Unsupervised Overlapping Feature Selection for Conditional Random Fields Learning in Chinese Word Segmentation

Ting-hao Yang
Institute of Information Science
Academia Sinica
tinghaoyang@iis.sinica.edu.tw

Tian-Jian Jiang
Department of Computer Science
National Tsing-Hua University
tmjiang@iis.sinica.edu.tw

Chan-hung Kuo
Institute of Information Science
Academia Sinica
laybow@iis.sinica.edu.tw

Richard Tzong-han Tsai
Department of Computer Science & Engineering
Yuan Ze University
thtsai@saturn.yzu.edu.tw

Wen-lian Hsu
Institute of Information Science
Academia Sinica
hsu@iis.sinica.edu.tw

Abstract

This work represents several unsupervised feature selections based on frequent strings that help improve conditional random fields (CRF) model for Chinese word segmentation (CWS). These features include character-based N-gram (CNG), Accessor Variety based string (AVS), and Term Contributed Frequency (TCF) with a specific manner of boundary overlapping. For the experiment, the baseline is the *6-tag*, a state-of-the-art labeling scheme of CRF-based CWS; and the data set is acquired from SIGHAN CWS bakeoff 2005. The experiment results show that all of those features improve our system's F_1 measure (F) and Recall of Out-of-Vocabulary (R_{OOV}). In particular, the feature collections which contain AVS feature outperform other types of features in terms of F, whereas the feature collections containing TCB/TCF information has better R_{OOV}.

Keywords: Word Segmentation, Unsupervised Feature Selection, Conditional Random Fields

1. Introduction

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), while a little difficult for Asian languages such as Chinese and Japanese.

1.1 Background

Chinese word segmentation (CWS) is an essential pre-work for Chinese text processing applications and it 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, conducted five word segmentation bakeoffs (Sproat and Emerson, 2003; Emerson, 2005; Levow, 2006; Jin and Chen, 2007; Zhao and Liu, 2010). After years of intensive researches, CWS has achieved high precision, but the issue of out-of-vocabulary word handling still remains.

1.2 The State of the Art of CWS

Traditional approaches for CWS adopted dictionary and rules to segment unlabeled texts (c.f. Ma and Chen, 2003). In recent years, the mainstream is to use statistical machine learning models, especially the Conditional Random Fields (CRF) (Lafferty *et al*, 2001), which shows a moderate performance for sequential labeling problem and achieves competitive results with character position based methods (Zhao *et al*., 2010).

1.3 Unsupervised CRF Feature Selections for CWS

For incorporating unsupervised feature selections into character position based CRF for CWS, Zhao and Kit (2006; 2007) tried strings based on Accessor Variety (AV), which was developed by Feng *et al.* (2004), and co-occurrence strings (COS). Jiang *et al*. (2010) applied a feature similar to COS, called Term Contributed Boundary (TCB). Tsai (2010) employ statistical association measures non-parametrically through a natural but novel feature representation scheme. Those unsupervised feature selection are based on frequent strings extracted automatically from unlabeled corpora. They are suitable for closed training evaluation that any external resource or extra information is not allowed. Without proper knowledge, the closed training evaluation of word segmentation can be difficult with Out-of-Vocabulary (OOV) words, where frequent strings collected from the test data may help.

According to Zhao and Kit (2008), AV-based string (AVS) is one of the most effective unsupervised feature selection for CWS by character position based CRF. This motivates us to seek for explanations for AVS's success. We suspect that AVS is designed to keep overlapping strings but COS/TCB is usually selected with its longest-first nature before integrated into CRF. Hence, we conduct a series of experiments to examine this hypothesis.

The remainder of the article is organized as follows. Section 2 briefly introduces CRF. Common unsupervised feature selections based on the concept of frequent strings are explained in Section 3. Section 4 discusses related works. Section 5 describes the design of labeling scheme, feature templates and 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 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 sequential labeling problem (Lafferty *et al.*, 2001). Given an input sequence (or observation sequence) $X = x_1 \dots x_T$ and label sequence $Y = y_1 \dots y_T$, a conditional probability of linear-chain CRF with parameters $\Lambda = \{\lambda_1, ..., \lambda_n\}$ can be defined as:

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

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 $y_{t-1} \to y_t$, and the entire observation sequence *X* centered at the current position *t*.

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

$$y^* = \arg\max_{Y} P_\Lambda(Y \mid X) \quad (2)$$

.

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

3. 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 boundary such as

Chinese, character-based N-gram (CNG) is usually insufficient. For example, consider the following sample texts in Chinese

- “自然科學的重要性” (the importance of natural science);
- “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 has been researched in different languages. The distortion of phrase boundaries and frequencies was first observed in the Vodis Corpus when the word-based bigram “RAIL ENQUIRIES” and word-based trigram “BRITISH RAIL ENQUIRIES” were estimated and reported (O’Boyle, 1993; 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” precede it. However, when “RAIL” is preceded by words other than “BRITISH,” “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 and Yu (2001) reported a similar problem and its corresponding solution in the sense of reduced N-gram of Chinese character. By excluding N-grams with their numbers of appearance that fully depend on other super-sequences, “然科” and “學的” from the sample texts in the previous sub-section are not candidates of string anymore. 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 we use the program of TCB/TCF for experiment within this study, 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 how likely a string is a Chinese word. Another measurement called Boundary Entropy or Branching Entropy (BE) exists in some works (Tung and Lee, 1994; Chang and Su, 1997; Cohen and Adams, 2001;

Cohen *et al.*, 2002; Huang and Powers, 2003; Tanaka-Ishii, 2005; Jin and Tanaka-Ishii, 2006; Cohen *et al.*, 2006). The basic idea behind those measurements is closely related to one particular perspective of N-gram and information theory as 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 then decided to adopt AVS as unsupervised feature selection for CRF-based CWS. We follow 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 Equation (3), L_{av}(s) and R_{av}(s) are defined as the number of distinct preceding and succeeding characters, respectively, except if the adjacent character has been absent because of sentence boundary, then the pseudo-character of sentence beginning or sentence ending will be accumulated indistinctly. 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 selection and for the sake of simplicity, those additional rules are dropped in this study.

4. Other Related Works

This section briefly describes the following three related works.

4.1 Frequent String Extraction Algorithm

Besides papers of TCB/TCF extraction (Sung *et al.*, 2008), Chinese frequent strings (Lin *et al.*, 2001) and reduced N-gram (Ha *et al.*, 2005) that are mentioned earlier, the article about a linear algorithm for Frequency of Substring Reduction (Lü and Zhang, 2005) also falls into this category. Most of them focused on the computational complexity of algorithms. More general algorithms for frequent string extraction are usually suffix array (Manber and Myers, 1993) and PAT-tree (Chien, 1997).

4.2 Unsupervised Word Segmentation Method

Zhao and Kit (2008) have explored several unsupervised strategies with their unified goodness measurement of logarithm ranking, including Frequency of Substring with Reduction, Description Length Gain (Kit and Wilks, 1999; Kit, 2000), Accessor Variety and Boundary/Branching Entropy. Unlike the technique described in this paper for incorporating

unsupervised feature selections into supervised CRF learning, those methods usually filter out word-alike candidates by their own scoring mechanism directly.

4.3 Overlapping Ambiguity Resolution

Subword-based tagging (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).

5. CRF Labeling Scheme

5.1 Character Position Based Labels

In this study, the *6-tag* approach (Zhao *et al*., 2010) is adopted as our formulation, which achieves a very competitive performance recently, and is one of the most fine-grained character-position-based labeling schemes. According to Zhao *et al.* (2010), since less than 1% Chinese words are longer than five characters in most corpora from SIGHAN CWS bakeoffs 2003, 2005, 2006 and 2007, the coverage of *6-tag* approach should be good enough. This configuration of CRF without any additional unsupervised feature selection is also the control group of the experiment. Table 1 provides a sample of labeled training data.

Table 1. A Sample of the 6-tag Labels

Character	**Label**
反	B_1
而	E
會	S
欲	B_1
速	B_2
則	B_3
不	M
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_1 stands for the beginning character of a word, while B_2 and B_3 represent for 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 the middle characters between B_3 and *E* is *M*. Finally, the tag *S* is reserved for single-character words specifically.

5.2 Feature Templates

Feature instances are generated from templates based on the work of Ratnaparkhi (1996). Table 2 explains their abilities.

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

C_{-1}, C_0 and C_1 stand for the input tokens bound to the prediction label at current position individually. For example in Table 1, if the current position is at the label *M*, 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 *M* to "則不," "不達" and "則達," respectively. According to Zhao *et al*. (2010), the context window size in three tokens is effective to catch parameters of 6-tag approach for most strings not longer than five characters. Our pilot test for this case, however, shows that context window size in two tokens would be sufficient without significant performance decreasing. We also intentionally avoid using feature templates that determine character types like alphabet, digit, punctuation, date/time and other non-Chinese characters, to stay with the strict protocol of closed training and unsupervised learning.

Unsupervised feature selections that will be introduced in the next sub-section are of course generated by the same template, except the binding target moves column by column as listed in tables of the next sub-section.

By default, CRF++ generates features not only for the prediction label at the current position, but also for combinations of the prediction label at both the previous and the current position, which should not be confused with the context window size mentioned above.

5.3 A Unified Feature Representation for CNG, AVS and TCB

To compare different types of overlapping strings as unsupervised feature selection systematically, we extend the work of Zhao and Kit (2008) into a unified representation of features. The representation accommodates both character position of a string and this string's likelihood ranked in logarithm. Formally, the ranking function for a string *s* with a score *x* counted by either CNG, AVS or TCB is defined as

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

.

The logarithm ranking mechanism in Equation (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 are then concatenated as feature tokens. To give the reader a clearer picture about what feature tokens look like, a sample representation for CNG, AVS or TCB is demonstrated and explained by Table 3.

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 $3B_1$ 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.

Table 3. A Sample of the Unified Feature Representation for Overlapping String

Input	**Unsupervised Feature Selection**					**Label**
	1 char	2 char	3 char	4 char	5 char	
反	*5S*	$3B_1$	$4B_1$	$0B_1$	$0B_1$	B_1
而	*6S*	*3E*	$4B_2$	$0B_2$	$0B_2$	*E*
會	*6S*	*0E*	*4E*	$0B_3$	$0B_3$	*S*
欲	*4S*	*0E*	*0E*	*0E*	*0M*	B_1
速	*4S*	*0E*	*0E*	*0E*	*0E*	B_2
則	*6S*	$3B_1$	*0E*	*0E*	*0E*	B_3
不	*7S*	*3E*	*0E*	*0E*	*0E*	*M*
達	*5S*	*3E*	*0E*	*0E*	*0E*	*E*

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, we simply inherit the above rules by Zhao and Kit (2008) for the sake of compatibility. In fact, we have done a pilot test with a more complicated representation like $3E\text{-}0B_1$ for "而" and $3E\text{-}3B_1$ for "不" to keep the overlapping information within each column, but the test result shows no significant differences in terms of performance. Since the statistics of the pilot test could be considerably redundant, they are omitted in this paper.

To make an informative comparison, we also apply the original version of non-overlapping COS/TCB feature that is selected by forward maximum matching algorithm and without ranks (Zhao and Kit, 2007; Jiang *et al*., 2010). The following table illustrates a sample representation of features for this case.

Table 4. A Sample of the Representation for Non-overlapping COS/TCB Strings

Input	Original COS/TCB Feature	Label
反	B_1	B_1
而	B_2	E
會	E	S
欲	*-1*	B_1
速	*-1*	B_2
則	*-1*	B_3
不	*-1*	M
達	*-1*	E

Note that there are several features encoded as -1 individually to represent that the desired string is unseen. For the family of reduced N-grams, such as COS or TCB, it means that either the string is always occupied by other super-strings or simply does not appear more than once.

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

6. Experiment

The version 0.54 of the CRF++ employs L-BFGS optimization and the tunable hyper-parameter, i.e. the Gaussian prior, set to 100 throughout the whole experiment.

6.1 Data Set

The corpora used for experiment are from SIGHAN CWS bakeoff 2005. It comes with four different standards including Academia Sinica (AS), City University of Hong Kong (CityU), Microsoft Research (MSR) and Peking University (PKU).

6.2 Unsupervised Feature Collection

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

6.3 Evaluation Metric

The evaluation metric of CWS task is adopted from SIGHAN bakeoffs, including test Precision (P), test Recall (R), F1 measure score (F) and test Recall of Out-of-Vocabulary (R_{OOV}). Their formulae are list as follows.

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

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

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

$$R_{OOV} = \frac{\text{the number of OOV words that are correctly segmented}}{\text{the number of OOV words in the gold standard}} \times 100\% \quad (8)$$

To estimate the differences of performance between configurations of CWS experiment, this work uses the confidence level, which has been applied since SIGHAN CWS bakeoff 2003 (Sproat *et al*., 2003), that assume 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, and R_{OOV} appear in Table 5 under the column C_P, C_R, and C_{Roov}, respectively, are calculated at the 95% confidence interval with the formula $\pm 2\sqrt{}([X(1-X)]/N)$. Two configurations of CWS experiment are then considered to be statistically different at a 95% confidence level if one of their C_P, C_R, or C_{Roov} is different.

6.4 Experiment 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 those non-Chinese characters that may violate rules of closed training evaluation, a 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 manually, since there are no urgent expectations of correct segmentation on non-Chinese text. Table 5 lists the statistics after the post-processing. Further discussions are mainly based on this post-processed result without loss of generality. Numbers in bold face and bold-italic style indicate the best and the second-best results of a certain evaluation metric, respectively.

Statistics show clear trends that the feature collections which contain AVS outperforms other types of unsupervised feature selections on *F*, and the feature collections containing

TCB/TCF information usually has better R_{OOV}.

Table 5. Performance Comparison After Post-processing

Corpus	Feature	C_P	C_R	F	R_{OOV}	C_{Roov}
AS	6-tag	±0.00125	±0.00114	.955	.726	±0.01164
	CNG	±0.00124	±0.00113	.955	.730	±0.01159
	AVS	±0.00120	±0.00109	**.958**	.738	±0.01147
	TCF	±0.00126	±0.00117	.953	**.760**	±0.01114
	TCB	±0.00123	±0.00113	***.956***	.740	±0.01145
	AVS+TCF	±0.00123	±0.00113	***.956***	***.751***	±0.01128
	AVS+TCB	±0.00120	±0.00109	**.958**	.739	±0.01147
CityU	6-tag	±0.00219	±0.00221	.948	.738	±0.01536
	CNG	±0.00207	±0.00215	.953	.760	±0.01493
	AVS	±0.00199	±0.00203	***.957***	.766	±0.01480
	TCF	±0.00208	±0.00214	.953	.767	±0.01478
	TCB	±0.00209	±0.00214	.953	***.770***	±0.01470
	AVS+TCF	±0.00197	±0.00200	**.959**	**.777**	±0.01455
	AVS+TCB	±0.00207	±0.00213	.953	.771	±0.01469
MSR	6-tag	±0.00100	±0.00105	.971	.776	±0.01405
	CNG	±0.00100	±0.00104	***.972***	.784	±0.01387
	AVS	±0.00099	±0.00099	**.973**	.764	±0.01432
	TCF	±0.00099	±0.00104	***.972***	.786	±0.01384
	TCB	±0.00099	±0.00104	***.972***	***.787***	±0.01381
	AVS+TCF	±0.00107	±0.00114	.967	**.793**	±0.01367
	AVS+TCB	±0.00101	±0.00102	***.972***	.769	±0.01422
PKU	6-tag	±0.00139	±0.00159	.939	.680	±0.01140
	CNG	±0.00139	±0.00160	.938	.671	±0.01149
	AVS	±0.00132	±0.00146	**.947**	***.740***	±0.01072
	TCF	±0.00138	±0.00155	***.941***	.701	±0.01119
	TCB	±0.00139	±0.00159	.939	.688	±0.01133
	AVS+TCF	±0.00137	±0.00155	***.941***	.709	±0.01110
	AVS+TCB	±0.00132	±0.00147	**.947**	**.743**	±0.01067

It has been observed that using any of the unsupervised feature selections could create short patterns for CRF learner, which might break more English words than using the *6-tag* approach solely. 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 selection without rank and overlapping information, TCB successfully recognizes "依

靠 / 单位 / 的 / 纽带 / 来 / 维持," while the *6-tag* approach see this phrase incorrectly as "依靠 / 单位 / 的 / 纽 / 带来 / 维持." TCB also saves more factoids, such as "一二九·九 / 左右" (around 129.9) 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-alike candidate should be an important factor for unsupervised feature selection injected CRF learner. Relatively speaking, CNG probably brings in too much noise. Non-overlapping COS/TCB seems to be a moderate choice with a lower training cost of CRF than those of other overlapping features. This confirms our hypothesis at the end of Section 1.3 that, including overlapping information as an unsupervised feature selection may help improving CWS performance of supervised labeling scheme of CRF.

7. Conclusion and Future Works

This paper provides a study about CRF-based CWS integrated with unsupervised and overlapping feature selections. The experiment results show that the feature collections which contain AVS obtains better performance in terms of F_1 measure score, and TCB/TCF enhances the 6-tag approach on the Recall of Out-of-Vocabulary. In the future, we will search for a hybrid method that utilizes information both inside and outside Chinese words simultaneously.

Acknowledgement

This research was supported in part by the National Science Council under grant NSC 100-2631-S-001-001, and the research center for Humanities and Social Sciences under grant IIS-50-23. The authors would like to thank anonymous reviewers for their constructive criticisms.

References

[1] Peter O'Boyle, "A Study of an N-Gram Language Model for Speech Recognition", PhD Thesis, Queen's University Belfast , 1993.

[2] Cheng-Huang Tung and His-Jian Lee, "Identification of Unknown Words from Corpus", *Computational Proceedings of Chinese and Oriental Languages*, vol.8, pp.131-145, 1994.

[3] Jing-Shin Chang and Keh-Yih Su, "An Unsupervised Iterative Method for Chinese New Lexicon Extraction", *Computational Linguistics and Chinese Language Processing*, vol.2, no.2, pp.97-148, 1997.

[4] Maosong Sun, Changning Huang, Benjamin K.Tsou, "Using Character Bigram for Ambiguity Resolution In Chinese Word Segmentation (In Chinese)", *Computer*

Research and Development, vol.34, no.5, pp.332-339, 1997.

[5] Lee-Feng Chien, "PAT-tree-based Keyword Extraction for Chinese Information Retrieval", in *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.50-58, 1997.

[6] John Lafferty, Andrew McCallum, Fernando Pereira, "Conditional Random Fields Probabilistic Models for Segmenting and Labeling Sequence Data", in *Proceedings of International Conference on Machine Learning*, pp.591-598, 2001.

[7] Yih-Jeng Lin, Ming-Shing Yu, "Extracting Chinese Frequent Strings without a Dictionary from a Chinese Corpus and its Applications", *Journal of Information Science and Engineering 17*, pp.805-824, 2001.

[8] Andreas Stolcke, "*SRILM - An Extensible Language Modeling Toolkit"*, in *the Proceedings of Spoken Language Processing*, pp.901-904, 2002.

[9] Mu Li, Jianfeng Gao, Chang-Ning Huang, "Unsupervised Training for Overlapping Ambiguity Resolution in Chinese Word Segmentation", in *the Proceedings of The Second SIGHAN Workshop on Chinese Language Processing*, pp.1-7, 2003.

[10] Richard Sproat, Thomas Emerson, "The First International Chinese Word Segmentation Bakeoff", in *the Proceedings of The Second SIGHAN Workshop on Chinese Language Processing, Sapporo, July 11-12*, 2003, pp.113-143.

[11] Wei-Yun Ma, Keh-Jiann Chen, "Introduction to CKIP Chinese Word Segmentation System for the First International Chinese Word Segmentation Bakeoff", in *the Proceedings of Second SIGHAN Workshop on Chinese Language Processing*, pp.168-171, 2003.

[12] Haodi Feng, Kang Chen, Xiaotie Deng, and Wiemin Zheng, "*Accessor Variety Criteria for Chinese Word Extraction"*, *Computational Linguistics*, vol.30, no.1, pp.75-93, 2004.

[13] Hanna M. Wallach, "Conditional Random Fields An Introduction", Department of Computer and Information Science, University of Pennsylvania, Tech. Rep. MS-CIS-04-21, 2004.

[14] Le Quan Ha, Rowan Seymour, Philip Hanna and Francis J. Smith, "Reduced N-Grams for Chinese Evaluation", *Computational Linguistics and Chinese Language Processing*, vol.10, no.1, pp.19-34, 2005.

[15] Thomas Emerson, "The Second International Chinese Word Segmentation Bakeoff", in *the Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pp.123-133, 2005.

[16] Xueqiang Lü, Le Zhang, "Statistical Substring Reduction in Linear Time", in *the Proceedings of the 1st International Joint Conference on Natural Language Processing*, pp.320-327, 2005.

[17] Gina-Anne Levow, "The Third International Chinese Language Processing Bakeoff Word Segmentation and Named Entity Recognition", in *the Proceedings of The Fifth*

SIGHAN Workshop on Chinese Language Processing, pp.108-117, 2006.

[18] Ruiqiang Zhang, Genichiro Kikui, Eiichiro Sumita, "Subword-based Tagging for Confidence-dependent Chinese Word Segmentation", in *the Proceedings of COLING/ACL*, pp.961-968, 2006.

[19] Guangjin Jin, Xiao Chen, "The Fourth International Chinese Language Processing Bakeoff : Chinese Word Segmentation, Named Entity Recognition and Chinese POS Tagging", in *the Proceedings of The Sixth SIGHAN Workshop on Chinese Language Processing*, pp.69-81, 2007.

[20] Wei Qiao, Maosong Sun, Wolfgang Menzel, "Statistical Properties of Overlapping Ambiguities in Chinese Word Segmentation and a Strategy for Their Disambiguation", in *the Proceedings of Text, Speech and Dialogue*, pp.177-186, 2008.

[21] Cheng-Lung Sung, Hsu-Chun Yen, Wen-Lian Hsu, "Compute the Term Contributed Frequency", in *the Proceedings of The Eighth International Conference on Intelligent Systems Design and Applications*, pp.325–328, 2008.

[22] Hai Zhao, Chunyu Kit, "Exploiting Unlabeled Text with Different Unsupervised Segmentation Criteria for Chinese Word Segmentation", in *the Proceedings of The Ninth International Conference on Intelligent Text Processing and Computational Linguistics*, pp.17-23, 2008.

[23] Hai Zhao, Chang-Ning Huang, Mu Li, Lu, Bao-Liang Lu, "A Unified Character-Based Tagging Framework for Chinese Word Segmentation", *ACM Transactions on Asian Language Information Processing*, vol.9, no.2, 2010.

[24] Hai Zhao, Qun Liu, "The CIPS-SIGHAN CLP2010 Chinese Word Segmentation Backoff", in *the Proceedings of The First CIPS-SIGHAN Joint Conference on Chinese Language Processing*, pp.199-209, 2010.

[25] Richard Tzong-Han Tsai, "Chinese text segmentation: A hybrid approach using transductive learning and statistical association measures", *Expert Systems with Applications*, vol.37, no.5, pp.3553-3560, 2010.

[26] Tian-Jian Jiang, Shih-Hung Liu, Cheng-Lung Sung and Wen-Lian Hsu, "Term Contributed Boundary Tagging by Conditional Random Fields for SIGHAN 2010 Chinese Word Segmentation Bakeoff", in *the Proceedings of the CIPS-SIGHAN Joint Conference on Chinese Language Processing, Beijing, China*, pp.266-269, August 28-29, 2010.