Contextual-and-Semantic-Information-Based Domain-Adaptive Chinese Word Segmentation

Jing Zhang, Degen Huang, and Deqin Tong

School of Computer Science and Technology, Dalian University of Technology,
Dalian 116023, P.R. China
huangdg@dlut.edu.cn,
zhangjingqf@mail.dlut.edu.cn,
tongdeqin@gmail.com

Abstract. This paper presents a new domain-adaptive Chinese Word Segmentation (CWS) method. Considering the characteristics of the territorial Out-of – Vocabularies (OOVs), both the contextual information table and the semantic information are utilized based on Conditional Random Fields (CRFs) model to recall more OOVs and promote the performance of the CWS. This method is evaluated by the simplified domain-adaptive Chinese testing data from SIGHAN Bakeoff 2010. The experimental results show that the F-value and the recall of OOVs of the testing data in Computer, Medicine and Finance domain are higher than the best performance of SIGHAN Bakeoff 2010 participants, with the recall of OOVs of 84.3%, 79.0% and 86.2%, respectively.

Keywords: domain-adaptive CWS, Conditional Random Fields (CRFs), contextual variable table, semantic resources.

1 Introduction

CWS (Chinese Word Segmentation) is a fundamental task in Chinese Language processing. In recent years, widespread attention has been paid to CWS. Researchers in this field have made significant breakthrough with the rise of machine learning methods. Meanwhile, the Chinese word segmentation evaluations organized by SIGHAN (Special Internet Group of the Association for Computational Linguistics) play a prominent role in promoting the development of CWS, providing researchers with uniform training and testing data to compare their different methods of CWS in the same test platform. In previous SIGHAN Bakeoff, most of the systems with high-performance are based on machine learning methods to implement sequence labeling [1-2]. Among those methods, the character-based labeling machine learning methods [3-5] has got more and more attention and become the mainstream technology of CWS. However, Refs. [6-8] employed another sequence tagging based machine learning methods, namely, a word-based segmentation strategy, which is also based on the character-base sequence annotation.

With the development of the Internet, an increasing number of non-standard text, containing lots of new words, has been generated, which has brought many challenges

M. Zhou et al. (Eds.): NLPCC 2012, CCIS 333, pp. 110–120, 2012.

to the CWS. Although many methods have shown impressive results in some segmentation evaluation tasks, they are limited to corpus on specific area. Their accuracy will obviously decrease when used in a different domain. In practical applications, it is impossible for a CWS system to train all types of text beforehand. Additionally, the vast majority of the texts, which need to be segmented, do not have feature tags, such as Source, Subject, Part-of-speech, and so on. It is when it deals with the corpus which is different from the training data, or has a large number of OOVs that the CWS system can contribute the maximum value [9]. Therefore, SIGHAN-CIPS has set up to examine the ability of the cross-domain word segmentation since 2010. In that task, participants are demanded to test the corpus from four different domains, including computer, medical, financial and literary. The CWS systems need to be adaptive to different domains by training on only one domain corpus, namely, the so-called cross-domain CWS. One important thing the Cross-domain CWS should take into account is that there are many common-used words and terminologies in a specific area, and those words, a big inevitable challenge for CWS systems, are usually regarded as OOVs in other areas. Different from common OOVs, most of those territoriality OOVs belong to a specific area, and usually appear several times in the context of their respective areas. No matter how large the vocabulary of the segmentation system is, it is unable to include all the new words, thus a good cross-domain CWS should have a great ability to identify OOVs. Ref. [6] proposed a new cross-domain segmentation method based on a joint decoding approach which combined the character-based and word-based CRF models, made good use of the chapter information and fragments of the words, and achieved an impressive result. In the evaluation of SIGHAN Bakeoff 2010, some other excellent cross-domain word segmentation systems emerged. Among those systems, Ref. [10] introduced a multi-layer CWS system based on CRFs, integrating the outputs of the multi-layer CWS system and the conditional probability of all possible tags as the features by SVM-hmm. This system achieved the best performance in the opening tests, while it is a little bit complicated. In Ref. [11], the hidden Markov model HMM (Hidden Markov Models) was used to revise substrings whose marginal probability was low, and achieved high performance in both closed and open tasks, but its recall of OOV was not outstanding. Ref. [12] proposed a new CWS approach using the cluster of Self-Organizing Map networks and the entropy of N-gram as features, training on a large scale of unlabeled corpus, and it obtained an excellent performance. However, most of the participating systems are dealing with the OOVs, which have their own distinct territorial characteristics, as the general ones instead of the cross-domain ones on the basis of ensuring the overall performance of the CWS. However, most of the participating systems are dealing with the OOVs, which have their own distinct territorial characteristics, as the general ones instead of the cross-domain ones on the basis of ensuring the overall performance of the CWS.

According to the characteristics of the territoriality OOVs, we propose a new statistic variable, the Contextual Variable table, which records the contextual information of a candidate word and can affect the cost factor of the candidate words. Those candidate words are selected by the character-based CRFs. At the same time, we utilize the information of the synonym in the system dictionary instead of the

information of the OOVs in the candidate words, because of the similarity of syntax and context in the sentence environment. Moreover, we put all the candidate words into a set, which is called the word-lattice, and then we complete the word-lattice taking full advantage of the contextual information and the synonym information mentioned above. At last, we use the word-based CRFs to label the words in the word-lattice and select the best path as the final segmentation results.

The rest of this paper is organized as follows. Section 2 presents the machine learning models that we utilize in our experiments. In Section 3, we describe the Cross-Domain CWS algorithm. Section 4 shows the experimental results. Finally, some conclusions are given in Section 5.

2 Machine Learning Models

Conditional random fields (CRFs), a statistical model for sequence labeling, was first introduced by Lafferty et al in Ref. [2]. It is the undirected graph theory that CRFs mainly use to achieve global optimum sequence labeling. It is good enough to avoid label bias problem by using a global normalization.

2.1 Character-Based and Word-Based CRFs

In previous labeling task of character-based CRFs, the number of the characters in the observed sequence is as same as the one in the annotation sequence. However, for CWS task, the input of n-character will generate the output of m-word sequence on such a condition that m is not larger than n. But this problem can be well solved by word-lattice based CRFs, because the conditional probability of the output sequence depends no longer on the number of the observed sequence, but the words in the output path. For a given input sentence, its possible paths may be various and the word-lattice can well represent this phenomenon. A word-lattice can not only express all possible segmentation paths, but also reflect the different attributes of all possible words in the path. Refs. [13-14] have successfully used the word lattice in Japanese lexical analysis.

Our paper adopt the word-lattice based CRFs that combines the character-based CRFs and the word-based CRFs, and specifically, we put the candidate words selected by the character-based CRFs into a word-lattice, and then label all the candidate words in the word-lattice using word-based CRFs model. When training the word-lattice based CRFs model, the maximum likelihood estimation is used in order to avoid overloading. In the end, Viterbi algorithm is utilized in the decoding process which is similar with Ref. [6].

2.2 Feature Templates

The character-based CRFs in our method adopt a 6-tag set in Ref. [15] and its feature template comes from Ref. [11], including C_{-1}, C_0, C_1, $C_{-1}C_0$, C_0C_1, $C_{-1}C_1$ and $T_{-1}T_0T_1$, in which C stands for a character and T stands for the type of characters, and the

subscripts -1, 0 and 1 stand for the previous, current and next character, respectively. Four categories of character sets are predefined as: Numbers, Letters, Punctuation and Chinese characters. Furthermore, the Accessor Variety in Ref. [16] (AV) is applied as global feature.

Two kinds of features are selected for the word-based CRFs, like Ref. [6]: unigram features and bigram features. The unigram ones only consider the attributes information of current word, and bigram ones are also called compound features, which utilize contextual information of multiple words. Theoretically, the current word's context sliding window can be infinitely large, but due to efficiency factors, we define the sliding window as 2. The specific features are W_0, T_0, W_0T_0, W_0T_1, T_0T_1, W_0W_1, where W stands for the morphology of the word, T stands for the part-of-speech of the words, and subscript 0 and subscript 1, respectively, stand for the former and the latter of two adjacent words.

3 Cross-Domain CWS algorithm

The recognition of the OOVs will be limited, because the construction of the word-lattice depends on the dictionary. That can be solved by adding all the candidate words selected by the N-Best paths of the character-based CRFs into the word-lattice, so there could exit more OOVs in the word-lattice. What is more, the words in the dictionary and the OOVs can be treated equally by the character-based CRFs, which is of great help to recall OOVs. In our experiment, we finally choose 3-Best paths, because too many incorrect candidate words will be added into the word-lattice if we chose more than 3-Best paths, which not only put bad impact on the performance of the segmentation, but also affect the efficiency. When we choose less than 3-Best paths, the segmentation system does not work well on recalling the OOVs.

In the process of building the word-lattice, if the POS and the Cost of the words can not get from the system dictionary, then it will be treated as one of four different categories: Chinese characters, letters , numbers and punctuation, whose POS is, respectively, conferred as a noun, strings, numbers, punctuation. Additionally, the cost of the words equals the average of the costs of the words with the same POS in the dictionary.

Taking the characteristics of the territorial OOVs into account, we apply the contextual information and semantic information to improve the recall of the cross-domain OOVs.

3.1 Contextual Information

The territorial OOVs may repeatedly emerge in the specific domain, but it is hard to segment them correctly every time. As a result, we propose the contextual information to record the some useful information about the out-of-vocabulary candidate words. This approach is mainly based on the following assumptions:

Assumption 1: The occurrence of a word will increase the possibility of emerging of the word in the same chapter.

In other words, if a string of characters is regarded as a candidate word in multiple contexts, then it is probably a word, in that case, the Contextual Variable is proposed to quantify this assumption. The Contextual Variable consists of the morphology of the word (*w*), part of speech (*t*), the difficulty of the emerging of a candidate word (*Cost*), the frequency of being a candidate word (*Frequency*), the frequency of being the node in the final segmentation path (*rNum*).

The acquisition of the contextual information is throughout the entire segmentation algorithm, and the specific process is as follows:

Firstly, put all the candidate words w included by 3-Best paths into the set S (w_1 , w_2 , …, w_n). Secondly, search for each word *w* in set S from the system dictionary, if exists, then the information in the dictionary, such as the POS, the cost and so on, of the word *w* will be copied into the contextual information table. Otherwise, the contextual information table will be searched, and if there exits the candidate word *w*, then the *Frequency* in the table of the word increases by 1, and if not neither, then we will deal with the word as one of the four classification of the OOVs mentioned above. At last, repeat these steps until the last word w_n in set S has been searched.

It can be seen from the above process that the higher the frequency of the candidate word is, the more likely it tends to be a word. Considering that the *Frequency* and the *rNum* can affect the *Cost*, we adjust the *Cost* of the word *w*, according to Eq. (1), where $cost_0(w)$ stands for the original cost of the words.

$$cost'(w) = \begin{cases} \frac{1.0}{rNum+1} \times cost_0(w) & rNum>0 \\ \left(\frac{0.2}{\log(frequency+2)} + 0.8 \right) \times cost_0(w) & rNum=0 \end{cases} \quad (1)$$

3.2 Semantic Information

The number of Chinese words is tens of millions, while the types of semantic relations are limited, so we utilize the synonym relations, one kind of semantic information, to identify the OOVs, considering the similarity in syntax and grammar in the sentence environment. When building the word-lattice, we propose the synonym information to obtain the property and cost of the candidate words selected by the character-based CRFs via selecting the 3-Best paths, because the property and the cost of OOVs can not be found in the system dictionary, but can be substituted by the information of their synonyms.

To illustrate, the word fragment "劳模", an Out-of-vocabulary, is in the word-lattice, but not in the system dictionary. So we can not get the information of the candidate word such as the POS, the cost and so on. In this case, the synonym forest is very useful if it includes a synonym which is also in the system dictionary. For this example, the information of the word "模范", a synonym of the candidate word "劳模", can take the place of the information of "劳模".

Al 05A 01= 模范标兵表率榜样师表轨范楷范英模典型丰碑
Al 05A 02= 劳模劳动模范

The semantic resources we used in this paper is synonym forest (extended version), containing a total of 77,343 items, which have been organized into tree-like hierarchical structure and divided into three categories. In the expanded version of the synonym forest with five-level coding, for each word information, there is a eight bit semantic encoding, which can represent each single word in the synonym forest. From left to right, the encoding is expressed like this: the first level with capital letters, the second level with lowercase letters, the third level with two bytes of decimal integer, the fourth level with capital letters, and the fifth level with two bytes of decimal coding, the end with the sign of "=", "#" and "@". The specific coding rules are shown in Table 1:

Table 1. The Rule of Word Coding

Code Bit	1	2	3	4	5	6	7	8
Example	D	a	1	5	B	0	2	=\#\@
Signification	General class	Middle class	Sub-class		Word group	Atomic Word group		
Level	1	2	3		4	5		

Except for the synonym and the classification information, the synonym forest also includes some self-governed words, which do not have any synonyms. In order to enhance the search efficiency, we delete those self-governed words. Because the closer the distance of two synonym sets are, the more similar their meanings are, we follow the principle of proximity when search for the synonym of the candidate words.

The search process is as follows: first, find the synonym set of the candidate word, and then look up each synonym of that synonym set into the system dictionary to find if the synonym exists. If there it is, then we will replace the candidate word with the synonym and the information of it, and if not, then the fifth level of the synonym sets will be searched, and if not neither, then the fourth level. If the fourth level does not contain the synonym of the candidate, then we would like to stop looking up rather than search further. There are two reasons, one is the efficiency factor, the other one is that if the set of the word is too far away, the meaning of the words in two different sets will be much different, so we would rather giving it up than using it and bringing a negative impact.

3.3 Word Segmentation Process

With the contextual information and synonyms information added, the cross-domain word segmentation process is as follows:

Step1. Put all the candidate words in 3-Best paths selected by the character-based CRFs model into the word-lattice.

Step2. To build the word-lattice, in other word, give properties and costs to each node, the candidate words selected by character-based CRFs in Step1, in the word-lattice, which is divided into four cases to deal with: ①If the candidate words are in the system dictionary, then assign the properties and cost of the words in the system

dictionary directly to the candidate words in the word-lattice. ②If the candidate words are not in the system dictionary, but in the dictionary of contextual information, then the properties of the words in the contextual information dictionary will be assiged to the candidate words, and a weight value, calculated by Eq. (1), will be added to the cost of the candidate words. ③If the candidate words is not in the system dictionary, neither in the contextual information dictionary, then we will search the synonyms forest to find a synonym of the candidate words. If the synonym exits in the system dictionary, we'd like to replace the candidate word with it. ④If the above cases are not suitable for the candidate words, then the candidate words will be classified according to the classification mentioned above.

Step3. To find the optimal path, the least costly path of word segmentation, in the word-lattice using the Viterbi algorithm according to Eq. (4), and the values of $TransCost(t_i,t_{i+1})$ and $Cost(w_i)$ can be calculated by Eq. (2) and Eq. (3), respectively. Since all feature functions are binary ones, the cost of the word is equal to the sum of all the weight of the unigram features about the word, and the transition cost is equal to the sum of all bigram features about the two parts of speech.

$$Cost(w) = -factor * \sum_{f_k \in U(w)} \lambda_{f_k} \tag{2}$$

$$TransCost(t_1,t_2) = -factor * \sum_{f_k \in B(t_1,t_2)} \lambda_{f_k} \tag{3}$$

Where $U(w)$ is the unigram feature set of the current word, $B(t_1, t_2)$ is the bigram feature set of the adjacent words t_1 and t_2. λ_{fk} is the weight of the corresponding feature f_k and factor is the amplification coefficient.

$$Score(Y) = \sum_{i=0}^{Y^{\#}} \{ TransCost(t_i,t_{i+1}) + Cost(w_i^i) \} \tag{4}$$

It can be seen from the above process that the factors of recognizing the territorial words are considered in Step2. Contextual information as well as synonym information is used to adjust the cost and the properties of the candidate words in the path, which can contribute to the follow-up Step3 to select the best path.

4 Experimental Results and Analysis

4.1 Data Set

Our method is tested on the simplified Chinese domain-adaptive testing data from SIGHAN Bakeoff 2010. And it accords with the rules of the open test, since only a system dictionary and synonym forest is used in our method, without using any other manually annotated corpus resources. Thus, the experiment results are evaluated by P (Precise), R (Recall) and F-value. The system dictionary we used is extracted from the

People's Daily from January to June, in 2000, containing 85000 words, with the POS being the Peking University POS system. The word-based CRFs model is trained by the corpus with POS tag provided by the evaluation, which is from the People's Daily of January, in 1998).

4.2 Experimental Results

In order to prove the effect of the contextual information and semantic information described above, we have conducted four groups of experiments. Experiment 1 is the base experiment that does not include these two types of information. Experiment 2 is the +CV experiment with only contextual information added. Experiment 3 is the +CiLin experiment that add only synonyms information. Experiment 4 is the experiment with both two types of information added.

Table 2~5 give the segmentation results of four groups of experiments, respectively, in four different fields, including computer, medicine, finance and literary. It can be clearly seen from Table 2 to Table 5 that the performance in F-value and R_{oov} improves after the introduction of context and synonyms information, separately. And the improvement is more considerable when adding both of the two information simultaneously, with R_{oov} increasing by 1.6 to 5.6 percentage.

The following sentence fragments can help us analyze the impact of contextual information on the CWS:

"日本金融特任大臣①亀井静香 (Shizuka Kamei) 周五 (3月19日) 发表讲话……②亀井静香此前就一直呼吁推出新一轮的大规模经济刺激计划……③亀井静香表示，昨日发布的土地价格调查报告显示……④亀井静香还呼吁日本央行直接买入国债来为政府赤字提供融资……金融市场对⑤亀井静香的评论应该不会有太大反应……".

In the above five sentence fragments, the word "亀井静香"(name) appears five times totally in the context. If not bring the contextual information in the segmentation system, only three times that the word "亀井静香" is segmented correctly, while it is cut correctly all five times after adding the contextual information. Therefore, the contextual information is very helpful to identify such candidate words that repeat in a chapter, because its probability will be affected by the impact of the frequency of occurrence in the previous paragraph.

Table 2. The P, R and F value of computer

Computer	F	R	P	Roov
Base	0.9507	0.9562	0.9452	0.8233
+CV	0.9530	0.958	0.9481	0.8342
+CiLin	0.9515	0.9568	0.9462	0.83
++Both	0.9553	0.9591	0.9516	0.8428

Table 3. The P, R and F value of medicine

Medicine	F	R	P	Roov
Base	0.9424	0.946	0.9388	0.7563
+CV	0.9437	0.947	0.9404	0.7693
+CiLin	0.944	0.9473	0.9408	0.7788
++Both	0.9463	0.9492	0.9435	0.79

Table 4. The P, R and F value of finance

Finance	F	R	P	Roov
Base	0.9605	0.9585	0.9626	0.8458
+CV	0.962	0.9608	0.9631	0.852
+CiLin	0.9608	0.9592	0.9623	0.8517
++Both	0.9625	0.9609	0.9641	0.8618

Table 5. The P, R and F value of literature

Literature	F	R	P	Roov
Base	0.9421	0.9385	0.9458	0.6504
+CV	0.9433	0.9393	0.9473	0.6649
+CiLin	0.9437	0.9394	0.948	0.6839
++Both	0.946	0.9418	0.9506	0.7073

Table 6. Comparison with the open test results of Bakeoff

Corpora	Participants	F	R_{OOV}
Computer	1[10]	0.95	0.82
	2[12]	0.947	0.812
	3[11]	0.939	0.735
	ours	**0.955**	**0.843**
Medicine	1[12]	0.938	0.787
	2[10]	0.938	0.768
	3[11]	0.935	0.67
	ours	**0.946**	**0.790**
Finance	1[10]	0.96	0.847
	2[11]	0.957	0.763
	3[12]	0.951	0.853
	ours	**0.963**	**0.862**
Literature	1[11]	0.955	0.655
	2[17]	0.952	0.814
	3[12]	0.942	0.702
	ours	**0.946**	**0.707**

Table 6 shows the results of our method compared with the top three outstanding systems of the SIGHAN Bakeoff 2010 evaluation in F-value and R_{oov}. The experimental results show that the performance of our system in both F-value and R_{oov} is better than the best results of the SIGHAN Bakeoff 2010 evaluation in the three areas of the computer, medicine and finance.

5 Conclusions

In this paper, a new cross-domain CWS method is proposed. Due to the recurrences of the territorial OOVs in their specific areas, we bring up the contextual variable table to record the contextual information of the candidate words which are selected by the character-based CRFs, including the morphology of the word, the part-of-speech, the difficulty degree of appearing, the frequency as a candidate, and the frequency as the word node in the final segmentation path. Additionally, in order to approximate the cost of the candidate word in the entire path, we replace the property information and the cost of OOVs with their synonyms. As we know, the closer the sets of two synonyms are, the more similar their meanings are. Therefore, when we search for the synonym of a candidate word in the synonym forest, we follow the principle of proximity. At first, we get the 3-best paths with the help of character-based CRFs, and add all the words included by the 3-best paths into the word-lattice. And then, we make use of the contextual and semantic information to construct the word-lattice to recall more OOVs. At last, the word-based CRFs are utilized to select the least costly path from the word-lattice as the final segmentation results.

Our method not only take full advantage of character-based CRFs model to generate more OOVs, but also make good use of the lexical information of the territorial words. Our method is evaluated by the simplified Chinese domain-adaptive testing data from SIGHAN Bakeoff 2010. The experimental results show that the F-value and the recall of OOVs of the testing data in Computer, Medicine and Finance domain are higher than the best performance of SIGHAN Bakeoff 2010 participants, with the recall of OOVs of 84.3%, 79.0% and 86.2%, respectively.

Acknowledgements. This work has been supported by the National Natural Science Foundation of China (No.61173100, No.61173101), Fundamental Research Funds for the Central Universities (DUT10RW202). The authors wish to thank Jiang Zhenchao, Xu Huifang and Zhang Jing for their useful suggestions, comments and help during the design and editing of the manuscript.

References

1. Xue, N.: Chinese word segmentation as character tagging. J. Computational Linguistics 8(1), 29–48 (2003)
2. Lafferty, J., McCallum, A., Pereira, F.: Conditional Random Fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of ICML 2001, pp. 282–289. Morgan Kaufmann, San Francisco (2001)

3. Tseng, H., Chang, P., Andrew, G., Jurafsky, D., Manning, C.: A conditional random field word segmenter for SIGHAN bakeoff 2005. In: Proc. of the 4th SIGHAN Workshop on Chinese Language Processing, pp. 168–171. ACL, Jeju Island (2005)
4. Peng, F., Feng, F., McCallum, A.: Chinese segmentation and new word detection using conditional random fields. In: Proc. of COLING 2004, pp. 562–568. Morgan Kaufmann, San Francisco (2004)
5. Low, J.K., Ng, H.T., Guo, W.: A maximum entropy approach to Chinese word segmentation. In: Proc. of the 4th SIGHAN Workshop on Chinese Language Processing, pp. 161–164. ACL, Jeju Island (2005)
6. Huang, D., Tong, D.: Context Information and Fragments Based Cross-Domain Word Segmentation. J. China Communications 9(3), 49–57 (2012)
7. Zhang, R., Kikui, G., Sumita, E.: Subword-based tagging by conditional random fields for Chinese word segmentation. In: Proc. of HLT-NAACL 2006, pp. 193–196. ACL, Morristown (2006)
8. Huang, D., Jiao, S., Zhou, H.: Dual-Layer CRFs Baesed on Subword for Chinese Word Segmentation. Journal of Computer Research and Development 47(5), 962–968 (2010); 黄德根, 焦世斗, 周惠巍. 基于子词的双层CRFs中文分词. J. 计算机研究与发展 47(5), 962–968 (2010)
9. Huang, C.-R.: Bottleneck _ challenges _ turn for the better _new ideas of the Chinese word segmentation. In: Computational Linguistics Research Frontier 2007-2009, pp. 14–19. Chinese Information Processing Society, Beijing (2009); 黄居仁.黄居仁. 瓶颈_挑战_与转机_中文分词研究的新思维. 中国计算机语言学研究前沿进展(2007-2009), 14–19. 中国中文信息学会, 北京 (2009)
10. Gao, Q., Vogel, S.: A Multi-layer Chinese Word Segmentation System Optimized for Out-of-domain Tasks. In: Proc. of CIPS-SIGHAN Joint Conference on Chinese Processing, pp. 210–215. ACL, Beijing (2010)
11. Huang, D., Tong, D., Luo, Y.: HMM Revises Low Marginal Probability by CRF for Chinese Word Segmentation. In: Proc. of CIPS-SIGHAN Joint Conference on Chinese Processing, pp. 216–220. ACL, Beijing (2010)
12. Zhang, H., Gao, J., Mo, Q., et al.: Incporating New Words Detection with Chinese Word Segmentation. In: Proc. of CIPS-SIGHAN Joint Conference on Chinese Processing, pp. 249–251. ACL, Beijing (2010)
13. Zhang, C., Chen, Z., Hu, G.: A Chinese Word Segmentation System Based on Structured Support Vector Machine Utilization of Unlabeled Text Corpus. In: Proc. of CIPS-SIGHAN Joint Conference on Chinese Processing, pp. 221–227. ACL, Beijing (2010)
14. Nakagawa, T.: Chinese and Japanese word segmentation using word-level and character-level information. In: Proc. of COLING 2004, pp. 466–472. ACL, Geneva (2004)
15. Kudo, T., Yamamoto, K., Matsumoto, Y.: Applying conditional random fields to Japanese morphological analysis. In: Proc. of EMNLP 2004, pp. 230–237. ACL, Barcelona (2004)
16. Zhao, H., Huang, C., Li, M., et al.: Effective tag set selection in Chinese word segmentation via Conditional Random Field modeling. In: PACLIC 2006, pp. 87–94. ACL, Wuhan (2006)
17. Feng, H., Chen, K., Deng, X., Zheng, W.: Accessor variety criteria for Chinese word extraction. J. Computational Linguistics 30(1), 75–93 (2004)