Chinese Chunking based on Conditional Random Fields

Wenliang Chen, Yujie Zhang, Hitoshi Isahara
Computational Linguistics Group
National Institute of Information and Communications Technology
3-5 Hikari-dai, Seika-cho, Soraku-gun, Kyoto 619-0289, Japan
{chenwl, yujie, isahara}@nict.go.jp

Abstract

In this paper, we proposed an approach for Chinese chunking based on the Conditional Random Fields model (CRFs). For sequence labeling, CRFs has advantages over generative models. Furthermore, Chinese chunking is a difficult sequence labeling task. This paper describes how to use CRFs for Chinese chunking via capturing the arbitrary and overlapping features. We defined different types of features for the model, and then studied their effects on the data set of the UPENN Chinese TreeBank-4(CTB4). For comparison, we also applied the other models to the task on the same data set. The experimental results show that the proposed approach can achieve better performance than the other models.

1 Introduction

Chunking identifies the non-recursive cores of various types of phrases in text, possibly as a precursor to full parsing or information extraction[1]. Steven P. Abney[1] was the first person to introduce chunks for parsing. Ramshaw and Marcus[2] first represented base noun phrase recognition as a machine learning problem. In 2000, CoNLL-2000 introduced a shared task to tag many kinds of phrases besides noun phrase in English[3]. Additionally, many machine learning approaches, such as Memory-based Learning (MBL), Transformation-based Learning (TBL), and Hidden Markov Models (HMMs), have been applied to text chunking[3, 4].

Chinese chunking is a difficult sequence labeling task, and much work has been done on this topic[5, 6, 7, 8]. However, comparing the performance of the approaches in Chinese chunking is very difficult because of different chunk definitions and different data sets. With some machine learning approaches like Conditional Random Fields(CRFs) and Support Vector machines (SVMs), how to represent the data is an important issue. Megyesi[9] gained the best result using only the features of Part-Of-Speech tags, while some researchers use a combination of lexical and Part-Of-Speech information.

The aim of this study is, in particular, to find out what combinations of linguistic information are the most appropriate for Chinese chunking. Conditional Random Fields is a powerful sequence labeling model[10, 11], which is combining the advantages both the generative model and the classification model. Sha and Pereira[11] showed that CRFs can gain the state-of-the-art results in English chunking. In this paper, we build a Chinese chunker based on CRFs. For comparison, we conduct experiments with our Chinese chunker, MBL, and TBL on the dataset of UPENN Chinese Treebank-4 (CTB4).

2 Chinese Chunks

There are some different Chinese chunk definitions, which are often derived from different data sets. For instance, Li et al.[6] defined the chunks from the MSRA corpus, while Zhang and Zhou [8] defined the chunks from their own corpus. In this paper, we define the Chinese chunks according to the data set of CTB4, similar to the chunk definition by Tan et al.[7]. Here we define 12 types of chunks[1]: ADJP, ADVP, CLP, DNP, DP, DVP, LCP, LST, NP, PP, QP, VP. Table 1 shows the explanation of these chunks.

Table 1: Explanation of Chunks

Type	Explanation
ADJP	Adjective Phrase
ADVP	Adverbial Phrase
CLP	Classifier Phrase
DNP	DEG Phrase
DP	Determiner Phrase
DVP	DEV phrase
LCP	Localizer Phrase
LST	List Marker
NP	Noun Phrase
PP	Prepositional Phrase
QP	Quantifier Phrase
VP	Verb Phrase

To represent the chunks clearly, we represent the data using an IOB-based model, in which every word is to be tagged with a chunk label extended with I(inside a chunk), O (outside a chunk) or B (inside a chunk, but it is the first word of the chunk). Each chunk type could be extended with I or B tags. For instance, NP could be considered as two types of tags, B-NP or I-NP. We have 25 types of chunk tags based on IOB model, and every word in the sentence will be tagged with one of these chunk tags. For

[1] There are 15 types of chunks in the Upenn Chinese TreeBank-4. The other chunk types are FRAG, PRN, and UCP.

instance, a sentence "他-NR(he) /到达-VV(reached) /北京-NR(Beijing) /机场-NN(airport) /。/" will be tagged as follows:
Example 1:
S1: [NP 他][VP 到达][NP 北京/机场][O 。]
S2: B-NP 他/B-VP 到达/B-NP 北京/I-NP 机场/O 。/
Where, S1 denotes that the sentence is tagged with chunk types, and S2 denotes that the sentence is tagged with chunk tags based on IOB model.

After the data representation, the problem of Chinese chunking can be regarded as a sequence tagging task. Given a sequence of tokens, $x = x_1x_2...x_n$, we need to generate a sequence of chunk tags, $y = y_1y_2...y_n$.

3 Conditional Random Fields

3.1 The CRFs model

Conditional Random Fields(CRF), a statistical sequence modeling framework, was first introduced by Lafferty et al[12]. The model has been used for English chunking[11]. We only describe the model briefly since full details are presented in the paper[12].

For our sequence tagging problem, we create a linear-chain CRF based on an undirected graph $G = (V, E)$, where V is the set of random variables $Y = \{Y_i|1 \leq i \leq n\}$, for each of n tokens in an input sentence and $E = \{(Y_{i-1}, Y_i)|1 \leq i \leq n\}$ is the set of $n-1$ edges forming a linear chain. For each sentence x, we define two non-negative factors:
$exp(\sum_{k=1}^{K} \lambda_k f_k(y_{i-1}, y_i, x))$ for each edge
$exp(\sum_{k=1}^{K'} \lambda_k^{'} f_k^{'}(y_i, x))$ for each node
where f_k is a binary feature function, and K and K' are the number of features defined for edges and nodes respectively. Following Lafferty et al[12], the conditional probability of a sequence of tags y given a sequence of tokens x is:

$$P(y|x) = \frac{1}{Z(x)} exp(\sum_{i,k} \lambda_k f_k(y_{i-1}, y_i, x) + \sum_{i,k} \lambda_k^{'} f_k^{'}(y_i, x)) \tag{1}$$

where $Z(x)$ is the normalization constant. Given the training data D, a set of sentences (words with their part-of-speech tags), the parameters of the model are trained to maximize the conditional log-likelihood. When testing, given a sentence x in the test data, the tagging sequence y is given by $Argmax_{y'} P(y'|x)$.

CRFs allow us to utilize a large number of observation features as well as different state sequence based features and other features we want to add.

3.2 CRFs for Chinese chunking

Our CRFs-based chunker has a second-order Markov dependency between chunk tags.

In our experiments, we do not use feature selection and all features are used in training and testing. We use the following feature functions:

$$f(y_{i-1}, y_i, x, i) = p(x, i)q(y_{i-1}, y_i) \tag{2}$$

where $p(x, i)$ is a predicate on the input sequence x and current position i and $q(y_{i-1}, y_i)$ is a predicate on pairs of labels. For instance, $p(x, i)$ might be "the word at position i is 和(and)" or "the POS tags at positions $i-1$, i are NR, CC."

3.3 The Features

To obtain a good-quality estimation of the conditional probability of the event tag, the observations should be based on features that represent the difference of the two events. In this paper, we utilize both lexical and Part-Of-Speech(POS) information.

Tan et al.[7] applied the CRFs model on Chinese chunking with the features of lexical and POS information at the current position. In contrast to their simple approach, we use the lexical and POS information within a fixed window. We also consider different combinations of these. The features are listed as follows:

- WORD: uni-gram and bi-grams of words in an n window.
- POS: uni-gram and bi-grams of POS in an n window.
- Mixed:word-POS pairs in an n window. In Example 1, Mixed feature might be "the word-POS pair at position 3 is 北京-NR".

where n is a predefined number to denote window size.

4 Experiments

4.1 Experimental Setting

The UPENN Chinese Treebank-4(CTB4)[2] consists of 838 files. We use a tool[3] to generate the chunks tags(See section 2) from CTB4.

In the experiments, we used the first 728 files (FID from chtb_001.fid to chtb_899.fid) as training data, and the other 110 files as testing data. Table 2 shows the corpus information. To investigate the chunker sensitivity to the size of the training set, we generated different sizes of training sets, including 1%, 2%, 5%, 10%, 20%, 50%, 100% of the total training data.

Table 2: The CTB4 Corpus

	Training	Test
Num of Files	728	110
Num of Sentences	9,878	5,290
Num of Words	238,906	165,862
Num of Phrases	141,426	101,449

[2] More detailed information is at http://www.cis.upenn.edu/ chinese/

[3] Tool is available at http://www.nlplab.cn/chenwl/tools/chunklinkctb.txt

We used MALLET (V0.3.2)[13] to implement the CRF model, and we used all the default parameter settings of the package in our experiments.

The performance of the algorithm is measured with two scores: precision P and recall R. Precision measures how many chunks found by the algorithm are correct and the recall rate contains the percentage of chunks defined in the corpus that were found by the chunking program. The two rates can be combined in one measure:

$$F_1 = \frac{2 \times P \times R}{R + P} \qquad (3)$$

In this paper, we report the results with F_1 score.

4.2 Experiment 1: Effect of the features

In this experiment, we compared the performance of different combinations of features, including WORD, POS, WORD+ POS (Use WORD and POS at once), Mixed and ALL (WORD+ POS+ Mixed) (See section 3.3). We also investigated the effects of different sizes of training data.

First, we only used the features at the current position. Figure 1 shows the experimental results, where xtics denotes the size of the training data. We can see from the figure that POS, WORD+POS and ALL yield better performance than WORD and Mixed. When the size of training data was small, POS provided the best performance among all. However WORD+POS performed best when the size of training data increased. WORD+POS provided 80.48% F_1 on 100% of training data, while POS yielded 78.66% F_1.

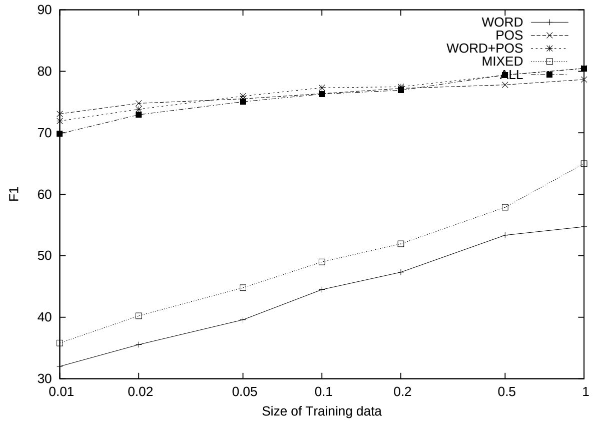

Figure 1: Results of different features

Then we investigated the effect of different window sizes using the features WORD+POS. Figure 2 shows the experimental results, where WIN0 denotes only that information of the current position was used, WIN1 denotes the features within 1-window-size, and so on. When the size of the training set was smaller than 10%, WIN1 performed better than the others. Otherwise, WIN2 had the best results.

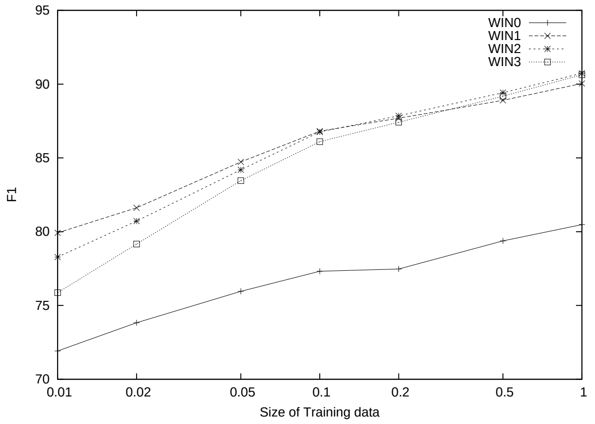

Figure 2: Effect of window-size

4.3 Experiment 2: Comparison with the other Approaches

In this experiment, we provide empirical evidence to prove that CRFs provides the state-of-the-art results for Chinese chunking. We apply two machine learning models based on MBL and TBL. Here we use TiMBL V5.1[4] toolkits to implement the MBL model, fnTBL V1.0[5] for the TBL model. In the experiments, we used the feature WORD+POS and set the window size as 2.

Table 3 shows the comparative results of the approaches. We can find that the CRFs approach is superior to both the MBL model and the TBL model. The CRF approach yielded 0.99% and 2.28% higher than the TBL model and the MBL model respectively.

Table 3: Comparative Results of Approaches

	CRFs	TBL	MBL
ADJP	84.55	84.21	79.09
ADVP	82.74	83.32	77.35
CLP	0.00	0.00	3.85
DNP	99.64	99.66	99.60
DP	99.40	99.70	99.25
DVP	92.89	99.61	99.41
LCP	99.85	99.80	99.80
LST	68.25	59.13	65.22
NP	89.79	89.66	87.48
PP	99.66	99.66	99.56
QP	96.53	96.56	96.16
VP	88.50	85.05	81.98
All	**90.74**	89.75	87.46

[4] TiMBL is available at http://ilk.uvt.nl/timbl/

[5] fnTBL is available at http://nlp.cs.jhu.edu/ rflorian/fntbl/index.html

4.4 Discussion

In this section, we make some detailed analysis about the chunking results to know what problems may still occurred.

Noun-Noun compounds: Compounds formed by more than two neighboring nouns are very common in Chinese. For instance, 世界(world)/ 和平(peace)/ 事业(work)/ (The cause of world peace). It is very difficult to distinguish whether it is {{世界/和平}/事业} or {世界/ {和平/ 事业}}. We also can see another example: ”教育(Education)/ 心理(Psychological)/ 系(Department)/” VS ”金融(Financial)/ 系(Department)/”. They are tagged as ”教育B-NP/ 心理I-NP/ 系B-NP/” and ”财金B-NP/ 系I-NP/”.

Coordination ambiguities: These problems are related to the conjunctions ”和(and) 与(and) 或(or) 暨(and)”. They can be divided into two types: chunks with conjunctions and without conjunctions. For instances, ”香港(HongKong)/ 和(and)/ 澳门(Macau)/” is an NP chunk (香港B-NP/ 和I-NP/ 澳门I-NP/), while in ”最低(least)/ 工资(salary)/ 和(and)/ 生活费(living maintenance)/” it is difficult to tell whether ”最低” is a shared modifier or not, even for people.

5 Conclusion and Future work

In this paper, we have described how to apply Conditional Random Fields model in Chinese chunking. The model can combine the features of lexical and part-of-speech tags. We also investigated the effect of different sizes of training data.

The experimental results showed that the CRFs model is a competitive approach for Chinese chunking. It can combine the advantages of different types of features, and so outperforms the MBL model and the TBL model. From the results, we also found that part-of-speech tags play an important role in Chinese chunking. When the size of training data was small, POS performed better than other combinations. The combination of lexical and part-of-speech (WORD+POS) performed best when the size of training data increased.

In our future work, we will study the effects of linguistic knowledge in the CRF framework, rather than using the information generated from training corpus. Another future work is to explore an unlabeled corpus to improve the performance of Chinese chunking.

References

[1] S. P. Abney. Parsing by chunks. In Robert C. Berwick, Steven P. Abney, and Carol Tenny, editors, *Principle-Based Parsing: Computation and Psycholinguistics*, pages 257–278. Kluwer, Dordrecht, 1991.

[2] Lance Ramshaw and Mitch Marcus. Text chunking using transformation-based learning. In David Yarovsky and Kenneth Church, editors, *Proceedings of the Third Workshop on Very Large Corpora*, pages 82–94, Somerset, New Jersey, 1995. Association for Computational Linguistics.

[3] Erik F. Tjong Kim Sang and Sabine Buchholz. Introduction to the conll-2000 shared task: Chunking. In *Proceedings of CoNLL-2000 and LLL2000*, pages 127–132, Lisbin, Portugal, 2000.

[4] James Hammerton, Miles Osborne, Susan Armstrong, and Walter Daelemans. Introduction to special issue on machine learning approaches to shallow parsing. *JMLR*, 2(3):551–558, 2002.

[5] Heng Li, Jonathan J. Webster, Chunyu Kit, and Tianshun Yao. Transductive hmm based chinese text chunking. In *Proceedings of IEEE NLP-KE2003*, pages 257–262, Beijing, China, 2003.

[6] H Li, CN Huang, J Gao, and X Fan. Chinese chunking with another type of spec. In *The Third SIGHAN Workshop on Chinese Language Processing*, 2004.

[7] Yongmei Tan, Tianshun Yao, Qing Chen, and Jingbo Zhu. Applying conditional random fields to chinese shallow parsing. In *Proceedings of CICLing-2005*, pages 167–176, Mexico City, Mexico, 2005. Springer.

[8] Yuqi Zhang and Qiang Zhou. Chinese base-phrases chunking. 2002.

[9] B. Megyesi. Shallow parsing with pos taggers and linguisitc features. *JMLR*, 2(3):639–668, 2002.

[10] Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. Identifying sources of opinions with conditional random fields and extraction patterns. In *HLT/EMNLP2005*, 2005.

[11] F. Sha and F. Pereira. Shallow parsing with conditional random fields, 2003.

[12] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning (ICML01)*, 2001.

[13] Andrew Kachites McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.