Chinese Chunking Algorithm Based on Cascaded Conditional Random Fields

Guang-Lu Sun[1] Yuan-Chao Liu[2] Pei-Li Qiao[1] Fei Lang[3]

[1] School of Computer Science and Technology, Harbin University of Science and Technology, Harbin, China {bati_sun@hit.edu.cn}
[2] School of Computer Science and Technology, Harbin Institute of Technology, China
[3] Department of Foreign Languages Teaching, Harbin Science and Technology, China

Abstract

This paper presents a new Chinese chunking algorithm based on cascaded conditional random fields. Conditional random fields solve the tagging problems well, while the cascaded models restrain the affection of part-of-speech errors. The experimental results show that this approach achieves impressive accuracy in terms of the F-score: 85.06%.

Keywords: Chinese chunking, part-of-speech tagging, cascaded conditional random fields

1. Introduction

With the advent of the Internet era, many technologies of language processing are widely used to process the amount of electronic data from web. As an alternative to full parsing, text chunking is a useful step among these technologies for the web applications. Since Abney firstly proposed the definition of chunk as a preprocessing to full parsing in 1991 [1], sentences are divided by chunks which are labeled, non-overlapping sequences of words on the basis of lexical information.

The chunking task involves two research issues. The first is the chunking specification. The English chunking specification was unified in the CoNLL-2000 conference [2]. Part of the sparkle projects for Chinese chunking focused on how to define the appropriate specification [3][4].

The other is the chunking algorithm. For the complexity of the chunking task, it makes an excellent benchmark problem for evaluating machine learning algorithms, such as weighted voting of eight Support Vector Machines [5], Memory-based learning [6], conditional random fields (CRF) [7] and Winnow [8].

In the realistic applications, it is difficult to tag text manually. Chinese texts from the web need to be segmented and part-of-speech (POS) tagged automatically. Relative to English chunking which is based on automatic POS tags, the research of Chinese chunking is always based on gold standard POS tags. Because the errors of automatic POS tags affect the performance of Chinese chunking seriously, it is essential to solve Chinese chunking method based on automatic POS tags.

CRF models are confirmed to achieve fairly high performance in the chunking task, so it is applied to our Chinese chunking task. A common method is processing the text orderly, including segmentation, POS tagging and chunking. The problem of this method is the affections of cascaded errors that the performance of current step is reduced by the errors of the steps before. For depressing the affection of POS errors and improving the performance of chunking, with gold standard segmentation, a new Chi-

nese chunking algorithm is proposed based on cascaded conditional random fields (CCRF). The present work focuses on Chinese chunking with the corpus of Microsoft Research Asia. In open tests, the chunking model obtained the F-score of 85.06% with the automatic POS tags based on CCRF models.

Section 2 describes POS tagging based on CRF models in detail. Section 3 presents the Chinese chunking algorithm using CCRF models. Section 4 presents the experiment results of the Chinese chunking system. Finally, we draw some conclusions respectively.

2. Part-of-speech tagging based on conditional random fields

In this section, CRF models are firstly introduced in brief. Then POS tagging method is present based on CRF models.

2.1. Conditional Random Fields

CRF models are the conditional probability model. Comparing with generative models, one of the main advantages of CRF models is its ability to incorporate various features; the other is modeling on the objective sequence. Comparing with other discriminative models like maximum entropy model (MEM) and MEMM, CRF models overcome the label bias problem which hurts all the non-probabilistic sequence tagging models with independently trained next-state classifiers.

Corresponding to the POS sequence P and the context sequence W, the conditional formula of CRF models is:

$$p(T/W)=\frac{1}{Z(W)}exp\left\{\sum_k l_k f_k\left(t_{i-1},t_i,W,i\right)\right\} \quad (1)$$

Where $Z(W)$ is the normalization constant.[9]

$$Z(W)=\sum_{t\in T} exp\left\{\sum_k l_k f_k\left(t_{i-1},t_i,W,i\right)\right\} \quad (2)$$

$f_k\left(t_{i-1},t_i,W,i\right)$ is the usual feature function of CRF models. It can be decomposed as two types of definitions. One is the edge feature, or transition feature:

$$r_{t',t}\left(t_{i-1},t_i,w_i\right)=\begin{cases}1 & if \quad t_{i-1}=t',t_i=t \quad and \quad w_i=w\\ 0 & else\end{cases} \quad (3)$$

The other is the node feature, or state feature:

$$s_{t,w}\left(t_i,w_i\right)=\begin{cases}1 & if \quad t_i=t \quad and \quad w_i=w\\ 0 & else\end{cases} \quad (4)$$

From the formula (3) and (4), formula (1) can be computed as:

$$p(T/W)=\frac{1}{Z(W)}\bullet exp\left\{\sum_k\left(\mu_k r_k\left(t_{i-1},t_i,W,i\right)+\varsigma_k s_k\left(t_i,W,i\right)\right)\right\} \quad (5)$$

μ_k and ς_k is the estimation weights of the transition feature and state feature respectively. L-BFGS algorithm are used to estimate the weights of the features:

$$L_\Lambda=\sum_{i=1}^{N} log(P_\Lambda(t_i/w_i))-\sum_{k=1}^{K}\frac{l_k^2}{2\sigma_k^2} \quad (6)$$

The second part of formula (6) is the Gaussian prior value of feature for smoothing. Dynamic programming algorithm is finally utilized to search the best sequence T*:

$$T^*=\underset{T}{argmax}\left\{p_l\left(T/W\right)\right\} \quad (7)$$

2.2. POS tagging using CRF models

In the training module of POS tagging models, features are extracted from the chunking corpus which has been tagged with POS tags and chunk tags. The features are composed of observation values and corresponding tags that are used to train the estimation parameters of value. In the test module, input sentence is used for feature extraction, and put into tagging model. The POS tagging results are put into practice based on the computation of features' weights.

Because CRF models highly depend on features, features selection is an important part of the model construction. The

left and right two context words are selected for features extraction. We utilized the lexical information of the five words and the affix of the current word. Table 1 shows the features using in the POS tagging model.

Table 1: Features template using in the POS tagging models

Atomic features	W_i, W_{i-1}, W_{i-2}, W_{i+1}, W_{i+2}, AF_i
Combined features	$W_{i-1}W_i$, W_iW_{i+1}, $W_{i-1}W_{i+1}$, $W_{i-2}W_{i-1}W_i$, $W_iW_{i+1}W_{i+2}$

We selected the features of which the frequency is higher than 5 for the purpose of the beneficial effect of the features.

3. Chinese chunking algorithm based on cascaded conditional random fields

In this section, CCRF models are described in detail. Then, chunking model based on CCRF models is introduced.

3.1. Cascaded conditional random fields

There are two kinds of methods to build multi-layers machine learning model. One connects the sub-models by linear combination; the other looks at the bottom model as the input of the topper model. The latter method which CCRF models belong to is selected to build the cascaded models because it has better combination characteristics.

CCRF models are the two-stage models which are constructed by two CRF models. It is adaptive to solve Chinese POS tagging and chunking problem.

3.2. Chunking method with CCRF models

As shown in Fig. 1, the lower CRF models are used to tag POS with the lexical information of words. The N-best tagging results are transferred to the upper CRF models as the input of POS information. The upper CRF models are utilized to do chunking with the features which consist of words and automatic POS tags.

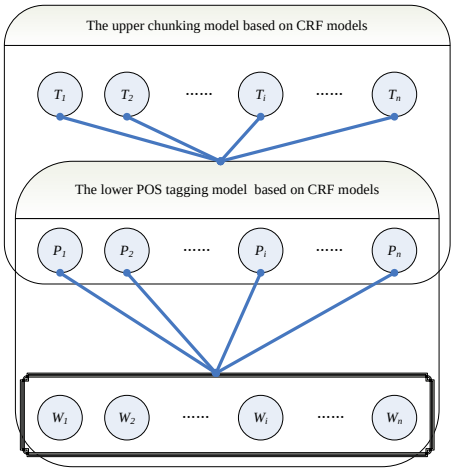

Fig. 1: Chinese chunking algorithm based on cascaded conditional random fields

The formula of cascaded conditional random fields:

$$T^* = \lambda_1 \underset{W}{argNmax}\, p(P/W) + (1-\lambda_1) \underset{P_i^* \in P^*}{argNmax}\, p(T/P_i^*, W) \quad (8)$$

P^* is N-best results of POS tagging, T^* is the best chunking result based on P^*. λ_1 is the parameter and set as 0.35.

The final results consist of the best chunking result and its corresponding POS tagging result.

4. Experimental results

In this section, we describe the dataset using in the experiment. Then, we give the experimental results and some discussions.

4.1. MSRA Chinese chunking Corpus

The MSRA chunking dataset is based on the Peking University corpus, which has been segmented, POS tagged, and chunk annotated manually. Forty-two types of

POS tags and forty-three types of chunk tags occurred in the data set. The overview of the chunking dataset can be seen in Table 2.

Table 2: MSRA Chinese chunking corpus

Chunking corpus	Chunks	words
Training dataset	*229,989*	*444,777*
Test dataset	*13,879*	*28,382*

The performance is measured with three rates: precision (P), recall (R) and F-score (F).

4.2. Experimental results and discussion

The training module of CCRF chunking model are on the basis of MSRA Chinese chunking corpus which is divided into training and test datasets. The Chinese chunking results is shown in Table 3 when the performance of POS tagging is 93.5% with CRF models:

Table 3: The chunking result based on cascaded conditional random fields

Model	P （%）	R （%）	F （%）
POS tagging + Chunking	82.01	81.07	81.50
Cascaded chunking model	85.80	84.33	85.06

Based on CCRF models, N-best POS tagging results are better than one POS tagging result. The improvement is 2.5% in accuracy. The cascaded chunking model restrains the affection of POS tagging errors. Therefore, the chunking performance increases to 85.06%. The improvement is 3.56%.

5. Conclusions

This paper describes a new algorithm for Chinese chunking based on CCRF models. The lower CRF models solve POS tagging problem. The upper CRF models solve chunking problem. The cascaded model restrains the affection of automatic POS tagging errors.

The experiments on the MSRA Chinese chunking corpus show that the new model achieves 85.06% in F-score. It performs considerably better than the chunking model based on the common method on the same task. The overall improvements are 3.56%.

Acknowledgement

This work is supported by National Natural Science Foundation 60673037, the High Technology Research and Development Program of China grant 2007AA01Z172, the Heilongjiang provincial Science Foundation F2007-06, Heilongjiang provincial special fund to support information industrialization.

6. References

[1] S. Abney, “Parsing by Chunks”, Principle-Based Parsing, Kluwer Academic Publishers, Dordrecht, pp 257-278, 1991.

[2] E. Tjong Kim Sang and S.Buchholz, “Introduction to the CoNLL-2000 Shared Task: Chunking”, *Proceeding of CoNLL-2000 and LLL-2000*, Lisbon, Portugal, pp. 127-132, 2000.

[3] G. Sun, C. Huang, X. Wang and Z. Xu. “Chinese Chunking Based on Maximum Entropy Markov Models.” International Journal of Computational Linguistics and Chinese Language Processing, 2006, 11(2): 115-136.

[4] Wenliang Chen, Yujie Zhang and Hitoshi Isahara. 2006. “An Empirical Study of Chinese Chunking.” In Pro-

ceedings of the 44th Annual Meeting of ACL, pages 97–104.
[5] T. Kudo and Y. Matsumoto, “Use of Support Vector Learning for Chunk Identification”, Proceeding of CoNLL-2000 and LLL-2000, Lisbon, Portugal, pp. 142-144, 2000.
[6] S. B. Park and B. T. Zhang, “Text Chunking by Combining Hand-crafted Rules and Memory-based Learning”, Proceedings of the 41st Annual Meeting of ACL, Sapporo, Japan, pp. 497-504, 2003.
[7] F. Sha and F. Pereira. “Shallow parsing with conditional random fields.” Proceedings of Human Language Technology Coference’2003, Edmonton, Canada, May 27-June 1, 2003, 134-141.
[8] T. Zhang, F. Damerau, D. Johnson. “Text chunking based on a generalization of winnow.” Journal of Machine Learning Research, 2(2002)3, 615-637.
[9] J. Lafferty, A. McCallum, and F. Pereira. “Conditional random fields: Probabilistic models for segmenting and labeling sequence data.” In Proc. ICML-01, pages 282-289, 2001.