An Algorithm Combining Statistics-based and Rules-based for Chunk Identification of Chinese Sentences

Wang Rongbo, Chi Zheru

School of Computer, Hangzhou Dianzi University, Xiasha Higher Education Park, Hangzhou, Zhejiang, China
Department of Electronic and Information Engineering, Hong Kong Polytechnic University
wangrbo@163.com

Abstract. Natural language processing (NLP) is a very hot research domain. One important branch of it is sentence analysis, including Chinese sentence analysis. However, currently, no mature deep analysis theories and techniques are available. An alternative way is to perform shallow parsing on sentences which is very popular in the domain. The chunk identification is a fundamental task for shallow parsing. The purpose of this paper is to characterize a chunk boundary parsing algorithm, using a statistical method combining adjustment rules, which serves as a supplement to traditional statistics-based parsing methods. The experimental results show that the model works well on the small dataset. It will contribute to the sequent processes like chunk tagging and chunk collocation extraction under other topics etc.

1. Introduction

Shallow parsing, also called partial parsing or chunk parsing, has become an interesting alternative to full parsing [Abney, 1991]. The main goal of a shallow parser is to divide a sentence into segments which correspond to certain syntactic units. Shallow parsing including chunking and tagging can be done by dividing a non-restricted sentence into non-overlapping segments in an efficient and reliable way although the detailed information from a full parsing is lost.

In recent years, many methods are proposed and applied into shallow parsing of different languages. In reference [Antonio Molina, Ferran Pla, 2002], a specialized HMMs is proposed for English sentences to solve different shallow parsing tasks as a tagging problem. At the same time, many methods are proposed for Chinese chunk parsing [Li Heng, etc., 2004a; Li Sujian, etc., 2002; Zhou Qiang, etc., 1996] and Chinese chunk identification [Li Heng, etc., 2004b; Liu Fang, etc., 2000; Xi Chenhai and Sun Maosong, 2002].

In this paper, we propose a chunk description scheme of Chinese sentences and a method combining statistics-based and rule-based on how to identify all chunks in a Chinese sentence. In our scheme, the essence of chunk definition of chunks is the relative meaning independency which is different from other definitions as a new view point. Relative meaning independency requires a chunk to present a relative independent meaning. For example, sentence “我/r 很/d 不/d 喜欢/v 看/v 电视/n” (I do not like to watch TV very much.) can be segmented into three chunks with relative independent meaning, “我”(I), “很不喜欢看” (do not like to watch at all) and “电视” (TV).

The main idea of our algorithm can be described as belows.

(1) Perform manually chunks identification of Chinese sentences in the training corpus.

(2) Obtain the statistical information which will be used in the proposed statistics-based model.

(3) Use the model to perform the chunk pre-identification.

(4) Obtain the wrong identified instances based on which the error-driven rules learning is performed and a decision tree of rules is constructed.

(5) Obtain the right identified instances based on which the information contained in the decision tree is updated.

(6) Based on the decision tree, the results of chunk pre-identification of the input sentences is checked and corrected.

2. Chinese Chunk Definition

At present, there is no standard definition on Chinese chunk similar to no standard definition on Chinese parts of speech which is due to the inherent complexity of language. Researchers presented different definitions of Chinese chunk according to their special topics and applications.

In our definition of Chinese chunks, we present our emphasis on the following factors. Some changes which are different from the other's chunk scheme have been made.

1. Relatively independent meaning. It means that the every chunk has relatively independent meaning expressed by a core constituent and its adjunctive constituents.
2. Non-recursive. It means that no any two kinds of chunks will contain each other.
3. No-nesting. It means that any kind of chunk will not be contained in others. This factor together with the factor 2 ensures that all chunks in a sentence are in the same level.

Table 1. The categories of Chinese chunks in our scheme.

Types of Chunks	*Description*	*Types of Chunks*	*Description*
NC	Noun Chunk	NQC	Numeral Quantity Chunk
VC	Verb Chunk	LC	Location Chunk
PC	Preposition Chunk	TC	Time Chunk
ADJC	Adjective Chunk	ADVC	Adverb Chunk (no need)
NOTC	Not Chunk	CC	Conjunction Chunk
OC	Punctuation Chunk		

According to detailed instances in the sentence database, we give the main definitions for every category chunk. We present the main categories not all cases because of the language complexity, especially Chinese. At the same time, some combinations of parts of speech which form Chinese chunks will not be enumerated one by one.

3. Statistical Model Design

Supposed that $S =< W, T >$ is the input sentence performed word segmentation and tagging. $W = w_1, w_2, ..., w_n$ is the Chinese word sequence of this sentence. $T = t_1, t_2, ..., t_n$ is the part of speech sequence corresponding the word sequence.

So，the chunk identification problem can be considered as to find a segmentation point sequence C' which makes:

$$C' = \arg\max_{C' \in \{C\}} P(C \mid S) \quad (1)$$

A given assumption is that the boundary identifications of two sequent chunks are independent. So Eq. (1) can be redefined as Eq. (2).

$$C' = \arg\max_{C' \in \{C\}} \prod_{i=1}^{NSeg-1} P(w_{c_{i-1}+1} \ldots w_{c_i}) \quad (2)$$

In Eq. (2), $NSeg$ is the number of the segmentations in a sentence which larger than the number of chunks by one. $C = c_0, c_1, c_2, \ldots, c_{NSeg-1}$ is the boundary identification situation of every chunk identified of a sentence. c_0 is the beginning of every sentence which means there is always a boundary segmentation at the beginning of a Chinese sentence ($C_0 = 0$). Similarly, c_{Nseg-1} is the ending of every sentence which means there is always a boundary segmentation at the ending of a Chinese sentence ($C_{NSeg-1} = n$).

Based on the main idea of our proposed method, we consider the part of speech sequence of a sentence and identify the one which has most probability to be identified chunks.

4. Error-driven Adjustment Rules Learning

4.1 The Hierarchy of the Rules

The chunk pre-identification results are obtained by making use of our proposed statistical model. For better identification performance, the next step is to check and correct such preliminary results using the rules which are learned form the wrong identified instances. It is necessary to consider the context of a segmentation boundary of the wrong identifications. The obtained rules are organized in a tree structure whose style is as following.

<Wrong Identification> :: <Right Identification>. For example,

<d v | v n>:: <d v v | n>

The above style is transferred into a decision tree structure for storing learning rules. There are three kinds of nodes in the tree, including root node, internal node and leaf node. For root node, it is the ancestor of all other nodes which contains property “NumRules” to count the total number of learned rules. The internal nodes include properties “cntC”, “cntE”, “rate” and “TagPair”. The “cntC” and “cntE” are used to count the number of correct instances and wrong instances respectively. The “rate” is the error rate of wrong instances which is the ratio of “cntE” and the summation of “cntC” and “cntE”. The “TagPair” contains the content of generated node. The leaf nodes contain the property “cnt” and content correction chunk sequence. The correction chunk sequence is the chunk sequence which is used to replace wrong chunk identification.

4.2 Learning Algorithm

The learning process can be divided into two phases.

(1) Learning phase

In this phase, the rules used to correct wrong chunk pre-identification results are retrieved and stored in a XML file corresponding to the tree structure. The procedures can be described as following steps.

A. Obtain the chunk pre-identification results. They are compared with the correct ones which are identified manually and the wrong identified instances are obtained.

B. Construct the decision tree. The internal nodes and leaf nodes are generated according to the difference between the right instances and wrong ones. The addition operations are performed to the decision tree. At the same time, the counter of error instances contained in every node is increased by one when one instance passes it and the counter will be set the default value 1 when the node is generated.

(2) Validating phase

A. Obtain the right instances of chunk identification of sentences. That is the corresponding training set which is performed chunk identification manually.

B. Update the decision tree. Corresponding to every independent chunk and every chunk sequence or subsequence with different lengths, the nodes are generated by the generation process and the longest path is found in the decision tree. Then all counters of nodes in the path are increased by one.

After all generated nodes of the tag sequences corresponding to the sentences performed chunk identification manually are passed through the decision tree, all error rates of nodes are computed and updated which will be used to judge whether the context in the path will be replaced by the correction chunk sequence or not. If the rate is larger than the preset threshold, we will replace the context in the path with its correction chunk sequence which is contained in its leaf node. In our experiment, we use 0.5 as the preset threshold.

Then the rules set can be used to check and correct the pre-identified results. The check and correct process can be described as following steps.

(1) Obtain a pre-identified chunk sequence and consider it as an input node. If it can be found in the first level of the rules decision tree and its rate is larger than 0.5, then retrieve the correction chunk sequence of its leaf node to replace the pre-identification result.

(2) If it can not be found in the decision tree, then the new chunk sequence trimmed the final chunk is considered as a new input node. This process is iterated until the length of chunk sequence is equal to 2.

(3) If the chunk sequence is equal to 2, then all its nodes are generated by generation process and used to search in the decision tree. If the following three conditions can be met, then the replacement operation is performed by using the correction chunk sequence in the leaf node with maximal probability.

A) All generated nodes can be found in the decision tree;

B) The final reached node has leaf node;

C) The error rate is larger than the threshold, that is 0.5.

(4) If any one of the three conditions in the (3) can not be met, then the first chunk of the chunk sequence is considered as an input node. If it can be found in the decision tree and the rate of the node is larger than the threshold, then the correction operation is performed. Otherwise, no correction operation is executed to the first chunk which means it is identified correctly.

(5) From the next un-chunked pre-identified chunk, go through the step (1)-(5) until all pre-identified chunks are checked.

5. Experiment Results and Analysis

5. 1 Corpus

In our experiment, we use the sentences corpus provided by Tsinghua university. There are 2030 sentences in this corpus which contains totally about 9420 Chinese words. The average length of sentence is about 5 Chinese words.

5.2 Experimental Results Analysis

The following parameters are used to evaluate the system's performance.

(1) Precision

It denotes the rate of chunks identified correctly and the total number of chunks identified.

(2) Recall

It denotes the rate of chunks identified correctly and the total number of correct chunks in the corpus.

(3) Error Adjustment Rate (*EAR*)

It denotes the rate of adjusted incorrectly using the retrieved rules and the total number of adjustment operations including correct and incorrect adjustment.

In our experiment, the corpus is divided into two sets, training set and test set. There are 1800 tag sequences of Chinese sentences in the training set and 230 tag sequences of sentences in the test set. The sentences in the training set are performed word segmentation and tagging firstly, then chunk identification is performed manually which are used to train the proposed model and construct the rules decision tree.

The experiment results are listed in Table 2. SM (statistical model) is the experiment result of the statistical model used only. SM+RA (rules adjustment) is the result after the adjustment by learned rules. The threshold of the error rate of all rules used in this experiment is 0.5.

Table 2. The Experiment Result of the System

	Close Test		Open Test	
	SM	SM+RA	SM	SM+RA
Precision (%)	88.09	93.17	64.55	85.87
Recall (%)	81.68	92.84	60.97	89.06
EAR (%)	/	3.27	/	5.94

From Table 2, we can know that the precision and recall of close test are high especially after the automatic adjustment is performed using the retrieved rules. Although the precision and recall are not high before the automatic adjustment is performed in the open test, the performance can be improved greatly after the automatic rules adjustment operation. In our proposed chunk identification, the performance of chunk per-identification only using the statistical model is not good enough. That means the statistical model can only give an elementary identification result. However, the final system can give a good enough performance.

The wrong identifications, especially rules adjustment phase, can be mainly classified into the following cases.

(1) No adjustment is performed. This is mainly because that the error rate of the nodes corresponding to the wrong identification is not larger than 0.5 which is used as the threshold. If the error rate is not larger than the threshold, then no adjustment operation is performed. This case brings on about 56% errors.

(2) The adjustment rule with larger ratio is chosen to replace the wrong identification. In the rules adjustment phase, the system will always choose the correction chunk sequence with larger count frequency to correct the wrong identification. This case brings on about 22 percent errors.

6. Conclusion

In this paper, we proposed a statistical model which combines learned rules for Chinese chunk identification. From the elementary promising experiment results, we can know it is feasible on the dataset.

Reference:

1. Abney, Steven. Chunks and dependencies: Bring processing evidence to bear syntax. In computational linguistics and the foundation of linguistic theory. CSLI. 1995.
2. Antonio Molina, Ferran Pla. Shallow Parsing using Specialized HMMs. Journal of Machine Learning Research 2 (2002) pp. 595-613.
3. Li Heng, Zhu Jingbo, Yao Tianshun. SVM Based Chinese Text Chunking. Journal of Chinese Information Processing. Vol.18, No.2, 2004a.pp. 1-7.
4. Li Heng, Tan Yongmei, Zhu Jingbo, Yao Tianshun. Recognition of Chinese Chunk. Journal of Northeast University (Natural Science), China. Vol.25, No.2, 2004b. pp.114-117.
5. Li Sujian, Liu Qun, Bai Suo. Chinese Chunking Parsing using Rule-based and Statistics-based Methods. Journal of Computer Research and Development. Vol.39, No.4, 2002. pp.385-391.
6. Liu Fang, Zhao Tiejun, Yu Hao, Yang Muyun, Fang Gaolin. Statistics-based Chinese Chunk Parsing. Journal of Chinese Information Processing. Vol.14, No.6, 2000. pp. 28-32,39.
7. Xi Chenhai, Sun Maosong. Automatic Prediction of Chinese Phrase Boundary Location with Neural Networks. Journal of Chinese Information Processing. Vol.16, No.2. 2002. pp. 20-26.
8. Zhou Qiang. A Model for Automatic Prediction of Chinese Phrase Boundary Location. Journal of Software. 1996, Vol 7, Supplement. pp.315-322.