Combining Character-Based and Subsequence-Based Tagging for Chinese Word Segmentation

Jiangde Yu, Chuan Gu, Wenying Ge
School of Computer and Information Engineering, Anyang Normal University, Anyang 455002, China
jiangde_yu@tom.com, {jkx-20,ligepw}@163.com

Abstract

Chinese word segmentation is the initial step for Chinese information processing. The performance of Chinese word segmentation has been greatly improved by character-based approaches in recent years. This approach treats Chinese word segmentation as a character-word-position-tagging problem. With the help of powerful sequence tagging model, character-based method quickly rose as a mainstream technique in this field. This paper presents our segmentation system for evaluation of CIPS-SIGHAN 2010 in which method combining character-based and subsequence-based tagging is applied and conditional random fields (CRFs) is taken as sequence tagging model. We evaluated our system in closed and open tracks on four corpuses, namely *Literary*, *Computer science*, *Medicine* and *Finance,* and reported our evaluation results.

1 Introduction

In Chinese information processing, word is the minimum unit to be used independently and meaningfully. But, Chinese sentences are written as string of characters without clear delimiters. Therefore, the first step in Chinese information processing is to identify the sequence of words in a sentence, namely Chinese word segmentation. It's the foundation of syntax analysis, semantic analysis and discourse comprehension, and also the important section of machine translation, question answering, information retrieval and information extraction(Jiang Wei, et al., 2007; Liu Qun, et al., 2004).

The research of Chinese word segmentation has been advancing rapidly and has gained many exciting achievements in recent years(Huang Changning, Zhao Hai. 2007; Song Yan, et al., 2009), especially after the First International Chinese Word Segmentation Bakeoff held in 2003. In this field, character-based tagging attracts more eyes and almost all excellent systems in evaluations has adopted this technology thought(Huang Changning, Zhao Hai. 2007; Zhao Hai, Jie Chunyu. 2007). In 2002, Xue presented the first paper about character-based tagging on the 1st international workshop of special interest group on Chinese language processing, SIGHAN. He segmented Chinese words with four character tags: LL, RR, MM and LR, depending on its position within a word using a maximum entropy tagger(Xue N W, Converse S P. 2002). Huang et al. implemented character-based segmentation system with conditional random fields, six word-position tags: B, B2, B3, M, E, S, and TMPT-6 and achieved very excellent results(Huang Changning, Zhao Hai. 2006; Huang Changning, Zhao Hai. 2007). On this base, Zhao hai presented an effective subsequence-based tagging for Chinese word segmentation(Zhao Hai, Jie Chunyu. 2007). All these references considered Chinese segmentation as character or subsequence tagging problem and implemented with statistical language models.

The evaluation for Chinese word segmentation in CIPS-SIGHAN 2010 has two subtasks: word segmentation for simplified Chinese text and for traditional Chinese text. The simplified Chinese corpuses are offered by Institute of Computing Technology(ICT) and Peking University(PKU), and the traditional Chinese corpuses are offered City University of Hong Kong(CityU). The corpuses involved four domains: *literary*, *computer science*, *medicine* and *finance*. Considering plenty of abbreviations, numeric and other non-Chinese strings, our segmentation system adopted a method combining character-based and subsequence-based tagging, and took CRFs as sequence tagging model. CRFs is a kind of conditional probability model for sequence tagging presented by Lafferty et al. in 2001(Lafferty J, et al., 2001). In our experiment, the CRF++0.53 toolkit[1] is used. CRF++ is a simple, customizable, and open source implementation of CRFs for segmenting sequential data. This paper described our system participating CIPS-SIGHAN 2010 and presented our word-position tag set and feature template set and their change in open tracks. Finally, we report the results of our evaluation.

2 Combining character-based and subsequence-based tagging for Chinese word segmentation

In character-based tagging approach to Chinese word segmentation, it tags the word-position of non-Chinese characters, such as punctuation, letter words and numeric, just like what to do with Chinese characters. This method works well when there is a small quantity of these characters. But plenty of these characters will cut down the segmentation performance, especially some abbreviation and programming statement in computer science domain. Considering this, we used a method combining character-based and subsequence-based tagging that is to take an English word or programming statement as a subsequence to tag its word-position. The correct tag for one-character word is S.

2.1 Word-position tag set

In the closed track of traditional Chinese and simplified Chinese, four word-position tag set is used: B (Beginning of word), M(Middle of word), E(End of word) and S(one-character word). The tag set is also used in open tracks of traditional Chinese. And we used six word-position tag set for open tracks of simplified Chinese: B(Beginning of word), B2(2nd character of word), B3(3rd character of word) M(Middle of word), E(End of word) and S(one-character word).

2.2 Feature templates

To define the relationship of some specific language components or information in context and some being forecasted things is the main function of feature template. It is generally considered that feature template is abstracted from a group of context features on same attributes.

In CRF++0.53 toolkit, there are two kind of templates: Unigram template and Bigram template. In word-position tagging Chinese segmentation, available features are rather limited. The main feature needed to take into account is character feature, which includes current character, previous and next character. Jiang, Wang and Guan (2007) abstracted the character features into six templates according different distances from current character. They are Unigram templates. The type and meaning of these templates are presented in table 1. When training with CRFs model, these templates will be extended to thousands of features and every feature has a group of corresponding feature functions. All these functions are very important to CRFs model learning. Seen from table 1, Bigram feature has only one template: $T_{-1}T_0$ which describes the word-position transfer feature of two adjacent characters or subsequences. This feature extends limited features in training. Take four-WORD-POSITION-tag for instance, it can be extended into sixteen features. In our tracks, open or closed one, the seven templates in table 1: C_{-1}, C_0, C_1, $C_{-1}C_0$, C_0C_1, $C_{-1}C_1$, $T_{-1}T_0$ are used.

Table 1 List of feature templates

Type of template	template	Meaning of template
Unigram	C_{-1}	previous character
	C_0	current character
	C_1	next character
	$C_{-1}C_0$	String of current character and previous one
	C_0C_1	String of current character and next one
	$C_{-1}C_1$	String of previous and next character
Bigram	$T_{-1}T_0$	Word-position transfer feature of two adjacent character

[1] Download from this website: http://crfpp.sourceforge.net/

3 Experiments and results

3.1 Data set

Our training and test corpuses are gained from evaluation conference. The training and test corpuses of simplified Chinese are offered by ICT and PKU, while traditional Chinese by CityU. These corpuses involved in four domains: *literary(A), computer science(B), medicine(C), finance(D)*. In addition, we also use the CityU2005 training corpuses which gained from the Bakeoff2005 for open track.

3.2 Evaluation metrics

Five evaluation metrics: precision(*P*), recall(*R*), f-measue(*F1*), out-of-vocabulary words recall rate (*OOV RR*) and In-vocabulary words recall rate (*IV RR*) are used in our evaluation experiments.

3.3 Experiments and results

We adopted combining character-based and subsequence-based tagging for Chinese word segmentation, and conducted closed track experiments on these corpuses. Four word-position tag set(B, M, E, S) and seven templates(C_{-1}, C_0, C_1, $C_{-1}C_0$, C_0C_1, $C_{-1}C_1$, $T_{-1}T_0$) are adopted in closed tracks of simplified and traditional Chinese. Our results of the closed tracks are described in Table 2.

In our open tracks of simplified Chinese, we used six word-position tag set: B, B2, B3, M, E, S and seven templates same with closed tracks. Tag set and templates used in open tracks of traditional Chinese are same with closed tracks, too. In open tracks of traditional Chinese, we trained the combination of CityU2005 and corpus from this conference with CRFs model. The results of open tracks are shown in Table 3.

Talbe 2 Our results of closed tracks

corpuses	domains	*R*	*P*	*F1*	*OOV RR*	*IV RR*
simplified	Literature(A)	0.908	0.918	0.913	0.556	0.935
	Computer science(B)	0.89	0.908	0.899	0.592	0.943
	Medicine(C)	0.902	0.907	0.904	0.633	0.935
	Finance(D)	0.925	0.938	0.931	0.664	0.95
traditional	Literature(A)	0.888	0.905	0.896	0.728	0.904
	Computer(B)	0.908	0.931	0.919	0.684	0.931
	Medicine(C)	0.905	0.924	0.914	0.725	0.919
	Finance(D)	0.891	0.912	0.901	0.676	0.907

Table 3 Our results of open tracks

corpuses	domains	*R*	*P*	*F1*	*OOV RR*	*IV RR*
simplified	Literature(A)	0.908	0.916	0.912	0.535	0.936
	Computer science(B)	0.893	0.908	0.9	0.607	0.944
	Medicine(C)	0.904	0.906	0.905	0.635	0.937
	Finance(D)	0.925	0.937	0.931	0.669	0.95
traditional	Literature(A)	0.905	0.9	0.902	0.775	0.918
	Computer(B)	0.911	0.924	0.918	0.698	0.933
	Medicine(C)	0.903	0.903	0.903	0.729	0.917
	Finance(D)	0.903	0.916	0.91	0.721	0.916

4 Conclusion

As a fundamental task in Chinese information processing, Chinese segmentation gained more eyes in recent years and character-based tagging becomes the main segmentation technology. This paper describes our Chinese word segmentation system for CIPS-SIGHAN 2010. Then we present our word-position tag set and feature templates used in closed tracks and change of these parameters in open tracks. Finally, we report the results of the evaluation.

Acknowledgments

We would like to thank the anonymous reviewers for their helpful comments and suggestions.

References

Huang Changning, Zhao Hai. 2007. *Chinese word segmentation: A decade review. Journal of Chinese Information Processing*, 2007, 21(3):8-19.

Huang Changning, Zhao Hai. 2006. *Character-based tagging: A new method for Chinese word segmentation.* In *Proceedings of Chinese Information Processing Society 25 Annual Conference*. Beijing, China: Tsinghua University Press, 2006:53-63.

Jiang Wei, Wang Xiaolong, Guan Yi. 2007. *Research on Chinese Lexical Analysis System by Fusing Multiple Knowledge Sources. Chinese Journal of Computers*, 2007，30(1):137-145.

Liu Qun, Zhang Huaping, Yu Hongkui. 2004. *Chinese lexical analysis using cascaded hidden Markov model. Journal of Computer Research and Development*, 2004, 41(8):1421-1429.

Lafferty J, Pereira F, McCallum A. 2001. *Conditional random fields: probabilistic models for segmenting and labeling sequence data. In Proceedings of 18th International Conference on Machine Learning*, 2001:282-289.

Song Yan, Cai Dongfeng, Zhang Guiping. 2009. *Approach to Chinese word segmentation based on character-word joint decoding. Journal of Software*, 2009,20(9):2366-2375.

Xue N W, Converse S P. 2002. *Combining classifiers for Chinese word segmentation. In Proceedings of the First SIGHAN Workshop on Chinese Language Processing*. Taipei , Taiwan, China: AS Press, 2002：20-27.

Zhao Hai, Jie Chunyu. 2007. *Effective subsequence-based tagging for Chinese word segmentation. Journal of Chinese Information Processing*, 2007,21(5):8-13.