Improving Chinese Word Segmentation with Description Length Gain

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

Abstract *Supervised and unsupervised learning has seldom joined with and thus lend strength to each other in the field of Chinese word segmentation (CWS). This paper presents a novel approach to CWS that utilizes description length gain (DLG), an empirical goodness measure for unsupervised word discovery, to enhance the segmentation performance of conditional random field (CRF) learning. Specifically, we attempt to integrate the lexical information acquired from the unsupervised DLG segmentation into the supervised CRF learning of character tagging for CWS. Our experimental results show that the CRF learning can be further improved on top of its state-of-the-art performance in the field by making good use of DLG information.*

Keywords: Chinese word segmentation, description length gain, conditional random fields

1 Introduction

The task of Chinese word segmentation (CWS) is to segment an input text into words. It is a special case of tokenization in natural language processing (NLP) shared by many other languages that have no explicit word delimiters such as spaces. Researchers in the field have been pursuing various machine learning approaches for further performance enhancement since Bakeoff-2003[1] [19]. Segmentation via character tagging is a simple but effective formulation of the problem suitable for various competitive supervised machine learning models [25, 13, 12, 23, 29].

So far, however, supervised and unsupervised learning has been working as two disjointed categories of techniques for CWS. The former relies on a pre-segmented corpus as training data or, at least, on a predefined lexicon; whereas the latter applies without such resources [1, 5, 9, 14, 6]. Sophisticated technologies have been developed in both categories. A very important issue to explore now is how the two can join together effectively, in particular, how the latter can be integrated into the former for performance enhancement.

Towards this goal, this paper is intended to explore the plausibility of integrating the unsupervised approach to word discovery by description length gain (DLG) [9, 7] into the conditional random fields (CRFs) model [10], a supervised learning approach popularly applied to CWS in recent years. Our experiments show that DLG information is effective in further enhancing the state-of-the-art performance of CRF model on CWS via character tagging.

The rest of the paper is organized as follows. The next section introduces the DLG measure for identifying word candidates in raw text input. Section 3 formulates the integration of DLG information into CRF learning of character tagging for CWS. Then, our experimental results will be presented in Section 4. Finally, the paper is concluded with a summary of our contribution in Section 5.

2 Description Length Gain

An unsupervised strategy for word segmentation has to follow some predefined criterion to recognize individual words. An early investigation in this direction using mutual information (MI) can be found in [20]. Many successive works applied MI and other statistical methods [3, 21, 5, 26, 27].

Description length gain (DLG), as proposed in [9, 7] to measure the compression effect of extracting a substring as a word from a corpus, is another empirical criterion for unsupervised word discovery.

[1]The First International Chinese Word Segmentation Bakeoff, http://www.sighan.org/bakeoff2003/.

Theoretically, it is rooted in the minimum description length (MDL) principle [15, 16]. A pilot study of its potentials to detect out-of-vocabulary (OOV) words for CWS is presented in [8].

The DLG from extracting all occurrences of a substring $x_i x_{i+1} ... x_j$ (also denoted as $x_{i..j}$) from a corpus $X = x_1 x_2 ... x_n$ (with a vocabulary V) as a word is defined as

$$DLG(x_{i..j}) = L(X) - L(X[r \rightarrow x_{i..j}] \oplus x_{i..j}) \quad (1)$$

where $X[r \rightarrow x_{i..j}]$ represents the resultant corpus from replacing all instances of $x_{i..j}$ with a new symbol r throughout X and $\oplus$ denotes the concatenation of two strings. $L(\cdot)$ is the empirical description length of a corpus in bits that can be estimated by the Shannon-Fano code or Huffman code as below, following classic information theory [18, 4].

$$L(X) \doteq -|X| \sum_{x \in V} \hat{p}(x) \log_2 \hat{p}(x) \quad (2)$$

where $|\cdot|$ denotes the length of a string and $\hat{p}(x)$ is x's frequency in X.

Unsupervised segmentation with DLG is to infer the optimal sequence of substrings with the greatest sum of DLG for an input. In principle, a substring with a negative DLG should not be given up in word discovery, for it can be one among the optimal sequence [9]. For the simplicity in integrating DLG into CRF modeling for CWS, however, we only consider substrings with a positive DLG value in this study.

3 CRF Modeling

3.1 CWS as Character Tagging

CWS was first formulated as character tagging in [25]. The basic idea is to carry out segmentation via tagging each character in terms of its position in a word. A MaxEnt model trained for such a character tagging task was first reported in [25].

Conditional random fields (CRFs) [10] are a statistical sequence modeling framework with a number of advantages over other popular models such as MaxEnt [17]. CRF learning was first applied to CWS in [13], treating CWS as a binary decision task for each Chinese character in the input: is it the beginning of a word?

The probability assigned to a label sequence y for a particular character sequence s by a CRF model is given by the equation below:

$$P_\lambda(y|s) = \frac{1}{Z(s)} \exp(\sum_{c \in C} \sum_k \lambda_k f_k(y_c, s, c)), \quad (3)$$

where λ_k and f_k are, respectively, the model parameter and feature function for feature k, c indexes into characters in the sequence being labeled, and the normalization term

$$Z(s) = \sum_y \exp(\sum_{c \in C} \sum_k \lambda_k f_k(y_c, s, c)). \quad (4)$$

Our implementation of character tagging for CWS uses the CRF++ package by Taku Kudo[2].

3.2 Tag Set and Feature Templates

It is shown in [30] that a 6-tag set enables CRF learning to achieve a better segmentation performance than other tag sets. Thus, we opt for this tag set, and its n-gram feature templates, as the baseline for evaluation. This allows us to compare our work with the state-of-the-art performance in the field.

A 2-tag set (namely, B vs. non-B) was used in [13] and a 4-tag set $\{B, M, E, S\}$ in [25], where the tags B, M and E are for the beginning, middle and end of a word, respectively, and S for a single character as a word. The 6-tag set is extended from the 4-tag set by extending B into B_1, B_2 and B_3 to accommodate more specific information about word prefix. Table 1 illustrates how words of different length are tagged with this tag set.

The feature templates for the baseline are listed in Table 2, with the subscript 0 for the current character. Five special character types are defined for template c, namely, numerical characters, temporal characters for date and time, foreign letters, punctuations, and other special characters. They provide critical latent information about word boundaries to support decision making concerning such characters.

3.3 Feature Templates from DLG

The basic idea of integrating the unsupervised DLG segmentation into a supervised leaner such as CRF for CWS is to inform the supervised leaner of the substrings identified by DLG as word candidates. Such candidates are marked with the 6 tags for CRF training in the same way as the true words in the pre-segmented training data.

Word candidates of different length are available as DLG feature templates. But, for the sake of efficiency, we only allow four specific feature templates generated by DLG, namely, the n-grams for $n = 2, 3, 4$ and 5. Unigrams are excluded, for our tagging cliques are exactly individual characters.

[2] http://chasen.org/ taku/software/CRF++/

Table 1: Tagging words of various length

Length	Tag sequence for a word
1	S
2	$B_1\ E$
3	$B_1\ B_2\ E$
4	$B_1\ B_2\ B_3\ E$
5	$B_1\ B_2\ B_3\ M\ E$
6	$B_1\ B_2\ B_3\ M\ M\ E$
≥7	$B_1\ B_2\ B_3\ M\ \cdots M\ E$

S $B_1\ E$ $B_1\ B_2\ B_3\ E$ $B_1\ B_2\ B_3\ M\ E$
她 来 自 澳 大 利 亚 新 南 威 尔 斯
She comes from New South Wales, Australia

Table 2: Baseline feature templates

Type		Feature templates
a	Unigram	C_{-1}, C_0, C_1
b	Bigram	$C_{-1}C_0$, C_0C_1, $C_{-1}C_1$
c	Special	$T_{-1}T_0T_1$

4 Experimental Results

To test the effectiveness of our approach, a number of word segmentation experiments are performed on all four corpora for Bakeoff-2006[3] [11]. The size of these corpora in number of words is presented in Table 3. Conventionally, segmentation performance is measured by F-measure

$$F = \frac{2PR}{P+R} \tag{5}$$

where the precision P and recall R are the proportion of the correctly segmented words to all words in a segmenter's output and to all those in the gold-standard segmentation, respectively. All our experimental results are presented in terms of F scores. The DLG features used in these experiments are acquired from the training and test corpora, both without any annotation.

Our work will be compared against those in the closed test of the Bakeoff. The rule for the closed test is that no additional information beyond a training corpus is allowed, while an open test allows any resource. In a sense, the former is more a competition of methodology with the same resource support, whereas the latter more a competition of the wealth of resources in proper use. Undoubtedly, our research is aimed specifically at the advancement of methodology.

[3]The Third International Chinese Language Processing Bakeoff was an international evaluation proceeding on CWS and named entity recognition, sponsored by SIGHAN, the special interest group for Chinese language processing of the Association of Computational Linguistics (ACL). Bakeoff-2006 is the last of the three Bakeoffs since 2003. All corpora used in this study are accessible from the official Bakeoff-2006 website at http://www.sighan.org/bakeoff2006.

Table 3: Bakeoff-2006 corpus size

Corpus	AS	CityU	CTB	MSRA
Training (M)	5.45	1.64	0.5	1.26
Test (K)	91	220	154	100

In addition to the best in the latest Bakeoff, we also consider two baselines of CRF modeling for comparison, one with n-gram feature templates a and b as defined in Table 2, and the other plus feature template c in addition[4]. Feature template c, or similar ones consisting of character types, can lead to performance enhancement in CWS, as reported in [12]. The price for this, however, is to bring into the closed test some extra information beyond a given training corpus. Therefore, we have to differentiate between the two in order to have a fair comparison of performance in terms of different experimental settings.

4.1 Comparing to the Best

A summary of the best results in the closed test of Bakeoff-2006 is presented in Table 4, with site IDs in parentheses. All participants with at least a third best performance are shown in this table [2, 22, 24, 28, 29, 31]. The closed test was so competitive that the top results are very close one another.

The performance comparison of the CRF+DLG approach, the baselines and the best in Bakeoff-2006 closed test is presented in Table 5. Two observations can be drawn from these results. One is that DLG brings in a significant improvement upon both CRF baselines in three of the four tracks of close test, confirming the effectiveness of incorporating DLG information into CRF learning. The other is that the best performance of the CRF+DLG approach goes beyond the best of Bakeoff-2006 closed in a greater degree than the best beyond the second, indicating the significance of the enhancement. Notice that the enhancement is achieved on top of the state of the art.

[4]Some participants in Bakeoff-2006 closed test used feature templates of this kind [22, 29, 31], while others did not.

Table 4: Best F-scores (%) in Bakeoff-2006 closed

Participant	AS	CityU	CTB	MSRA
Zhu (1)	94.4	96.8	92.7	95.6
Carpenter (9)	94.3	96.1	90.7	95.7
Tsai (15)	95.7	**97.2**	–	95.5
Zhao (20)	**95.8**	97.1	**93.3**	–
Zhang (26)	94.9	96.5	92.6	95.7
Wang (32)	95.3	97.0	93.0	**96.3**
Best closed	95.8	97.2	93.3	96.3
Best - 2nd	+0.1	+0.1	+0.3	+0.6

Table 5: Performance comparison

Corpus	AS	CityU	CTB	MSRA
Best closed	95.8	97.2	93.3	96.3
BL: a+b	95.4	96.9	93.2	96.1
+DLG	95.6	97.2	94.0	96.1
BL: a+b+c	95.9	97.2	93.4	96.1
+DLG	96.0	97.4	94.1	96.2
DLG - Best	+0.2	+0.2	+0.8	-0.1

BL: baseline

4.2 Discussion

DLG information stems from word candidates in unsupervised learning. One might argue that a lexicon extracted from the training corpus or some other reliable linguistic resources could be more helpful. Unfortunately, this is not always true. Existing work did show that a proper external lexicon could bring forth performance improvement to a MaxEnt model [12]. Nevertheless, we had quite a bit of experience counter to this while working with CRF modeling: A lexicon extracted from the training corpus caused performance loss rather than any gain. A key issue with CRF learning in utilizing lexical information for CWS is thus to find a lexicon with a nice fit to an input text, instead of simply using an external lexicon as large as possible. The technique we have developed here is a method to infer a good supplementary lexicon to CRF training that can lead to further performance enhancement.

To our knowledge, assembling unsupervised and supervised segmentation together for performance enhancement is a brand-new research area for CWS, in which successful work is yet to be reported. A simple divide-and-conquer strategy was attempted in [8] to integrate DLG-based segmentation and example-based learning. DLG was only applied to recognize new words among the mono-character sequences in the output of the latter. Unfortunately, however, this attempt was not particularly successful, although the potentials of the DLG measure for OOV detection were exemplified.

In practice, CRF modeling of character tagging for CWS is inevitably a heavy burden of computation for many current hardware settings. This situation could be much worse if more tags were introduced into a CRF model in the hope of performance gains. Thus, the tradeoff between computational efficiency and performance is also an important issue with this learning framework. A particularly good thing with the proposed method to integrate DLG information into CRF learning is that, according to our experiments, it adds only some minor computational cost beyond that for the baselines.

5 Conclusion

In this paper, we have presented a novel approach to improving supervised learning for Chinese word segmentation, by integrating unsupervised segmentation outcomes into a supervised learning model. Our experiments on the latest Bakeoff data sets provide evidence to show that the character-based CRF modeling for CWS can make most of DLG information to achieve a better performance than the best records in the field.

Furthermore, the ensemble strategy we have applied allows DLG information to be used in exactly the same way as local features in CRF modeling. In this regard, our approach is straightforward, easy to implement, and highly adaptable, in addition to its effectiveness and efficiency.

Acknowledgements

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

References

[1] M. R. Brent. An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, 34:71–105, February 1999.

[2] B. Carpenter. Character language models for Chinese word segmentation and named entity recognition. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 169–172, Sydney, Australia, July 2006. Association for Computational Linguistics.

[3] L.-F. Chien. PAT-tree-based keyword extraction for Chinese information retrieval. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 50–58, Philadelphia, 1997.

[4] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley and Sons, Inc., New York, 1991.

[5] X. Ge, W. Pratt, and P. Smyth. Discovering Chinese words from unsegmented text. In *SIGIR '99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 271–272, Berkeley, CA, USA, August 15-19, 1999. ACM.

[6] S. Goldwater, T. L. Griffiths, and M. Johnson. Contextual dependencies in unsupervised word segmentation. In *COLING-ACL 2006*, pages 673–670, Sidney, Australia, 2006.

[7] C. Kit. *Unsupervised Lexical Learning as Inductive Inference*. PhD thesis, University of Sheffield, 2000.

[8] C. Kit and X. Liu. An example-based Chinese word segmentation system for CWSB-2. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pages 146–149, Jeju Island, Korea, 2005.

[9] C. Kit and Y. Wilks. 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, 1999.

[10] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML'01: Proceedings of the 18th International Conference on Machine Learning*, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[11] G.-A. Levow. 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, Sidney, Australia, 2006.

[12] J. K. Low, H. T. Ng, and W. Guo. A maximum entropy approach to Chinese word segmentation. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pages 161–164, Jeju Island, Korea, 2005.

[13] F. Peng, F. Feng, and A. McCallum. Chinese segmentation and new word detection using conditional random fields. In *COLING 2004*, pages 562–568, Geneva, Switzerland, 2004.

[14] F. Peng and D. Schuurmans. Self-supervised Chinese word segmentation. In *The Forth International Symposium on Intelligent Data Analysis*, pages 238–247, Lisbon, Portugal, September 2001.

[15] J. Rissanen. Modelling by shortest data description. *Automatica*, 14:465–471, 1978.

[16] J. Rissanen. *Stochastic Complexity in Statistical Inquiry*. World Scientific, N.J., 1989.

[17] B. Rosenfeld, R. Feldman, and M. Fresko. A systematic cross-comparison of sequence classifiers. In *Proceedings of the Sixth SIAM International Conference on Data Mining (SDM06)*, pages 563–567, Bethesda, Maryland, 2006.

[18] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–656, July, October 1948.

[19] R. Sproat and T. Emerson. The first international Chinese word segmentation bakeoff. In *The Second SIGHAN Workshop on Chinese Language Processing*, pages 133–143, Sapporo, Japan, 2003.

[20] R. Sproat and C. Shih. A statistical method for finding word boundaries in Chinese text. *Computer Processing of Chinese and Oriental Languages*, 4(4):336–351, 1990.

[21] M. Sun, D. Shen, and B. K. Tsou. Chinese word segmentation without using lexicon and hand-crafted training data. In *COLING-ACL'98*, volume 2, pages 1265–1271, Montreal, Quebec, Canada, 1998.

[22] R. T.-H. Tsai, H.-C. Hung, C.-L. Sung, H.-J. Dai, and W.-L. Hsu. On closed task of Chinese word segmentation: An improved CRF model coupled with character clustering and automatically generated template matching. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 108–117, Sidney, Australia, 2006.

[23] H. Tseng, P. Chang, G. Andrew, D. Jurafsky, and C. Manning. A conditional random field word segmenter for SIGHAN bakeoff 2005. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pages 168–171, Jeju Island, Korea, 2005.

[24] X. Wang, X. Lin, D. Yu, H. Tian, and X. Wu. Chinese word segmentation with maximum entropy and n-gram language model. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 138–141, Sidney, Australia, 2006.

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

[26] M. Yamamoto and K. W. Church. Using suffix arrays to compute term frequency and document frequency for all substrings in a corpus. *Computational Linguistics*, 27(1):1–30, 2001.

[27] J. Zhang, J. Gao, and M. Zhou. Extraction of Chinese compound words – an experimental study on a very large corpus. In *Proceedings of the*

Second Chinese Language Processing Workshop, pages 132–139, Hong Kong, China, 2000.

[28] M. Zhang, G.-D. Zhou, L.-P. Yang, and D.-H. Ji. Chinese word segmentation and named entity recognition based on a context-dependent mutual information independence model. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 154–157, Sidney, Australia, 2006.

[29] H. Zhao, C.-N. Huang, and M. Li. An improved Chinese word segmentation system with conditional random field. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 162–165, Sidney, Australia, 2006.

[30] H. Zhao, C.-N. Huang, M. Li, and B.-L. Lu. Effective tag set selection in Chinese word segmentation via conditional random field modeling. In *Proceedings of PACLIC-20*, pages 87–94, Wuhan, China, 2006.

[31] M.-H. Zhu, Y.-L. Wang, Z.-X. Wang, H.-Z. Wang, and J.-B. Zhu. Designing special post-processing rules for SVM-based Chinese word segmentation. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 217–220, Sidney, Australia, 2006.