Leveraging Rich Linguistic Features for Cross-domain Chinese Segmentation

Guohua Wu, Dezhu He, Keli Zhong, Xue Zhou and Caixia Yuan
School of Computer,
Beijing University of Posts and Telecommunications,
China, 100876
trustwugh@gmail.com, hedezhu@bupt.edu.cn, zhk@126.com
bupt.zhouxue@gmail.com, yuancx@bupt.edu.cn

Abstract

This paper describes the system that we use for Chinese segmentation task in the 3rd CIPS-SIGHAN bakeoff. We use character sequence labeling method for segmentation, and in order to improve segmentation accuracy over multi-domain, we present a CRF-based Chinese segmentation system integrating supervised, unsupervised and lexical features. We firstly preliminarily segment the target data using CRF model trained over three types of features mentioned above, from the result of which new words are detected and absorbed into the lexicon. To generalize across different domains, we then execute the second segment with the updated lexicon. The OOV recognition is further promoted with refined post processing. All the features we used share a unified feature template trained by CRF. Our system achieves a competitive F score of 0.9730 for this bakeoff.

1 Introduction

Word is the fundamental unit in natural language understanding. Since people do not retain the boundary information between words in practical use, Chinese Word Segmentation (CWS) is the very first step in Chinese information processing. A considerable amount of research has shown that using character sequence labeling is a simple but effective formulation of Chinese word segmentation task (Xue and others, 2003; Peng et al., 2004; Low et al., 2005; Zhao et al., 2006a), among which the method using sequence labeling based on CRF (Lafferty et al., 2001) is widely used with attractive performance. However, most of the existing segmentation systems greatly rely on data that the model was trained over. The segmentation performance tends to would reduce significantly when the test data differs greatly from the training data in phraseology and vocabulary. Exploiting corpora in multi-domain for model learning can solve the problem above directly, whereas labeling corpora manually costs a lot, so that it is unrealistic to label mass corpora.

So far there are two ways to improve the performance of cross-domain word segmentation system. The first way is proposed in (Zhao and Kit, 2007; Zhao and Kit, 2008; Zhao and Kit, 2011), in which they put forward a unified framework that integrated supervised and unsupervised segmentation together, where they could take full advantage of unsupervised segmentation to discover new word from untagged corpora and obtain the ability of supervised segmentation to recognize the known words at the same time. The segmentation system is generalized to some extent. The second way is to build a segmentation system with multi-layers. The first layer is a set of distinctive word segmentation subsystems, who might has an outstanding performance on specific domain. And the second layer combines all the outputs of these subsystems, determining the most possible segmentation boundaries on test dataset. Gao and Vogel (2010) used this method achieved top performance in three test domains out of the four during Bakeoff-2010 (Zhao and Liu, 2010). In this paper we follow the first method to improve the performance of cross-domain segmentation, meanwhile add some of the effective features that mentioned in method two. And the performance of handling OOV is improved by adding lexical feature and new words discovery.

In Section 2, we describe the features we adopted in our system. Section 3 represents how we discover new words from preliminary segmentation results and how we expand the lexicon to update lexical feature before we segment test data again to improve the segmentation performance.

Word length	Tag sequence for a word
1	S
2	BE
3	BB_2E
4	BB_2B_3E
5	BB_2B_3ME
≥ 6	BB_2B_3M$\cdots$ME

Table 1: Illustration of character tagging

The experimental result that tested on Bakeoff dataset compared with the best official result is provided in Section 4. Section 5 leads to the conclusion.

2 System Description

We formulate Chinese word segmentation task into a sequence labeling problem and use CRF to train the segmentation model. Our implementation of CRF-based CWS system uses the CRF++[1] package by Taku Kudo. We regard "，", "。", "？", "！", "；" as the boundary of a sentence and both the training and testing corpora are segmented by these boundaries.

Zhao et al. (2006b) prove that CRF segmentation performance using 6-tag set for training is better than other tag set, so we adopt 6-tag (B, B_2, B_3, M, E, S) set labeling the characters in words. Table 1 explains how to label the characters in words with different length. We follow six n-gram character features that are used in (Zhao et al., 2006b; Zhao and Kit, 2008), as C_{-1}, C_0, C_1, $C_{-1}C_0$, C_0C_1 and $C_{-1}C_1$ respectively, in which C represents the character, subscript -1, 0 and 1 means the previous character, the current character and the next character. With respect to the other features in our system, the similar six n-gram feature template is also applied to them.

2.1 Character Type Features

We simply classify all the characters by its Unicode code point into 5 classes: Chinese character (C), English character (E), number[2] (N), punctuation (P) and others (O). Denote character type feature as CTF, and define the feature template as CTF_{-1}, CTF_0, CTF_1, $CTF_{-1}CTF_0$, CTF_0CTF_1 and $CTF_{-1}CTF_1$.

[1] http://crfpp.googlecode.com/svn/trunk/doc/index.html

[2] Numbers including Arabic numerals and its Chinese version accordingly.

2.2 Conditional Entropy Feature

Gao and Vogel (2010) improve the segmentation performance on 2010 Bakeoff (Zhao and Liu, 2010) dataset by using conditional entropy feature. The forward conditional entropy for specific character C is the entropy that combines all the entropy of characters which might appear in the following position after C throughout the corpora, recorded as $H_f(C)$, while the backward conditional entropy consists of all the entropy of characters that might appear in the next position after C throughout the corpora, denoted as $H_b(C)$. We could mix unlabeled corpora in multi-domain to calculate forward and backword conditional entropy, which makes this feature more domain adaptive. Forward and backward conditional entropy can be efficiently carried out with the aid of Statistical *bigram matrixes*.

Continuous values of conditional entropy can be mapped into discrete numeric values by means of the method proposed by Gao and Vogel (2010) as following: $[0, 1.0) \mapsto 0$, $[1.0, 2.0) \mapsto 1$, $[2.0, 3.5) \mapsto 2$, $[3.5, 5.0) \mapsto 4$, $[5.0, 7.0) \mapsto 5$, $[7.0, +\infty) \mapsto 6$. The template is similar to character feature template, and forward conditional entropy template is in accordance with the backward one. Here, the forward conditional entropy feature templates are given: $H_f(C_{-1})$, $H_f(C_0)$, $H_f(C_1)$, $H_f(C_{-1})H_f(C_0)$, $H_f(C_0)H_f(C_1)$, $H_f(C_{-1})H_f(C_1)$.

2.3 Lexical Feature

Appropriately using of lexical feature has shown some improvement in Segmentation, and hence we adopt the definition of lexical feature from (Gao and Vogel, 2010). Feature $L_{begin}(C)$ represents the maximum length of words begin with character C in the lexicon via forward maximum matching from character C in the current sentence, and $L_{end}(C)$ represents the maximum length of words end with character C in the lexicon via backward maximum matching from character C. When processing forward and backward maximum matching, we only deal with the word with length equal or greater than 2, furthermore, the lexical feature value will be 0 where matching failed. Especially when feature value is equal or greater than 6, we set these feature values to 6. We hope to increase the performance by using a large-scale cross-domain lexicon. Six feature templates are defined for

$L_{begin}(C)$: $L_{begin}(C_{-1})$, $L_{begin}(C_0)$, $L_{begin}(C_1)$, $L_{begin}(C_{-1})L_{begin}(C_0)$, $L_{begin}(C_0)L_{begin}(C_1)$ and $L_{begin}(C_{-1})L_{begin}(C_1)$. As six feature templates of $L_{end}(C)$ could be inferred from above.

2.4 Accessor variety feature

Accessor variety (AV) proposed by Feng et al. (2004) could be used to measure the possibility of whether a substring is a Chinese word. Zhao and Kit (2007) thought that the method above is agreed with the method proposed by Harris (1970), in which morpheme could be found in unfamiliar language. Zhao and Kit (2008)'s experiments proved that AV feature improves the performance of CRF segmentation model on dataset in Bakeoff-2003, Bakeoff-2005 and Bakeoff-2006 (Sproat and Emerson, 2003; Emerson, 2005; Levow, 2006) while achieved the best performance on close test in Bakeoff-2008 (Chen and Jin, 2008). Therefore in this paper, AV feature is employed and we make further improvement of the performance by making better use of AV feature method. As to substring s, AV feature is defined as follow:

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

in which $L_{av}(s)$ and $R_{av}(s)$ represent the number of different characters before s and after s respectively, while the sign in the begin or the end of sentence would be double counted.

How we use AV is similar to (Zhao and Kit, 2008; Yang et al., 2011), considering the AV value of substrings with length is equal or less than 5 in sentence and designing several feature templates accordingly. We used the formula below to discrete AV value of substring s:

$$f(s) = t, \text{ if } 2^t \leq AV(s) < 2^{t+1}$$

Discrete value t is regarded as the feature value. The difference between our method and the method above is that for substring s, we marked the feature value of s on the first character of s, not on every character of s. Representation of lexical feature mentioned in Section 2.3 was used for reference because we believed labeling this way could highlight boundary information between words. Table 2 shows the differences in detail. For instance, consider all the substring consist of 4 characters. In this case, we have a substring "在我心中 (in the middle of my heart)" with AV feature value $t = 1$. So that we updated

In	Accessor Variety Feature Selection										T
	1 char		2 char		3 char		4 char		5 char		
而	9	9	5	5	2	2	0	0	0	0	S
在	10	10	5	5	2	1	1	1	1	1	S
我	9	9	5	3	2	2	1	0	1	0	S
心	8	8	5	5	2	2	1	0	1	0	B
中	9	9	8	8	2	0	1	0	1	0	E
,	11	11	8	0	2	0	0	0	1	0	S

Table 2: Comparison of how to use AV feature

feature values in "4 char" row. The left row indicates that for every character "在", "我", "心", "中", feature values should be set to 1 according to method (Zhao and Kit, 2008; Yang et al., 2011). The right row indicates the feature values in our method, in which only the first character "在" is given feature value of 1. We created 6 templates similar to character feature template for each row in Table 2.

In order to prove the effectivity of improved AV feature in our method, we continued to use the experiment setting of (Zhao and Kit, 2008; Yang et al., 2011) and and had experiment on the dataset of Bakeoff-2005 (Emerson, 2005) and the simplified Chinese dataset of Bakeoff-2010 (Zhao and Liu, 2010). OldAV stands for their AV feature while our feature named as NewAV. 6 n-gram character features and character type feature mentioned in Section 2.1 were used in each experiment. Evaluation indicator F score equals $F = 2RP/(R+P)$, in which R is the recall and P stands for precision. After combined corresponding training and test dataset of Bakeoff-2005 together without segmentation marks, statistical AV features were created. Then the training corpus, unlabeled corpus and test corpus of Bakeoff-2010 were combined together without segmentation marks to count AV features. The experiment results in Table 3 indicates that our improvement in AV feature is effective due to the performance is better than other old methods. These experiment results were not post-processed so as to compare segmentation performance easily.

2.5 Post-processing

Post-processing aimed at handling segmentation error in English word, Arabic numeric string and URL. Faced with this situation, these characters should be regarded as a whole segment unit, but out system might make segmentation errors. In

Bakeoff-2005		AS	CityU	MSRA	PKU
Baseline	F	0.954	0.955	0.971	0.950
	R_{OOV}[1]	0.700	0.798	0.772	0.778
OldAV	F	0.957	0.961	0.973	0.952
	R_{OOV}	0.688	0.807	0.747	0.770
NewAV	F	0.957	0.964	0.973	0.954
	R_{OOV}	0.688	0.822	0.743	0.773
Bakeoff-2010		A	B	C	D
Baseline	F	0.921	0.93	0.918	0.953
	R_{OOV}	0.629	0.773	0.72	0.853
OldAV	F	0.933	0.94	0.935	0.956
	R_{OOV}	0.656	0.784	0.77	0.848
NewAV	F	0.935	0.945	0.936	0.956
	R_{OOV}	0.659	0.807	0.763	0.843

[1] Recall of out-of-vocabulary (OOV) words.

Table 3: Comparion experiment on AV feature, n-gram feature and character type feature were used for each experiment

Table 4 we have an example of URL segmented incorrectly, and raw represents the original sentence; result shows the result after segmentation; final stands for the result after post-processing. To deal with this kind of problem, we have to make sure that when we take gaps away from the segmented sentence, it should be in correspondences with original characters in sentence. Here is a quick procedure of how we restored URL segmentation error. First, we put the original sentence in a string; then saved the segmented result in to a list. Every element in the list is a word with subscript starts from 0.

1. Use regular expression to find the start and the end position of the original sentence. In case `http://t.cn/aBPxzO`, the start and end index is 4 and 22 respectively.

2. Accumulating word length in the word list from left to right, we can get the start index of URL is 2 and end index is 3 according to word list.

3. Combine the 2nd and 3rd word in the word list as one word.

English word and Arabic numeric string can be handled in the same way.

raw	点击网址http://t.cn/aBPxzO
result	点击 网址 http://t.cn/aB PxzO
final	点击 网址 http://t.cn/aBPxzO

Table 4: Post-processing of particular string (URL)

3 Improve The Segmentation Performance of New Words

The segmentation system that we described in Section 2 was not very stable when it comes to new words. New words with some sort of context can be segmented correctly while other context might lead to mistake. For example, the word "涅维拉济莫夫 (涅维拉济莫夫)" with context "文官涅维拉济 莫夫在起草一封贺信 (civil officer Nie Vilage is making a draft of congratulatory letter)" can be segmented correctly, but the sentence "于是涅维拉济莫夫开始绞尽脑汁 (hence Nie Vialge began to rack his brain)" was wrongly segmented. To solve this sort of problem, we tried to find these new words by rules, then added new words to the lexicon, re-calculated the lexical features of test corpora, segmented test corpora again in the end. Let 's mark the lexicon used for extracting lexical features when training segmentation model as $Lexicon_{train}$, and count the Bigram statistical information on segmented corpora of People's Daily 1998 and 2000 as PKU_{bigram} without smoothing. For the preliminary segmentation result, if word w meets the following conditions, we deemed w as a new word:

1. (w with length between 2 to 6) or (w with length greater than 6 and w is a foreign name at the same time (en dash • exists in w)),

2. w does not exist in $Lexicon_{train}$,

3. w is not a Chinese name,

4. w can not be the concatenation of w_{-1} and w_0 for $\forall(w_{-1}, w_0) \in PKU_{bigram}$.

We checked every word in result after segmentation so that we have a new version of new words list named $Lexicon_{test}$. If $Lexicon_{test}$ has two words with inclusion relation, we only reserved the word with longer length. Combine $Lexicon_{train}$ and $Lexicon_{test}$ together then we have a new word list named $Lexicon_{new}$. This new word list could be used for calculating lexical feature of the test corpora to update segmentation result.

Name	Features	Lexicon
Baseline	CF,CTF	None
Closed	CF,CTF,EF,AV	None
Open[1]	CF,CTF,EF,AV	Webdict
Refined[2]	CF,CTF,EF,AV	Webdict

[1] Webdict were used to calculate lexical feature for both testing and training.

[2] Webdict were used to calculate lexical feature for training, then the method mentioned in Section 3 was used for performance improvement.

Table 5: Feature combination: CF represents 6 ngram features of character, CTF represents character type feature, EF represents conditional entropy feature and AV represents Accessor variety feature

4 Experiment

In order to prove the performance of our method, we considered four kinds of feature combination demonstrated in Table 5, in which *Closed* means closed test, *Open* means open test in which we used a cross-domain lexicon — Webdict[3]. *Refined* represents that we added new words' process proposed in Section 3 on the basis of *Open*. For *Refined*, we needed corpora to create statistical Bigram information and a lexicon for training. Because of the limited scale of labeled data and we have merely sufficient simplified Chinese training data and lexicon, we didn't process both the AS and CityU of Bakeoff-2005 for *Refined*. All the experiments in this section were linked to post-processing mentioned in Section 2.5. We tested our system on Bakeoff-2005 and Bakeoff-2010 dataset with major measure index F score.

Table 6 shows the experiment result on Bakeoff-2005. When computing conditional entropy feature and AV feature, corresponding test corpus and training corpus should be mixed together, wiping off of the segmentation boundaries before the feature extraction. "Best closed" indicates the best result on closed test of Bakeoff-2005 and "Best open" stands for the best open test of official outcome. Our closed test outcome fully exceeded the "Best closed", and open test outcome exist a slight achieves a slightly lower F scores compared with "Best open" only on PKU test set, which might due to the deficiency of corpora and might be improved only by enlarging the amount of training corpora.

[3] https://github.com/ling0322/webdict

Bakeoff-2005		AS	CityU	MSRA	PKU
Best closed	F	0.952	0.943	0.964	0.95
	R_{OOV}	0.696	0.698	0.717	0.636
Baseline	F	0.955	0.956	0.971	0.950
	R_{OOV}	0.708	0.806	0.772	0.779
Closed	F	0.957	0.963	0.974	0.954
	R_{OOV}	0.705	0.817	0.739	0.770
Open	F	0.958	0.965	0.977	**0.962**
	R_{OOV}	0.700	0.811	0.751	0.765
Refined	F	-	-	**0.976**	0.962
	R_{OOV}	-	-	0.751	0.766
Best open	F	0.956	0.962	0.972	0.969
	R_{OOV}	0.684	0.806	0.59	0.838

Table 6: Test result on Bakeoff-2005 dataset

Table 7 shows the test result on Bakeoff-2010 simplified Chinese dataset. When computing conditional entropy feature and AV feature, we needed to combine all of the simplified Chinese corpus together without segmentation boundaries of Bakeoff-2010 corpora to create the statistical feature values. "Best closed" and "Best open" shows the best result on official closed test and open test. Our closed test result on test set A differs greatly from "Best closed", yet the result is closer to "Best closed" on other test sets. The performance on *Closed* improves a lot comparing to the baseline. In addition, our method exceeded "Best open" on dataset C, D in open test, while slightly poorer results than the best on dataset A and B but the differences are not significant.

From the *Refined* results of both Table 6 and Table 7, we can observe that our strategy on detecting new words provide improvements over all the R_{OOV} compared to all the Open system in general. Meanwhile, our *Refined* model provide more balanced F scores among all the dataset.

It is proved on two Bakeoff datasets that our *Open* feature combination and *Refined* feature combination are effective. On account of lacking training corpus of this Bakeoff, Open data test is required. Hence we used *Open* and *Refined* feature combination in Table 5. With purpose of making model to be more cross-domain adaptive, we made use of a large number of unlabeled corpora to extract conditional entropy feature and AV feature. Web crawler was used to get totally 1.5G corpora in 5 domains, including finance, literature,

Bakeoff-2010		A	B	C	D
Best closed	F	0.946	0.951	0.939	0.959
	R_{OOV}	0.816	0.827	0.75	0.827
Baseline	F	0.921	0.933	0.918	0.954
	R_{OOV}	0.629	0.781	0.72	0.86
Closed	F	0.935	0.949	0.936	0.958
	R_{OOV}	0.658	0.819	0.763	0.853
Open	F	0.95	0.949	0.943	0.963
	R_{OOV}	0.509	0.766	0.571	0.879
Refined	F	0.95	0.949	0.943	0.963
	R_{OOV}	**0.519**	**0.768**	**0.572**	**0.883**
Best open	F	0.955	0.95	0.938	0.96
	R_{OOV}	0.655	0.82	0.768	0.847

Table 7: Test result on Bakeoff-2010 dataset

news, microblog and novel. The data we used is explained as followed:

- PKU-Corpus: labeled People's Daily corpus in year 1998 and 2000.
- PKU-Raw: PKU-Corpus without segmentation boundaries.
- Web-Corpus: combines all the unlabeled corpora from web crawler.
- Sample-Corpus: randomly select 15% from Web-Corpus.
- Entropy-Corpus: PKU-Raw together with Web-Corpus.
- AV-Corpus: PKU-Raw together with Sample-Corpus.

Finally we used PKU-Corpus as training data, and extracted from Entropy-Corpus to extract conditional entropy feature while making use of AV-Corpus to extract AV features, together with character feature and character type feature to train CRF word segmentation model. Our results on this bakeoff are showed in Table 8, which achieves a competitive F score of 0.9730. From this table, we can catch that *Refined* feature combination outperforms *Open*, which further confirms that the new word detection is critical for cross-domain Chinese segmentation.

5 Conclusion

In this paper we attempted to implement a word segmentation system with the ability to handle the situation of cross domain. We combined supervised and unsupervised global features together and improved the ability to recognize OOV through adding cross-domain lexical feature. Discovering new words from target test set then recomputing the lexical feature to refine the segmentation results makes the model more domain adaptive.

	Precision	Recall	F Score
Open	0.9673	0.9776	0.9724
Refined	0.9681	0.9779	0.9730

Table 8: Results on Bakeoff-2014 dataset

Yet our system still have many deficiencies which can be improved from three aspects. First of all, we only used one kind of unsupervised feature and there might be other unsupervised features or feature combination that could achieve better performance. Next, we coined all the feature into one set of template mainly due to its simplicity in practice. However, there might exist a more fitting feature template for different features. At last, our rule-based method to discover new words could be changed into automatic discovery.

Acknowledgments

This work was partially supported by National Natural Science Foundation of China Grant NO.61202248. Many thanks to our colleagues participating in this work. We also thank Huiming Duan and Zhifang Sui for their excellent organization.

References

Xiao Chen and Guangjin Jin. 2008. The fourth international chinese language processing bakeoff: Chinese word segmentation named entity recognition and chinese pos tagging. In *Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing*.

Thomas Emerson. 2005. The second international chinese word segmentation bakeoff. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, volume 133.

Haodi Feng, Kang Chen, Xiaotie Deng, and Weimin Zheng. 2004. Accessor variety criteria for chinese word extraction. *Computational Linguistics*, 30(1):75–93.

Qin Gao and Stephan Vogel. 2010. A multi-layer chinese word segmentation system optimized for out-of-domain tasks. In *Proceedings of CIPS-SIGHAN*

Joint Conference on Chinese Language Processing (CLP2010), pages 210–215.

Zellig S Harris. 1970. *Morpheme boundaries within words: Report on a computer test*. Springer.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Gina-Anne Levow. 2006. The third international chinese language processing bakeoff: Word segmentation and named entity recognition. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 108–117, Sydney, Australia, July. Association for Computational Linguistics.

Jin Kiat Low, Hwee Tou Ng, and Wenyuan Guo. 2005. A maximum entropy approach to chinese word segmentation. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, volume 1612164.

Fuchun Peng, Fangfang Feng, and Andrew McCallum. 2004. Chinese segmentation and new word detection using conditional random fields. In *Proceedings of the 20th international conference on Computational Linguistics*, page 562. Association for Computational Linguistics.

Richard Sproat and Thomas Emerson. 2003. The first international chinese word segmentation bakeoff. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, pages 133–143, Sapporo, Japan, July. Association for Computational Linguistics.

Nianwen Xue et al. 2003. Chinese word segmentation as character tagging. *Computational Linguistics and Chinese Language Processing*, 8(1):29–48.

Ting-hao Yang, Tian-Jian Jiang, Chan-hung Kuo, Richard Tzong-han Tsai, and Wen-lian Hsu. 2011. Unsupervised overlapping feature selection for conditional random fields learning in chinese word segmentation. In *Proceedings of the 23rd Conference on Computational Linguistics and Speech Processing*, pages 109–122. Association for Computational Linguistics.

Hai Zhao and Chunyu Kit. 2007. Incorporating global information into supervised learning for chinese word segmentation. In *10th Conference of the Pacific Association for Computational Linguistics*, pages 66–74.

Hai Zhao and Chunyu Kit. 2008. Unsupervised segmentation helps supervised learning of character tagging for word segmentation and named entity recognition. In *The Sixth SIGHAN Workshop on Chinese Language Processing*, pages 106–111.

Hai Zhao and Chunyu Kit. 2011. Integrating unsupervised and supervised word segmentation: The role of goodness measures. *Information Sciences*, 181(1):163–183.

Hongmei Zhao and Qun Liu. 2010. The cips-sighan clp 2010 chinese word segmentation bakeoff. In *Proceedings of the First CPS-SIGHAN Joint Conference on Chinese Language Processing*, pages 199–209.

Hai Zhao, Chang-Ning Huang, and Mu Li. 2006a. An improved chinese word segmentation system with conditional random field. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, volume 1082117. Sydney: July.

Hai Zhao, Chang-Ning Huang, Mu Li, and Bao-Liang Lu. 2006b. Effective tag set selection in chinese word segmentation via conditional random field modeling. In *Proceedings of PACLIC*, volume 20, pages 87–94.