Complete Syntactic Analysis Based on Multi-level Chunking

ZhiPeng Jiang and **Yu Zhao** and **Yi Guan** and
Chao Li and **Sheng Li**
School of Computer Science and Technology,
Harbin Institute of Technology,
150001, Harbin, China
xyf-3456@163.com; woshizhaoy@gmail.com
guanyi@hit.edu.cn; beyondlee2008@yahoo.cn
lisheng@hit.edu.cn

Abstract

This paper describes a complete syntactic analysis system based on multi-level chunking. On the basis of the correct sequences of Chinese words provided by CLP2010, the system firstly has a Part-of-speech (POS) tagging with Conditional Random Fields (CRFs), and then does the base chunking and complex chunking with Maximum Entropy (ME), and finally generates a complete syntactic analysis tree. The system took part in the Complete Sentence Parsing Track of the Task 2 Chinese Parsing in CLP2010, achieved the F-1 measure of 63.25% on the overall analysis, ranked the sixth; POS accuracy rate of 89.62%, ranked the third.

1 Introduction

Chunk is a group of adjacent words which belong to the same s-projection set in a sentence, whose syntactic structure is actually a tree (Abney, 1991), but apart from the root node, all other nodes are leaf nodes. Complete syntactic analysis requires a series of analyzing processes, eventually to get a full parsing tree. Parsing by chunks is proved to be feasible (Abney, 1994).

The concept of chunking was first proposed by Abney in 1991, who defined chunks in terms of major heads, and parsed by chunks in 1994 (Abney, 1994). An additional chunk tag set {B, I, O} was added to chunking (Ramshaw and Marcus, 1995), which limited dependencies between elements in a chunk, changed chunking into a question of sequenced tags, to promote the development of chunking. Chunking algorithm was extended to the bottom-up parser, which is trained and tested on the Wall Street Journal (WSJ) part of the Penn Treebank (Marcus, Santorini and Marcinkiewicz 1993), and achieved a performance of 80.49% F-measure, the results show that it performed better than a standard probabilistic context-free grammar, and can improve performance by adding the information of parent node (Sang, 2000).

On Chinese parsing, Maximum Entropy Model was first used to have a POS tagging and chunking, and then a full parsing tree was generated (Fung, 2004), training and testing in the Penn Chinese Treebank, which achieved 79.56% F-measure. The parsing process was divided into POS tagging, base chunking and complex chunking, having a POS tagging and chunking on a given sentence, and then looping the process of complex chunking up to identify the root node (Li and Zhou, 2009). This parsing method is the basis of this paper. In addition, we have the existing Chinese chunking system in laboratory, which ranked first in Task 2: Chinese Base Chunking of CIPS-ParsEval-2009, so we try to apply chunking to complete syntactic analysis in CLP2010, to achieve better results.

We will describe the POS tagging based on CRFs in Section 2, including CRFs, feature template selection and empirical results. Multi-level chunking based on ME will be expounded in Section 3, including ME, MEMM, base chunking and complex chunking. Finally, we will summarize our work in Section 4.

2 POS Tagging Based on CRFs

2.1 Conditional Random Fields

X is a random variable over data sequences to be labeled, and Y is a random variable over corresponding label sequences. All components Y_i of Y are assumed to range over a finite label alphabet. For example, X might range over natural language sentences and Y range over part-of-speech tags of those sentences, a finite label alphabet is the set of possible part-of-speech tags (Lafferty and McCallum and Pereira, 2001). CRFs is represented by the local feature vector f and the corresponding weight vector, f is divided into the state feature s (y, x, i) and transfer feature t (y, y', x, i), where y and y' are possible POS tags, x is the current input sentence, i is the position of current term (Jiang and Guan and Wang, 2006). Formalized as follows:

$$s\ (y, x, i) = s\ (y_i, x, i) \qquad (1)$$

$$t(y,x,i)=\begin{cases} t(y_{i-1}, y_i, x, i) & i>1 \\ 0 & i=1 \end{cases} \qquad (2)$$

By the local feature of the formula (1) and (2), the global features of x and y:

$$F(y,x)=\sum_i f(y,x,i) \qquad (3)$$

At this point of (X, Y), the conditional probability distribution of CRFs:

$$p_\lambda(Y\,|\,X) = \frac{\exp(\lambda\cdot F(Y,X))}{Z_\lambda(X)} \qquad (4)$$

where $Z_\lambda(x)=\sum_y \exp(\lambda\cdot F(y,x))$ is a factor for normalizing. For the input sentence x, the best sequence of POS tagging:

$$y^* = \arg\max_y p_\lambda(y\,|\,x)$$

2.2 Feature Template Selection

We use the template as a baseline which is taken by Yang (2009) in CIPS-ParsEval-2009, directly testing the performance, whose accuracy was 93.52%. On this basis, we adjust the feature template through the experiment, and improve the tagging accuracy of unknown words by introducing rules, in the same corpus for training and testing, accuracy is to 93.89%. Adjusted feature template is shown in Table 1, in which the term pre is the first character in current word, suf is the last character of current word, num is the number of characters of current word, pos_{-1} is the tagging results of the previous word.

Table 1: feature template

feature template
w_2,w_1,w_0,w_{-1},w_{-2},$w_{+1}w_0$,w_0w_{-1},pre_0, pre_0w_0,suf_0, w_0suf_0,num,pos_{-1}

2.3 Empirical Results and Analysis

We divide the training data provided by CLP2010 into five folds, the first four of which are train corpus, the last one is test corpus, on which we use the CRF++ toolkit for training and testing. Tagging results with different features are shown in table 2.

Table 2: tagging results with different features

Model	Explain	Accuracy
CRF	baseline	93.52%
CRF1	add w_{-1}, pos_{-1}	93.58%
CRF2	add num	93.66%
CRF3	add num, w_{-1}, pos_{-1}	93.68%
CRF4	add num, rules	93.80%
CRF5	add num, w_{-1}, pos_{-1}, rules	93.89%

Tagging results show that the number of character and POS information can be added to improve the accuracy of tagging, but in CLP2010, the tagging accuracy is only 89.62%, on the one hand it may be caused by differences of corpus, on the other hand it may be due to that we don’t use all the features of CRFs but remove the features which appear one time in order to reduce the training time.

3 Multi-level Chunking Based on ME

3.1 Maximum Entropy Models and Maximum Entropy Markov Models

Maximum entropy model is mainly used to estimate the unknown probability distribution whose entropy is the maximum under some existing conditions. Suppose h is the observations of context, t is tag, the conditional probability p (t | h) can be expressed as:

$$P(t\,|\,h)=\frac{\exp(\sum_i \lambda_i f_i(t,h))}{Z(h)}$$

where f_i is the feature of model, $Z(h)=\sum_t \exp(\sum_i \lambda_i f_i(t,h))$ is a factor for normalizing. λ_i is weigh of feature f_i, training is the process of seeking the value of λ_i.

Maximum entropy Markov model is the serialized form of Maximum entropy model (McCallum and Freitag and Pereira, 2000), for example, transition probabilities and emission probabilities are merged into a single conditional probability function $P(t_i \mid t_{i-1}, h)$ in binary Maximum entropy Markov model, $P(t_i \mid t_{i-1}, h)$ is turned to $p(t \mid h)$ to be solved by adding features which can express previously tagging information (Li and Sun and Guan, 2009).

3.2 Base Chunking

Following the method of multi-level chunking, we first do the base chunking on the sentences which are through the POS tagging, then loop the process of complex chunking until they can’t be merged. We use the existing Chinese base chunking system to do base chunking in laboratory, which marks boundaries and composition information of chunk with MEMM, and achieved 93.196% F-measure in Task 2: Chinese Base Chunking of CIPS-ParsEval -2009. The input and output of base chunking are as follows:

Input：中国/nS 传统/a 医学/n 是/v 中华/nR 民族/n 在/p 长期/n 的/uJDE 医疗/n 、/wD 生活/n 实践/vN 中/f ，/wP 不断/d 积累/v ，/wP 反复/d 总结/v 而/c 逐渐/d 形成/v 的/uJDE 具有/v 独特/a 理论/n 风格/n 的/uJDE 医学/n 体系/n 。/wE

Output：中国/nS [np 传统/a 医学/n] 是/v [np 中华/nR 民族/n] 在/p 长期/n 的/uJDE [np 医疗/n 、/wD 生活/n] 实践/vN 中/f ，/wP [vp 不断/d 积累/v] ，/wP [vp 反复/d 总结/v] 而/c [vp 逐渐/d 形成/v] 的/uJDE 具有/v [np 独特/a 理论/n] 风格/n 的/uJDE [np 医学/n 体系/n] 。/wE

3.3 Complex Chunking

We take the sentences which are through POS tagging and base chunking as input, using Li’s tagging method and feature template. Categories of complex chunk include xx_Start, xx_Middle, xx_End and Other, where xx is a category of arbitrary chunk. The process of complex chunking is shown as follows:

Step 1: input the sentences which are through POS tagging and base chunking, for example:

中国/nS [np 传统/a 医学/n] 的/uJDE [np 发生/vN 发展/vN] 及/c [np 学术/n 特点/n]

Step 2: if there are some category tags in the sentence, then turn a series of tags to brackets, for instance, if continuous cells are marked as xx_Start, xx_Middle, ..., xx_Middle, xx_End, then the combination of continuous cells is a complex chunk xx;

Step 3: determine the head words with the set of rules, and compress the sentence:

中国/nS [np 医学/n] 的/uJDE [np 发展/vN] 及/c [np 特点/n]

Step 4: if the sentence can be merged, mark the sentence with ME, then return step 2, else the analysis process ends:

中国/nS@np_Start [np 医学/n]@np_End 的/uJDE@Other [np 发展/vN]@np_Start 及/c@np_Middle [np 特点/n]@np_End

At last, the output is:

[np [np 中国/nS [np 传统/a 医学/n]] 的/uJDE [np [np 发生/vN 发展/vN] 及/c [np 学术/n 特点/n]]]

Following the above method, we first use the Viterbi decoding, but in the decoding process we encountered two problems:

1. Similar to the label xx_Start, whose back is only xx_Middle or xx_End, so the probability of xx_Start label turning to Other is 0, But, if only using ME to predict, the probability may not be 0.
2. Viterbi decoding can’t solve that all the labels of predicted results are Other, if all labels are Other, they can’t be merged, this result doesn’t make sense.

Solution:

For the first question, we add the initial transfer matrix and the end transfer matrix in decoding process, that is, the corresponding xx_Middle or xx_End of xx_Start is seted to 1 in the transfer matrix, the others are marked as 0, matrix multiplication is taken during the state transition. It can effectively avoid errors caused by probability to improve accuracy.

To rule out the second question, we use heuristic search approach to decode, and exclude all Other labels with the above matrix. In addition, we defined another ME classifier to do some pruning in the decoding process, the features of ME classifier are POS, the head word, the POS of head word. The pseudo-code of Heuristic search is:

```
While searching priority queue is not empty
    Take the node with the greatest priority in the
        queue;
    If the node’s depth = length of the chunking
        results
            Searching is over, reverse the search-
                ing path to get searching results;
    Else
            Compute the probability of all candi-
                date children nodes according to
                the current probability;
            Record searching path;
            Press it into the priority queue;
```

In addition, we found that some punctuation at the end of a sentence can’t be merged, probably due to sparseness of data, according to that the tone punctuation (period, exclamation mark, question mark) at the end of the sentence can be added to implement a complete sentence (zj) (Zhou, 2004), we carried out a separate deal with this situation, directly add punctuation at the end of the sentence, to form a sentence.

In training data provided by CLP2010 in subtask: Complete Sentence Parsing, the head words aren’t marked. We can’t use the statistical method to determine the head words, but only by rules. We take Li’s rule set as baseline, but the rule set was used to supplement the statistical methods, so some head words don’t appear in the rule set, resulting in many head words are marked as NULL, for this situation, we add some rules through experiment, Table 3 lists some additional rules.

Table 3: increasing part of rules

parent	head words
vp	vp, vB, vSB, vM, vJY, vC, v
ap	a, b, d
mp	qN, qV, qC, q
dj	vp, dj, ap, v, fj
dlc	vp
mbar	m, mp

3.4 Empirical Results and Analysis

We take the corpus which are through correct POS tagging and base chunking for training and testing, it is divided into five folds, the first four as training corpus, the last one as testing corpus, using the existing ME toolkit to train and test model in laboratory. Table 4 shows the results on Viterbi decoding and Heuristic Search method, where head words are determined by rules.

Table 4: results with different decoding

Decoding	Accuracy	Recall	Fmeasure
Viterbi	84.87%	84.47%	84.67%
Heuristic Search	85.62%	85.19%	85.40%

The system participated in the Complete Sentence Parsing of CLP2010, results are shown in Table 5 below. Because we can’t determine the head words by statistical method on the corpus provided by CLP2010, resulting in the accuracy decreasing, creating a great impact on results.

Table 5: the track results

Training mode	Model use	F-measure	POS Accuracy
Closed	Single	63.25%	89.62%

4 Conclusions

In this paper, we use CRFs to have a POS tagging, and increase the tagging accuracy by adjusting the feature template; multi-level chunking is applied to complete syntactic analysis, we do the base chunking with MEMM to recognize boundaries and components, and make the complex chunking with ME to generate a full parsing tree; on decoding, we add transfer matrix to improve performance, and remove some features with a ME classifier to reduce training time.

As the training data are temporarily changed, our system’s training on the Event Description Sub-sentence Analysis of CLP2010 isn’t completed, and head words are marked in the training corpus of this task, so our next step will be to complete training and testing of this task, compare the existing evaluation results, and use ME classifier to determine head words, analyze impact of head words on system. On the POS tagging, we will retain all features to train and compare tagging results.

Acknowledgement

We would like to thank XingJun Xu and BenYang Li for their valuable advice to our work in Complete Sentence Parsing of CLP2010. We also thank JunHui Li, XiaoRui Yang and HaiLong Cao for paving the way for our work.

References

S. Abney (1991) Parsing by Chunks. Kluwer Academic Publishers, Dordrecht, 257-278

Lance A. Ramshaw, Mitchell P. Marcus (1995) Text Chunking Using Transformation-Based Learning. In Proceeding of the Third ACL Workshop on Very Large Corpora, USA, 87-88

Erik F. Tjong Kim Sang (2001) Transforming a Chunker to a Parser. Computational Linguistics in the Netherlands 2000, 6-8

YongSheng Yang, BenFeng Chen (2004) A Maximum-Entropy Chinese Parser Augmented by Transformation-Based Learning. ACM Transactions on Asian Language Information Processing, 4-8

John Lafferty, Andrew McCallum, and Fernando Pereira (2001) Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Proceedings of the Eighteenth International Conference on Machine Learning, 282-289

Junhui Li, Guodong Zhou (2009) Soochow University Report for the 1st China Workshop on Syntactic Parsing. CIPS-ParsEval-2009, 5-8

Wei Jiang, Yi Guan, and Xiaolong Wang (2006) Conditional Random Fields Based POS Tagging.Computer Engineering and Applications, 14-15

Xiaorui Yang, Bingquan Liu, Chengjie Sun, and Lei Lin (2009) InsunPOS: a CRF-based POS Tagging System. CIPS-ParsEval-2009, 4-6

A. McCallum, D. Freitag, and F. Pereira (2000) Maximum Entropy Markov Models for Information Extraction and Segmentation. Proceedings of ICML-2000, Stanford University, USA, 591-598

Chao Li, Jian Sun, Yi Guan, Xingjun Xu, Lei Hou, and Sheng Li (2009) Chinese Chunking With Maximum Entropy Models. CIPS-ParsEval-2009, 2-4

Qiang Zhou (2004) Annotation Scheme for Chinese Treebank. Journal of Chinese Information Processing, 4-5