Description of the NCU Chinese Word Segmentation and Named Entity Recognition System for SIGHAN Bakeoff 2006

Yu-Chieh Wu
Dept. of Computer Science and Information Engineering
National Central University
Taoyuan, Taiwan
bcbb@db.csie.ncu.edu.tw

Jie-Chi Yang
Graduate Institute of Network Learning Technology
National Central University
Taoyuan, Taiwan
yang@cl.ncu.edu.tw

Qian-Xiang Lin
Dept. of Computer Science and Information Engineering
National Central University
Taoyuan, Taiwan
93522083@cc.ncu.edu.tw

Abstract

Asian languages are far from most western-style in their non-separate word sequence especially Chinese. The preliminary step of Asian-like language processing is to find the word boundaries between words. In this paper, we present a general purpose model for both Chinese word segmentation and named entity recognition. This model was built on the word sequence classification with probability model, i.e., conditional random fields (CRF). We used a simple feature set for CRF which achieves satisfactory classification result on the two tasks. Our model achieved 91.00 in F rate in UPUC-Treebank data, and 78.71 for NER task.

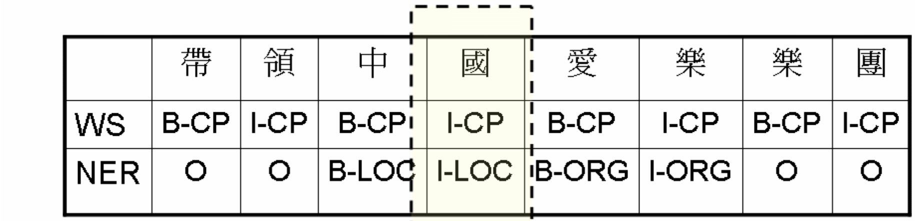
1 Introduction

With the rapid expansion of text media sources such as news articles, technical reports, there is an increasing demand for text mining and processing. Among different cultures and countries, the Asian languages are far from the other languages, there is not an explicit boundary between words, for example Chinese. Similar to English, the preliminary step of most natural language processing is to "tokenize" each word. In Chinese, the word tokenization is also known as word segmentation or Chinese word tokenization.

To support the above targets, it is necessary to detect the boundaries between words in a given sentence. In tradition, the Chinese word segmentation technologies can be categorized into three types, (heuristic) rule-based, machine learning, and hybrid. Among them, the machine learning-based techniques showed excellent performance in many research studies (Peng et al., 2004; Zhou et al., 2005; Gao et al., 2004). This method treats the word segmentation problem as a sequence of word classification. The classifier online assigns either "boundary" or "non-boundary" label to each word by learning from the large annotated corpora. Machine learning-based word segmentation method is quite similar to the word sequence inference techniques, such as part-of-speech (POS) tagging, phrase chunking (Wu et al., 2006a) and named entity recognition (Wu et al., 2006b).

In this paper, we present a prototype for Chinese word segmentation and named entity recognition based on the word sequence inference model. Unlike previous researches (Zhou et al., 2005; Shi, 2005), we argue that without using the word segmentation information, Chinese named entity recognition task can also be viewed as a variant word segmentation technique. Therefore, the two tasks can be accomplished without adapting the word sequence inference model. The preliminary experimental result show that in the word segmentation task, our method can achieve 91.00 in F rate for the UPUC Chinese Treebank data, while it at-

Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, pages 209–212,
Sydney, July 2006. ©2006 Association for Computational Linguistics



	帶	領	中	國	愛	樂	樂	團
WS	B-CP	I-CP	B-CP	I-CP	B-CP	I-CP	B-CP	I-CP
NER	O	O	B-LOC	I-LOC	B-ORG	I-ORG	O	O

CP: Chinese word phrase LOC: Location ORG: Organization O: Non-named entity word

Figure 1: Sequence of word classification model

tends 78.76 F rate for the Microsoft Chinese named entity recognition task.

The rest of this paper is organized as follows. Section 2 describes the word sequence inference model and the used learner. Experimental result and evaluations are reported in section 3. Finally, in section 4, we draw conclusion and future remarks.

2 System Description

In this section, we firstly describe the overall system architecture for the word segmentation and named entity recognition tasks. In section 2.2, the employed classification model- conditional random fields (CRF) is then presented.

2.1 Word Sequence Classification

Similar to English text chunking (Ramshaw and Marcus, 1995; Wu et al., 2006a), the word sequence classification model aims to classify each word via encoding its context features. An example can be shown in Figure 1. In Figure1, the model is classifying the Chinese character "國" (country). The second row in Figure 1 means the corresponding category of each in the word-segmentation (WS) task, while the third row indicates the class in the named entity recognition (NER) task. For the WS task, there are only two word types, B-CP (Begin of Chinese phrase) and I-CP (Interior of Chinese phrase). In contrast, the word types in the NER task depend on the pre-defined named class. For example, both in MSR and CityU datasets, person, location, and organization should be identified. In this paper, we used the similar IOB2 representation style (Wu et al., 2006a) to express the Chinese word structures.

By encoding with IOB style, both WS and NER problems can be viewed as a sequence of word classification. During testing, we seek to find the optimal word type for each Chinese character. These types strongly reflect the actual word boundaries for Chinese words or named entity phrases.

To effect classify each character, in this paper, we employ 13 feature templates to capture the context information of it. Table 1 lists the adopted feature templates.

Table 1: Feature template used for both Chinese word segmentation and named entity recognition tasks

Feature Type	Examples	Feature Type	Examples
W_{-2}	領	$W_0 + W_{+1}$	國+愛
W_{-1}	中	$W_{+1} + W_{+2}$	愛+樂
W_0	國	$W_{+1} + W_{+2}$	愛+樂
W_{+1}	愛	$W_{-2}+W_{-1}+W_0$	領+中+國
W_{+2}	樂	$W_{-1}+W_0+W_{+1}$	中+國+愛
$W_{-2} + W_{-1}$	領+中	$W_0+W_{+1}+W_{+2}$	國+愛+樂
$W_{-1} + W_0$	中+國		

2.2 Conditional Random Fields

Conditional random field (CRF) was an extension of both Maximum Entropy Model (MEMs) and Hidden Markov Models (HMMs) that was firstly introduced by (Lafferty et al., 2001). CRF defined conditional probability distribution P(Y|X) of given sequence given input sentence where Y is the "class label" sequence and X denotes as the observation word sequence.

A CRF on (X,Y) is specified by a feature vector F of local context and the corresponding feature weight λ. The F can be treated as the combination of state transition and observation value in conventional HMM. To determine the optimal label sequence, the CRF uses the following equation to estimate the most probability.

$$y = \arg\max_{y} P(y \mid x, \lambda) = \arg\max_{y} \lambda F(y, x)$$

The most probable label sequence y can be efficiently extracted via the Viterbi algorithm. However, training a CRF is equivalent to estimate the parameter set λ for the feature set. In this paper, we directly use the quasi-Newton L-BFGS[1] method (Nocedal and Wright, 1999) to iterative update the parameters.

3 Evaluations and Experimental Result

3.1 Dataset and Evaluations

We evaluated our model in the close track on UPUC Chinese Treebank for Chinese word segmentation task, and CityU corpus for Chinese NER task. Both settings are the same for the two tasks. The evaluations of the two tasks were mainly measured by the three metrics, namely recall, precision, and f1-measurement. However, the evaluation style for the NER and WS is quite different. In WS, participant should reformulate the testing data into sentence level whereas the NER was evaluated in the token-level. Table 2 lists the results of the two tasks with our preliminary model.

Table 2: Official results on the word segmentation and named entity recognition tasks

	Dataset	F1-measure
Word segmentation	UPUC	91.00
Named entity recognition	CityU	78.71

Table 3: Experimental results for the three Chinese word segmentation datasets

Closed Task	CityU	MSR	UPUC
Recall	0.958	0.940	0.917
Precision	0.926	0.906	0.904
F-measure	0.942	0.923	0.910

3.2 Experimental Result on Word Segmentation Task

To explore the effectiveness of our method, we go on extend our model to the other three tasks for the WS track, namely CityU, MSR. Table3 shows the experimental results of our model in the all close WS track except for CKIP corpus. These results do not officially provided by the SIGHAN due to the time limitation.

3.3 Experimental Result on Named Entity Recognition Task

In the second experiment, we focus on directly adapting our method for the NER track. Table 4 lists the experimental result of our method in the CityU and MSR datasets. It is worth to note that due to the different evaluation style in NER tracks, our tokenization rules did not consistent with the SIGHAN provided testing tokens. Our preliminary tokenization rules produced 371814 characters for the testing data, while there are 364356 tokens in the official provided testing set. Such a big trouble deeply earns the actual performance of our model. To propose a reliable and actual result, we directly evaluate our method in the official provided testing set again. As shown in Table 4, the our method achieved 0.787 in F rate with non-correct version. In contrast, after correcting the Chinese tokenization rules as well as SIGHAN official provided tokens, our method significantly improved from 0.787 to 0.868. Similarly, our method performed very on the MSR track which reached 0.818 in F rate.

Table 4: Experimental results for MSR and City closed NER tasks

Closed Task	City (official result)	City (correct)	MSR
Recall	0.697	0.931	0.752
Precision	0.935	0.814	0.896
F-measure	0.787	**0.868**	**0.818**

4 Conclusions and Future Work

Chinese word segmentation is the most important foundations for many Chinese linguistic technologies such as text categorization and information retrieval. In this paper, we present simple Chinese word segmentation and named entity recognition models based on the conventional sequence classification technique. The main focus of our work is to provide a light-weight and simple model that could be easily ported to different domains and languages. Without any prior knowledge and rules, such a simple technique shows satisfactory results on both word segmentation and named entity recognition tasks. To reach state-of-the-art this model still needs to employed more detail feature engines and analysis. In the future, one of the main directions is to extend this model toward full unsuper-

[1] http://www-unix.mcs.anl.gov/tao/

vised learning from large un-annotated text. Mining from large unlabeled data have been showed benefits to improve the original accuracy. Thus, not only the more stochastic feature analysis, but also adjust the learner from unlabeled data are important future remarks.

References

Lafferty, J., McCallum, A., and Pereira, F. 2001. Conditional Random Field: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the International Conference on Machine Learning.

Gao, J., Wu, A., Li, M., Huang, C. N., Li, H., Xia, X., and Qin, H. 2004. Adaptive Chinese word segmentation. In Proceedings the 41st Annual Meeting of the Association for Computational Linguistics, pp. 21-26.

Lance A. Ramshaw and Mitchell P. Marcus. 1995. Text chunking using transformation-based learning. In Proceedings of the 3rd Workshop on Very Large Corpora, pages 82-94.

Nocedal, J., and Wright, S. 1999. Numerical optimization. Springer.

Peng, F., Feng, F., and McCallum, A. 2004. Chinese segmentation and new word detection using conditional random fields. In Porceedings of the Computational Linguistics, pp. 562-568.

Shi, W. 2005. Chinese Word Segmentation Based On Direct Maximum Entropy Model. In Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing.

Wu, Y. C., Chang, C. H. and Lee, Y. S. 2006a. A general and multi-lingual phrase chunking model based on masking method. Lecture Notes in Computer Science (LNCS): Computational Linguistics and Intelligent Text Processing, 3878: 144-155.

Wu, Y. C., Fan, T. K., Lee Y. S. and Yen, S. J. 2006b. Extracting named entities using support vector machines," Lecture Notes in Bioinformatics (LNBI): Knowledge Discovery in Life Science Literature, (3886): 91-103.

Wu, Y. C., Lee, Y. S., and Yang, J. C. 2006c. The Exploration of Deterministic and Efficient Dependency Parsing. In Proceedings of the 10th Conference on Natural Language Learning (CoNLL).

Zhou, J., Dai, X., Ni, R., Chen, J. 2005. .A Hybrid Approach to Chinese Word Segmentation around CRFs. In Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing.