# Unsupervised Segmentation Helps Supervised Learning of Character Tagging for Word Segmentation and Named Entity Recognition

**Hai Zhao and Chunyu Kit**
Department of Chinese, Translation and Linguistics
City University of Hong Kong
Tat Chee Ave., Kowloon, Hong Kong
Email: {haizhao, ctckit}@cityu.edu.hk

## Abstract

This paper describes a novel character tagging approach to Chinese word segmentation and named entity recognition (NER) for our participation in Bakeoff-4.[1] It integrates unsupervised segmentation and conditional random fields (CRFs) learning successfully, using similar character tags and feature templates for both word segmentation and NER. It ranks at the top in all closed tests of word segmentation and gives promising results for all closed and open NER tasks in the Bakeoff. Tag set selection and unsupervised segmentation play a critical role in this success.

## 1 Introduction

A number of recent studies show that character sequence labeling is a simple but effective formulation of Chinese word segmentation and name entity recognition for machine learning (Xue, 2003; Low et al., 2005; Zhao et al., 2006a; Chen et al., 2006). Character tagging becomes a prevailing technique for this kind of labeling task for Chinese language processing, following the current trend of applying machine learning as a core technology in the field of natural language processing. In particular, when a full-fledged general-purpose sequence learning model such as CRFs is involved, the only work to do for a given application is to identify an ideal set of features and hyperparameters for the purpose of achieving the best learning model that we can with available training data. Our work in this aspect provides a solid foundation for applying an unsupervised segmentation criterion to enrich the supervised CRFs learning for further performance enhancement on both word segmentation and NER.

This paper is intended to present the research for our participation in Bakeoff-4, with a highlight on our strategy to select character tags and feature templates for CRFs learning. Particularly worth mentioning is the simplicity of our system in contrast to its success. The rest of the paper is organized as follows. The next section presents the technical details of the system and Section 3 its evaluation results. Section 4 looks into a few issues concerning character tag set, unsupervised segmentation, and available name entities (NEs) as features for open NER test. Section 5 concludes the paper.

## 2 System Description

Following our previous work (Zhao et al., 2006a; Zhao et al., 2006b; Zhao and Kit, 2007), we continue to apply the order-1 linear chain CRFs (Lafferty et al., 2001) as our learning model for Bakeoff-4. Specifically, we use its implementation CRF++ by Taku Kudo[2] freely available for research purpose. We opt for a similar set of character tags and feature templates for both word segmentation and NER.

In addition, two key techniques that we have explored in our previous work are applied. One is to introduce more tags in the hope of utilizing more precise contextual information to achieve more pre-

[1] The Fourth International Chinese Language Processing Bakeoff & the First CIPS Chinese Language Processing Evaluation, at http://www.china-language.gov.cn/bakeoff08/bakeoff-08_basic.html.

[2] http://crfpp.sourceforge.net/

Table 1: An exmaple of NE tagging for a character sequence

| Characters | 爱 | 乐 | 乐 | 团 | 到 | 沪 | 访 | 问 | 演 | 出 |
|---|---|---|---|---|---|---|---|---|---|---|
| Tags | B-ORG | $B_2$-ORG | $B_3$-ORG | E-ORG | O | S-LOC | O | O | O | O |

Table 2: Illustration of character tagging

| Word length | Tag sequence for a word |
|---|---|
| 1 | S |
| 2 | B E |
| 3 | B $B_2$ E |
| 4 | B $B_2$ $B_3$ E |
| 5 | B $B_2$ $B_3$ M E |
| $\geqslant 6$ | B $B_2$ $B_3$ M $\cdots$ M E |

cise labeling results. This also optimizes the active features for the CRFs training. The other is to integrate the unsupervised segmentation outputs into CRFs as features. It assumes no word boundary information in the training and test corpora for NER.

### 2.1 Tag Set

Our previous work shows that a 6-tag set enables the CRFs learning of character tagging to achieve a better segmentation performance than others (Zhao et al., 2006a; Zhao et al., 2006b). So we keep using this tag set for Bakeoff-4. Its six tags are B, $B_2$, $B_3$, M, E and S. Table 2 illustrates how characters in words of various lengths are tagged with this tag set.

For NER, we need to tell apart three types of NEs, namely, *person*, *location* and *organization* names. Correspondingly, the six tags are also adapted for characters in these NEs but distinguished by the suffixes -PER, -LOC and -ORG. For example, a character in a person name may be tagged with either B-PER, $B_2$-PER, $B_3$-PER, M-PER, E-PER, or S-PER. Plus an additional tag "O" for none NE characters, altogether we have 19 tags for NER. An example of NE tagging is illustrated in Table 1.

### 2.2 Feature Templates

We use not only a similar tag set but also the same set of feature templates for both the word segmentation and NER closed tests in Bakeoff-4. Six n-gram templates, namely, $C_{-1}$, $C_0$, $C_1$, $C_{-1}C_0$, $C_0C_1$, $C_{-1}C_1$, are selected as features, where C stands for a character and the subscripts -1, 0 and 1 for the previous, current and next character, respectively.

In addition to these n-gram features, unsupervised segmentation outputs are also used as features, for the purpose of providing more word boundary information via global statistics derived from all unlabeled texts of the training and test corpora. The basic idea is to inform a supervised leaner of which substrings are recognized as word candidates by a given unsupervised segmentation criterion and how likely they are to be true words in terms of that criterion (Zhao and Kit, 2007; Kit and Zhao, 2007).

We adopt the *accessor variety* (AV) (Feng et al., 2004a; Feng et al., 2004b) as our unsupervised segmentation criterion. It formulates an idea similar to linguist Harris' (1955; 1970) for segmenting utterances of an unfamiliar language into morphemes to facilitate word extraction from Chinese raw texts. It is found more effective than other criteria in supporting CRFs learning of character tagging for word segmentation (Zhao and Kit, 2007). The AV value of a substring $s$ is defined as

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

where the left and right AV values $L_{av}(s)$ and $R_{av}(s)$ are defined, respectively, as the numbers of its distinct predecessor and successor characters.

In our work, AV values for word candidates are derived from an unlabeled corpus by substring counting, which can be efficiently carried out with the aid of the *suffix array* representation (Manber and Myers, 1993; Kit and Wilks, 1998). Heuristic rules are applied in Feng et al.'s work to remove insignificant substrings. We do not use any such rule.

Multiple feature templates are used to represent word candidates of various lengths identified by the AV criterion. For the sake of efficiency, all candidates longer than five characters are given up. To accommodate the word likelihood information, we need to extend the feature representation in (Zhao and Kit, 2007), where only the candidate substrings are used as features for word segmentation. Formally put, our new feature function for a word can-

Table 3: Training corpora for assistant learners

| Track | CityU NER | MSRA NER |
|---|---|---|
| Ass. Seg. | CityU (Bakeoff-1 to 4) | MSRA (Bakeoff-2) |
| ANER-1 | CityU(Bakeoff-3) | CityU(Bakeoff-3) |
| ANER-2 | MSRA(Bakeoff-3) | CityU(Bakeoff-4) |

Table 4: NE lists from Chinese Wikipedia

| Category | Number |
|---|---|
| Place name suffix | 85 |
| Chinese place name | 6,367 |
| Foreign place name | 1,626 |
| Chinese family name | 573 |
| Most common Chinese family name | 109 |
| Foreign name | 2,591 |
| Chinese university | 515 |

Table 5: Segmentation results for previous Bakeoffs

| Bakeoff-1 | | AS | CityU | CTB | PKU |
|---|---|---|---|---|---|
| –AV | F | .9727 | .9473 | .8720 | .9558 |
| | $R_{OOV}$[a] | .7907 | .7576 | .7022 | .7078 |
| +AV | F | .9725 | .9554 | .9023 | .9612 |
| | $R_{OOV}$ | .7597 | .7616 | .7502 | .7208 |
| **Bakeoff-2** | | **AS** | **CityU** | **MSRA** | **PKU** |
| –AV | F | .9534 | .9476 | .9735 | .9515 |
| | $R_{OOV}$ | .6812 | .6920 | .7496 | .6720 |
| +AV | F | .9570 | .9610 | .9758 | .9540 |
| | $R_{OOV}$ | .6993 | .7540 | .7446 | .6765 |
| **Bakeoff-3** | | **AS** | **CityU** | **CTB** | **MSRA** |
| –AV | F | .9538 | .9691 | .9322 | .9608 |
| | $R_{OOV}$ | .6699 | .7815 | .7095 | .6658 |
| +AV | F | .9586 | .9747 | .9431 | .9660 |
| | $R_{OOV}$ | .6935 | .8005 | .7608 | .6620 |

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

didate $s$ with a score $AV(s)$ is defined as

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

where $t$ is an integer to logarithmize the score. This is to alleviate the sparse data problem by narrowing down the feature representation involved. Note that $t$ is used as a feature value rather than a parameter for the CRFs training in our system. For an overlap character of several word candidates, we only choose the one with the greatest AV score to activate the above feature function for that character. It is in this way that the unsupervised segmentation outcomes are fit into the CRFs learning.

### 2.3 Features for Open NER

Three extra groups of feature template are used for the open NER beyond those for the closed.

The first group includes three segmentation feature templates. One is character type feature template $T(C_{-1})T(C_0)T(C_1)$, where $T(C)$ is the type of character $C$. For this, five character types are defined, namdely, number, foreign letter, punctuation, date and time, and others. The other two are generated respectively by two assistant segmenters (Zhao et al., 2006a), a maximal matching segmenter based on a dictionary from Peking University[3] and a CRFs segmenter using the 6-tag set and the six n-gram feature templates for training.

The second group comes from the outputs of two assistant NE recognizers (ANERs), both trained with a corresponding 6-tag set and the same six n-gram feature templates. They share a similar feature representation as the assistant segmenter. Table 3 lists the training corpora for the assistant CRFs segmenter and the ANERs for various open NER tests.

The third group consists of feature templates generated from seven NE lists acquired from Chinese Wikipedia.[4] The categories and numbers of these NE items are summarized in Table 4.

## 3 Evaluation Results

The performance of both word segmentation and NER is measured in terms of the F-measure $F = 2RP/(R + P)$, where $R$ and $P$ are the recall and precision of segmentation or NER.

We tested the techniques described above with the previous Bakeoffs' data[5] (Sproat and Emerson, 2003; Emerson, 2005; Levow, 2006). The evaluation results for the closed tests of word segmentation are reported in Table 5 and those for the NER on two corpora of Bakeoff-3 are in the upper part of Table 7. '+/–AV' indicates whether AV features are applied.

For Bakeoff-4, we participated in all five closed tracks of word segmentation, namely, CityU, CKIP, CTB, NCC, and SXU, and in all closed and open NER tracks of CityU and MSRA.[6] The evaluation

[3]It consists of about 108K words of one to four character-slong, available at http://ccl.pku.edu.cn/doubtfire/Course/Chinese%20Information%20Processing/Source_Code/Chapter_8/Lexicon_full.zip.

[4]http://zh.wikipedia.org/wiki/首页

[5]http://www.sighan.org

[6]We declare that our team has never been exposed to the

Table 6: Evaluation results of word segmentation on Bakeoff-4 data sets

| Feature | Data | F | P | R | $F_{IV}$[a] | $P_{IV}$ | $R_{IV}$ | $F_{OOV}$ | $P_{OOV}$ | $R_{OOV}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| –AV ($n$-gram) | CityU | .9426 | .9410 | .9441 | .9640 | .9636 | .9645 | .7063 | .6960 | .7168 |
| | CKIP | .9421 | .9387 | .9454 | .9607 | .9581 | .9633 | .7113 | .7013 | .7216 |
| | CTB | **.9634** | .9641 | .9627 | .9738 | .9761 | .9715 | .7924 | .7719 | .8141 |
| | NCC | .9333 | .9356 | .9311 | .9536 | .9612 | .9461 | .5678 | .5182 | .6280 |
| | SXU | .9552 | .9559 | .9544 | .9721 | .9767 | .9675 | .6640 | .6223 | .7116 |
| +AV*[b] | CityU | .9510 | .9493 | .9526 | .9667 | .9626 | .9708 | .7698 | .7912 | .7495 |
| | CKIP | .9470 | .9440 | .9501 | .9623 | .9577 | .9669 | .7524 | .7649 | .7404 |
| | CTB | **.9589** | .9596 | .9583 | .9697 | .9704 | .9691 | .7745 | .7761 | .7730 |
| | NCC | .9405 | .9407 | .9402 | .9573 | .9583 | .9562 | .6080 | .5984 | .6179 |
| | SXU | .9623 | .9625 | .9622 | .9752 | .9764 | .9740 | .7292 | .7159 | .7429 |

[a]F-score for in-vocabulary (IV) words.
[b]Henceforth the official evaluation results in Bakeoff-4 are marked with "*".

Table 7: NER evaluation results

| Track | Setting | $F_{PER}$ | $F_{LOC}$ | $F_{ORG}$ | $F_{NE}$ |
|---|---|---|---|---|---|
| Bakeoff-3 | | | | | |
| CityU | –AV | .8849 | .9219 | .7905 | .8807 |
| | +AV | .9063 | .9281 | .7981 | .8918 |
| MSRA | –AV | .7851 | .9072 | .8242 | .8525 |
| | +AV | .8171 | .9139 | .8164 | .8630 |
| Bakeoff-4 | | | | | |
| CityU | –AV | .8222 | .8682 | .6801 | .8092 |
| | +AV* | .8362 | .8677 | .6852 | .8152 |
| | $Open_1$* | .9125 | .9216 | .7862 | .8869 |
| | $Open_2$ | .9137 | .9214 | .7853 | .8870 |
| MSRA | –AV | .9221 | .9193 | .8367 | .8968 |
| | +AV* | .9319 | .9219 | .8414 | .9020 |
| | Open* | 1.000 | .9960 | .9920 | .9958 |
| | $Open_1$[a] | .9710 | .9601 | .9352 | .9558 |
| | $Open_2$[b] | .9699 | .9581 | .9359 | .9548 |

[a]For our official submission to Bakeoff-4, we also used an ANER trained on the MSRA NER training corpus of Bakeoff-3. This makes our official evaluation results extremely high but trivial, for a part of this corpus is used as the MSRA NER test corpus for Bakeoff-4. Presented here are the results without using this ANER.
[b]$Open_2$ is the result of $Open_1$ using no NE list feature.

results of word segmentation and NER for our system are presented in Tables 6 and 7, respectively.

For the purpose of comparison, the word segmentation performance of our system on Bakeoff-4 data using the 2- and 4-tag sets and the best corresponding n-gram feature templates as in (Tsai et al., 2006; Low et al., 2005) are presented in Table 8.[7] This comparison reconfirms the conclusion in (Zhao et al., 2006b) about tag set selection for character tagging for word segmentation that the 6-tag set is more effective than others, each with its own best corresponding feature template set.

Table 8: Segmentation F-scores by different tag sets

| AV | Tags | CityU | CKIP | CTB | NCC | SXU |
|---|---|---|---|---|---|---|
| – | 2 | .9303 | .9277 | .9434 | .9198 | .9454 |
| | 4 | .9370 | .9348 | .9481 | .9280 | .9512 |
| | 6 | .9426 | .9421 | .9634 | .9333 | .9552 |
| + | 2 | .9382 | .9319 | .9451 | .9239 | .9485 |
| | 4 | .9482 | .9423 | .9527 | .9356 | .9593 |
| | 6 | .9510 | .9470 | .9589 | .9405 | .9623 |

## 4 Discussion

### 4.1 Tag Set and Computational Cost

Using more labels in CRFs learning is expected to bring in performance enhancement. Inevitably, however, it also leads to a huge rise of computational cost for model training. We conducted a series of experiments to study the computational cost of CRFs training with different tag sets using Bakeoff-3 data. The experimental results are given in Table 9, showing that the 6-tag set costs nearly twice as much time as the 4-tag set and about three times as the 2-tag set. Fortunately, its memory cost with the six n-gram feature templates remains very close to that of the 2- and 4-tag sets with the n-gram feature template sets from (Tsai et al., 2006; Xue, 2003).

However, a 2-tag set is popular in use for word segmentation and NER for the reason that CRFs training is very computationally expensive and a large tag set would make the situation worse. Cer-

CityU data sets in any other situation than the Bakeoff.

[7]The templates for the 2-tag set, adopted from (Tsai et al., 2006), include $C_{-2}$, $C_{-1}$, $C_0$, $C_1$, $C_{-3}C_{-1}$, $C_{-2}C_0$, $C_{-2}C_{-1}$, $C_{-1}C_0$, $C_{-1}C_1$ and $C_0C_1$. Those for the 4-tag set, adopted from (Xue, 2003) and (Low et al., 2005), include $C_{-2}$, $C_{-1}$, $C_0$, $C_1$, $C_2$, $C_{-2}C_{-1}$, $C_{-1}C_0$, $C_{-1}C_1$, $C_0C_1$and $C_1C_2$.

Table 9: Comparison of computational cost

| Tags | Templates | AS | CityU | CTB | MSRA |
|---|---|---|---|---|---|
| Training time (Minutes) | | | | | |
| 2 | Tsai | 112 | 52 | 16 | 35 |
| 4 | Xue | 206 | 79 | 28 | 73 |
| 6 | Zhao | 402 | 146 | 47 | 117 |
| Feature numbers ($\times 10^6$) | | | | | |
| 2 | Tsai | 13.2 | 7.3 | 3.1 | 5.5 |
| 4 | Xue | 16.1 | 9.0 | 3.9 | 6.8 |
| 6 | Zhao | 15.6 | 8.8 | 3.8 | 6.6 |
| Memory cost (Giga bytes) | | | | | |
| 2 | Tsai | 5.4 | 2.4 | 0.9 | 1.8 |
| 4 | Xue | 6.6 | 2.8 | 1.1 | 2.2 |
| 6 | Zhao | 6.4 | 2.7 | 1.0 | 2.1 |

tainly, a possible way out of this problem is the computer hardware advancement, which is predicted by Moore's Law (Moore, 1965) to be improving at an exponential rate in general, including processing speed and memory capacity. Specifically, CPU can be made twice faster every other year or even 18 months. It is predictable that computational cost will not be a problem for CRFs training soon, and the advantages of using a larger tag set as in our approach will be shared by more others.

### 4.2 Unsupervised Segmentation Features

Our evaluation results show that the unsupervised segmentation features bring in performance improvement on both word segmentation and NER for all tracks except CTB segmentation, as highlighted in Table 6. We are unable explain this yet, and can only attribute it to some unique text characteristics of the CTB segmented corpus. An unsupervised segmentation criterion provides a kind of global information over the whole text of a corpus (Zhao and Kit, 2007). Its effectiveness is certainly sensitive to text characteristics.

Quite a number of other unsupervised segmentation criteria are available for word discovery in unlabeled texts, e.g., boundary entropy (Tung and Lee, 1994; Chang and Su, 1997; Huang and Powers, 2003; Jin and Tanaka-Ishii, 2006) and description-length-gain (DLG) (Kit and Wilks, 1999). We found that among them AV could help the CRFs model to achieve a better performance than others, although the overall unsupervised segmentation by DLG was slightly better than that by AV. Combining any two of these criteria did not give any further performance improvement. This is why we have opted for AV for Bakeoff-4.

### 4.3 NE List Features for Open NER

We realize that the NE lists available to us are far from sufficient for coping with all NEs in Bakeoff-4. It is reasonable that using richer external NE lists gives a better NER performance in many cases (Zhang et al., 2006). Surprisingly, however, the NE list features used in our NER do not lead to any significant performance improvement, according to the evaluation results in Table 7. This is certainly another issue for our further inspection.

## 5 Conclusion

Without doubt our achievements in Bakeoff-4 owes not only to the careful selection of character tag set and feature templates for exerting the strength of CRFs learning but also to the effectiveness of our unsupervised segmentation approach. It is for the sake of simplicity that similar sets of character tags and feature templates are applied to two distinctive labeling tasks, word segmentation and NER. Relying on little preprocessing and postprocessing, our system simply follows the plain training and test routines of machine learning practice with the CRFs model and achieves the best or nearly the best results for all tracks of Bakeoff-4 in which we participated. Simple is beautiful, as Albert Einstein said, "Everything should be made as simple as possible, but not one bit simpler." Our evaluation results also provide evidence that simple can be powerful too.

## Acknowledgements

The research described in this paper was supported by the Research Grants Council of Hong Kong S.A.R., China, through the CERG grant 9040861 (CityU 1318/03H) and by City University of Hong Kong through the Strategic Research Grant 7002037. Dr. Hai Zhao was supported by a Postdoctoral Research Fellowship in the Department of Chinese, Translation and Linguistics, City University of Hong Kong.

## References

Jing-Shin Chang and Keh-Yih Su. 1997. An unsupervised iterative method for Chinese new lexicon ex-

traction. *Computational Linguistics and Chinese Language Processing*, 2(2):97–148.

Wenliang Chen, Yujie Zhang, and Hitoshi Isahara. 2006. Chinese named entity recognition with conditional random fields. In *SIGHAN-5*, pages 118–121, Sydney, Australia, July 22-23.

Thomas Emerson. 2005. The second international Chinese word segmentation bakeoff. In *SIGHAN-4*, pages 123–133, Jeju Island, Korea, October 14-15.

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

Haodi Feng, Kang Chen, Chunyu Kit, and Xiaotie Deng. 2004b. Unsupervised segmentation of Chinese corpus using accessor variety. In *First International Joint Conference on Natural Language Processing* (IJCNLP-04), pages 255–261, Sanya, Hainan Island, China, March 22-24. Also in K. Y. Su, J. Tsujii, J. H. Lee & O. Y. Kwong (eds.), *Natural Language Processing - IJCNLP 2004*, LNAI 3248, pages 694-703. Springer.

Zellig Sabbetai Harris. 1955. From phoneme to morpheme. *Language*, 31(2):190–222.

Zellig Sabbetai Harris. 1970. Morpheme boundaries within words. In *Papers in Structural and Transformational Linguistics*, page 68 - 77.

Jin Hu Huang and David Powers. 2003. Chinese word segmentation based on contextual entropy. In Dong Hong Ji and Kim-Ten Lua, editors, *PACLIC - 17*, pages 152–158, Sentosa, Singapore, October, 1-3. COLIPS Publication.

Zhihui Jin and Kumiko Tanaka-Ishii. 2006. Unsupervised segmentation of Chinese text by use of branching entropy. In *COLING/ACL–2006*, pages 428–435, Sidney, Australia, July 17-21.

Chunyu Kit and Yorick Wilks. 1998. The virtual corpus approach to deriving n-gram statistics from large scale corpora. In Changning Huang, editor, *Proceedings of 1998 International Conference on Chinese Information Processing Conference*, pages 223–229, Beijing, Nov. 18-20.

Chunyu Kit and Yorick Wilks. 1999. Unsupervised learning of word boundary with description length gain. In M. Osborne and E. T. K. Sang, editors, *CoNLL-99*, pages 1–6, Bergen, Norway.

Chunyu Kit and Hai Zhao. 2007. Improving Chinese word segmentation with description length gain. In *2007 International Conference on Artificial Intelligence* (ICAI'07), Las Vegas, June 25-28.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML'2001*, pages 282–289, San Francisco, CA.

Gina-Anne Levow. 2006. The third international Chinese language processing bakeoff: Word segmentation and named entity recognition. In *SIGHAN-5*, pages 108–117, Sydney, Australia, July 22-23.

Jin Kiat Low, Hwee Tou Ng, and Wenyuan Guo. 2005. A maximum entropy approach to Chinese word segmentation. In *SIGHAN-4*, pages 161–164, Jeju Island, Korea, October 14-15.

Udi Manber and Gene Myers. 1993. Suffix arrays: A new method for on-line string searches. *SIAM Journal on Computing*, 22(5):935–948.

Gordon E. Moore. 1965. Cramming more components onto integrated circuits. *Electronics*, 3(8), April 19.

Richard Sproat and Thomas Emerson. 2003. The first international Chinese word segmentation bakeoff. In *SIGHAN-2*, pages 133–143, Sapporo, Japan.

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. In *SIGHAN-5*, pages 108–117, Sydney, Australia, July 22-23.

Cheng-Huang Tung and His-Jian Lee. 1994. Identification of unknown words from corpus. *Computational Proceedings of Chinese and Oriental Languages*, 8:131–145.

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

Suxiang Zhang, Ying Qin, Juan Wen, and Xiaojie Wang. 2006. Word segmentation and named entity recognition for SIGHAN Bakeoff3. In *SIGHAN-5*, pages 158–161, Sydney, Australia, July 22-23.

Hai Zhao and Chunyu Kit. 2007. Incorporating global information into supervised learning for Chinese word segmentation. In *PACLING-2007*, pages 66–74, Melbourne, Australia, September 19-21.

Hai Zhao, Chang-Ning Huang, and Mu Li. 2006a. An improved Chinese word segmentation system with conditional random field. In *SIGHAN-5*, pages 162–165, Sydney, Australia, July 22-23.

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 *PACLIC-20*, pages 87–94, Wuhan, China, November 1-3.