other superstrings, "然科" and "學的" from the sample texts in the previous sub-section are no longer candidates of the string. Zhao and Kit (2007) described the same concept briefly as *co-occurrence string* (COS). Sung *et al.* (2008) invented a specific data structure for suffix array algorithm to calculate exact boundaries of phrase-alike string and their frequencies called *term-contributed boundaries* (TCB) and *term-contributed frequencies* (TCF), respectively, to analogize similarities and differences with the term frequencies. Since this work uses the program of TCB and TCF (namely YASA, yet another suffix array) for experiments, the family of *reduced n-gram* will be referred as TCB hereafter for convenience.

3.3 Uncertainty of Succeeding Character

Feng *et al.* (2004) proposed *accessor variety* (AV) to measure the likelihood a substring is a Chinese word. Another measurement, called *boundary entropy* or *branching entropy* (BE), exists in some works (Chang & Su, 1997; Cohen *et al.*, 2007; Huang & Powers, 2003; Tanaka-Ishii, 2005; Tung & Lee, 1994). The basic idea behind those measurements is closely related to one particular perspective of *n*-gram and information theory, *cross-entropy* or *perplexity*. According to Zhao and Kit (2007), AV and BE both assume that the border of a potential Chinese word is located where the uncertainty of successive character increases. They believe that AV and BE are the discrete and continuous version, respectively, of a fundamental work of Harris (1970), and they decided to adopt AVS as an unsupervised feature for CRF-based CWS. This work follows their choice in hope of producing a comparable study. AV of a string s is defined as:

$$AV(s) = \min\{L_{av}(s), R_{av}(s)\} \quad (3)$$

In (3), $L_{av}(s)$ and $R_{av}(s)$ are defined as the number of distinct preceding and succeeding characters, respectively, except, when the adjacent character is absent because of a sentence boundary, the pseudo-character of sentence beginning or sentence ending will be accumulated. Feng *et al.* (2004) also developed more heuristic rules to remove strings that contain known words or adhesive characters. For the strict meaning of unsupervised feature and for the sake of simplicity, these additional rules are dropped in this study.

Since a recent work of Sun and Xu (2011) used both $L_{av}(s)$ and $R_{av}(s)$ as features of CRF, this work will apply a similar approach, which is denoted as LRAVS, to make a thorough comparison.

4. Other Related Works

4.1 Frequent String Extraction Algorithm

Besides previous works of TCB and TCF extraction (Sung *et al.*, 2008), Chinese frequent strings (Lin & Yu, 2001), and *reduced n-gram* (Ha *et al.*, 2005), which have already been mentioned, the article about a linear algorithm for *frequency of substring with reduction* (Lü & Zhang, 2005) also falls into this category. Most of these projects focused on the computational complexity of algorithms. Broader algorithms for frequent string extraction are suffix array (Manber & Myers, 1993) and PAT-tree (Chien, 1997).

4.2 Unsupervised Word Segmentation Method

Zhao and Kit have explored several unsupervised strategies with their unified goodness measurement of logarithm ranking (Zhao & Kit, 2007), including *frequency of substring with reduction* (Lü & Zhang, 2005), *description length gain* (Kit & Wilks, 1999), *accessor variety* (Feng *et al*., 2004), and *boundary/branching entropy* (Chang & Su, 1997; Cohen *et al.*, 2007; Huang & Powers, 2003; Tanaka-Ishii, 2005; Tung & Lee, 1994). Unlike the technique described in this paper for incorporating unsupervised features into supervised CRF learning, those methods usually filter out word-alike candidates using their own scoring mechanism directly as unsupervised word segmentation.

4.3 Overlapping Ambiguity Resolution

Subword based tagging of Zhang *et al.* (2006) utilizes confidence measurement. Other overlapping ambiguity resolution approaches are Naïve Bayesian classifiers (Li *et al.*, 2003), mutual information, difference of *t*-test (Sun *et al.*, 1997), and sorted table look-up (Qiao *et al.*, 2008). These works concentrate on overlapping of words according to some (supervised) standard, rather than overlapping of substrings from unsupervised selection.

5. CRF Labeling Scheme

5.1 Character Position Based Labels

In this study, the CRF label set for CWS prediction adopts the *6-tag* approach of Zhao *et al.* (2010), which achieves very competitive performance and is one of the most fine-grained character position based labeling schemes. According to Zhao *et al.* (2010), since less than 1% of Chinese words are longer than five characters in most corpora from SIGHAN CWS bakeoffs 2003, 2005, 2006, and 2008, the coverage of a *6-tag* approach should be sufficient. This configuration of CRF without additional unsupervised features is also the control group of the experiment. Table 1 provides a sample of labeled training data.

Table 1. Sample of the 6-tag labels.

Character	Label
反	*B*
而	*E*
會	*S*
欲	*B*
速	*C*
則	*D*
不	*I*
達	*E*

For the sample text "反而 (contrarily) / 會 (make) / 欲速則不達 (more haste, less speed)" (on the contrary, haste makes waste), the tag *B* stands for the beginning character of a word, while *C* and *D* represent the second character and the third character of a word, respectively. The ending character of a word is tagged as *E*. Once a word consists of more than four characters, the tag for all of the middle characters between *D* and *E* is *I*. Finally, the tag *S* is reserved specifically for single-character words.

5.2 Feature Templates

Feature instances are generated from templates based on the work of Ratnaparkhi (1996). Table 2 explains their abilities. C_{-1}, C_0, and C_1 stand for the input tokens individually bound to the prediction label at the current position. For example, in Table 1, if the current position is at the label *I*, features generated by C_{-1}, C_0, and C_1 are "則," "不," and "達," respectively. Meanwhile, for window size 2, $C_{-1}C_0$, C_0C_1, and $C_{-1}C_1$ expands features of the label *I* to "則不," "不達," and "則達," respectively. One may argue that the feature template should expand to five tokens to cover the whole range of the 6-tag approach; however, according to Zhao *et al.* (2010), the context window size in three tokens is effective to catch parameters of the *6-tag* approach for most strings that do not exceed five characters. Our pilot test for this case also showed that context window size in two tokens would be sufficient without a significant decrease in performance (Jiang *et al.*, 2010).

Unsupervised features that will be introduced in the next subsection are generated by the same template, except the binding target moves column by column, as listed in tables of the next subsection.

Table 2. Feature template

Feature	Function
C_{-1}, C_0, C_1	Previous, current, or next token
$C_{-1}C_0$	Previous and current tokens
C_0C_1	Current and next tokens
$C_{-1}C_1$	Previous and next tokens

5.3 Unified Feature Representation of CNG/AVS/TCF/TCB

To our knowledge, TCF, which is designed to fulfill a symmetrical comparison between the properties of inner pattern (CNG, TCF, or COS/TCB) vs. outer pattern (AVS) and between overlapping string (CNG, AVS, or TCF) vs. maximally matched string (COS/TCB), has not been evaluated in any previous work. In short, while the original version of COS/TCB selects the maximally matched string (*i.e.*, *non-overlapping* string) as the feature (Feng *et al.*, 2004; Jiang *et al.*, 2010; Zhao & Kit, 2007), TCF collects features of *reduced n-gram* from every character position with additional rank of likelihood converted from *term-contributed frequency*, as its name implies. To compare different types of overlapping strings as unsupervised features systematically, this work extends the previous work of Zhao and Kit (2007) into a unified representation of features. The representation accommodates both character position of a string and the string's likelihood ranked in the logarithm. Formally, the ranking function for a string *s* with a score *x* counted by CNG, AVS, or TCF is defined as:

$$f(s) = r, if\, 2^r \leq x < 2^{r+1} \quad (4)$$

The logarithm ranking mechanism in (4) is inspired by Zipf's law with the intention to alleviate the potential data sparseness problem of infrequent strings. The rank *r* and the corresponding character positions of a string then are concatenated as feature tokens. To give the reader a clearer picture about what feature tokens look like, a sample representation, which is denoted in regex as "[*0-9*]+[*B*|*C*|*D*|*I*|*E*|*S*]" for rank and character position, of CNG, AVS, or TCF is demonstrated and explained by Figure 1 and Table 3.

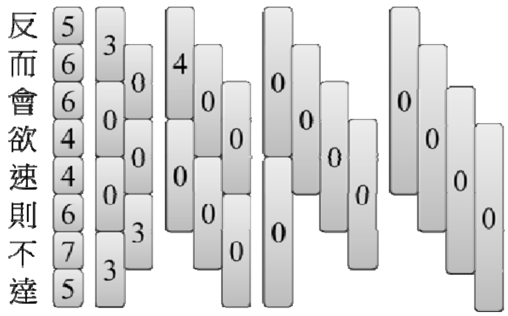

Figure 1. Example of overlapping strings with ranks.

Table 3. Sample of the unified feature representation for overlapping strings.

Input	**Unsupervised Feature**					**Label**
	1 char	2 char	3 char	4 char	5 char	
反	*5S*	*3B*	*4B*	*0B*	*0B*	*B*
而	*6S*	*3E*	*4C*	*0C*	*0C*	*E*
會	*6S*	*0E*	*4E*	*0D*	*0D*	*S*
欲	*4S*	*0E*	*0E*	*0E*	*0I*	*B*
速	*4S*	*0E*	*0E*	*0E*	*0E*	*C*
則	*6S*	*3B*	*0E*	*0E*	*0E*	*D*
不	*7S*	*3E*	*0E*	*0E*	*0E*	*I*
達	*5S*	*3E*	*0E*	*0E*	*0E*	*E*

For example, judging by strings with two characters, one of the strings “反而” gets rank r = 3; therefore, the column of two-character feature tokens has “反” denoted as *3B* and “而” denoted as *3E*. If another two-character string “而會” competes with “反而” at the position of “而” with a lower rank r = 0, then *3E* is selected for feature representation of the token at a certain position.

Note that, when the string “則不” conflicts with the string “不達” at the position of “不” with the same rank r = 3, the corresponding character position with rank of the leftmost string, which is *3E* in this case, is applied arbitrarily.

Although those are indeed common situations of overlapping strings, this work simply implements the above rules by Zhao and Kit (2007) for the sake of compatibility. In fact, pilot tests have been done with a more complicated representation, like *3E-0B* for “而” and *3E-3B* for “不,” to keep the overlapping information within each column, but the test result shows no significant differences in terms of accuracy and OOV recognition. Since the statistics of the pilot tests could be redundant, they are omitted in this paper.

To make an informative comparison, this work also applies the original version of *non-overlapping* COS/TCB features that is without ranks and is selected by the forward maximum matching algorithm (Feng *et al.*, 2004; Jiang *et al.*, 2010; Zhao & Kit, 2007). Table 4 illustrates a sample representation of features in this case. Notably, there are several features encoded as *-1* individually to represent that the desired string is unseen. For the *non-overlapping* siblings of the *reduced n-grams* family, such as COS/TCB, either the string is always occupied by other superstrings or it simply does not appear more than once.

Table 4. Sample of the unified feature representation for Non-overlapping COS/TCB strings.

Input	Original COS/TCB Feature	Label
反	*B*	*B*
而	*C*	*E*
會	*E*	*S*
欲	*-1*	*B*
速	*-1*	*C*
則	*-1*	*D*
不	*-1*	*I*
達	*-1*	*E*

The length of a string is limited to five characters for the sake of efficiency and consistency with the *6-tag* approach.

6. Experiments

CRF++ 0.54 (http://crfpp.sourceforge.net/) employs L-BFGS optimization and the tunable hyper-parameter (CRF++ training function argument "-c"), *i.e.*, the Gaussian prior, set to 100 throughout the whole experiment.

6.1 Data Set

The corpora used for the experiment are from the SIGHAN CWS bakeoff 2005 (Emerson, 2005) and SIGHAN CWS bakeoff 2010 (Zhao & Liu, 2010). SIGHAN 2005 comes with four different standards, including Academia Sinica (AS), City University of Hong Kong (CityU), Microsoft Research (MSR), and Peking University (PKU). SIGHAN 2010 provides a Traditional Chinese corpus and a Simplified Chinese corpus. Each corpus has training/test sets of four domains, including literature, computers, medicine, and finance, that are denoted as domains A, B, C, and D, respectively. For comparison, statistics on most corpora of SIGHAN 2003, 2006, and 2008 that have been obtained are listed in the appendix.

6.2 Unsupervised Feature Selection

Unsupervised features are collected according to pairs of corresponding training/test corpora. CNG and AVS are arranged with the help from SRILM (Stolcke, 2002). TCB strings and their ranks converted from TCF are calculated by YASA (Sung *et al.*, 2008). To distinguish the ranked and overlapping features of TCB/TCF from those of the original version of non-overlapping COS/TCB-based features, the former are denoted as TCF to indicate the score source of frequency for ranking, and the abbreviation of the later remains as TCB.

6.3 Evaluation Metrics

The evaluation metrics of CWS task are adopted from SIGHAN bakeoffs, including test *precision* (P), test *recall* (R), and their harmonic average *F_1 measure score* (F), as (5), (6), and (7), respectively. For performance of OOV, formulae that are similar to P/R/F are employed. To estimate the differences of performance between configurations of CWS experiments, this work uses the confidence level, which has been applied since SIGHAN CWS bakeoff 2003 (Sproat & Emerson, 2003). The confidence level assumes that the *recall (*or *precision) X of accuracy (or OOV recognition)* represents the probability that a word (or OOV word) will be identified from N words in total and that a binomial distribution is appropriate for the experiment. Confidence levels of P, R, P_{OOV}, and R_{OOV} appear in Tables 5-10 under the columns C_P, C_R, C_{Poov}, and C_{Roov}, respectively, and they are calculated at the 95% confidence interval with the formula $\pm 2 \sqrt{([X(1\text{-}X)] / N)}$. Two configurations of CWS experiments then are considered to be statistically different at a 95% confidence level if ***one of*** their C_P, C_R,

C_{Poov}, or C_{Roov} is different.

$$P = \frac{\text{the number of words thatare correctly segmented}}{\text{the number of words that are segmented}} \times 100\% \quad (5)$$

$$R = \frac{\text{the number of words that are correctly segmented}}{\text{the number of words in the gold standard}} \times 100\% \quad (6)$$

$$F = \frac{2 \times P \times R}{P + R} \quad (7)$$

6.4 Experimental Results

The most significant type of error is unintentionally segmented alphanumeric sequences, such as English words or factoids in Arabic numerals. Rather than developing another set of feature templates for non-Chinese characters that may violate the rules of closed training evaluation, post-processing, which is mentioned in the official report of SIGHAN CWS bakeoff 2005 (Emerson, 2005), has been applied to remove spaces between non-Chinese characters in the gold standard data of the AS corpus manually, since there are no urgent expectations of correct segmentation on non-Chinese text. In SIGHAN 2005 and 2006, however, some participants used character types, such as digits, date/time specific Chinese characters, English letters, punctuation, and others (Chinese characters) as extra features, which triggered a debate of closed training criteria (Zhao *et al.*, 2010). Consequently, SIGHAN 2010 decided to allow four types of characters, distinguished as Chinese characters, English letters, digits, and punctuation. This work provides preliminary tests on non-Chinese patterns extracted from SIGHAN 2010 unlabeled training corpora A and B, extra features of character types (in character based trigram, $T_{-1}T_0T_1$, where T can be E, D, P, or C for alphabets, digits, punctuations, or Chinese characters, respectively), and their combinations to verify the performance impact of these special treatments, as shown in Table 5 –Table 8. On the one hand, the statistics indicate that the character types perform well and stably on most of the corpora. On the other hand, the features, such as AVS and TCF, may still need help from non-Chinese patterns of unlabeled training corpora A and B. As a matter of fact, our other preliminary test suggests that SIGHAN 2010 test corpora contain a lot of OOV and inconsistent segments from non-Chinese text (for example, inconsistency of usage on full-width or half-width non-Chinese characters, some English words and factoids being segmented but some of them not, *etc.*), which only can be memorized from the non-Chinese patterns. Consequently, the experimental results of SIGHAN 2010 corpora involve non-Chinese treatment based on the combination of the extra character type features and the non-Chinese patterns, but the experimental results of SIGHAN 2005 corpora do not.

Table 5. Non-Chinese treatment on SIGHAN'10 simplified Chinese corpora.

Domain	Feature	***P***	***C_P***	***R***	***C_R***	***F***
A	Original 6-tag	92.16	±0.002869	91.63	±0.002956	91.89
	+(Non-Chinese Pattern)	92.32	±0.002842	91.27	±0.003013	91.79
	+(Character Type)	*92.70*	±0.002777	**92.33**	±0.002840	*92.51*
	+(Non-Chinese Pattern, Character Type)	**92.71**	±0.002775	**92.33**	±0.002841	**92.52**
B	Original 6-tag	77.44	±0.004558	86.72	±0.003701	81.82
	+(Non-Chinese Pattern)	89.85	±0.003294	83.62	±0.004036	86.62
	+(Character Type)	91.68	±0.003013	**93.58**	±0.002673	**92.62**
	+(Non-Chinese Pattern, Character Type)	**92.93**	±0.002795	*91.19*	±0.003091	*92.05*
C	Original 6-tag	89.61	±0.003466	90.64	±0.003309	90.12
	+(Non-Chinese Pattern)	90.87	±0.003272	89.77	±0.003443	90.32
	+(Character Type)	*91.11*	±0.003233	**92.02**	±0.003078	**91.56**
	+(Non-Chinese Pattern, Character Type)	**91.54**	±0.003161	*91.29*	±0.003203	*91.42*
D	Original 6-tag	89.82	±0.003367	91.24	±0.003148	90.52
	+(Non-Chinese Pattern)	*93.48*	±0.002749	91.06	±0.003176	92.25
	+(Character Type)	92.35	±0.002960	**93.99**	±0.002646	*93.16*
	+(Non-Chinese Pattern, Character Type)	**93.97**	±0.002650	*93.61*	±0.002723	**93.79**

Table 6. Non-Chinese treatment OOV on SIGHAN'10 simplified Chinese corpora.

Domain	Feature	***R_{OOV}***	***C_{Roov}***	***P_{OOV}***	***C_{Poov}***	***F_{OOV}***
A	Original 6-tag	55.52	±0.019647	52.00	±0.019752	53.71
	+(Non-Chinese Pattern)	53.71	±0.019714	52.34	±0.019746	53.01
	+(Character Type)	**62.42**	±0.019149	*58.86*	±0.019455	**60.59**
	+(Non-Chinese Pattern, Character Type)	*61.77*	±0.019212	**59.24**	±0.019427	*60.48*
B	Original 6-tag	36.06	±0.014105	20.49	±0.011855	26.13
	+(Non-Chinese Pattern)	41.38	±0.014467	52.17	±0.014673	46.16
	+(Character Type)	**76.27**	±0.012496	*71.40*	±0.013274	**73.76**
	+(Non-Chinese Pattern, Character Type)	*67.49*	±0.013759	**76.28**	±0.012495	*71.62*
C	Original 6-tag	59.69	±0.016736	49.40	±0.017059	54.06
	+(Non-Chinese Pattern)	58.80	±0.016793	54.76	±0.016982	56.71
	+(Character Type)	**68.14**	±0.015898	*59.69*	±0.016736	**63.64**
	+(Non-Chinese Pattern, Character Type)	*66.03*	±0.016159	**60.54**	±0.016677	*63.17*
D	Original 6-tag	48.79	±0.018869	35.90	±0.018109	41.36
	+(Non-Chinese Pattern)	53.98	±0.018815	55.56	±0.018757	54.76
	+(Character Type)	**68.81**	±0.017487	*57.73*	±0.018648	*62.79*
	+(Non-Chinese Pattern, Character Type)	*68.64*	±0.017514	**66.30**	±0.017844	**67.45**

Table 7. Non-Chinese treatment on SIGHAN'10 traditional Chinese corpora.

Domain	Feature	*P*	C_P	*R*	C_R	*F*
A	Original 6-tag	90.63	±0.003065	88.72	±0.003326	89.66
	+(Non-Chinese Pattern)	90.73	±0.003049	88.58	±0.003344	89.64
	+(Character Type)	**92.95**	±0.002691	*92.16*	±0.002826	*92.55*
	+(Non-Chinese Pattern, Character Type)	*92.94*	±0.002693	**92.20**	±0.002819	**92.57**
B	Original 6-tag	94.52	±0.002248	93.28	±0.002474	93.90
	+(Non-Chinese Pattern)	94.12	±0.002325	91.32	±0.002781	92.70
	+(Character Type)	**96.15**	±0.001902	**95.53**	±0.002042	**95.84**
	+(Non-Chinese Pattern, Character Type)	*95.63*	±0.002019	*94.22*	±0.002307	*94.92*
C	Original 6-tag	92.95	±0.002479	91.42	±0.002712	92.18
	+(Non-Chinese Pattern)	92.69	±0.002521	90.77	±0.002803	91.72
	+(Character Type)	**94.72**	±0.002167	**93.95**	±0.002308	**94.33**
	+(Non-Chinese Pattern, Character Type)	*94.62*	±0.002186	*93.77*	±0.002341	*94.19*
D	Original 6-tag	94.06	±0.002199	93.39	±0.002312	93.72
	+(Non-Chinese Pattern)	93.85	±0.002236	92.73	±0.002416	93.28
	+(Character Type)	**95.50**	±0.001928	**95.51**	±0.001926	**95.51**
	+(Non-Chinese Patter, Character Type)	*95.48*	±0.001933	*95.34*	±0.001961	*95.41*

Table 8. Non-Chinese treatment OOV on SIGHAN'10 traditional Chinese corpora.

Domain	Feature	R_{OOV}	C_{Roov}	P_{OOV}	C_{Poov}	F_{OOV}
A	Original 6-tag	72.50	±0.015297	57.20	±0.016951	63.95
	+(Non-Chinese Pattern)	71.62	±0.015446	57.04	±0.016959	63.50
	+(Character Type)	*75.45*	±0.014745	*67.72*	±0.016017	*71.38*
	+(Non-Chinese Pattern, Character Type)	**75.60**	±0.014715	**68.44**	±0.015923	**71.84**
B	Original 6-tag	*76.46*	±0.014455	71.38	±0.015399	73.83
	+(Non-Chinese Pattern)	68.49	±0.015828	65.20	±0.016229	66.80
	+(Character Type)	**80.44**	±0.013514	**81.81**	±0.013143	**81.12**
	+(Non-Chinese Pattern, Character Type)	74.07	±0.014931	*76.40*	±0.014466	*75.22*
C	Original 6-tag	73.48	±0.015336	58.33	±0.017128	65.03
	+(Non-Chinese Pattern)	69.69	±0.015968	56.31	±0.017232	62.29
	+(Character Type)	**76.91**	±0.014641	**68.87**	±0.016087	**72.67**
	+(Non-Chinese Pattern, Character Type)	*75.97*	±0.014843	*68.18*	±0.016181	*71.87*
D	Original 6-tag	78.54	±0.013963	66.01	±0.016110	71.73
	+(Non-Chinese Pattern)	75.53	±0.014622	63.69	±0.016355	69.11
	+(Character Type)	**81.58**	±0.013184	**76.99**	±0.014315	**79.22**
	+(Non-Chinese Pattern, Character Type)	*80.64*	±0.013438	*76.22*	±0.014481	*78.37*

This empirical decision implies that CWS benchmarking corpus should be prepared more carefully to avoid unpredictable side effects from non-Chinese text. Note that the treatment does not use unlabeled training corpora A and B separately. Further discussions are mainly based on this treatment, hopefully without loss of generality and of interest for comparative studies. Numbers in bold face and italic style indicate the best and the second best results of a certain evaluation metric, respectively, except for the topline and the best record from each year of SIGHAN bakeoffs. Configurations with the same values of confidence level on *P* or *R* are underlined, but only records that have the same confidence level on ***both*** *P* and *R* should be considered as statistically insignificant, and this phenomenon did not occur in our experiment results.

Unlike the previous work, which showed a relatively clearer trend of feature selection (Jiang *et al.*, 2011), CWS performance may vary between different CWS standards and domains in this study. Considering either the best or second best records in terms of F, feature combinations consisting of LRAVS or AVS usually outperform, except on MSR of SIGHAN 2005 corpora. Nevertheless, in terms of F_{OOV}, feature combinations consisting of TCF or TCB consistently increase in performance on every corpus. Similar situations also can be recognized from the experiments on some of the SIGHAN 2003, 2006, and 2008 corpora; please refer to the appendix for details. This complicated phenomenon indicates that, since CWS studies usually struggle with incremental and small improvements, different CWS standards and/or domains can make comparative research difficult and cause experimental results of related works to be incompatible. For equipping supervised CWS with unsupervised feature selection from unlabeled data, the experimental results of this work suggests that using LRAVS+TCF with more careful non-Chinese text treatments and CRF parameter tuning (*e.g.*, more cross-validations to find a specific hyper-parameter of Gaussian prior) would be a very good choice. Nevertheless, it is still worth noting that the best performance of this work in terms of *F* is found on the best official records on traditional Chinese domain B (Computer) of SIGHAN 2010 corpora and all of the SIGHAN 2005 corpora except the PKU corpus. This is especially true when this work does not apply any special treatment of character type and non-Chinese text that many other related works do on SIGHAN 2005 corpora. Note that “Our Baseline/Topline” in the following tables indicates where official baseline/topline suffered from official release script for maximum matching malfunctions on data in UTF-8 encoding and/or some uncertain incompatibilities between obtained corpora and official ones that caused inconsistent statistics during experiment reproductions.

Table 9. Performance comparison of accuracy on SIGHAN 2005 AS corpus.

Configuration	***P***	C_P	***R***	C_R	***F***
6-tag	94.50	±0.001308	95.74	±0.001159	95.12
CNG	95.12	±0.001236	95.53	±0.001186	95.32
AVS	95.14	±0.001234	95.86	±0.001143	95.50
TCB	94.48	±0.001311	95.73	±0.001160	95.10
TCF	94.86	±0.001267	95.92	±0.001135	95.39
AVS+TCB	*95.21*	±0.001226	95.96	±0.001130	*95.58*
AVS+TCF	**95.27**	±0.001218	**96.02**	±0.001121	**95.65**
LRAVS	94.88	±0.001265	95.91	±0.001136	95.39
LRAVS+TCB	95.03	±0.001247	**96.02**	±0.001122	95.52
LRAVS+TCF	95.00	±0.001251	96.01	±0.001124	95.50
2005 Best	95.10	±0.001230	*95.20*	±0.001220	95.20
2005 Baseline	85.70	±0.002000	90.90	±0.001643	88.20
Our Baseline	86.40	±0.001967	91.15	±0.001629	88.71
2005 Topline	98.50	±0.000694	97.90	±0.000819	98.20
Our Topline	98.64	±0.000665	97.97	±0.000809	98.30

Table 10. Performance comparison of OOV on SIGHAN 2005 AS corpus.

Configuration	R_{OOV}	C_{Roov}	P_{OOV}	C_{Poov}	F_{OOV}
6-tag	66.09	±0.012356	61.85	±0.012678	63.90
CNG	67.39	±0.012235	66.81	±0.01229	67.10
AVS	68.93	±0.012078	70.73	±0.011875	69.82
TCB	66.16	±0.012349	64.02	±0.012668	64.02
TCF	**70.27**	±0.011929	63.89	±0.012536	66.93
AVS+TCB	69.31	±0.012037	**71.49**	±0.011783	**70.38**
AVS+TCF	69.59	±0.012006	*70.94*	±0.011850	*70.26*
LRAVS	66.31	±0.012336	67.07	±0.012266	66.69
LRAVS+TCB	67.33	±0.012241	67.91	±0.012184	67.62
LRAVS+TCF	*69.82*	±0.011981	66.15	±0.012350	67.94
2005 Best	69.60	±0.012005	N/A	N/A	N/A
2005 Baseline	0.40	±0.001647	N/A	N/A	N/A
Our Baseline	1.41	±0.003080	3.08	±0.004512	1.94
2005 Topline	99.60	±0.001647	N/A	N/A	N/A
Our Topline	99.59	±0.001677	95.48	±0.005420	97.49

Table 11. Performance comparison of accuracy on SIGHAN 2005 CityU corpus.

Configuration	***P***	C_P	***R***	C_R	***F***
6-tag	94.82	±0.002207	94.64	±0.002245	94.73
CNG	**95.55**	±0.002055	94.39	±0.002292	94.97
AVS	95.27	±0.002115	94.93	±0.002185	95.10
TCB	95.21	±0.002129	94.93	±0.002186	95.07
TCF	95.30	±0.002107	94.96	±0.002180	95.13
AVS+TCB	95.34	±0.002100	95.13	±0.002145	95.23
AVS+TCF	95.39	±0.002088	95.15	±0.002140	95.27
LRAVS	95.35	±0.002099	95.08	±0.002155	95.21
LRAVS+TCB	*95.45*	±0.002077	**95.21**	±0.002127	**95.33**
LRAVS+TCF	95.41	±0.002085	*95.20*	±0.002130	*95.30*
2005 Best	94.60	±0.002230	94.10	±0.002330	94.30
2005 Baseline	79.00	±0.004026	88.20	±0.003189	83.30
Our Baseline	83.84	±0.003667	90.81	±0.002877	87.19
2005 Topline	99.10	±0.000934	98.80	±0.001076	98.20
Our Topline	99.24	±0.000867	98.90	±0.001040	99.07

Table 12. Performance comparison of OOV on SIGHAN 2005 CityU corpus.

Configuration	R_{OOV}	C_{Roov}	P_{OOV}	C_{Poov}	F_{OOV}
6-tag	69.15	±0.016141	65.54	±0.016609	67.30
CNG	69.68	±0.016063	69.41	±0.016104	69.55
AVS	70.48	±0.015942	71.90	±0.015709	71.18
TCB	*71.83*	±0.015721	70.12	±0.016236	70.12
TCF	**72.39**	±0.015624	68.76	±0.016198	70.53
AVS+TCB	71.14	±0.015836	72.70	±0.01557	*71.91*
AVS+TCF	70.97	±0.015863	72.77	±0.015556	71.86
LRAVS	69.78	±0.016048	72.09	±0.015676	70.92
LRAVS+TCB	70.57	±0.015926	*73.06*	±0.015505	71.80
LRAVS+TCF	71.17	±0.015831	**73.22**	±0.015475	**72.18**
2005 Best	69.80	±0.016046	N/A	N/A	N/A
2005 Baseline	0.00	±0.000000	N/A	N/A	N/A
Our Baseline	16.22	±0.012882	33.91	±0.016544	21.94
2005 Topline	99.70	±0.001911	N/A	N/A	N/A
Our Topline	99.74	±0.001794	98.82	±0.003771	99.28

Table 13. Performance comparison of accuracy on SIGHAN 2005 MSR corpus.

Configuration	***P***	***C_P***	***R***	***C_R***	***F***
6-tag	*97.29*	±0.000998	97.03	±0.001042	*97.16*
CNG	97.02	±0.001045	96.87	±0.001069	96.95
AVS	97.24	±0.001007	96.91	<u>±0.001063</u>	97.07
TCB	**97.32**	±0.000993	**97.09**	±0.001033	**97.20**
TCF	97.02	±0.001044	96.70	±0.001097	96.86
AVS+TCB	97.16	±0.001020	96.91	<u>±0.001063</u>	97.04
AVS+TCF	97.25	±0.001005	97.00	±0.001049	97.12
LRAVS	97.20	±0.001014	97.01	±0.001046	97.10
LRAVS+TCB	97.21	±0.001012	*97.05*	±0.001040	97.13
LRAVS+TCF	*97.29*	±0.000997	96.43	±0.001139	96.86
2005 Best	96.60	±0.001110	96.20	±0.001170	96.40
2005 Baseline	91.20	±0.001733	95.50	±0.001268	93.30
Our Baseline	91.74	±0.001691	95.69	±0.001247	93.67
2005 Topline	99.20	±0.000545	99.10	±0.000578	99.10
Our Topline	99.31	±0.000510	99.10	±0.000580	99.20

Table 14. Performance comparison of OOV on SIGHAN 2005 MSR corpus.

Configuration	***R_{OOV}***	***C_{Roov}***	***P_{OOV}***	***C_{Poov}***	***F_{OOV}***
6-tag	72.22	±0.015108	60.52	±0.016487	65.85
CNG	71.37	±0.015247	62.08	±0.016365	66.40
AVS	69.88	±0.015474	61.96	±0.016375	65.68
TCB	72.96	±0.014982	**66.73**	±0.016414	66.73
TCF	**73.81**	<u>±0.014830</u>	58.68	±0.016608	65.38
AVS+TCB	70.41	±0.015395	62.11	±0.016362	66.00
AVS+TCF	71.12	±0.015286	62.54	±0.016325	66.56
LRAVS	70.91	±0.015319	63.02	±0.016283	*66.73*
LRAVS+TCB	71.05	±0.015297	*63.49*	±0.016239	**67.06**
LRAVS+TCF	**73.81**	<u>±0.014830</u>	59.28	±0.016571	65.75
2005 Best	71.70	±0.015194	N/A	N/A	N/A
2005 Baseline	0.00	±0.000000	N/A	N/A	N/A
Our Baseline	2.47	±0.005240	16.71	±0.012582	4.31
2005 Topline	99.80	±0.001507	N/A	N/A	N/A
Our Topline	99.79	±0.001552	99.37	±0.002676	99.58

Table 15. Performance comparison of accuracy on SIGHAN 2005 PKU corpus.

Configuration	***P***	C_P	***R***	C_R	***F***
6-tag	93.73	±0.001512	92.70	±0.001623	93.21
CNG	**94.36**	±0.001438	**93.57**	±0.001530	**93.96**
AVS	94.21	±0.001457	93.24	±0.001566	93.72
TCB	93.97	±0.001485	92.76	±0.001616	93.36
TCF	93.94	±0.001488	92.81	±0.001611	93.37
AVS+TCB	94.33	<u>±0.001443</u>	93.31	±0.001559	93.81
AVS+TCF	94.25	±0.001451	93.44	<u>±0.001544</u>	93.85
LRAVS	*94.34*	±0.001441	*93.48*	±0.001540	*93.91*
LRAVS+TCB	94.32	<u>±0.001443</u>	93.44	<u>±0.001544</u>	93.88
LRAVS+TCF	93.91	±0.001492	92.20	±0.001672	93.05
2005 Best	94.60	±0.001400	95.30	±0.001310	95.00
2005 Baseline	83.60	±0.002292	90.40	±0.001824	86.90
Our Baseline	84.29	±0.002269	90.68	±0.001813	87.37
2005 Topline	98.80	±0.000674	98.50	±0.000752	98.70
Our Topline	98.96	±0.000634	98.62	±0.000726	98.79

Table 16. Performance comparison of OOV on SIGHAN 2005 PKU corpus.

Configuration	R_{OOV}	C_{Roov}	P_{OOV}	C_{Poov}	F_{OOV}
6-tag	57.48	±0.012083	48.04	±0.012211	52.33
CNG	**65.58**	±0.011612	**57.87**	±0.012068	**61.48**
AVS	62.69	±0.011821	55.60	±0.012144	58.93
TCB	60.07	±0.011970	54.87	<u>±0.012220</u>	54.87
TCF	60.39	±0.011954	50.41	<u>±0.012220</u>	54.95
AVS+TCB	64.02	±0.011730	56.97	±0.012101	60.29
AVS+TCF	63.80	±0.011746	56.06	±0.012130	59.68
LRAVS	65.02	±0.011656	57.31	±0.012089	60.92
LRAVS+TCB	*65.42*	±0.011625	*57.60*	±0.012079	*61.26*
LRAVS+TCF	60.42	±0.011952	48.92	±0.012218	54.07
2005 Best	63.60	±0.011760	N/A	N/A	N/A
2005 Baseline	5.90	±0.005759	N/A	N/A	N/A
Our Baseline	6.86	±0.006178	6.10	±0.005850	6.46
2005 Topline	99.40	±0.001888	N/A	N/A	N/A
Our Topline	99.37	±0.001938	97.72	±0.003645	98.54

Table 17. Non-Chinese treatment performance comparison of accuracy on SIGHAN 2010 simplified Chinese domain A (Literature) corpus.

Configuration	***P***	C_P	***R***	C_R	***F***
6-tag	92.83	±0.002754	92.37	±0.002833	92.60
CNG	*93.69*	±0.002595	91.94	±0.002906	92.81
AVS	93.47	±0.002638	92.89	±0.002744	93.18
TCB	93.12	±0.002702	92.56	±0.002801	92.84
TCF	93.18	±0.002690	92.52	±0.002808	92.85
AVS+TCB	93.68	±0.002596	92.99	±0.002726	93.33
AVS+TCF	93.67	±0.002600	93.10	±0.002705	*93.38*
LRAVS	93.55	±0.002623	93.08	±0.002709	93.31
LRAVS+TCB	93.56	±0.002620	*93.11*	±0.002703	93.33
LRAVS+TCF	**93.72**	±0.002589	**93.28**	±0.002673	**93.50**
2010 Best	94.60	±0.002390	94.50	±0.002410	94.60
2010 Baseline	86.20	±0.003648	91.70	±0.002919	88.90
Our Baseline	86.24	±0.003676	91.67	±0.002949	88.88
2010 Topline	99.00	±0.001053	98.60	±0.001243	98.80
Our Topline	99.02	±0.001052	98.57	±0.001268	98.79

Table 18. Non-Chinese treatment performance comparison of OOV on SIGHAN 2010 simplified Chinese domain A (Literature) corpus.

Configuration	R_{OOV}	C_{Roov}	P_{OOV}	C_{Poov}	F_{OOV}
6-tag	62.62	±0.019128	59.98	±0.01937	61.27
CNG	**65.36**	±0.018812	62.81	±0.019109	64.06
AVS	64.80	±0.018882	66.63	±0.018643	65.70
TCB	64.48	±0.018921	63.35	±0.019164	63.35
TCF	65.00	±0.018858	62.36	±0.019155	63.65
AVS+TCB	*65.04*	±0.018853	67.43	±0.018528	66.22
AVS+TCF	64.96	±0.018863	*67.60*	±0.018502	*66.26*
LRAVS	63.67	±0.019015	66.71	±0.018632	65.15
LRAVS+TCB	64.35	±0.018936	67.09	±0.018578	65.69
LRAVS+TCF	64.92	±0.018868	**68.48**	±0.018368	**66.65**
2010 Best	81.60	±0.015320	N/A	N/A	N/A
2010 Baseline	15.60	±0.014346	N/A	N/A	N/A
Our Baseline	15.69	±0.014378	30.61	±0.01822	20.74
2010 Topline	99.60	±0.002495	N/A	N/A	N/A
Our Topline	99.60	±0.002505	96.48	±0.007282	98.02

Table 19. Non-Chinese treatment performance comparison of accuracy on SIGHAN 2010 simplified Chinese domain B (Computer) corpus.

Configuration	***P***	C_P	***R***	C_R	***F***
6-tag	90.95	±0.003129	92.46	±0.002880	91.70
CNG	91.45	±0.003050	92.36	±0.002898	91.90
AVS	91.25	±0.003081	*92.72*	±0.002833	91.98
TCB	91.21	±0.003087	92.53	±0.002867	91.87
TCF	90.86	±0.003143	92.62	±0.002852	91.73
AVS+TCB	91.60	±0.003026	92.67	±0.002842	92.13
AVS+TCF	90.81	±0.003151	92.16	±0.002932	91.48
LRAVS	*91.71*	±0.003007	92.61	±0.002854	*92.16*
LRAVS+TCB	**91.97**	±0.002963	**92.76**	±0.002826	**92.37**
LRAVS+TCF	91.28	±0.003077	92.60	±0.002856	91.93
2010 Best	95.00	±0.002320	95.30	±0.002250	95.10
2010 Baseline	63.20	±0.005132	85.60	±0.003736	72.70
Our Baseline	63.26	±0.005258	85.68	±0.003820	72.78
2010 Topline	99.30	±0.000887	99.10	±0.001005	99.20
Our Topline	99.25	±0.000940	99.06	±0.001052	99.16

Table 20. Non-Chinese treatment performance comparison of OOV on SIGHAN 2010 simplified Chinese domain B (Computer) corpus.

Configuration	R_{OOV}	C_{Roov}	P_{OOV}	C_{Poov}	F_{OOV}
6-tag	70.62	±0.013380	67.66	±0.013740	69.11
CNG	70.38	±0.013412	65.17	±0.013994	67.67
AVS	69.85	±0.013479	66.16	±0.013898	67.96
TCB	71.23	±0.013297	**69.66**	±0.013684	69.66
TCF	**72.01**	±0.013187	66.02	±0.013913	68.89
AVS+TCB	70.25	±0.013429	67.22	±0.013788	68.70
AVS+TCF	69.63	±0.013507	63.73	±0.014123	66.55
LRAVS	71.25	±0.013294	68.25	±0.013673	*69.72*
LRAVS+TCB	*71.81*	±0.013216	*69.47*	±0.013528	**70.62**
LRAVS+TCF	70.92	±0.013340	66.13	±0.013902	68.44
2010 Best	82.70	±0.011111	N/A	N/A	N/A
2010 Baseline	16.30	±0.010850	N/A	N/A	N/A
Our Baseline	16.65	±0.010944	6.39	±0.007185	9.24
2010 Topline	99.00	±0.002923	N/A	N/A	N/A
Our Topline	99.00	±0.002930	98.08	±0.004028	98.54

Table 21. Non-Chinese treatment performance comparison of accuracy on SIGHAN 2010 simplified Chinese domain C (Medicine) corpus.

Configuration	***P***	***C_P***	***R***	***C_R***	***F***
6-tag	91.27	±0.003207	91.96	±0.003089	91.61
CNG	92.84	±0.002928	92.07	±0.003069	92.46
AVS	92.40	±0.003011	92.89	±0.002919	92.64
TCB	91.55	±0.003159	92.19	±0.003048	91.87
TCF	91.62	±0.003147	92.21	±0.003045	91.91
AVS+TCB	92.73	±0.002949	92.90	±0.002917	92.82
AVS+TCF	92.82	±0.002933	93.07	±0.002885	92.94
LRAVS	*93.12*	±0.002876	*93.22*	±0.002856	*93.17*
LRAVS+TCB	**93.12**	±0.002875	**93.33**	±0.002834	**93.23**
LRAVS+TCF	93.07	±0.002884	93.20	±0.002859	93.14
2010 Best	93.60	±0.002760	94.20	±0.002630	93.90
2010 Baseline	77.40	±0.004714	88.60	±0.003582	82.60
Our Baseline	77.46	±0.004746	88.64	±0.003604	82.68
2010 Topline	99.10	±0.001064	98.90	±0.001176	99.00
Our Topline	99.18	±0.001025	98.97	±0.001146	99.08

Table 22. Non-Chinese treatment performance comparison of OOV on SIGHAN 2010 simplified Chinese domain C (Medicine) corpus.

Configuration	***R_{OOV}***	***C_{Roov}***	***P_{OOV}***	***C_{Poov}***	***F_{OOV}***
6-tag	66.70	±0.016081	61.15	±0.016630	63.80
CNG	70.90	±0.015498	70.46	±0.015567	70.68
AVS	71.02	±0.015479	69.61	±0.015692	70.31
TCB	66.41	±0.016115	60.67	±0.016667	63.41
TCF	66.44	±0.016112	60.65	±0.016668	63.41
AVS+TCB	70.10	±0.015621	69.00	±0.015780	69.54
AVS+TCF	69.66	±0.015685	69.11	±0.015765	69.38
LRAVS	**71.62**	±0.015382	**70.91**	±0.015497	**71.26**
LRAVS+TCB	71.45	±0.015410	70.39	±0.015576	70.92
LRAVS+TCF	*71.56*	±0.015392	*70.53*	±0.015556	*71.04*
2010 Best	75.00	±0.014774	N/A	N/A	N/A
2010 Baseline	12.30	±0.011206	N/A	N/A	N/A
Our Baseline	12.33	±0.011218	15.34	±0.012294	13.67
2010 Topline	98.00	±0.004777	N/A	N/A	N/A
Our Topline	98.21	±0.004519	97.21	±0.005623	97.71

Table 23. Non-Chinese treatment performance comparison of accuracy on SIGHAN 2010 simplified Chinese domain D (Finance) corpus.

Configuration	P	C_P	R	C_R	F
6-tag	93.01	±0.002838	93.74	±0.002697	93.38
CNG	*94.40*	±0.002561	93.66	±0.002714	94.02
AVS	93.54	±0.002736	*94.30*	±0.002581	93.92
TCB	93.35	±0.002774	94.14	±0.002614	93.74
TCF	93.10	±0.002822	93.88	±0.002669	93.49
AVS+TCB	**94.56**	±0.002526	**94.49**	±0.002540	**94.53**
AVS+TCF	94.05	±0.002633	94.10	±0.002624	94.08
LRAVS	94.30	±0.002582	94.13	±0.002616	94.21
LRAVS+TCB	94.36	±0.002568	94.16	±0.002611	94.26
LRAVS+TCF	94.36	±0.002569	94.19	±0.002604	*94.28*
2010 Best	96.00	±0.002160	95.90	±0.002180	95.90
2010 Baseline	80.30	±0.004377	91.40	±0.003085	85.50
Our Baseline	80.26	±0.004431	91.41	±0.003119	85.48
2010 Topline	99.50	±0.000776	99.40	±0.000850	99.40
Our Topline	99.56	±0.000734	99.47	±0.000810	99.52

Table 24. Non-Chinese treatment performance comparison of OOV on SIGHAN 2010 simplified Chinese domain D (Finance) corpus.

Configuration	R_{OOV}	C_{Roov}	P_{OOV}	C_{Poov}	F_{OOV}
6-tag	67.60	±0.017666	61.28	±0.018388	64.28
CNG	*73.53*	±0.016655	67.77	±0.017642	70.53
AVS	71.10	±0.017111	64.17	±0.018101	67.46
TCB	70.58	±0.017201	66.44	±0.018250	66.44
TCF	70.13	±0.017277	61.19	±0.018396	65.35
AVS+TCB	**73.80**	±0.016598	**70.79**	±0.017166	**72.26**
AVS+TCF	70.76	±0.017172	67.73	±0.017648	69.21
LRAVS	71.66	±0.017012	68.54	±0.017528	70.07
LRAVS+TCB	72.63	±0.016831	*69.82*	±0.017328	*71.20*
LRAVS+TCF	72.38	±0.016878	69.40	±0.017396	70.86
2010 Best	82.70	±0.014279	N/A	N/A	N/A
2010 Baseline	23.30	±0.015958	N/A	N/A	N/A
Our Baseline	23.32	±0.015963	14.15	±0.013157	17.61
2010 Topline	99.50	±0.002663	N/A	N/A	N/A
Our Topline	99.72	±0.001985	99.34	±0.003047	99.53

Table 25. Non-Chinese treatment performance comparison of accuracy on SIGHAN 2010 traditional Chinese domain A (Literature) corpus.

Configuration	***P***	C_P	***R***	C_R	***F***
6-tag	93.06	±0.002672	92.31	±0.002802	92.68
CNG	93.66	±0.002562	91.16	±0.002985	92.39
AVS	93.61	±0.002572	92.78	±0.002721	93.19
TCB	93.21	±0.002646	92.33	±0.002798	92.77
TCF	93.33	±0.002623	92.58	±0.002756	92.95
AVS+TCB	93.61	±0.002572	92.85	±0.002709	93.23
AVS+TCF	93.68	±0.002559	92.98	±0.002685	93.33
LRAVS	*93.77*	±0.002542	*93.04*	±0.002676	*93.40*
LRAVS+TCB	**93.77**	±0.002541	**93.06**	±0.002673	**93.41**
LRAVS+TCF	93.65	±0.002564	92.92	±0.002697	93.28
2010 Best	94.20	±0.002450	94.20	±0.002450	94.20
2010 Baseline	78.80	±0.004286	86.30	±0.003606	82.40
Our Baseline	78.83	±0.004295	86.39	±0.003605	82.44
2010 Topline	98.80	±0.001142	98.10	±0.001432	98.50
Our Topline	98.83	±0.001130	98.11	±0.001430	98.47

Table 26. Non-Chinese treatment performance comparison of OOV on SIGHAN 2010 traditional Chinese domain A (Literature) corpus.

Configuration	R_{OOV}	C_{Roov}	P_{OOV}	C_{Poov}	F_{OOV}
6-tag	75.89	±0.014654	68.68	±0.015889	72.11
CNG	74.12	±0.015004	69.46	±0.015780	71.71
AVS	75.10	±0.014816	73.34	±0.015148	74.21
TCB	**77.19**	±0.014376	69.27	±0.015807	73.01
TCF	*77.10*	±0.014395	69.82	±0.015727	73.28
AVS+TCB	75.54	±0.014727	73.46	±0.015127	74.48
AVS+TCF	75.60	±0.014715	73.92	±0.015042	74.75
LRAVS	75.42	±0.014751	*74.93*	±0.014848	*75.18*
LRAVS+TCB	75.66	±0.014703	**75.12**	±0.014810	**75.39**
LRAVS+TCF	75.27	±0.014780	74.44	±0.014944	74.85
2010 Best	78.80	±0.014003	N/A	N/A	N/A
2010 Baseline	4.10	±0.006793	N/A	N/A	N/A
Our Baseline	4.10	±0.006791	8.93	±0.009769	5.62
2010 Topline	99.80	±0.001531	N/A	N/A	N/A
Our Topline	99.82	±0.001439	99.33	±0.002804	99.57

Table 27. Non-Chinese treatment performance comparison of accuracy on SIGHAN 2010 traditional Chinese domain B (Computer) corpus.

Configuration	P	C_P	R	C_R	F
6-tag	95.15	±0.002122	93.20	±0.002487	94.17
CNG	95.60	±0.002027	93.16	±0.002494	94.36
AVS	*95.67*	±0.002012	*93.83*	±0.002378	*94.74*
TCB	95.21	±0.002111	93.25	±0.002480	94.22
TCF	95.28	±0.002095	93.42	±0.002450	94.34
AVS+TCB	95.62	±0.002023	93.72	±0.002398	94.66
AVS+TCF	**95.74**	±0.001996	93.83	±0.002378	**94.77**
LRAVS	95.57	±0.002034	93.79	±0.002384	94.67
LRAVS+TCB	95.63	±0.002020	**93.85**	±0.002373	94.73
LRAVS+TCF	95.55	±0.002038	93.81	±0.002381	94.67
2010 Best	95.70	±0.001950	94.80	±0.002130	95.20
2010 Baseline	70.10	±0.004390	87.30	±0.003193	77.80
Our Baseline	70.15	±0.004522	87.33	±0.003286	77.80
2010 Topline	99.10	±0.000906	98.80	±0.001044	99.00
Our Topline	99.38	±0.000778	98.85	±0.001055	99.11

Table 28. Non-Chinese-Pattern performance comparison of OOV on SIGHAN 2010 traditional Chinese domain B (Computer) corpus.

Configuration	R_{OOV}	C_{Roov}	P_{OOV}	C_{Poov}	F_{OOV}
6-tag	58.79	±0.016769	68.17	±0.015871	63.14
CNG	*61.77*	±0.016556	70.16	±0.015589	65.70
AVS	60.59	±0.016649	72.29	±0.015248	65.93
TCB	59.09	±0.016751	68.81	±0.015784	63.58
TCF	59.34	±0.016735	69.21	±0.015727	63.89
AVS+TCB	60.89	±0.016626	72.24	±0.015257	66.08
AVS+TCF	61.35	±0.01659	72.90	±0.015143	66.63
LRAVS	61.67	±0.016564	72.84	±0.015155	*66.79*
LRAVS+TCB	**61.82**	±0.016552	**73.07**	±0.015113	**66.98**
LRAVS+TCF	61.55	±0.016574	*72.94*	±0.015135	66.76
2010 Best	66.60	±0.016069	N/A	N/A	N/A
2010 Baseline	1.00	±0.003390	N/A	N/A	N/A
Our Baseline	1.03	±0.003445	0.55	±0.002515	0.72
2010 Topline	99.60	±0.002150	N/A	N/A	N/A
Our Topline	99.34	±0.002765	99.41	±0.002609	99.37

Table 29. Non-Chinese treatment performance comparison of accuracy on SIGHAN 2010 traditional Chinese domain C (Medicine) corpus.

Configuration	***P***	C_P	***R***	C_R	***F***
6-tag	94.70	±0.002170	93.83	±0.002331	94.26
CNG	95.35	±0.002039	93.35	±0.002414	94.34
AVS	95.28	±0.002055	94.37	±0.002232	94.82
TCB	94.76	±0.002158	93.87	±0.002324	94.31
TCF	94.88	±0.002135	94.05	±0.002291	94.46
AVS+TCB	95.33	±0.002044	94.49	±0.002209	94.91
AVS+TCF	95.33	±0.002043	94.44	±0.002219	94.88
LRAVS	**95.52**	±0.002003	**94.60**	±0.002190	**95.06**
LRAVS+TCB	95.36	±0.002038	*94.51*	±0.002206	*94.93*
LRAVS+TCF	*95.42*	±0.002025	94.42	±0.002224	94.91
2010 Best	95.70	±0.001950	95.30	±0.002030	95.50
2010 Baseline	81.00	±0.003764	88.60	±0.003049	84.60
Our Baseline	80.98	±0.003801	88.63	±0.003075	84.64
2010 Topline	98.90	±0.001001	98.40	±0.001204	98.60
Our Topline	98.91	±0.001006	98.38	±0.001223	98.64

Table 30. Non-Chinese treatment performance comparison of OOV on SIGHAN 2010 traditional Chinese domain C (Medicine) corpus.

Configuration	R_{OOV}	C_{Roov}	P_{OOV}	C_{Poov}	F_{OOV}
6-tag	74.79	±0.015086	67.98	±0.016209	71.22
CNG	77.16	±0.014586	71.22	±0.015730	74.07
AVS	76.13	±0.014810	74.80	±0.015083	75.46
TCB	75.60	±0.014922	68.64	±0.016119	71.95
TCF	75.79	±0.014883	69.29	±0.016026	72.39
AVS+TCB	76.72	±0.014683	*75.75*	±0.014890	76.23
AVS+TCF	77.22	±0.014572	75.69	±0.014903	76.44
LRAVS	**78.65**	±0.014237	**76.37**	±0.014759	**77.49**
LRAVS+TCB	77.75	±0.014451	75.54	±0.014934	76.63
LRAVS+TCF	*78.03*	±0.014385	75.65	±0.014911	*76.82*
2010 Best	79.80	±0.013949	N/A	N/A	N/A
2010 Baseline	2.70	±0.005631	N/A	N/A	N/A
Our Baseline	2.71	±0.005639	4.34	±0.007082	3.34
2010 Topline	99.20	±0.003095	N/A	N/A	N/A
Our Topline	99.16	±0.003171	98.73	±0.003891	98.94

Table 31. Non-Chinese treatment performance comparison of accuracy on SIGHAN 2010 traditional Chinese domain D (Finance) corpus.

Configuration	***P***	***C_P***	***R***	***C_R***	***F***
6-tag	95.52	±0.001925	95.46	±0.001937	95.49
CNG	**96.13**	±0.001794	95.04	±0.002020	95.58
AVS	95.99	±0.001825	95.79	±0.001868	95.89
TCB	95.55	±0.001918	95.51	±0.001927	95.53
TCF	95.61	±0.001907	95.57	±0.001915	95.59
AVS+TCB	95.93	±0.001839	95.77	±0.001874	95.85
AVS+TCF	95.99	±0.001825	**95.88**	±0.001850	**95.93**
LRAVS	96.02	±0.001820	95.73	±0.001881	95.87
LRAVS+TCB	*96.04*	±0.001814	*95.82*	±0.001862	*95.93*
LRAVS+TCF	95.94	±0.001836	95.71	±0.001885	95.83
2010 Best	96.20	±0.001760	96.40	±0.001720	96.30
2010 Baseline	82.60	±0.003492	88.80	±0.002905	85.50
Our Baseline	82.56	±0.003531	88.77	±0.002937	85.55
2010 Topline	98.60	±0.001082	98.10	±0.001258	98.40
Our Topline	98.63	±0.001081	98.10	±0.00127	98.36

Table 32. Non-Chinese treatment performance comparison of OOV on SIGHAN 2010 traditional Chinese domain D (Finance) corpus.

Configuration	***R_{OOV}***	***C_{Roov}***	***P_{OOV}***	***C_{Poov}***	***F_{OOV}***
6-tag	80.45	±0.013488	76.61	±0.014398	78.48
CNG	**82.96**	±0.012787	78.16	±0.014053	80.49
AVS	81.33	±0.013253	81.28	±0.013267	81.30
TCB	80.99	±0.013346	77.44	±0.014216	79.17
TCF	80.92	±0.013363	77.26	±0.014255	79.05
AVS+TCB	80.99	±0.013346	81.55	±0.013193	81.27
AVS+TCF	80.99	±0.013346	81.96	±0.013077	81.47
LRAVS	*82.62*	±0.012889	*82.10*	±0.013038	**82.36**
LRAVS+TCB	82.18	±0.013016	**82.44**	±0.012942	*82.31*
LRAVS+TCF	81.86	±0.013105	82.04	±0.013054	81.95
2010 Best	81.20	±0.013288	N/A	N/A	N/A
2010 Baseline	0.60	±0.002627	N/A	N/A	N/A
Our Baseline	0.60	±0.002618	2.28	±0.005078	0.95
2010 Topline	99.70	±0.001860	N/A	N/A	N/A
Our Topline	99.69	±0.001902	98.54	±0.004076	99.11

It has been observed that using any of the unsupervised features could create short patterns for the CRF learner, which might break more English words than using the *6-tag* approach alone. AVS, TCF, and TCB, however, resolve more overlapping ambiguities of Chinese words than the *6-tag* approach and CNG. Interestingly, even for the unsupervised feature without rank or overlapping information, TCB/TCF successfully recognizes “依靠 / 单位 / 的 / 纽带 / 来 / 维持,” while the *6-tag* approach sees this phrase incorrectly as “依靠 / 单位 / 的 / 纽 / 带来 / 维持.” TCB/TCF also saves more factoids, such as “一二九・九 / 左右” (129.9 / around) from scattered tokens, such as “一二九 / ・ / 九 / 左右” (129 / point / 9 / around).

The above observations suggest that the quality of a string as a word-like candidate should be an important factor for the unsupervised feature injected CRF learner. Relatively speaking, CNG probably brings in too much noise. Feature combinations of LRAVS and TCF usually improve F and F_{OOV}, respectively. Improvements are significant in terms of C_R, C_P, C_{Roov}, and C_{Poov},, which confirms the hypothesis mentioned at the end of Section 1.3 that, combining information from the outer pattern of a substring (*i.e.*, LRAVS) with information from the inner pattern of a substring (*i.e.*, TCF) into a compound of unsupervised feature could help improving CWS performance of supervised labeling scheme of CRF. Nevertheless, since AVS or TCB sometimes gain better results, fine-tuning of feature engineering according to different corpora and segmentation standards is necessary.

7. Conclusion and Future Work

This work provides a unified view of CRF-based CWS integrated with unsupervised features via frequent string, and it reasons that, since LRAVS comes with inner structure and TCF comes with outer structure of overlapping string, utilizing their compound features could be more useful than applying one of them solely. The thorough experimental results show that the compound features of LRAVS and TCF usually obtain competitive performance in terms of F and F_{OOV}, respectively. Sometimes, AVS and TCB may contribute more, but generally combining the outer pattern of a substring (*i.e.*, LRAVS or AVS) with the inner pattern of a substring (*i.e.*, TCF or TCB) into a compound of unsupervised features could help improve CWS performance of a supervised labeling scheme of CRF. Recommended future investigation is unknown word extraction and named entity recognition using AVS (Li *et al.*, 2010) and TCF/TCB(Chang & Lee, 2003; Zhang *et al.*, 2010) as features for more complicated CRF (Sun & Nan, 2010).

Reference

Chang, J.-S., & Su, K.-Y. (1997). An Unsupervised Iterative Method for Chinese New Lexicon Extraction. in *Proc. Computational Linguistics and Chinese Language Processing,* 2(2), 97-148.

Chang, T.-H., & Lee, C.-H. (2003). Automatic Chinese unknown word extraction using small-corpus-based method. in *Proc. International Conference on Natural Language Processing and Knowledge Engineering*, 459-464.

Chien, L.-F. (1997). PAT-tree-based Keyword Extraction for Chinese Information Retrieval. in *Proc. 20th Annnual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 50-58.

Cohen, P., Adams, N., & Heeringa, B. (2007). Voting Experts: An Unsupervised Algorithm for Segmenting Sequences. *Intelligent Data Analysis,* 11(6), 607-625.

Emerson, T. (2005). The Second International Chinese Word Segmentation Bakeoff. in *Proc. 4th SIGHAN Workshop on Chinese Language Processing*.

Feng, H., Chen, K., Deng, X., & Zheng, W. (2004). Accessor Variety Criteria for Chinese Word Extraction. *Computational Linguistics,* 30(1), 75-93.

Ha, L. Q., Seymour, R., Hanna, P., & Smith, F. J. (2005). Reduced N-Grams for Chinese Evaluation. *Computational Linguistics and Chinese Language Processing,* 10(1), 19-34.

Harris, Z. S. (1970). Morpheme Boundaries within Words. Paper presented at the *Structural and Transformational Linguistics*.

Huang, J. H., & Powers, D. (2003). Chinese Word Segmentation based on contextual entropy. in *Proc. 17th Asian Pacific Conference on Language, Information and Computation*, 152-158.

Jiang, T.-J., Hsu, W.-L., Kuo, C.-H., & Yang, T.-H. (2011). Enhancement of Unsupervised Feature Selection for Conditional Random Fields Learning in Chinese Word Segmentation. in *Proc. 7th IEEE International Conference on Natural Language Processing and Knowledge Engineering,* 382-389.

Jiang, T.-J., Liu, S.-H., Sung, C.-L., & Hsu, W.-L. (2010). Term Contributed Boundary Tagging by Conditional Random Fields for SIGHAN 2010 Chinese Word Segmentation Bakeoff. in *Proc. 1st CIPS-SIGHAN Joint Conf. on Chinese Language Processing*, Beijing, China.

Jin, G., & Chen, X. (2007). The Fourth International Chinese Language Processing Bakeoff : Chinese Word Segmentation, Named Entity Recognition and Chinese POS Tagging. in *Proc. 6th SIGHAN Workshop on Chinese Language Processing*, 69-81.

Kit, C., & Wilks, Y. (1999). Unsupervised learning of word boundary with description length gain. in *Proc. CoNLL-99*, 1-6.

Lü, X., & Zhang, L. (2005). Statistical Substring Reduction in Linear Time. in *Proc. 1st Internal Joint Conference on Natural Language Processing*.

Lafferty, J., McCallum, A., & Pereira, F. C. N. (2001). Conditional Random Fields

Probabilistic Models for Segmenting and Labeling Sequence Data. in *Proc. ICML*. 282-289.

Levow, G.-A. (2006). The Third International Chinese Language Processing Bakeoff Word Segmentation and Named Entity Recognition. in *Proc. 5th SIGHAN Workshop on Chinese Language Processing,* 108-117.

Li, L., Li, Z., Ding, Z., & Huang, D. (2010). A Hybrid Model Combining CRF with Boundary Templates for Chinese Person Name Recognition. *International Journal Advanced Intelligent,* 2(1), 73-80.

Li, M., Gao, J., Huang, C., & Li, J. (2003). Unsupervised Training for Overlapping Ambiguity Resolution in Chinese Word Segmentation. in *Proc. 2nd SIGHAN Workshop on Chinese Language Processing*, 17, 1-7.

Lin, Y.-J., & Yu, M.-S. (2001). Extracting Chinese Frequent Strings without a Dictionary from a Chinese Corpus and its Applications. *J. Information Science and Engineering,* 17, 805-824.

Ma, W.-Y., & Chen, K.-J. (2003). Introduction to CKIP Chinese Word Segmentation System for the First International Chinese Word Segmentation Bakeoff. in *Proc. 2nd SIGHAN Workshop on Chinese Language Processing*, 17, 168-171.

Manber, U., & Myers, G. (1993). Suffix arrays: a new method for on-line string searches. *SIAM J. Computing,* 22(5), 935-948.

O'Boyle, P. (1993). *A Study of an N-Gram Language Model for Speech Recognition.* (Ph.D.), Queen's University Belfast.

Qiao, W., Sun, M., & Menzel, W. (2008). Statistical Properties of Overlapping Ambiguities in Chinese Word Segmentation and a Strategy for Their Disambiguation. in *Proc. Text, Speech and Dialogue*, 177-186.

Ratnaparkhi, A. (1996). A Maximum Entropy Model for Part-of-Speech Tagging. in *Proc. Empirical Methods in Natural Language Processing*, 133-142.

Sproat, R., & Emerson, T. (2003). The First International Chinese Word Segmentation Bakeoff. in *Proc. 2nd SIGHAN Workshop on Chinese Language Processing*, 17, 133-143.

Stolcke, A. (2002). SRILM - An Extensible Language Modeling Toolkit. in *Proc. Spoken Language Processing*, 901-904.

Sun, M., Huang, C. N., Lu, F., & Shen, D. Y. (1997). Using Character Bigram for Ambiguity Resolution In Chinese Word Segmentation (In Chinese). *Computer Research and Development,* 34(5), 332-339.

Sun, W., & Xu, J. (2011). Enhancing Chinese Word Segmentation Using Unlabeled Data. in *Proc. Empirical Methods in Natural Language Processing,* 970-979.

Sun, X., & Nan, X. (2010). Chinese base phrases chunking based on latent semi-CRF model. in *Proc. International Conference on Natural Language Processing and Knowledge Engineering*, 1-7.

Sung, C.-L., Yen, H.-C., & Hsu, W.-L. (2008). Compute the Term Contributed Frequency. in

Proc. 8th Int. Conference Intelligent System Design and Application, 2, 325-328.

Tanaka-Ishii, K. (2005). Entropy as an Indicator of Context Boundaries: An Experiment Using a Web Search Engine. in *Proc. Internal Joint Conference on Natural Language Processing*, 93-105.

Tung, C.-H., & Lee, H.-J. (1994). Identification of Unkown Words from Corpus. *Computational Proc. Chinese and Oriental Languages,* 8, 131-145.

Wallach, H. M. (2004). *Conditional Random Fields An Introduction*. (MS-CIS-04-21).

Zhang, H., Huang, H., Zhu, C., & Shi, S. (2010). A pragmatic model for new Chinese word extraction. in *Proc. International Conference on Natural Language Processing and Knowledge Engineering*, 1-8.

Zhang, R., Kikui, G., & Sumita, E. (2006). Subword-based Tagging for Confidence-dependent Chinese Word Segmentation. in *Proc. COLING/ACL*, 961-968.

Zhao, H., Huang, C.-N., Li, M., & Lu, B.-L. (2010). A Unified Character-Based Tagging Framework for Chinese Word Segmentation. *ACM Trans. on Asian Language Information Processing,* 9(2).

Zhao, H., & Kit, C. (2007). Incorporating Global Information into Supervised Learning for Chinese Word Segmentation. in *Proc. 10th PACLIC*, 66-74.

Zhao, H., & Liu, Q. (2010). The CIPS-SIGHAN CLP2010 Chinese Word Segmentation Backoff. in *Proc. 1st CIPS-SIGHAN Joint Conf. on Chinese Language Processing*, 199-209.

Appendix

Table 33. Performance comparison of accuracy on SIGHAN 2003 AS corpus.

Configuration	P	C_P	R	C_R	F
6-tag	**97.18**	±0.003024	97.23	<u>±0.002998</u>	**97.21**
CNG	97.05	±0.003091	97.16	±0.003033	97.11
AVS	97.06	±0.003086	97.23	<u>±0.002998</u>	97.14
TCB	*97.16*	±0.003037	97.18	±0.003024	97.17
TCF	97.15	±0.003042	97.11	±0.003059	97.13
AVS+TCB	97.04	±0.003098	*97.24*	±0.002994	97.14
AVS+TCF	97.07	±0.003081	**97.30**	±0.002958	*97.19*
LRAVS	96.89	±0.003172	97.15	±0.003042	97.02
LRAVS+TCB	97.03	±0.003103	97.20	±0.003011	97.12
LRAVS+TCF	96.94	±0.003147	*97.24*	±0.002994	97.09
2003 Best	95.60	±0.003700	96.60	±0.003300	96.10
2003 Baseline	91.20	±0.005175	91.70	±0.005040	91.50
Our Baseline	91.23	±0.005168	91.74	±0.005029	91.48
2003 Topline	99.30	±0.001523	99.00	±0.001818	99.20
Our Topline	99.30	±0.001526	99.02	±0.001804	99.16

Table 34. Performance comparison of OOV on SIGHAN 2003 AS corpus.

Configuration	R_{OOV}	C_{Roov}	P_{OOV}	C_{Poov}	F_{OOV}
6-tag	*77.13*	±0.052294	75.09	±0.053848	**76.10**
CNG	73.64	±0.054857	75.10	±0.053845	74.36
AVS	70.93	<u>±0.056540</u>	77.22	±0.052227	73.94
TCB	76.74	±0.052603	74.44	±0.054316	*75.57*
TCF	**77.91**	±0.051658	71.02	±0.056486	74.31
AVS+TCB	70.93	<u>±0.056540</u>	*77.54*	±0.051960	74.09
AVS+TCF	70.93	<u>±0.056540</u>	**77.87**	±0.051687	74.24
LRAVS	69.77	±0.057185	76.27	±0.052971	72.87
LRAVS+TCB	69.38	±0.057391	76.50	±0.052797	72.76
LRAVS+TCF	70.16	±0.056975	76.37	±0.052894	73.13
2003 Best	36.40	±0.059910	N/A	N/A	N/A
2003 Baseline	0.00	±0.000000	N/A	N/A	N/A
Our Baseline	0.00	±0.000000	0.00	±0.000000	0.00
2003 Topline	98.80	±0.013558	N/A	N/A	N/A
Our Topline	98.84	±0.013348	97.33	±0.020079	98.08

Table 35. Performance comparison of accuracy on SIGHAN 2003 CityU corpus.

Configuration	***P***	***C_P***	***R***	***C_R***	***F***
6-tag	94.77	±0.002381	94.79	±0.002377	94.78
CNG	**95.24**	±0.002278	**95.48**	±0.002222	**95.36**
AVS	95.13	±0.002302	95.20	±0.002286	95.17
TCB	94.84	±0.002367	94.87	±0.002360	94.85
TCF	94.78	±0.002380	94.77	±0.002382	94.77
AVS+TCB	*95.18*	±0.002291	95.24	±0.002278	95.21
AVS+TCF	95.08	±0.002313	95.19	±0.002288	95.14
LRAVS	95.00	±0.002332	95.21	±0.002284	95.10
LRAVS+TCB	*95.18*	±0.002292	*95.33*	±0.002256	*95.26*
LRAVS+TCF	95.00	±0.002330	95.27	±0.002271	95.14
2003 Best	93.40	±0.002700	94.70	±0.002400	94.00
2003 Baseline	83.00	±0.004018	90.80	±0.003092	86.70
Our Baseline	82.97	±0.004021	90.77	±0.003097	86.69
2003 Topline	99.10	±0.001010	98.60	±0.001257	98.90
Our Topline	99.10	±0.001009	98.62	±0.001249	98.86

Table 36. Performance comparison of OOV on SIGHAN 2003 CityU corpus.

Configuration	***R_{OOV}***	***C_{Roov}***	***P_{OOV}***	***C_{Poov}***	***F_{OOV}***
6-tag	75.80	±0.017149	66.07	±0.018969	70.60
CNG	**77.25**	±0.016796	**73.25**	±0.017735	75.20
AVS	75.16	±0.017311	71.79	±0.018030	73.44
TCB	76.20	±0.017061	66.63	±0.018891	71.10
TCF	*76.28*	±0.017041	66.38	±0.018927	70.99
AVS+TCB	75.44	±0.017245	72.06	±0.017977	*73.71*
AVS+TCF	74.88	±0.017376	71.66	±0.018055	73.23
LRAVS	74.12	±0.017548	72.01	±0.017987	73.05
LRAVS+TCB	74.88	±0.017376	*72.92*	±0.017804	**73.89**
LRAVS+TCF	74.32	±0.017503	72.23	±0.017943	73.26
2003 Best	62.50	±0.019396	N/A	N/A	N/A
2003 Baseline	3.70	±0.007563	N/A	N/A	N/A
Our Baseline	3.69	±0.007555	5.20	±0.008896	4.32
2003 Topline	99.60	±0.002529	N/A	N/A	N/A
Our Topline	99.60	±0.002533	98.65	±0.004626	99.12

Table 37. Performance comparison of accuracy on SIGHAN 2003 PKU corpus.

Configuration	P	C_P	R	C_R	F
6-tag	92.98	±0.003897	93.67	±0.003713	93.32
CNG	94.35	±0.003521	94.70	±0.003417	94.53
AVS	**94.39**	±0.003510	94.70	±0.003417	*94.54*
TCB	93.14	±0.003856	93.69	±0.003709	93.41
TCF	93.43	±0.003780	93.58	±0.003739	93.50
AVS+TCB	**94.43**	±0.003498	**94.84**	±0.003376	94.63
AVS+TCF	94.32	±0.003529	*94.83*	±0.003377	**94.58**
LRAVS	94.18	±0.003572	94.71	±0.003415	94.44
LRAVS+TCB	94.26	±0.003548	94.81	±0.003383	94.53
LRAVS+TCF	94.04	±0.003611	94.62	±0.003441	94.33
2003 Best	94.00	±0.003600	96.20	±0.002900	95.10
2003 Baseline	82.90	±0.005743	90.90	±0.004387	86.70
Our Baseline	82.96	±0.005735	90.87	±0.004392	86.74
2003 Topline	99.60	±0.000963	99.50	±0.001076	99.50
Our Topline	99.63	±0.000930	99.45	±0.001125	99.54

Table 38. Performance comparison of OOV on SIGHAN 2003 PKU corpus.

Configuration	R_{OOV}	C_{Roov}	P_{OOV}	C_{Poov}	F_{OOV}
6-tag	60.22	±0.028389	49.69	±0.029	54.45
CNG	67.70	±0.027122	63.24	±0.027966	65.39
AVS	66.36	±0.027405	64.94	±0.027676	65.64
TCB	61.14	±0.028271	51.49	±0.028988	55.90
TCF	63.58	±0.027910	54.74	±0.028870	58.83
AVS+TCB	68.54	±0.026932	**66.31**	±0.027414	**67.41**
AVS+TCF	*68.29*	±0.026990	65.22	±0.027624	*66.72*
LRAVS	67.12	±0.027249	64.56	±0.027743	65.81
LRAVS+TCB	**68.46**	±0.026952	*64.91*	±0.027681	66.64
LRAVS+TCF	66.95	±0.027284	63.02	±0.028	64.93
2003 Best	61.65	±0.025928	N/A	N/A	N/A
2003 Baseline	5.00	±0.012641	N/A	N/A	N/A
Our Baseline	4.96	±0.012596	5.12	±0.01278	5.04
2003 Topline	100.00	±0.000000	N/A	N/A	N/A
Our Topline	100.00	±0.000000	99.92	±0.001681	99.96

Table 39. Performance comparison of accuracy on SIGHAN 2003 CTB corpus.

Configuration	*P*	C_P	*R*	C_R	*F*
6-tag	87.30	±0.003334	86.83	±0.003385	87.06
CNG	**89.61**	±0.003054	**88.66**	±0.003175	**89.13**
AVS	89.38	±0.003085	88.06	±0.003246	88.71
TCB	87.46	±0.003315	86.86	±0.003382	87.16
TCF	87.18	±0.003347	86.45	±0.003426	86.81
AVS+TCB	89.31	±0.003092	88.08	±0.003244	88.69
AVS+TCF	*89.39*	±0.003082	88.17	±0.003233	*88.78*
LRAVS	89.30	±0.003094	*88.21*	±0.003228	88.75
LRAVS+TCB	89.37	±0.003086	88.09	±0.003243	88.72
LRAVS+TCF	89.31	±0.003093	88.07	±0.003244	88.68
2003 Best	87.50	±0.003300	86.60	±0.003200	88.10
2003 Baseline	66.30	±0.004731	80.00	±0.004004	72.50
Our Baseline	66.33	±0.004730	80.01	±0.004003	72.53
2003 Topline	98.80	±0.001090	98.20	±0.001331	98.50
Our Topline	98.84	±0.001072	98.19	±0.001333	98.52

Table 40. Performance comparison of OOV on SIGHAN 2003 CTB corpus.

Configuration	R_{OOV}	C_{Roov}	P_{OOV}	C_{Poov}	F_{OOV}
6-tag	69.85	±0.010805	62.24	±0.011415	65.83
CNG	**71.79**	±0.010596	**71.31**	±0.010650	**71.55**
AVS	70.59	±0.010728	69.61	±0.010830	70.09
TCB	70.23	±0.010766	62.51	±0.011398	66.14
TCF	69.49	±0.010841	61.91	±0.011434	65.48
AVS+TCB	70.73	±0.010714	70.05	±0.010785	70.39
AVS+TCF	*70.95*	±0.010690	69.80	±0.010811	70.37
LRAVS	70.35	±0.010753	69.98	±0.010793	70.16
LRAVS+TCB	70.58	±0.010730	*70.49*	±0.010739	*70.53*
LRAVS+TCF	70.24	±0.010765	70.05	±0.010785	70.15
2003 Best	70.50	±0.010738	N/A	N/A	N/A
2003 Baseline	6.20	±0.005678	N/A	N/A	N/A
Our Baseline	6.24	±0.005694	8.36	±0.006516	7.14
2003 Topline	99.00	±0.002343	N/A	N/A	N/A
Our Topline	99.02	±0.002324	97.46	±0.003703	98.23

Table 41. Performance comparison of accuracy on SIGHAN 2006 AS corpus.

Configuration	*P*	C_P	*R*	C_R	*F*
6-tag	94.57	±0.001499	95.76	±0.001333	95.16
CNG	95.13	±0.001424	96.16	±0.001271	95.64
AVS	95.25	±0.001407	96.18	±0.001267	95.71
TCB	94.74	±0.001477	95.87	±0.001316	95.30
TCF	94.80	±0.001468	95.85	±0.001319	95.32
AVS+TCB	95.32	±0.001398	96.23	±0.001260	*95.77*
AVS+TCF	*95.33*	±0.001395	96.21	±0.001263	*95.77*
LRAVS	95.24	±0.001408	*96.25*	±0.001256	95.74
LRAVS+TCB	**95.34**	±0.001394	**96.31**	±0.001247	**95.82**
LRAVS+TCF	95.12	±0.001424	95.97	±0.001300	95.55
2006 Best	95.50	±0.001371	96.10	±0.00128	95.80
2006 Baseline	87.00	±0.002224	91.50	±0.001844	89.20
Our Baseline	87.03	±0.002222	91.47	±0.001848	89.19
2006 Topline	98.70	±0.000749	98.00	±0.000926	98.30
Our Topline	98.68	±0.000754	97.98	±0.00093	98.33

Table 42. Performance comparison of OOV on SIGHAN 2006 AS corpus.

Configuration	R_{OOV}	C_{Roov}	P_{OOV}	C_{Poov}	F_{OOV}
6-tag	65.19	±0.015339	60.36	±0.015751	62.68
CNG	67.68	±0.01506	71.51	±0.014533	69.54
AVS	66.90	±0.015152	73.68	±0.01418	70.13
TCB	65.86	±0.015268	61.53	±0.015666	63.62
TCF	67.47	±0.015085	62.17	±0.015616	64.71
AVS+TCB	67.31	±0.015104	*74.18*	±0.014092	70.58
AVS+TCF	67.94	±0.015028	**74.33**	±0.014065	*70.99*
LRAVS	67.73	±0.015054	72.89	±0.014314	70.21
LRAVS+TCB	*68.25*	±0.014989	73.34	±0.014238	70.70
LRAVS+TCF	**69.62**	±0.014808	73.89	±0.014143	**71.69**
2006 Best	70.20	±0.014727	N/A	N/A	N/A
2006 Baseline	3.00	±0.005493	N/A	N/A	N/A
Our Baseline	2.98	±0.005476	5.86	±0.00756	3.95
2006 Topline	99.70	±0.001761	N/A	N/A	N/A
Our Topline	99.64	±0.001936	97.17	±0.005341	98.39

Table 43. Performance comparison of accuracy on SIGHAN 2006 CityU corpus.

Configuration	***P***	C_P	***R***	C_R	***F***
6-tag	96.92	±0.000736	96.88	±0.000741	96.90
CNG	97.26	±0.000696	97.21	±0.000701	97.23
AVS	97.31	±0.000690	**97.34**	±0.000686	97.32
TCB	96.95	±0.000733	96.89	±0.000740	96.92
TCF	96.96	±0.000732	96.90	±0.000739	96.93
AVS+TCB	97.32	±0.000689	97.32	±0.000689	97.32
AVS+TCF	**97.35**	±0.000685	97.32	<u>±0.000688</u>	*97.33*
LRAVS	**97.35**	±0.000684	97.32	<u>±0.000688</u>	**97.34**
LRAVS+TCB	*97.34*	±0.000686	*97.33*	±0.000687	**97.34**
LRAVS+TCF	97.23	±0.000700	97.26	±0.000696	97.24
2006 Best	97.20	±0.000703	97.30	±0.000691	97.20
2006 Baseline	88.20	±0.002134	93.00	±0.001687	90.60
Our Baseline	88.22	±0.001374	93.06	±0.001083	90.57
2006 Topline	98.50	±0.000804	98.20	±0.000879	98.40
Our Topline	98.55	±0.00051	98.19	±0.000568	98.37

Table 44. Performance comparison of OOV on SIGHAN 2006 CityU corpus.

Configuration	R_{OOV}	C_{Roov}	P_{OOV}	C_{Poov}	F_{OOV}
6-tag	78.35	±0.008738	69.60	±0.009759	73.72
CNG	79.66	±0.008540	76.97	±0.008932	78.29
AVS	79.27	±0.008600	78.08	±0.008777	78.67
TCB	78.55	±0.008708	69.97	±0.009725	74.01
TCF	78.94	±0.008651	69.94	±0.009728	74.17
AVS+TCB	79.31	±0.008595	77.93	±0.008798	78.61
AVS+TCF	79.70	±0.008533	78.30	±0.008745	78.99
LRAVS	**79.84**	±0.008512	78.32	±0.008742	*79.07*
LRAVS+TCB	*79.82*	±0.008514	*78.57*	±0.008706	**79.19**
LRAVS+TCF	79.48	±0.008568	**77.93**	±0.008798	78.70
2006 Best	78.70	±0.008686	N/A	N/A	N/A
2006 Baseline	0.90	±0.002004	N/A	N/A	N/A
Our Baseline	0.95	±0.002053	2.47	±0.003293	1.37
2006 Topline	99.30	±0.001769	N/A	N/A	N/A
Our Topline	99.31	±0.001752	95.22	±0.004526	97.22

Table 45. Performance comparison of accuracy on SIGHAN 2006 PKU corpus.

Configuration	***P***	C_P	***R***	C_R	***F***
6-tag	92.51	±0.001338	93.79	±0.001227	93.14
CNG	*93.54*	±0.001250	94.38	±0.001170	93.96
AVS	93.43	±0.001259	*94.41*	±0.001167	93.92
TCB	92.54	±0.001335	93.75	±0.001230	93.14
TCF	92.54	±0.001335	93.72	±0.001233	93.13
AVS+TCB	93.43	±0.001259	94.37	±0.001171	93.90
AVS+TCF	93.42	±0.001260	94.32	±0.001176	93.87
LRAVS	**93.59**	±0.001245	**94.44**	±0.001164	**94.01**
LRAVS+TCB	*93.54*	±0.001250	94.40	±0.001168	*93.97*
LRAVS+TCF	93.40	±0.001262	94.30	±0.001178	93.85
2006 Best	92.60	±0.001330	94.00	±0.001207	93.30
2006 Baseline	79.00	±0.002694	86.90	±0.002231	82.80
Our Baseline	79.04	±0.002069	86.87	±0.001717	82.77
2006 Topline	97.60	±0.001012	96.10	±0.00128	96.80
Our Topline	97.59	±0.000779	96.08	±0.000986	96.83

Table 46. Performance comparison of OOV on SIGHAN 2006 PKU corpus.

Configuration	R_{OOV}	C_{Roov}	P_{OOV}	C_{Poov}	F_{OOV}
6-tag	70.51	±0.007834	70.70	±0.00782	70.60
CNG	74.97	±0.007442	**78.04**	±0.007112	*76.47*
AVS	74.57	±0.007481	77.78	±0.007142	76.14
TCB	70.73	±0.007817	70.90	±0.007804	70.81
TCF	70.96	±0.007799	70.19	±0.007859	70.57
AVS+TCB	74.51	±0.007487	77.68	±0.007154	76.06
AVS+TCF	74.14	±0.007522	77.13	±0.007215	75.61
LRAVS	**75.28**	±0.007411	*77.93*	±0.007125	**76.58**
LRAVS+TCB	*75.13*	±0.007427	77.68	±0.007154	76.38
LRAVS+TCF	74.53	±0.007486	77.03	±0.007226	75.76
2006 Best	70.70	±0.007819	N/A	N/A	N/A
2006 Baseline	1.10	±0.001792	N/A	N/A	N/A
Our Baseline	1.11	±0.001803	3.42	±0.003124	1.68
2006 Topline	98.90	±0.001792	N/A	N/A	N/A
Our Topline	98.94	±0.001762	92.56	±0.004507	95.65

Table 47. Performance comparison of accuracy on SIGHAN 2006 MSR corpus.

Configuration	***P***	C_P	***R***	C_R	***F***
6-tag	**96.44**	±0.001169	95.71	±0.001279	*96.08*
CNG	96.19	±0.001208	95.58	±0.001298	95.88
AVS	96.30	±0.001191	95.84	±0.001260	96.07
TCB	*96.40*	<u>±0.001177</u>	95.74	±0.001275	96.07
TCF	96.35	±0.001183	95.69	±0.001283	96.02
AVS+TC	96.38	±0.001180	*95.87*	±0.001256	**96.12**
AVS+TCF	96.40	<u>±0.001177</u>	95.73	±0.001276	96.06
LRAVS	96.22	±0.001203	95.85	<u>±0.001259</u>	96.04
LRAVS+TCB	96.24	±0.001200	**95.88**	±0.001255	96.06
LRAVS+TC	96.16	±0.001213	95.85	<u>±0.001259</u>	96.01
2006 Best	96.10	±0.001222	96.40	±0.001176	96.30
2006 Baseline	90.00	±0.001984	94.90	±0.001455	92.40
Our Baseline	90.03	±0.001891	94.94	±0.001384	92.42
2006 Topline	99.30	±0.000551	99.10	±0.000625	99.20
Our Topline	99.28	±0.000534	99.08	±0.000603	99.18

Table 48. Performance comparison of OOV on SIGHAN 2006 MSR corpus.

Configuration	R_{OOV}	C_{Roov}	P_{OOV}	C_{Poov}	F_{OOV}
6-tag	**66.57**	±0.016171	55.62	±0.017031	60.60
CNG	61.60	±0.016672	58.23	±0.016906	59.87
AVS	64.60	±0.016393	*60.83*	±0.016733	*62.66*
TCB	66.86	±0.016136	55.95	±0.017018	60.92
TCF	*66.42*	±0.016189	54.67	±0.017065	59.97
AVS+TCB	64.72	±0.016380	**61.19**	±0.016705	**62.91**
AVS+TCF	62.78	±0.016571	59.86	±0.016803	61.28
LRAVS	63.92	±0.016462	59.94	±0.016797	61.87
LRAVS+TCB	62.87	±0.016563	60.40	±0.016765	61.61
LRAVS+TCF	62.96	±0.016554	59.56	±0.016824	61.21
2006 Best	61.20	±0.016704	N/A	N/A	N/A
2006 Baseline	2.20	±0.005028	N/A	N/A	N/A
Our Baseline	2.17	±0.004999	11.13	±0.010780	3.64
2006 Topline	99.90	±0.001083	N/A	N/A	N/A
Our Topline	99.85	±0.001313	99.24	±0.002975	99.55

Table 49. Performance comparison of accuracy on SIGHAN 2008 AS corpus.

Configuration	***P***	***C_P***	***R***	***C_R***	***F***
6-tag	82.36	±0.002526	83.25	±0.002475	82.80
CNG	83.00	±0.002490	83.77	±0.002444	83.38
AVS	**83.09**	±0.002484	**83.83**	±0.002440	**83.46**
TCB	82.28	±0.002531	83.20	±0.002478	82.74
TCF	82.54	±0.002516	83.37	±0.002468	82.95
AVS+TCB	82.83	±0.002499	83.62	±0.002453	83.23
AVS+TCF	82.97	±0.002492	*83.80*	±0.002442	83.38
LRAVS	82.98	±0.002491	83.78	±0.002443	83.38
LRAVS+TCB	*83.03*	±0.002488	*83.80*	±0.002442	*83.42*
LRAVS+TCF	82.86	±0.002498	83.72	±0.002447	83.29
2008 Best	94.40	±0.001527	95.01	±0.001445	94.70
2008 Baseline	82.32	±0.002534	89.78	±0.002012	85.69
Our Baseline	80.99	±0.002601	89.29	±0.002050	84.93
2008 Topline	98.80	±0.000723	98.23	±0.000876	98.52
Our Topline	98.53	±0.000796	97.84	±0.000963	98.19

Table 50. Performance comparison of OOV on SIGHAN 2008 AS corpus.

Configuration	***R_{OOV}***	***C_{Roov}***	***P_{OOV}***	***C_{Poov}***	***F_{OOV}***
6-tag	62.85	±0.011258	55.49	±0.011580	58.94
CNG	63.78	±0.011199	**63.07**	±0.011245	**63.42**
AVS	63.38	±0.011225	*62.50*	±0.011280	*62.94*
TCB	62.42	±0.011285	55.61	±0.011576	58.82
TCF	63.61	±0.011210	56.22	±0.011560	59.69
AVS+TCB	*62.89*	±0.011256	60.88	±0.011371	61.87
AVS+TCF	63.60	±0.011211	61.80	±0.011321	62.68
LRAVS	63.30	±0.01123	62.19	±0.011298	62.74
LRAVS+TCB	63.34	±0.011228	62.27	±0.011294	62.80
LRAVS+TCF	**62.81**	±0.011261	61.71	±0.011326	62.25
2008 Best	74.04	±0.010215	76.49	±0.009881	75.24
2008 Baseline	2.08	±0.003325	6.78	±0.005858	3.19
Our Baseline	4.03	±0.004583	8.08	±0.006348	5.38
2008 Topline	99.32	±0.001915	96.42	±0.004329	97.84
Our Topline	99.40	±0.001795	96.41	±0.004337	97.88

Table 51. Performance comparison of accuracy on SIGHAN 2008 CTB corpus.

Configuration	P	C_P	R	C_R	F
6-tag	95.56	±0.001682	95.51	±0.001691	95.54
CNG	95.54	±0.001686	95.53	±0.001688	95.54
AVS	95.68	±0.001660	95.71	±0.001655	95.70
TCB	95.54	±0.001687	95.54	±0.001687	95.54
TCF	95.52	±0.001689	95.54	±0.001685	95.53
AVS+TCB	95.58	±0.001680	95.61	±0.001674	95.59
AVS+TCF	**95.98**	±0.001605	**95.96**	±0.001609	**95.97**
LRAVS	95.55	±0.001684	95.56	±0.001682	95.56
LRAVS+TCB	95.53	±0.001687	95.56	±0.001683	95.55
LRAVS+TCF	*95.69*	±0.001658	*95.72*	±0.001653	*95.71*
2008 Best	95.96	±0.001386	95.83	±0.001408	95.89
2008 Baseline	84.27	±0.002563	88.64	±0.002234	86.40
Our Baseline	84.05	±0.002991	88.86	±0.002570	86.39
2008 Topline	98.25	±0.000923	97.10	±0.001181	97.67
Our Topline	98.42	±0.001018	97.55	±0.001264	97.98

Table 52. Performance comparison of OOV on SIGHAN 2008 CTB corpus.

Configuration	R_{OOV}	C_{Roov}	P_{OOV}	C_{Poov}	F_{OOV}
6-tag	77.63	±0.014611	70.56	±0.01598	73.92
CNG	76.28	±0.014915	74.58	±0.015266	75.42
AVS	77.69	±0.014597	*75.87*	±0.015001	76.77
TCB	77.69	±0.014597	70.71	±0.015955	74.04
TCF	77.69	±0.014597	71.03	±0.015904	74.21
AVS+TCB	77.20	±0.014710	75.14	±0.015153	76.16
AVS+TCF	**78.86**	±0.014316	77.43	±0.014657	**78.14**
LRAVS	77.11	±0.014731	75.21	±0.015139	76.15
LRAVS+TCB	77.04	±0.014745	75.19	±0.015142	76.11
LRAVS+TCF	*78.15*	±0.014488	**76.50**	±0.014865	*77.32*
2008 Best	77.30	±0.014687	77.61	±0.014615	77.45
2008 Baseline	2.83	±0.005814	7.69	±0.009341	4.14
Our Baseline	1.54	±0.004313	3.34	±0.006298	2.10
2008 Topline	99.20	±0.003123	97.07	±0.005913	98.12
Our Topline	99.54	±0.002375	97.56	±0.005409	98.54

Table 53. Performance comparison of accuracy on SIGHAN 2008 NCC corpus.

Configuration	***P***	***C_P***	***R***	***C_R***	***F***
6-tag	93.55	±0.001259	93.09	±0.001300	93.32
CNG	**93.84**	±0.001232	**93.90**	±0.001226	**93.87**
AVS	93.69	±0.001246	93.72	±0.001243	93.71
TCB	93.60	±0.001254	93.14	±0.001295	93.37
TCF	93.46	±0.001267	93.11	±0.001298	93.28
AVS+TCB	*93.79*	±0.001237	93.78	±0.001238	93.78
AVS+TCF	93.75	<u>±0.001240</u>	93.81	±0.001235	93.78
LRAVS	93.76	<u>±0.001240</u>	93.83	±0.001233	93.79
LRAVS+TCB	93.78	±0.001238	*93.86*	±0.001230	*93.82*
LRAVS+TCF	93.73	±0.001242	93.81	±0.001235	93.77
2008 Best	94.07	±0.001210	94.02	±0.001214	94.05
2008 Baseline	87.16	±0.001714	92.00	±0.001390	89.51
Our Baseline	87.18	±0.001713	91.99	±0.001391	89.52
2008 Topline	98.17	±0.000687	97.35	±0.000823	97.76
Our Topline	98.17	±0.000687	97.35	±0.000823	97.76

Table 54. Performance comparison of OOV on SIGHAN 2008 NCC corpus.

Configuration	***R_{OOV}***	***C_{Roov}***	***P_{OOV}***	***C_{Poov}***	***F_{OOV}***
6-tag	*62.32*	±0.0114	51.51	±0.011758	56.40
CNG	60.43	±0.011504	*59.39*	±0.011554	*59.90*
AVS	59.76	±0.011537	57.86	±0.011617	58.79
TCB	**63.28**	±0.011341	52.30	±0.011751	57.27
TCF	62.86	±0.011367	52.73	±0.011745	57.35
AVS+TCB	60.30	±0.011511	58.43	±0.011595	59.35
AVS+TCF	59.91	±0.01153	58.64	±0.011586	59.27
LRAVS	60.08	±0.011522	59.31	±0.011557	59.69
LRAVS+TCB	60.32	±0.01151	**59.49**	±0.011549	**59.90**
LRAVS+TCF	60.23	±0.011514	59.21	±0.011562	59.72
2008 Best	61.79	±0.011431	59.84	±0.011533	60.80
2008 Baseline	2.73	±0.003834	18.58	±0.00915	4.76
Our Baseline	2.73	±0.003831	18.58	±0.009151	4.75
2008 Topline	99.33	±0.001919	92.03	±0.006372	95.54
Our Topline	99.34	±0.001911	92.04	±0.006368	95.55