Description of the NCU Chinese Word Segmentation and Part-of-Speech Tagging for SIGHAN Bakeoff 2007

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

Yue-Shi Lee
Dept. of Computer Science and Information Engineering
Ming Chuan University
Taoyuan, Taiwan
leeys@mcu.edu.tw

Abstract

In Chinese, most of the language processing starts from word segmentation and part-of-speech (POS) tagging. These two steps tokenize the word from a sequence of characters and predict the syntactic labels for each segmented word. In this paper, we present two distinct sequential tagging models for the above two tasks. The first word segmentation model was basically similar to previous work which made use of conditional random fields (CRF) and set of predefined dictionaries to recognize word boundaries. Second, we revise and modify support vector machine-based chunking model to label the POS tag in the tagging task. Our method in the WS task achieves moderately rank among all participants, while in the POS tagging task, it reaches very competitive results.

1 Introduction

With the rapid expansion of online text articles such as blog, web news, and research/technical reports, there is an increasing demand for text mining and management. Different from western-like languages, handling oriented languages is far more difficult since there is no explicit boundary symbol to indicate what a word is in the text. However the most important preliminary step for natural language processing is to tokenize words and separate them from the word sequence. In Chinese, the word tokenization is also known as word segmentation or Chinese word tokenization. The problem of the Chinese word segmentation is very critical for most Chinese linguistics because the error segmented words deeply affects the downstream purpose, like POS tagging and parsing. In addition tokenizing the unknown words is also an unavoidable problem.

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 recent 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 (Clark et al., 2003; Gimenez and Marquez, 2003), phrase chunking (Lee and Wu, 2007) and word dependency parsing (Wu et al., 2006, 2007).

In this paper, we present two prototype systems for Chinese word segmentation and POS tagging

tasks. The former was basically an extension of previous literatures (Ng and Low, 2004; Zhou et al., 2006), while the latter incorporates the unknown word and known word tagging into one step. The two frameworks were designed based on two variant machine learning algorithms, namely CRF and SVM. In our pilot study, the SVM showed better performance than CRF in the POS tagging task. To identify unknown words, we also encode the suffix and prefix features to represent the training example. The strategy was showed very effective for improving both known and unknown word chunking on both Chinese and English phrase chunking (Lee and Wu, 2007). In this year, the presented word segmentation method achieved moderate rank among all participants. Meanwhile, the proposed SVM-based POS tagging model reached very competitive accuracy in most POS tasks. For example, our method yields second best result on the CTB POS tagging track.

The rest of this paper is organized as follows. Section 2 describes employed machine learning algorithms, CRF and SVM. In section 3, we present the proposed word segmentation and POS tagging framework which used for the SIGHAN-bake-off this year. Experimental result and evaluations are reported in section 4. Finally, in section 5, we draw conclusion and future remarks.

2 Classification Algorithms

2.1 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 CRF++ (Kudo and Matsumoto, 2003) which included the quasi-Newton L-BFGS [1] method (Nocedal and Wright, 1999) to iterative update the parameters.

2.2 Support Vector Machines

Assume we have a set of training examples,

$(x_1, y_1), (x_2, y_2), ..., (x_n, y_n),\ x_i \in \Re^D, y_i \in \{+1, -1\}$

where x_i is a feature vector in *D*-dimension space of the *i*-th example, and y_i is the label of x_i either positive or negative. The training of SVMs involves minimizing the following object function (primal form, soft-margin (Vapnik, 1995)):

$$\text{minimize}: W(\alpha) = \frac{1}{2}\vec{W}\cdot\vec{W} + C\sum_{i=1}^{n} Loss(\vec{W}\cdot x_i, y_i) \quad (1)$$

The loss function indicates the loss of misclassification risk. Usually, the hinge-loss is used (Vapnik, 1995; Keerthi and DeCoste, 2005). The factor *C* in (1) is a parameter that allows one to trade off training error and margin size. To classify a given testing example *X*, the decision rule takes the following form:

$$y(X) = sign((\sum_{x_i \in SVs} \alpha_i y_i K(X, x_i)) + b) \quad (2)$$

α_i represents the weight of training example x_i which lies on the hyperplane, and *b* denotes as a bias threshold. *SVs* means the support vectors and obviously has the non-zero weights of α_i. $K(X, x_i) = \phi(X)\cdot\phi(x_i)$ is a pre-defined kernel function that might transform the original feature space from $\Re^D$ to $\Re^{D'}$ (usually $D<<D'$). In the linear kernel form, the $K(X, x_i)$ simply compute the dot products of the two variables. By introducing of the polynomial kernel, we re-write the decision function of (1) as:

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

$$y(X) = sign((\sum_{x_i \in SVs} \alpha_i y_i K(X, x_i)) + b)$$
$$= sign((\sum_{x_i \in SVs} \alpha_i y_i (1 + dot(X, x_i)^d) + b) \quad (3)$$

where

$$K(X, x_i) = (1 + dot(X, x_i))^d \quad (4)$$

and *d* is the polynomial kernel degree.

In many NLP problems, the training and testing examples are represented as bits of binary vectors. In this section, we focus on this case. Later, we present a general form without considering this constraint.

3 System Description

In this section, we first describe the problem settings for the word segmentation problems. In section 3.2, the proposed POS tagging framework is then presented.

3.1 Word Sequence Classification

Similar to English text chunking (Ramshaw and Marcus, 1995; Lee and Wu, 2007), the word sequence classification model aims to classify each word via encoding its context features.

By encoding with BIES (LMR tagging scheme) or IOB2 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.

As reported by (Zhou et al., 2006), the use of richer tag set can effectively enhance the performance. They extend the tag of "Begin of word" into "second-begin" and "third-begin" to capture more character types. However, there are some ambiguous problem to the 3-character Chinese words and 4-character Chinese words. For example, to encode "素還真" with his extended tag set, the first character can be encoded as "B" tag. But for the second character, we can use "second-begin" or "I" tag to represent the middle of word.

In order to make the extension clearer, in this paper, we explicitly extend the B tag and E tag with "after begin" (BI), and "before end" (IE) tags. Table 1 lists the difference between the traditional BIES and the proposed E-BIES encodings methods. Table 2 illustrates an example of how the BIES and E-BIES encode with different number of characters.

Table 1: BIES and E-BIES encoding strategies

	BIES	E-BIES
Begin of a word	B	B
After begin of a word	-	BI
Middle of a word	I	I
Before end of a word	-	IE
End of a word	E	E
Single word	S	S

Table 2: An example of the BIES and E-BIES encoding strategies

N-character word	BIES	E-BIES
看	S	S
中原	B,E	B,E
素還真	B,I,E	B,BI,E
玄子神功	B,I,I,E	B,BI,IE,E
一氣化三千	B,I,I,I,E	B,BI,I,IE,E

To effect classify each character, in this paper, we adopted most feature types to train the CRF (Kudo and Matsumoto, 2004). Table 3 lists the adopted feature templates. The dictionary flag is very similar to previous literature (Ng and Low, 2004) while we adding up English full-character into our dictionary.

Table 3: Feature template used for Chinese word segmentation task

Feature Type	Context Position	Description
Unigram	C_{-2},C_{-1},C_0,C_1,C_2	Chinese character feature
Nearing Bi-gram	(C_{-2},C_{-1})(C_{-1},C_0) (C_1,C_0)(C_1,C_2)	Bi-character feature
Jump Bigram	(C_{-1},C_1)	Non-continuous character feature
Dictionary Flag	C_0	Date, Digital, English letter or punctuation
Dictionary Flag *N*-gram	(C_{-1},C_0,C_1)	*N*-gram of the dictionary flags

3.2 Feature Codification for Chinese POS Tagging

As reported by (Ng, and Low, 2004; Clark et al., 2003), the pure POS tagging performance is no more than 92% in the CTB data and no more than

96.8% in English WSJ. The learner used in his literature is maximum entropy model. However the main limitation of his POS tagging strategy is that the unknown word classification problem was not resolved.

To circumvent this vita, we simply extend the idea of SVM-based chunker (Lee and Wu, 2007) and develop our own SVM-based POS tagger. Although CRF showed excellent performance in word segmentation task, in English POS tagging, the SVM is more effective than CRF. Also in our closed experiment, we had tried transformation-based error-driven learner (TBL), CRF, and SVM classifiers. The pilot experiment showed that the SVM outperformed the other two learners and achieved almost 94% accuracy in the CTB data. Meanwhile TBL reached the worst result than the other two classifiers (~88%).

Handling unknown word is very important to POS tagging problem. As pointed out by (Lee and Wu, 2007; Gimenez, and Marquez, 2003), the introduction of suffix features can effectively help to guess the unknown words for tagging and chunking. Different from (Gimenez and Marquez, 2003), we did not derive data for unknown word guessing. Instead, we directly encode all suffix- and prefix-features for each training instance. In training phase, the rich feature types are able to disambiguate not only the unknown word guessing, but also improve the known word classification. As reported by (Lee and Wu, 2007), the strategy did improve the English and Chinese chunking performance for both known and unknown words.

Table 4: Feature patterns used for Chinese POS tagging task

Feature Type	Context Position	Description
Unigram	$W_{-2},W_{-1},W_0,W_1,W_2$	Chinese word feature
Nearing Bigram	$(W_{-2},W_{-1})(W_{-1},W_0)$ $(W_1,W_0)(W_1,W_2)$	Bi-word feature
Jump Bi-gram	$(W_{-2},W_0)(W_{-1},W_1)$ $(W_2,W_0)(W_1,W_3)$	Non-continuous character feature
Possible tags	W_0	Possible POS tag in the training data
Prefix 3/2/1 characters	W_{-1},W_0,W_1	Pre-characters of word
Suffix 3/2/1 characters	W_{-1},W_0,W_1	Post-characters of word

The used feature set of our POS tagger is listed in Table 4. In this paper, we did not conduct the feature selection experiment for each tagging corpus, instead a unified feature set was used due to the time line. We trust our POS tagger could be further improved by removing or adding new feature set.

The learner used in this paper (SVM) is mainly developed by our own (Wu et al., 2007). The cost factor C is simply set as 0.15 for all languages. Furthermore, to remove rare words, we eliminate the words which appear no more than twice in the training data.

4 Evaluations and Experimental Result

4.1 Dataset and Evaluations

In this year, we mainly focus on the close track for WS and POS tagging tracks. The CTB, SXU, and NCC corpora were used for evaluated the presented word segmentation method, while all the released POS tagging data were tested by our SVM-based tagger, included CityU, CKIP, CTB, NCC, and PKU. Both settings of the two models were set as previously noted. The evaluation of the two tasks was mainly measured by the three metrics, namely, recall, precision, and f-measure. However, the evaluation process for the POS tagging track is somewhat different from WS. In WS, participant should reform the testing data into sentence level whereas in the POS tagging track the word had been correctly segmented. Thus the measurement of the POS tagging track is mainly accuracy-based (correct or incorrect).

4.2 Experimental Result on Word Segmentation Task

In this year, we only select the following three data to perform our method for the word segmentation task. They are CTB, NCC, and SXU where the NCC and SXU are fresh in this year. Table5 shows the experimental results of our model in the close WS track with except for CKIP and CityU corpora.

Table 5: Official results on the word segmentation task (closed-task)

	Recall	Precision	F-measure
CTB	0.9471	0.9500	0.9486
NCC	0.9236	0.9269	0.9252
SXU	0.9505	0.9515	0.9510

As shown above, our method in the CTB data showed 10th best out of 26 submissions. In the NCC and SXU datasets, our method achieved 19/26 and 18/30 rank. In overall, the presented extend-BIES scheme seems to work well on the CTB data and results in middle rank in comparison to the other participants.

4.3 Experimental Result on Part-of-Speech Tagging Task

In the second experiment, we focus on the designed POS tagging model. To measure the effectiveness, we apply our method to all the released dataset, i.e., CityU, CKIP, CTB, NCC, and PKU. Table 6 lists the experimental result of our method in this task.

Similar to WS task, our method is still very effective to CTB dataset. It turns out our method achieved second best in the CTB, while for the other corpora, it achieved 4th best among all the participants. We also found that our method was very close to the top 1 score about 1.3% (CKIP) to 0.09%. For the NCC, and PKU, our method was worse than the best system in 0.8% in overall accuracy. We conclude that by selecting suitable features and cost factor C to SVM, our method can be further improved. We left the work as future direction.

Table 6: Official results on the part-of-speech tagging task (closed-task)

	Riv	Roov	Rmt	Accuracy
CityU	0.9326	0.4322	0.8707	0.8865
CKIP	0.9504	0.5631	0.9065	0.9160
CTB	0.9554	0.7135	0.9183	0.9401
NCC	0.9658	0.5822	0.9116	0.9456
PKU	0.9591	0.5832	0.9173	0.9368

5 Conclusions and Future Work

Chinese word segmentation is the most important infrastructure for many Chinese linguistic technologies such as text categorization and information retrieval. In this paper, we present simple Chinese word segmentation and part-of-speech tagging models based on the conventional sequence classification technique. We treat the two tasks as two different learning framework and applying CRF and SVM as separated learners. Without any prior knowledge and rules, such a simple technique shows satisfactory results on both word segmentation and part-of-speech tagging tasks. In POS tagging task, our model shows very competitive results which merely spend few hours to train. To reach state-of-the-art, our method still needs to further select features and parameter tunings. In the future, one of the main directions is to extend this model toward full unsupervised learning from large un-annotated text. Mining from large unlabeled data have been showed benefits to improve the original accuracy. Thus, not only the stochastic feature analysis, but also adjust the learner from unlabeled data are important future remarks.

References

Clark, S., Curran, J. R., and Osborne, M. 2003. Bootstrapping POS Taggers Using Unlabeled data. In Proceedings of the 7th Conference on Natural Language Learning (CoNLL), pages 49-55.

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.

Giménez, J., and Márquez, L. 2003. Fast and accurate Part-of-Speech tagging: the SVM approach revisited. In Proceedings of the International Conference on Recent Advances in Natural Language Processing, pages 158-165.

Keerthi, S., and DeCoste, D. 2005. A Modified Finite Newton Method for Fast Solution of Large Scale Linear SVMs. Journal of Machine Learning Research, 6: 341-361.

Kudo, T., and Matsumoto, Y. 2004. Appliying Conditional Random Fields to Japanese Morphological Analysis. In Proceedings of the Empirical. Methods in Natural Language Processing (EMNLP), pages 230-237.

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.

Ramshaw, L. A., and Marcus, M. P. 1995. Text Chunking Using Transformation-based Learning.

In Proceedings of the 3rd Workshop on Very Large Corpora, pages 82-94.

Lee, Y. S. and Wu, Y. C. 2007. A Robust Multilingual Portable Phrase Chunking System. Expert Systems with Applications, 33(3): 1-26.

Ng, H. T., and Low, J. K. 2004. Chinese Part-of-Speech Tagging: One-at-a-time or All-at-once? Word-based or Character-based? In Proceedings of the Empirical. Methods in Natural Language Processing (EMNLP).

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.

Vapnik, V. N. 1995. The Nature of Statistical Learning Theory. Springer.

Wu, Y. C., Yang, J. C., and Lee, Y. S. 2007. An Approximate Approach for Training Polynomial Kernel SVMs in Linear Time. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, pages 65-68.

Wu, Y. C., Lee, Y. S., and Yang, J. C. 2007. Multilingual Deterministic Dependency Parsing Framework using Modified Finite Newton Method Support Vector Machines. In Proceedings of the Joint Conferences on Empirical Methods on Natural Language Processing and Conference on Natural Language Learning (EMNLP-CoNLL), pages 1175-1181.

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

Zhou, H., Huang, C. N., and Li, M. 2006. An Improved Word Segmentation System with Conditional Random Fields. In Proceedings of the SIGHAN Workshop on Chinese Language Processing Workshop, pages 162-165.

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.