A Hybrid Approach to Chinese Base Noun Phrase Chunking

Fang Xu **Chengqing Zong** **Jun Zhao**

National Laboratory of Pattern Recognition
Institute of Automation
Chinese Academy of Sciences, Beijing 100080,China
{fxu, cqzong, jzhao}@nlpr.ia.ac.cn

Abstract

In this paper, we propose a hybrid approach to chunking Chinese base noun phrases (base NPs), which combines SVM (Support Vector Machine) model and CRF (Conditional Random Field) model. In order to compare the result respectively from two chunkers, we use the discriminative post-processing method, whose measure criterion is the conditional probability generated from the CRF chunker. With respect to the special structures of Chinese base NP and complete analyses of the first two results, we also customize some appropriate grammar rules to avoid ambiguities and prune errors. According to our overall experiments, the method achieves a higher accuracy in the final results.

1 Introduction

Chunking means extracting the non-overlapping segments from a stream of data. These segments are called chunks (Dirk and Satoshi, 2003). The definition of base noun phrase (base NP) is simple and non-recursive noun phrase which does not contain other noun phrase descendants. Base NP chunking could be used as a precursor for many elaborate natural language processing tasks, such as information retrieval, name entity extraction and text summarization and so on. Many other problems similar to text processing can also benefit from base NP chunking, for example, finding genes in DNA and phoneme information extraction.

The initial work on base NP chunking is focused on the grammar-based method. Ramshaw and Marcus (1995) introduced a transformation-based learning method which considered chunking as a kind of tagging problem. Their work inspired many others to study the applications of learning methods to noun phrase chunking. (Cardie and Pierce, 1998, 1999) applied a scoring method to select new rules and a naive heuristic for matching rules to evaluate the results' accuracy.

CoNLL-2000 proposed a shared task (Tjong and Buchholz, 2000), which aimed at dividing a text in syntactically correlated parts of words. The eleven systems for the CoNLL-2000 shared task used a wide variety of machine learning methods. The best system in this workshop is on the basis of Support Vector Machines used by (Kudo and Matsumoto, 2000).

Recently, some new statistical techniques, such as CRF (Lafferty *et al*. 2001) and structural learning methods (Ando and Zhang, 2005) have been applied on the base NP chunking. (Fei and Fernando, 2003) considered chunking as a sequence labeling task and achieved good performance by an improved training methods of CRF. (Ando and Zhang, 2005) presented a novel semi-supervised learning method on chunking and produced performances higher than the previous best results.

The research on Chinese Base NP Chunking is, however, still at its developing stage. Researchers apply similar methods of English Base NP chunking to Chinese. Zhao and Huang (1998) made a strict definition of Chinese base NP and put forward a quasi-dependency model to analysis the structure of Chinese base NPs. There are some other methods to deal with Chinese phrase (no only base NP) chunking, such as HMM (Heng Li *et al.*, 2003), Maximum Entropy (Zhou Yaqian *et al.*, 2003), Memory-Based Learning (Zhang and Zhou, 2002) etc.

Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, pages 87–93,
Sydney, July 2006.

However, according to our experiments over 30,000 Chinese words, the best results of Chinese base NP chunking are about 5% less than that of English chunking (Although we should admit the chunking outcomes vary among different sizes of corpus and rely on the details of experiments). The differences between Chinese NPs and English NPs are summarized as following points: First, the flexible structure of Chinese noun phrase often results in the ambiguities during the recognition procedure. For example, many English base NPs begin with the determinative, while the margin of Chinese base NPs is more uncertain. Second, the base NPs begins with more than two noun-modifiers, such as "高(high)/JJ 新(new)/JJ 技术(technology)/NN", the noun-modifiers "高/JJ " can not be completely recognized. Third, the usage of Chinese word is flexible, as a Chinese word may serve with multi POS (Part-of-Speech) tags. For example, a noun is used as a verbal or an adjective component in the sentence. In this way the chunker is puzzled by those multi-used words. Finally, there are no standard datasets and elevation systems for Chinese base NP chunking as the CoNLL-2000 shared task, which makes it difficult to compare and evaluate different Chinese base NP chunking systems.

In this paper, we propose a hybrid approach to extract the Chinese base NPs with the help of the conditional probabilities derived from the CRF algorithm and some appropriate grammar rules. According to our preliminary experiments on SVM and CRF, our approach outperforms both of them.

The remainder of the paper is organized as follows. Section 2 gives a brief introduction of the data representations and methods. We explain our motivations of the hybrid approach in section 3. The experimental results and conclusions are introduced in section 4 and section 5 respectively.

2 Task Description

2.1 Data Representation

Ramshaw and Marcus (1995) gave mainly two kinds of base NPs representation — the open/close bracketing and IOB tagging. For example, a bracketed Chinese sentence,

[外商(foreign businessmen) 投资(investment)] 成为(become) [中国 (Chinese) 外贸(foreign trade)] [重要(important) 增长点(growth)] 。

The IOB tags are used to indicate the boundaries for each base NP where letter 'B' means the current word starts a base NP, 'I' for a word inside a base NP and 'O' for a word outside a NP chunk. In this case the tokens for the former sentence would be labeled as follows:

外商/B 投资/I 成为/V 中国/B 外贸/I 重要/B 增长点/O 。/O

Currently, most of the work on base NP identification employs the trainable, corpus-based algorithm, which makes full use of the tokens and corresponding POS tags to recognize the chunk segmentation of the test data. The SVM and CRF are two representative effective models widely used.

2.2 Chunking with SVMs

SVM is a machine learning algorithm for a linear binary classifier in order to maximize the margin of confidence of the classification on the training data set. According to the different requirements, distinctive kernel functions are employed to transfer non-linear problems into linear problems by mapping it to a higher dimension space.

By transforming the training data into the form with IOB tags, we can view the base NP chunking problem as a multi-class classification problem. As SVMs are binary classifiers, we use the pairwise method to convert the multi-class problem into a set of binary class problem, thus the I/O/B classifier is reduced into 3 kinds of binary classifier — I/O classifier, O/B classifier, B/I classifier.

In our experiments, we choose TinySVM[1] together with YamCha[2] (Kudo and Matsumoto, 2001) as the one of the baseline systems for our chunker. In order to construct the feature sets for training SVMs, all information available in the surrounding contexts, including tokens, POS tags and IOB tags. The tool YamCha makes it possible to add new features on your own. Therefore, in the training stage, we also add two new features according to the words. First, we give special tags to the noun words, especially the proper noun, as we find in the experiment the proper nouns sometimes bring on errors, such as base

[1] http://chasen.org/~taku/software/TinySVM/
[2] http://chasen.org/~taku/software/yamcha

NP “四川(Sichuan)/NR 盆地(basin)/NN”, containing the proper noun “四川/NR”, could be mistaken for a single base NP “盆地/NN”; Second, some punctuations such as separating marks, contribute to the wrong chunking, because many Chinese compound noun phrases are connected by separating mark, and the ingredients in the sentence are a mixture of simple nouns and noun phrases, for example,
“国家(National)/NN 统计局(Statistics Office)/NN，中国(Chinese)/NR 社会(Social Sciences)/NN 科学院(Academy)/NN 和(and)/CC 中科院(Chinese Academy of Sciences)/NN-SHORT”

The part of base NP – “中国/B 社会/I 科学院/I” can be recognized as three independent base NPs --“中国/B 社会/B 科学院/B”. The kind of errors comes from the conjunction “和(and)” and the successive sequences of nouns, which contribute little to the chunker. More information and analyses will be provided in Section 4.

2.3 Conditional Random Fields

Lafferty *et al.*(2001) present the Conditional Random Fields for building probabilistic models to segment and label sequence data, which was used effectively for base NP chunking (Sha & Pereira, 2003). Lafferty *et al.* (2001) point out that each of the random variable label sequences Y conditioned on the random observation sequence X. The joint distribution over the label sequence Y given X has the form

$$p(y \mid x, \lambda) = \frac{1}{Z(x)} \exp(\sum_j \lambda_j F_j(y, x))$$

$$F(y, x) = \sum_{i=1}^{n} f(y_{i-1}, y_i, x, i)$$

where $f_j(y_{i-1}, y_i, x, i)$ is either a transition feature function $s(y_{i-1}, y_i, x, i)$ or a state feature function $t(y_{i-1}, y_i, x, i)$; y_{i-1}, y_i are labels, x is an input sequence, i is an input position, $Z(x)$ is a normalization factor; λ_k is the parameter to be estimated from training data.

Then we use the maximum likelihood training, such as the log-likelihood to train CRF given training data $T = \{(x_k, y_k)\}$,

$$L(\lambda) = \sum_k \left[\log \frac{1}{Z(x_k)} + \lambda \cdot F(y_k, x_k) \right]$$

$L(\lambda)$ is minimized by finding unique zero of the gradient

$$\nabla L(\lambda) = \sum_k [F(y_k, x_k) - E_{p(Y|x_k, \lambda)} F(Y, x_k)]$$

$E_{p(Y|x_k, \lambda)} F(Y, x_k)$ can be computed using a variant of the forward-backward algorithm. We define a transition matrix as following:

$$M_i(y', y \mid x) = \exp(\sum_j \lambda_j f_j(y', y, x, i))$$

Then,

$$p(y \mid x, \lambda) = \frac{1}{Z(x)} \prod_{i=1}^{n+1} M_i(y_{i-1}, y_i \mid x)$$

and let * denote component-wise matrix product,

$$E_{p(Y|x_k, \lambda)} F(Y, x_k) = \sum_y p(Y = y \mid x_k, \lambda) F(y, x_k)$$

$$= \sum_i \frac{\alpha_{i-1}(f_i * M_i)\beta_i^T}{Z(x)}$$

$$Z(x) = a_n \cdot 1^T$$

Where α_i, β_i as the forward and backward state-cost vectors defined by

$$\alpha_i = \begin{cases} \alpha_{i-1} M_i & 0 < i \le n \\ 1 & i = 0 \end{cases}, \beta_i^T = \begin{cases} M_{i+1} \beta_{i+1}^T & 1 \le i < n \\ 1 & i = n \end{cases}$$

Sha & Pereira (2003) provided a thorough discussion of CRF training methods including preconditioned Conjugate Gradient, limited-Memory Quasi-Newton and voted perceptron. They also present a novel approach to model construction and feature selection in shallow parsing.

We use the software CRF++[3] as our Chinese base NP chunker baseline software. The results of CRF are better than that of SVM, which is the same as the outcome of the English base NP chunking in (Sha & Pereira, 2003). However, we find CRF products some errors on identifying long-range base NP, while SVM performs well in this aspect and the errors of SVM and CRF are of different types. In this case, we develop a combination approach to improve the results.

3 Our Approach

(Tjong *et al.*, 2000) pointed out that the performance of machine learning can be improved by combining the output of different systems, so they combined the results of different classifiers

[3] http://www.chasen.org/~taku/software/CRF++/

and obtained good performance. Their combination system generated different classifiers by using different data labels and applied respective voting weights accordingly. (Kudo and Matsumoto 2001) designed a voting arrangement by applying cross validation and VC-bound and Leave-One-Out bound for the voting weights.

The voting systems improve the accuracy, the choices of weights and the balance between different weights is based on experiences, which does not concern the inside features of the classification, without the guarantee of persuasive theoretical supports. Therefore, we developed a hybrid approach to combine the results of the SVM and CRF and utilize their advantages. (Simon, 2003) pointed out that the SVM guarantees a high generalization using very rich features from the sentences, even with a large and high-dimension training data. CRF can build efficient and robust structure model of the labels, when one doesn't have prior knowledge about data. Figure 1 shows the preliminary chunking and pos-processing procedure in our experiments

First of all, we use YamCha and CRF++ respectively to treat with the testing data. We got two original results from those chunkers, which use the exactly same data format; in this case we can compare the performance between CRF and SVM. After comparisons, we can figure out the same words with different IOB tags from the two former chunkers. Afterward, there exist two problems: how to pick out the IOB tags identified improperly and how to modify those wrong IOB tags.

To solve the first question, we use the conditional probability from the CRF to help determine the wrong IOB tags. For each word of the testing data, the CRF chunker works out a conditional probability for each IOB tag and chooses the most probable tag for the output. We bring out the differences between the SVM and CRF, such as "四川 (Sichuan)" in a base noun phrase is recognized as "I" and "O" respectively, and the distance between P(I| "四川") and P(O| "四川") is tiny. According to our experiment, about 80% of the differences between SVM and CRF share the same statistical characters, which indicate the correct answers are inundated by the noisy features in the classifier.

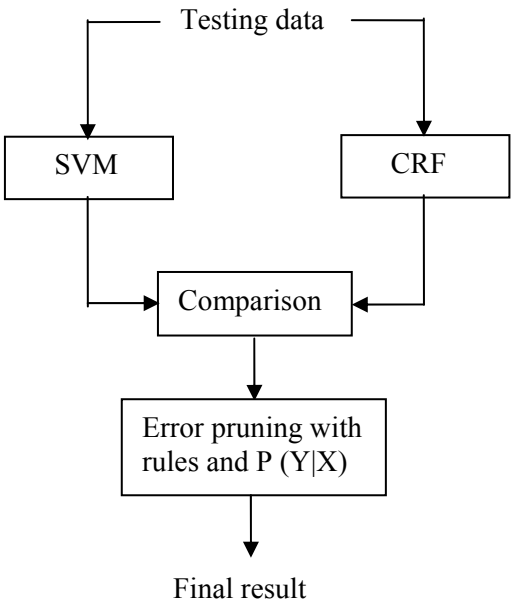

Figure 1 the Experiments' Procedure

Using the comparison between SVM and CRF we can check most of those errors. Then we could build some simple grammar rules to figure out the correct tags for the ambiguous words corresponding to the surrounding contexts. Then At the error pruning step, judging from the surrounding texts and the grammar rules, the base NP is corrected to the right form. We give 5 mainly representative grammar rules to explain how they work in the experiments.

The first simple sample of grammar rules is just like "BNP → NR NN", used to solve the proper noun problems. Take the "四川 (Sichuan)/NR/B 盆地 (basin)/NN/I" for example, the comparison finds out the base NP recognized as "四川 (Sichuan)/NR/I 盆地 (basin)/NN/B". Second, with respect to the base NP connecting with separating mark and conjunction words, two rules "BNP → BNP CC (BNP | Noun), BNP →BNP PU (BNP | Noun)" is used to figure out those errors; Third, with analyzing our experiment results, the CRF and SVM chunker recognize differently on the determinative, therefore the rule "BNP → JJ BNP", our combination methods figure out new BNP tags from the preliminary results according to this rule. Finally, the most complex situation is the determination of the Base NPs composed of series of nouns, especially the proper nouns. With figuring out the maximum length of this kind of noun phrase, we highlight the proper nouns and then separate the complex noun phrase to base noun phrases, and according to the our experiments, this

method could solve close to 75% of the ambiguity in the errors from complex noun phrases. Totally, the rules could solve about 63% of the found errors.

4 Experiments

The CoNLL 2000 provided the software[4] to convert Penn English Treebank II into the IOB tags form. We use the Penn Chinese Treebank 5.0[5], which is improved and involved with more POS tags, segmentation and syntactic bracketing. As the sentences in the Treebank are longer and related to more complicated structures, we modify the software with robust heuristics to cope with those new features of the Chinese Treebank and generate the training and testing data sets from the Treebank. Afterward we also make some manual adjustments to the final data.

In our experiments, the SVM chunker uses a polynomial kernel with degree 2; the cost per unit violation of the margin, C=1; and tolerance of the termination criterion, $\varepsilon = 0.01$.

In the base NPs chunking task, the evaluation metrics for base NP chunking include precision *P*, recall *R* and the F_β. Usually we refer to the F_β as the creditable metric.

$$P = \frac{\#\ of\ correct\ proposed\ baseNP}{\#\ of\ \ proposed\ baseNP} * 100\%$$

$$R = \frac{\#\ of\ correct\ proposed\ baseNP}{\#\ of\ corect\ baseNP} * 100\%$$

$$F_\beta = \frac{(\beta^2 + 1)RF}{\beta^2 R + F} \quad (\beta = 1)$$

All the experiments were performed on a Linux system with 3.2 GHz Pentium 4 and 2G memory. The total size of the Penn Chinese Treebank words is 13 MB, including about 500,000 Chinese words. The quantity of training corpus amounts to 300,000 Chinese words. Each word contains two Chinese characters in average. We mainly use five kinds of corpus, whose sizes include 30000, 40000, 50000, 60000 and 70,000 words. The corpus with an even larger size is improper according to the training corpus amount.

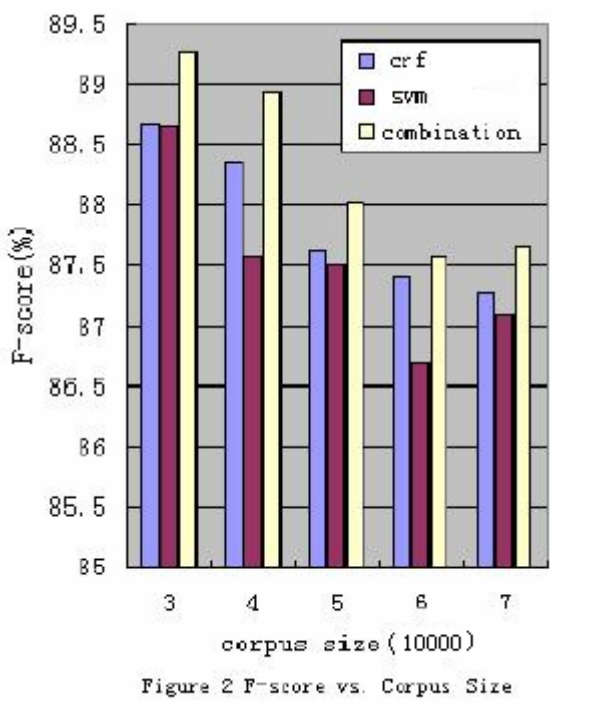

Figure 2 F-score vs. Corpus Size

From Figure 2, we can see that the results from CRF are better than that from SVM and the error-pruning performs the best. Our hybrid error-pruning method achieves an obvious improvement F-scores by combining the outcome from SVM and CRF classifiers. The test F-scores are decreasing when the sizes of corpus increase. The best performance with F-score of 89.27% is achieved by using a test corpus of 30k words. We get about 1.0% increase of F-score after using the hybrid approach. The F-score is higher than F-score 87.75% of Chinese base NP chunking systems using the Maximum Entropy method in (Zhou *et al.*, 2003),. Which used the smaller 3 MB Penn Chinese Treebank II as the corpus.

The Chinese Base NP chunkers are not superior to those for English. Zhang and Ando (2005) produce the best English base NP accuracy with F-score of 94.39+ (0.79), which is superior to our best results. The previous work mostly considered base NP chunking as the classification problem without special attention to the lexical information and syntactic dependence of words. On the other hand, we add some grammar rules to strength the syntactic dependence between the words. However, the syntactic structure derived from Chinese is much more flexible and complex than that from English. First, some Chinese words contain abundant meanings or play different syntactic roles. For example, "其中 (among which)/NN 重 庆 (Chongqing)/NR 地 区 (district)/NN" is recognized as a base NP. Actu-

[4] http://ilk.kub.nl/~sabine/chunklink/
[5] http://www.cis.upenn.edu/~chinese/

ally the Chinese word "其中/NN (among)" refers to the content in the previous sentence and "其中 (thereinto)" sometimes used as an adverb. Second, how to deal with the conjunctions is a major problem, especially the words "与 (and)" can appear in the preposition structure "与 …… 相关 (relate to)", which makes it difficult to judge those types of differences. Third, the chunkers can not handle with compact sequence data of chunks with name entities and new words (especially the transliterated words) satisfactorily, such as

"中国 (China) /NR 红十字会(Red Cross) /NR名誉 (Honorary) /NN 会长 (Chairman) /NN 江泽民(Jiang Ze-min) /NR"

As it points above, the English name entities sequences are connected with the conjunction such as "of, and, in". While in Chinese there are no such connection words for name entities sequences. Therefore when we use the statistical methods, those kinds of sequential chunks contribute slightly to the feature selection and classifier training, and are treated as the useless noise in the training data. In the testing section, it is close the separating margin and hardly determined to be in the right category. What's more, some other factors such as Idiomatic and specialized expressions also account for the errors. By highlighting those kinds of words and using some rules which emphasize on those proper words, we use our error-pruning methods and useful grammar rules to correct about 60% errors.

5 Conclusions

This paper presented a new hybrid approach for identifying the Chinese base NPs. Our hybrid approach uses the SVM and CRF algorithm to design the preliminary classifiers for chunking. Furthermore with the direct comparison between the results from the former chunkers, we figure out that those two statistical methods are myopic about the compact chunking data of sequential noun. With the intention of capturing the syntactic dependence within those sequential chunking data, we make use of the conditional probabilities of the chunking tags given the corresponding tokens derived from CRF and some simple grammar rules to modify the preliminary results.

The overall results achieve 89.27% precision on the base NP chunking. We attempt to explain some existing semantic problems and solve a certain part of problems, which have been discovered and explained in the paper. Future work will concentrate on working out some adaptive machine learning methods to make grammar rules automatically, select better features and employ suitable classifiers for Chinese base NP chunking. Finally, the particular Chinese base phrase grammars need a complete study, and the approach provides a primary solution and framework to process an analyses and comparisons between Chinese and English parallel base NP chunkers.

Acknowledgments

This work was partially supported by the Natural Science Foundation of China under Grant No. 60575043, and 60121302, the China-France PRA project under Grant No. PRA SI02-05, the Outstanding Overseas Chinese Scholars Fund of the Chinese Academy of Sciences under Grant No.2003-1-1, and Nokia (China) Co. Ltd, as well.

References

Claire Cardie and David Pierce. 1998. *Error-Driven Pruning of Treebank Grammars for Base Noun Phrase Identification.* Proceedings of the 36th ACL and COLING-98, 218-224.

Claire Cardie and David Pierce. 1999. *The role of lexicalization and pruning for base noun phrase grammars.* Proceedings of the 16th AAAI, 423-430.

Dirk Ludtke and Satoshi Sato. 2003. *Fast Base NP Chunking with Decision Trees —— Experiments on Different POS tag Settings.* CICLing 2003, 136-147. LNC S2588, Springer-Verlag Berlin Heidelberg.

Erik F. Tjong Kim Sang and Sabine Buchholz. 2000. *Introduction to the CoNLL-2000 Shared Task: Chunking*. Proceedings of CoNLL and LLL-2000, 127-132.

Erik F. Tjong Kim Sang, Walter Daelemans, Hervé Déean, Rob Koeling, Yuval Krymolowski, Vasin Punyakanok, and Dan Roth. 2000. *Applying system combination to base noun phrase identification.* Proceedings of COLING 2000, 857-863.

Fei Sha and Fernando Pereira. 2003. *Shallow Parsing with Conditional Random Fields.* Proceedings of HLT-NAACL 2003, 134-141.

Heng Li, Jonathan J. Webster, Chunyu Kit, and Tianshun Yao. 2003. *Transductive HMM based Chinese Text Chunking.* IEEE NLP-KE 2003, Beijing, 257-262.

Lance A. Ramshaw and Mitchell P. Marcus. 1995. *Text Chunking using Transformation-Based Learning.* Proceedings of the Third ACL Workshop on Very Large Corpora, 82–94.

Lafferty A. McCallum and F. Pereira. 2001. *Conditional random Fields.* Proceedings of ICML 2001, 282-289.

Rie Kubota Ando and Tong Zhang. 2004. *A framework for learning predictive structures from multiple tasks and unlabeled data.* RC23462. Technical report, IBM.

Rie Kubota Ando and Tong Zhang. 2005. *A High-Performance Semi-Supervised Learning Method for Text Chunking*. Proceedings of the 43rd Annual Meeting of ACL, 1-9.

Simon Lacoste-Julien. 2003. *Combining SVM with graphical models for supervised classification: an introduction to Max-Margin Markov Network.* CS281A Project Report, UC Berkeley.

Taku Kudo and Yuji Matsumoto. 2001. *Chunking with support vector machine*. Proceeding of the NAACL, 192-199.

Zhang Yuqi and Zhou Qiang. 2002. *Chinese Base-Phrases Chunking.* First SigHAN Workshop on Chinese Language Processing, COLING-02.

Zhao Jun and Huang Changling. 1998. *A Quasi-Dependency Model for Structural Analysis of Chinese BaseNPs.* 36th Annual Meeting of the ACL and 17th International Conference on Computational Linguistics.

Zhou Yaqian, Guo YiKun, Huang XuanLing and Wu Lide. 2003. *Chinese and English Base NP Recognition on a Maximum Entropy Model.* Vol140, No13. Journal of Computer Research and Development. (In Chinese)