Which Is Essential for Chinese Word Segmentation: Character versus Word

Chang-Ning Huang and Hai Zhao

Microsoft Research Asia,
49, Zhichun Road, Haidian District,
Beijing, China, 100080
{cnhuang, f-hzhao}@msrchina.research.microsoft.com

Abstract. This paper proposes an empirical comparison between word-based method and character-based method for Chinese word segmentation. In three Chinese word segmentation Bakeoffs, character-based method quickly rose as a mainstream technique in this field. We disclose the linguistic background and statistical feature behind this observation. Also, an empirical study between word-based method and character-based method are performed. Our results show that character-based method alone can work well for Chinese word segmentation without additional explicit word information from training corpus.

1 Introduction

Chinese text is written without natural delimiters, so word segmentation is an essential first step in Chinese language processing. In this aspect, Chinese is quite different from English in which sentences of words delimited by white spaces. Though it seems very simple, Chinese word segmentation (CWS) is not a trivial problem. Actually, it has been active area of research in computational linguistics for almost 20 years and has drawn more and more attention in the Chinese language processing community. To accomplish such a task, various technologies are developed [1][2].

In the early work of Chinese word segmentation, word-based method once played the dominant role, in which maximum matching algorithm is the most typical method. Here, the term, word, means those known words are shown in known lexicon or training corpus (also are called in-vocabulary(IV) words.). Explicit known word information was still important learning object even after statistical methods were introduced in CWS [1].

To give a comprehensive comparison of Chinese segmentation on common test corpora, three International Chinese Word Segmentation Bakeoffs were held in 2003, 2005, and 2006[1], and there were 12, 23 and 23 participants, respectively [3], [4], [5]. Four segmentation corpora were presented in each Bakeoff. Thus, twelve corpora are available from Bakeoff 2003, 2005, and 2006. A summary of these corpora is shown in Table 1.

In all of proposed methods, character-based tagging method [6], instead of traditional word-based one, quickly rose in Bakeoff-2005 as a remarkable one with state-of-the-art performance. Especially, two participants, Ng and Tseng, gave the best results

[1] In 2006, the name of the third Bakeoff has been changed into International Chinese Language Processing Bakeoff for the reason that named entity recognition task was added

in almost all tracks [7], [8]. In Bakeoff-2006, all participants whose system performance ranked first in a track at least used character-based method. Researchers turned to character-based method from traditional word-based method only with four years.

The success of Bakeoffs not only gave some public consistent segmentation standards, but also proposed a corpus-based segmentation standard representation, instead of the representation of known word lexicon and segmentation manual before. Thus Chinese word segmentation becomes more like corpus-based machine learning procedure in this sense.

With the supply of common segmentation standards of Bakeoffs, the comparison problem on word-based method and character-based method are still remained. Though most effective Chinese word segmentation techniques are turned to pure character-based methods, some researchers are still insisting that character-based method alone can not be superior to the method that combines both word information and character information [9] [10][11]. In this paper, we will briefly explore the linguistic background of such turnaround in Chinese word segmentation and give an empirical comparison of these methods.

Table 1. Corpora statistics of Bakeoff 2003, 2005 and 2006

Provider	Corpus	Encoding	#Training words	#Test words	OOV rate
Academia Sinica	AS2003	Big5	5.8M	12K	0.022
	AS2005	Big5	5.45M	122K	0.043
	AS2006	Big5	5.45M	91K	0.042
Hong Kong City University	CityU2003	Big5	240K	35K	0.071
	CityU2005	Big5	1.46M	41K	0.074
	CityU2006	Big5	1.64M	220K	0.040
University of Pennsylvania	CTB2003	GB	250K	40K	0.181
	CTB2006	GB	508K	154K	0.088
Microsoft Research Asia	MSRA2005	GB	2.37M	107K	0.026
	MSRA2006	GB	1.26M	100K	0.034
Peking University	PKU2003	GB	1.1M	17K	0.069
	PKU2005	GB	1.1M	104K	0.058

The remainder of the paper is organized as follows. The next section reviews the track of character-based method. We discuss the linguistic background of character-based features (especially for unigram feature) in Section 3. We evaluate unigram feature through CWS performance comparison in Section 4. In Section 5, the experimental results between word-based method and character-based method are demonstrated. We summarize our contribution in Section 6.

2 The Track of Character-based Method

Character-based tagging method is a classification technique for Chinese characters according to their positions occurring in Chinese words. This method was first conducted in [12], two classifiers were combined to perform Chinese word segmentation. First, a maximum entropy model was used to segment the text, and then an error driven transformation model was used to correct the word boundaries. This method was continuously improved in [6] and [13], where a unified maximum entropy model was used to perform character-based tagging task.

As mentioned above, two top participants, Tseng and Low, won the most outstanding success in Bakeoff-2005 with the similar character-based tagging method, though the former used conditional random field model while the latter still used maximum entropy model.

In Bakeoff-2006, all participants whose system performance ranked first in a track at least used character-based method. There are five participants ranked the first in one track at least [14][15][16][17][18], in which two participants used conditional random field, and the other three used maximum entropy as learning model. Especially, four participants directly or indirectly used the technique in [7].

3 Features of Character Classification for CWS

CWS is the primary processing in Chinese language processing. Thus it is difficult or even impossible to use derivative features like other Chinese language processing tasks. The basic features that we can use are characters themselves.

We perform a position frequency statistics of Chinese characters in MSRA2005 training corpus. All characters appearing in this corpus are counted. Six positions are distinguished, which are represented by a 6-tag set including B, E, S, B_2, B_3, and M [14]. Tag B and E stand for the first and the last position in a multi-character word, respectively. S stands up a single-character word. B_2 and B_3 stand for the second and the third position in a multi-character word, whose length is larger than two-character or three-character. M stands for the fourth or more rear position in a multi-character word, whose length is larger than four-character.

Let $T = \{B, E, S, B_2, B_3, M\}$, we calculate the *productivity*, $P_{C_i}(t_j)$, of each position of each character C_i:

$$P_{C_i}(t_j) = \frac{count(C_i, t_j)}{\sum_{t_j \in T} count(C_i, t_j)} \tag{1}$$

We count those characters whose productivity is larger than 0.5 for a certain tag. The results are shown in Table 2. There are 5,147 different characters in MSRA2005 training corpus. Our statistics shows that most characters, 76.16% of all, trend to have a stable position in the word. This is important for a character-based tagging method. However, there are still 1,227 characters without dominant tag. We regard these characters as free ones. The fact that no special positions are dominant for a character means that this character can occur in every possible positions in a word. That is, this character is free for word formation. In our threshold of productivity 0.5, 1/4 characters (precisely, 23.84%) in one of real corpora, MSRA2005, are free ones.

Table 2. The distribution of numbers of characters in each position

Tag	B	B_2	B_3	M	E	S	Total
Number of characters	1634	156	27	33	1438	632	3920
Percent(%)	31.74	3.03	0.52	0.64	27.94	12.28	76.16

We list ten top frequent characters and their tag distributions in Table 3 according to MSRA2005 training corpus.

Table 3. Top frequent characters and their tag distributions

Characters	Frequency	B	E	S	B_2	B_3	M
的	129132	0.001169	0.010338	**0.987679**	0.000519	0.000163	0.000132
一	40189	**0.540023**	0.058648	0.285650	0.086889	0.019408	0.009381
国	40091	0.310070	**0.468609**	0.020828	0.151206	0.024968	0.024320
在	32594	0.024821	0.099742	**0.869485**	0.003712	0.002178	0.000061
中	29762	**0.490558**	0.093609	0.315570	0.032424	0.032323	0.035515
了	29305	0.026480	0.052346	**0.919980**	0.000478	0.000682	0.000034
是	28020	0.015703	0.338829	**0.641113**	0.001642	0.002712	0.000000
人	27260	**0.355026**	0.304952	0.228833	0.023844	0.063243	0.024101
和	26328	0.047820	0.008356	**0.922440**	0.007710	0.001785	0.011888
有	26196	0.268133	0.313597	**0.376661**	0.018934	0.008207	0.014468

To demonstrate the distribution of characters with different productivity as single-character word, we count different types of characters in certain range, A bar figure is shown in 1. This figure further shows that most characters trend to be components of multi-character words, instead of single-character words. Especially, more than half of characters nearly never be a single-character word. This is another obvious statistical characteristic for word formation from character combination.

Another convenience for character-based method is that it can be more easily to handle out-of-vocabulary (OOV) words. As well known, the set of all Chinese characters is almost a closed set. 2,500 Chinese characters can cover 97.97% text that one can meet in his life, while 3,500 characters can cover more than 99.48% text[2]. We see that the OOV rate of word for MSRA2005 corpus is 2.6%, while the OOV rate of character is only 0.42% (12 OOV characters versus 2,837 characters in MSRA2005 test corpus). We see that the former is much larger than the latter. In addition, six of these OOV characters appear only once.

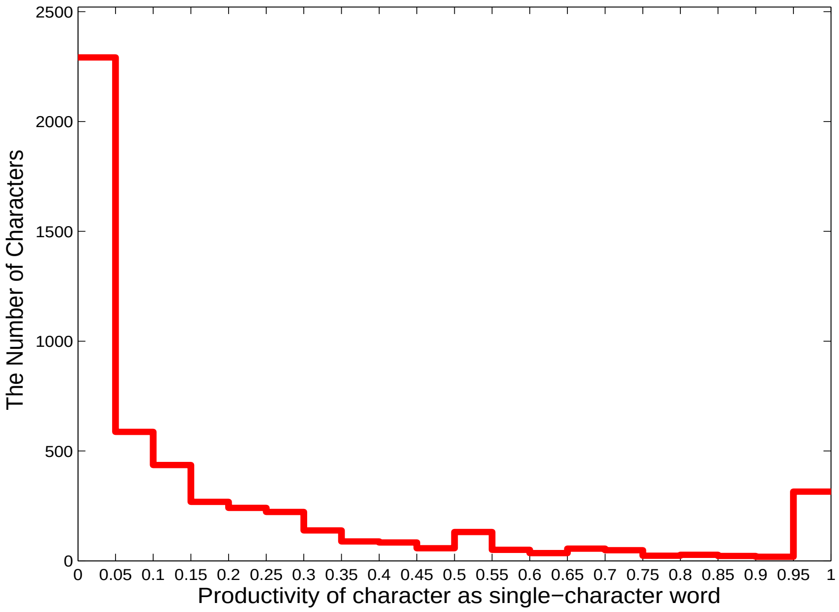

Fig. 1. Character distribution with different productivity as single-character word. All counting are performed when $P_{C_i}(S) \geq j * 0.05$ and $P_{C_i}(S) < 0.05 + j * 0.05$, where $j = 0, 1, ..., 20$.

The productivity of character is the concept of linguistics, while it is just the learning goal as the unigram feature for a sequence learning model. If the context is free or absent, then what a character itself should be a word alone is determined by the productivity of position "S", and what it should be the begging of a word is determined by the productivity of position "B", and so on. Note that segmentation is an operation to determine separation of sequence or not at the current character. The usefulness of productivity of character, or namely unigram feature in learning model, is obvious.

[2] *The introduction to modern Chinese character list in common use (*现代汉语常用字表说明*)*, published by the State Language Affairs Commission and the State Education Commission on January 26th, 1988.

Since most characters trend to be in stable position in word formation, it will be efficient for a character-based classification technique for CWS. One remained challenging thing is the task to determine those characters that can freely appear in each position of words without favoritism, whose percent is 23.84% in all kinds of characters. This leads to more strict context to perform the task to determine the classification of these free characters.

In a character sequence, the straightforward way to represent context is using adjacent characters. Actually, this means that more n-gram features are used. We explain this case in a real sentence, “葡萄是红的(The grape is red.)”. The final segmentation result will be “葡萄/是/红/的”. In a bigram sense, the reason of such segmentation is bigram probability of “葡-萄” to be a word is much higher than any other bigram probabilities of “萄-是”, “是-红” and “红-的”. Thus, “葡萄” is finally recognized as a word.

In most Chinese word segmentation systems, all possible n-gram features in a certain character-window of sequence are often used. The difference among them is the length of this character-window. Three-character window and five-character window centered by the current character are mostly adopted in existing work until now.

4 How Unigram Feature Affect the CWS Performance

We adopt the character-based CWS system that was described in [14] in this paper. The learning model is conditional random field [20], and tag set is 6-tag set as mentioned above. However, all none n-gram features in [14] are removed, and feature template list is shown in Table 4. The reason is to conform to the constraints of closed test in Bakeoff, and all features that are beyond provided training corpus are not allowed. All comparisons below will be performed in closed test settings for a consistent circumstance.

Table 4. Feature templates

Code	Type	Feature	Function
a	Unigram	$C_n, n = -1, 0, 1$	The previous (current, next) character
b	Bigram	$C_nC_{n+1}, n = -1, 0$	The previous (next) character and current character
		$C_{-1}C_1$	The previous character and next character

We explain these selected features from a real sentence, “我们在北京” (We are in Beijing). If the current character is “在”, then all active features will be “们”,“在”, “北”, “们在”, “在北”, and “们北”.

We give a performance comparison among different types of n-gram features and forward maximum matching (FMM) algorithm in MSRA2006 corpus. As for FMM algorithm, we use two dictionaries, one is extracted from training corpus, the other is extracted from both training corpus and test corpus. The results of two types of FMM algorithms will be regarded as the baseline result and topline result, respectively. The comparison is shown in Table 5.

We find that though character-based method with unigram feature only is not better than FMM/baseline in the whole F measure and IV F measure, it is much better than the latter in OOV word identification. Also, it is not surprising that bigram features make significant contributions on all aspects of segmentation performance.

Table 5. Performance comparison among different n-gram features and FMM algorithm in MSRA2006 corpus

Feature/Algorithm		FMM/Baseline	Character-based tagging method		FMM/Topline
			Unigram	Unigram+Bigram	
Total	F-measure	0.9243	0.8552	0.9609	0.9846
	Precision	0.9495	0.8608	0.9645	0.9827
	Recall	0.9004	0.8497	0.9572	0.9866
IV	F-measure	0.9422	0.8901	0.9746	0.9843
	Precision	0.9821	0.9226	0.9818	0.9821
	Recall	0.9054	0.8598	0.9676	0.9864
OOV	F-measure	0.0364	0.3182	0.6069	0.9940
	Precision	0.0217	0.2215	0.5589	0.9971
	Recall	0.1113	0.5649	0.6639	0.9909

5 Empirical Studies: Word-based Method versus Character-based Method

It is well known that there does not exists such a lexicon that can contains all possible words, this makes any known word lexicon is not efficient. The consequence is that all existing segmentation methods, including word-based ones, should carefully handle the case of out-of-vocabulary (OOV) words. In recent work, an OOV identification model were often designed in the segmentation system, in spite of word-based method or not. However, is it still necessary to integrate word-based technique into a character-based system? Or, is it better to combine word information and character information than character-based method alone? This has not been studied in previous work. Here, we will compare typical word-based or word/character based methods with a pure character-based method. The character-based system is the one described in Section 4.

5.1 Rule-based method with boundary modification learning

In [9], SVM learning algorithm was used to correct the boundaries determined by maximum matching algorithm. The experimental result comparison between this method and our method is shown in Table 6.

In detail, Goh's method is still a one-fold character-based classification one for Chinese word segmentation. However, they adopted FMM/BMM feature in learning except for character-based n-gram features. The dictionary that FMM/BMM algorithm used was extracted from training corpus. The idea behind this method is that there are always more known words than unknown words in a text, and then it is advantageous if those known words can be segmented beforehand. In another word, more information of known words than that of unknown ones was introduced into Goh's character-based system. Thus unknown words and known words are detected in unbalanced way. We attribute the reason that our system outperforms Goh's in this aspect, which handles known words and unknown ones in a more balance way determined by learning algorithm automatically.

Table 6. Comparisons of character-based tagging method and Goh's method in corpora of Bakeoff-2003

Tagging method	AS2003	CityU2003	PKU2003	CTB2003
Goh's	0.959	0.937	0.947	0.847
Character-based	0.973	0.947	0.956	0.872

5.2 Sub-word tagging method

In [10] and [11], a mixture method is used for Chinese word segmentation. Some top frequent multi-character words are also used in this tagging scheme, instead of character-only tagging method. All these tagging cliques were called as sub-words. To extract necessary sub-words from a raw sentence, a FMM/BMM like method was used in [10] and [11]. Again, we meet a character-based method that adopted known word information beforehand.

However, sub-words tagging method did not work very well alone. Thus another dictionary-based n-gram method was used to strengthen known words identification as an additional technique. A confidence measure was then defined in [10] and [11]. This measure was calculated from combining probability of sub-word tagging procedure. If this measure is less than a threshold, then the result of dictionary-based n-gram method was adopted, otherwise the results of sub-word tagging were adopted.

Though it is efficient in training for this method as declared in [10] and [11], it did not work better than a character-based only tagging method. The result comparison with our method is shown in Table 7.

We argue that known word information should be carefully introduced in processing. In our view, sub-word tagging method incorporated with dictionary-based method just tries to more carefully balance weights of known words and unknown words. However，the balance operation in Zhang’s algorithm was not quite successful than a character-based only method.

Table 7. Comparisons of character-based tagging method and sub-word-based tagging method in corpora of Bakeoff-2005

Tagging method	AS2005	CityU2005	PKU2005	MSRA2005
Sub-word	0.936	0.931	0.936	0.954
Sub-word+Dictionary	0.951	0.951	0.951	0.971
Character-based	0.953	0.948	0.952	0.974

5.3 Semi-CRF learning

Semi-Markov CRF (semi-CRF) is a modification version of standard CRF. It is different from standard CRF that semi-CRF permits labelling continuous cliques with the same tags. Thus it can been regarded as word-based CRF learning in some sense. We may introduce some additional features in semi-CRF that are intuitively very useful. Also, it has been shown that an order-M seimi-CRF is strictly more powerful than an order-M CRF.

However, the use of a semi-Markov CRF for Chinese word segmentation did not find significant gains over the standard CRF in Liang’s previous work [21]. The comparison with our method in three corpora of Bakeoff-2003 is shown in Table 8. Liang used two learning model, standard CRF and semi-CRF in the same features, the results did not support the superiority of semi-CRF, too.

A hybrid Markov/Semi-Markov CRF learning method was proposed in [22]. The traditional semi-CRF method was effectively improved since Liang’s work, and the F-measure of MSRA2005 corpus has been 0.9684, which is much better than the best results in Bakeoff-2005. However, we see that it is not better than character-based tagging method, which is 0.974 as seen in Table 7.

Table 8. Comparisons of CRF method and semi-CRF method in corpora of Bakeoff-2003

Tagging method	CityU2003	PKU2003	CTB2003
CRF learning in Liang’s method	0.937	0.941	0.879
Semi-CRF learning in Liang’s method	0.936	0.936	0.868
Character-based method	0.947	0.956	0.872

5.4 Comparison with the Best Existing Work

Comparisons between our results and best existing results in three Bakeoffs are shown in Table 9-11. There are two types of existing results for Bakeoff-2003 and 2005. One is the best F scores of Bakeoff 2003, 2005 for each corpus in closed test tracks. The other are the results of Peng and Tseng [19] [8]. As for Bakeoff-2006, our results is slightly lower than those in [14], because more feature templates were used in [14] than that in Table 4.

Table 9. Comparisons of best existing results and our results in the corpora of Bakeoff 2003

Participant	AS2003	CTB2003	CityU2003	PKU2003
Peng	0.956	0.849	0.928	0.941
Tseng	0.970	0.863	0.947	0.953
Best results of Bakeoff	0.961	0.881	0.940	0.951
Ours	0.973	0.872	0.947	0.956

Table 10. Comparisons of best existing results and our results in the corpora of Bakeoff 2005

Participant	AS2005	CityU2005	PKU2005	MSRA2005
Tseng	0.947	0.943	0.950	0.964
Best results of Bakeoff	0.952	0.943	0.95	0.964
Ours	0.953	0.948	0.952	0.974

Through Table 9 to 11, we could also see that the performance of CWS system today has been substantially improved compared with year 2003 and 2005, respectively.

6 Conclusion

In this paper, we have explored the linguistic background of character-based tagging method for Chinese word segmentation via productivity analysis of characters. Also, we gave an empirical comparison between existing word-based tagging method and a character-based tagging method in the similar learning models. The experimental results show that it is often difficult to effectively integrate word information within training corpus and character information into a segmentation system. In fact, character-based method can work better without explicit known word information within training

Table 11. Comparisons of best existing results and ours in the corpora of Bakeoff 2006

Participant	AS2006	CTB2006	CityU2006	MSRA2006
Other best results of Bakeoff	0.957	0.930	0.972	0.963
Other second best results of Bakeoff	0.957	0.927	0.972	0.957
Other third best results of Bakeoff	0.953	0.926	0.970	0.957
Ours	0.954	0.932	0.969	0.961

corpus. It is hard to say that a method that integrates known word information from training corpus will be never superior to a character-based only method. However, our comparisons show that it is not easy to realize such an effective integration for both word and character information.

References

1. Richard Sproat, Chilin Shih, William Gale, and Nancy Chang. 1996. A stochastic finite-state word-segmentation algorithm for Chinese. *Computational Linguistics*, Vol. 22(3):377-404.
2. Jianfeng Gao, Mu Li, Andi Wu and Chang-Ning Huang. 2005. Chinese Word Segmentation and Named Entity Recognition: A Pragmatic Approach. *Computational Linguistics*, Vol. 31(4): 531-574.
3. Richard Sproat and Thomas Emerson. 2003. The First International Chinese Word Segmentation Bakeoff. *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, 133-143. Sapporo, Japan.
4. Thomas Emerson. 2005. The Second International Chinese Word Segmentation Bakeoff. *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, 123-133. Jeju Island, Korea.
5. Gina-Anne Levow. 2006. The Third International Chinese Language Processing Bakeoff: Word Segmentation and Named Entity Recognition. *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, 108-117. Sidney, Australia.
6. Nianwen Xue. 2003. Chinese Word Segmentation as Character Tagging. *Computational Linguistics and Chinese Language Processing*, Vol. 8(1): 29-48.
7. Jin Kiat Low, Hwee Tou Ng and Wenyuan Guo 2005. A Maximum Entropy Approach to Chinese Word Segmentation. *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, 161-164. Jeju Island, Korea.
8. Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, Christopher Manning. 2005. A Conditional Random Field Word Segmenter for Sighan Bakeoff 2005. *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, 168-171. Jeju Island, Korea.
9. Chooi-Ling Goh, Masayuku Asahara and Yuji Matsumoto. 2005. Chinese Word Segmentatin by Classification of Characters. *Computational Linguistics and Chinese Language Processing*, Vol. 10(3): 381-396.
10. Ruiqiang Zhang, Genichiro Kikui and Eiichiro Sumita. 2006. Subword-based tagging by Conditional Random Fields for Chinese Word Segmentation. *Proceedings of Human Language Technology Conference/North American chapter of the Association for Computational Linguistics annual meeting (HLT/NAACL-2006)*, 193-196. New York.

11. Ruiqiang Zhang, Genichiro Kikui and Eiichiro Sumita. 2006. Subword-based tagging for Confidence dependent Chinese word segmentation. *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006)*, 961-968. Sidney, Australia.
12. Nianwen Xue and S. P. Converse. 2002. Combining Classifiers for Chinese Word Segmentation. *Proceedings of the First SIGHAN Workshop on Chinese Language Processing*, 57-63. Taipei, Taiwan.
13. Nianwen Xue and Libin Shen. 2003. Chinese Word Segmentation as LMR Tagging. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, in conjunction with ACL'03*, 176-179. Sapporo, Japan
14. Hai Zhao, Chang-Ning Huang and Mu Li. 2006. An Improved Chinese Word Segmentation System with Conditional Random Field. *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, 108-117. Sidney,Australia.
15. Xinhao Wang, Xiaojun Lin, Dianhai Yu, Hao Tian and Xihong Wu. 2006. Chinese Word Segmentation with Maximum Entropy and N-gram Language Model. *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, 108-117. Sidney, Australia.
16. Aaron J. Jacobs and Yuk Wah Wong. 2006. Maximum Entropy Word Segmentation of Chinese Text. *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, 108-117. Sidney, Australia.
17. Richard Tzong-Han Tsai, Hsieh-Chuan Hung, Cheng-Lung Sung, Hong-Jie Dai and Wen-Lian Hsu. 2006. On Closed Task of Chinese Word Segmentation: An Improved CRF Model Coupled with Character Clustering and Automatically Generated Template Matching. *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, 108-117. Sidney, Australia.
18. Wu Liu, Heng Li, Yuan Dong, Nan He, Haitao Luo and Haila Wang. 2006. France Telecom R&D Beijing Word Segmenter for Sighan Bakeoff 2006. *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, 108-117. Sidney, Australia.
19. Fuchun Peng, Fangfang Feng and Andrew McCallum. 2004. Chinese Segmentation and New Word Detection Using Conditional Random Fields. In *COLING 2004*, 562-568. August 23-27, 2004, Geneva, Switzerland.
20. John Lafferty, A. McCallum and F. Pereira. 2001. Conditional Random Field: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, 282-289. Williams College, Williamstown, MA, USA, June 28-July 01, 2001.
21. Percy Liang. 2005. Semi-supervised learning for natural language. Master's thesis, MAssachusetts Institute of Technology.
22. Galen Andrew. 2006. A Hybrid Markov/Semi-Markov Conditional Random Field for Sequence Segmentation. *Proceedings of the 2006 conference on Empirical Methods in Natural Language Processing (EMNLP-2006)*, 465-472. Sidney, Australia, July 2006.