ATLANTIS PRESS

Advances in Engineering Research, volume 122

7th International Conference on Applied Science, Engineering and Technology (ICASET 2017)

Application of Chunk Grammar in the Understanding Technology of Natural Language

Yu Zhenlei

Information Ministry of Library, Qilu University of Technology, Jinan 250100，China

Abstract. Chunk Grammar, as the basic research in natural language understanding technology, is very important in language research area. In this paper FSA(Finite State Automaton) technology was introduced to parse chunk in order to simplify the parsing procedure of the syntax and improve the accuracy of parsing .

Keywords：Chunk; Chunk Grammar; FSA; FST

Introduction

CG Chunk Grammar is a basic research in natural language understanding technology, the core issue is to construct chunk identified model through learning corpus, FSM(finite state machine) or FSA(finite state automaton) refers to a kind of calculation model which is abstracted for the study of calculation course with finite memory and some language class. In text editor and compiler, finite state is used to design various processor which is used to identify valid strings, such as identifier and figures, while FST(finite state transducer) is the basic form of FSM. This paper is trying to apply cascading FST concept to the parsing of chunk grammar.

General design of chunk grammar parsing system

Parsing is the foundation of the understanding of natural language. In view of the difficulties of parsing in parsing massive text, we can try to decompose a complete parsing problem into some easy sub-problems to lower the degree of difficulty of complete parsing and to improve the parsing efficiency.

Part-of-speech tagging (POS tagging)

The task of part-of-speech is to tag right part-of-speech according to the context of a certain sentence. We use an English POS tagging tool called postagger-1.0, which is based on the the maximum entropy model. The tool has 16 models and adopts Upenn Treebank tagset. The local context information take part-of-speech disambiguation on the basis of regulation by using the result of tagging of part-of-speech.

Chunk parsing Basic Model

Partial Parsing, which is also called chunk parsing, is to explain the sentences into smaller unit. Chunk is a kind of structure, which is a non-recursive phrase that can meet certain syntactic function.

Principle of chunk:

(I) Definition of non-recursive of chunk. Various types of chunk is equal on constitution. Any chunk must meet certain syntactic rule strictly and can not be constituted by other types of chunk.

(II) Overlapping can not be presented between chunks. As the definition of chunk is made of non-recursive characters(ie. POS tagging) with out nesting, and longest match principle shall be followed when there is any ambiguity, which can screen small chunk under the circumstances of consisting of big chunk, so the overlapping shall not be presented between chunks.

The paper usescfinite-state automaton to simplify the parsing of sentence on the basis of tagging of part-of-speech.[2] A reasonable basic chunk rule is made up of several POS tagging sequences that consisting chunk, which are added to basic chunk rule sets and matched according to model rules[3]. According to this principal, we use basic chunk rule to constitute FSA. A subset of basic chunk rule set is as follows:

Copyright © 2017, the Authors. Published by Atlantis Press.
This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

$$\text{Subj} \rightarrow [_{\text{NP}}\ n{=}\ \mathbf{D}?\ n{=}\ [_{\text{N1}}\ \mathbf{A}^*\ n{=}\ \mathbf{N}\ ]\ ]\ \mathbf{V}$$

In the figure, the subject chunk is made up of noun chunks and adjective chunks, we definite the core words of chunk as the end of chunk and create another chunk for the dependents of core words. The order of noun chunk is determiner, adjective and noun. It is in correspondence with FSA.

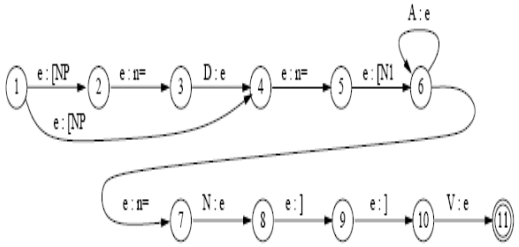

Figure 1. The flow chart of finite automata

subject rule set model change into finite state mechanism: “[NP”、“n=”、“[N1”、“]” stands for out symbol, D, A, N and V stands for part-of-speech of sentence. Scanning chunk result string form state 1, when meeting [, pop down it with chunk type symbol together. When meeting], pop up [and chunk type symbol, matching [with]. Chunk type symbol implies whether this kind of chunk can be created or not. If it can be created, it is used to pass the core words of it to the next higher level chunk to manage; while chunks can not be created, it used to pass the core words to the next higher level.

For example, identify of "the red ball", the tagging result is DT JJ NN, the automaton identify course is : match NN, JJ with N1 in N1 regular model (i.e., from state 9 to 6), match DT, N1 with NP in NP regular model(i.e., from state 10 to 1).

Analyse the input sentences. The analyse is a pattern matching course, during the course, if there is any conflict, then choose suitable model according to maximum matching principle.

L0 level is to input the result of part-of-speech tagging to FSA, searching for given matching model in L1 FSA, i.e., regular expressions. Then tagging sequence chunk to the part-of-speech which can meet the regular model, while the others will be come to the next level. The output of L1 level then become the input of L2 level.

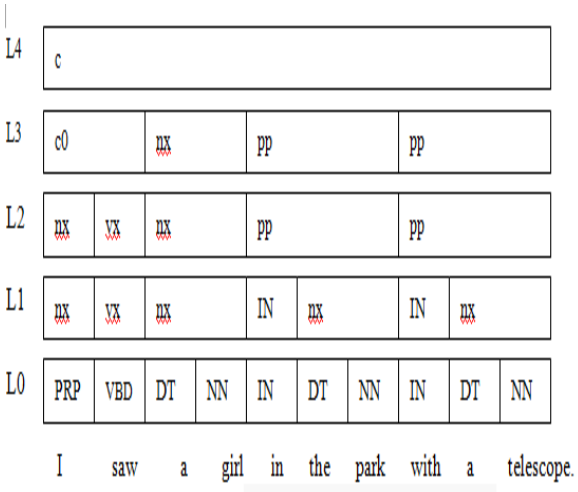

Figure 2. Cascading model matching

Regular based finite state group analysis

Introduction of regular model

Regular expression adopts REG(regular expression grammar). The quantity of regular model is total up to 223 pieces;

Generally it can be divided to 52 chunks.

The sort of chunk is introduced to four parts:

It is provided that the core part of each chunk is usually noun.

The first sort of chunk:

Noun chunk nx: the most basic noun phrase, it can be single noun, adjective plus noun phrase or noun plus conjunction plus noun plus ... paratactic noun phrase[4]. Wherein the conjunction also include comma, or word in apposition noun phrase constituted by noun plus word in apposition. Note: for the nouns connected without conjunction, each noun shall be considered as a noun chunk. The core words of this kind of chunk is the noun wherein.

Adjective chunk ax: adverb plus adjective.

The second sort of chunk:

Preposition chunk PP: preposition plus noun. The core words of this sort of chunk is the noun wherein.

Interrogative chunk src0、orc0: single interrogative pronoun or interrogative pronoun plus noun, for the latter, the core part is the noun. Wherein the former means that the interrogative pronoun is the subject, the latter means that the interrogative pronoun is the object.

The third sort of chunk:

subject-predicate chunk c0: equal to noun chunk plus verb.

The fourth kind of chunk:

Complete chunk:c: the biggest chunk, means a complete sentence.

The following is a noun chunk, see figure 3:

PRPPER stands for specific noun chunk, place means placename, for example, New York, we definite it as place chunk, name regular expression is name -> (nnp|nnps|i)+; wherein i means 26 capital letters and lowercase letters; COMMON is common noun chunk, wherein unit means measurement unit, such as meter, yard, etc. Unit means hourly basis, such as year, month, etc.

The above rule set constitute noun chunk nx.

```
PROPER   =  place | person | name | ci-st | doll;
COMMON   =  nn | nns | month | unit | units | tunit | tunits;
N        =  PROPER | COMMON | date;
ADVHD    =  rb | cdql | then | well;
ADV      =  ADVHD | rbr | more | rbs | ql;
VADVP    =  ADV* (ADVHD | only);
JX       =  ADV? (jj | jjr | jjs);
JXC      =  JX (cma JX)* (cc | cma) JX;
ADJ      =  JX | JXC | nx;
PTC      =  (ADV | rbr | more | rbs)? (vbn | vbg);
DET      =  dt | dtp | prp$ | (cdql | cdqlx)? (dt-a | dt-q | dtp-q);
NUM      =  cd | cdx;
MX       =  nx | units | tunits;

nx       →  DET? NUM? (ADJ | PTC)* (ADJ | N)* h=COMMON cd?
         |  DET NUM? (ADJ | PTC)* h=PROPER
         |  DET h=(jjr | jjs)
         |  cdql? h=dtp-q
         |  h=(  prp | cd | dtp | cd | dtp | qq | ex
                 | name | person | date | doll | ci-st | rbr | rbs
            )
         ;
```

Figure 3 Noun block diagram

Introduction of multilayer regular model

The process of chunk is to insert syntactic tagging, such as chunk border and chunk type. The process of chunk is carried out according to the grammar rules written by man. The chunk system include many layers and the analysis is carried out layer by layer. The chunk of each layer is carried out on the basis of the result of the next higher level. Analysing from the above, there are four sorts

which are totally 52 chunks. There are one or over one sort chunks on each layer, or other this layer only analyse these chunks. The rule set has 9 layers. The bottom is mainly to create lesser chunks, the top layer get a complete chunk c, that is an analysis of a sentence. During the chunk course of each layer, analysing from left to right of the sentence, according to part-of-speech or character to find whether it can meet the chunks that are allowed to present. One it is succeed to match, tag the chunks. During the model matching course, if there is any conflict, then choose suitable model according to maximum matching principle.

During the forming course of the chunks in the first layer, core words will be created and stored at the same time. The core words may need to deliver in the course of semantic analysis. For example, the multilayer model in table 1.

Table 1. multilayer model

Tagfixes	(word. Tag) tag
MesurePhrase	Dates and times, numeral, predetermines, dollar, amounts, city-state, names
Chunks	Noun, adjective, adverb, verb, infinitive chunks
N-mess	Unassembled noun chunk pieces
NP	Possessors
NG	Coordinated NP's
PP	Center embedding
RC	
C0	subj+pred: bleed Clause
Clause	

L0（level 0）: input the sentences that have been tagged the part-of-speech.
L1: use rules to identify the simple constitute group, including date, time and person name, etc.
date -> month cd (cma cd cma?)?;
L2: identify noun, verb, adjective overlay structure
nx -> such? DET? NUM? (ADJ | PTC)* (ADJ | N)* h=COMMON cd?
| DET? NUM? (ADJ | PTC)* h=PROPER
| DET h=(jjr | jjs | such)
| cdql? h=dtp-q
| h=(prp | cd | dtp | cd | dtp | qq | ex
| name | person | doll | ci-st | rbr | rbs)；
L3-L5: bigger noun chunks
ng -> h=NOM0 of k=NOM0 (of NOM0)*
| h=NOM0 (of k=NOM0 (of NOM0)*)?
(cc NOM
| (CONNECT NOM)+ cma? cc NOM) ;
L6: preposition chunks
pp -> f=PREP h=(NOM | tadvx);
L7-L9: chunks of sentences
c -> h=（s=NOM SUBJ-TAIL h=vx） o=NOM? ADV*;

Results and discussion

In the methods of the understanding of natural language, chunk is an usually method. Through the analysis of part-of-speech, it divided the sentence into many little meaningful group, which place a foundation of the understanding of the sentences. Focusing on the character that English sentences has certain rules, adopting chunk method is a better method. We have realized an autodecomposition system facing English complex sentence. In this paper, we use the texts which has been tagged in the Longman English Grammar about 3000 complex sentences(about 27000 words), using the method of rule analysis to analyse chunks to simplify the complex sentences. Wherein correct, training and perfect the rules to get valuable tagging texts. Meanwhile using the system to analyse the 100(about 3500 words) complex sentences as test and evaluation text. Comparing the result with tests and

evaluation texts, adopt the following indicators: associated words tagging, accuracy of argument tagging, recall rate and F value to measure the performance of the system of the paper. See table 2 for the test result.

Table 2 test result

	Accuracy	Recall rate	F-Measure	Time
Part-of-speech tagging	97.1%	100%	99%	0.5
associated words tagging	95%	100%	97%	1.5
Chunk analysis	73%	74%	73%	1.5

There are certain deviations of the result of chunk analysis, the errors include subjective factors and objective factors. We have the person to check up, and the course has the subjective, the experience of the person also will have some influence on the chunk tagging. The chunk base need to expand and check, improving in the range and consistency. For the rules, also with the problem of training texts, the rule set may have the problem of less expansion and less consistency.

So, the tagging of associated word and argument need to conclude more rule method from massive texts management.

Conclusion

Recently many domestic scientific and research institutes have done many research on the identification of Chinese chunk: Li Sujian(2002) has introduced the current research status of chunk analysis and the two technical route of chunk analysis, raised the importance and feasibility of Chinese chunk analysis tasks; Li Yan(2004) has raised HMM model based on the Gain-dBi to complete chunk analysis; Zhang Yuqi(2002) has used MBL based on examples to complete the auto identification of various Chinese basic phrases. The paper has introduced FSA technology to analyse chunks to simplify the parsing and improve the performance of parsing. We can also use the same way to analyse Chinese chunk.

But it is difficult to manage and research the Chinese information, we need to consider the management and research of Chinese information as a long term task, which shall not be considered to be completed in one day. Basis research is especially need to be strengthen. For Chinese, theories and technologies such as model and arithmetic can both learn from the other languages, while only knowledge base shall be created by oneself which shall not be copied from the others[5]. So the research of Chinese chunk shall first create its own texts base, then use the research productions of the other languages for reference to develop the research the Chinese chunk

Reference

[1] http://baike.baidu.com/view/157853.htm?fr=ala0.

[2] Steven Abney. Partial Parsing via Finite-State Cascades[J]. Natural Language Engineering,1996, 0404(2): 337-344.

[3] Wei-Chuan li, Tzusheng Pei, Bing-Huang Lee, Chuei-Peng Chiou. Parsing Long English Sentences With Pattern Rules[J]. Proc. of COLING-90, 1990, 410-412.

[4] L.G.Alexander. Longman English Grammar [M]. Lei Hang, Gan Meihua, Tian Luyi, Wang Chunli, reaslation, Beijing: Foreign Language Teaching and Research Publishing Houses, Publish year: 1991:2.

[5] editor in chief: Xubo. Some important issues of the management of Chinese information.[M]. Beijing: Science publishing Company, 2003:128