Statistics Based Hybrid Approach to Chinese Base Phrase Identification

Tie-jun ZHAO, Mu-yun YANG, Fang LIU, Jian-min YAO, Hao YU

Department of Computer Science and Engineering, Harbin Institute of Technology
{tjzaho, ymy, liufang, james, yu}@mtlab.hit.edu.cn

ABSTRACT

This paper extends the base noun phrase(BNP) identification into a research on Chinese base phrase identification. After briefly introducing some basic concepts on Chinese base phrase, this paper presents a statistics based hybrid model for identifying 7 types of Chinese base phrases in view. Experiments show the efficiency of the proposed method in simplifying sentence structure. Significance of the research lies in it provides a solid foundation for the Chinese parser.

Keywords: Chinese base phrase identification, parsing, statistical model

1 Introduction

Decomposing syntactic analysis into several phases so as to decrease its difficulty is a new stream in NLP research. The successful POS tagging has encouraged researchers to explore further possibility for resolving sub-problems in parsing(Zhou, et al, 1999). The typical examples are the recognition of BaseNP in English and Chinese.

In English BNP (base noun phrase) is defined as simple and non-nesting noun phrases, i.e. noun phrases that do not contain other noun phrase descendants (Church, 1988). After that researches on BNP identification reports promising results for such task in English. Observing that the Chinese BNP is different form English, (Zhao & Huang, 1999) puts forward the definition of Chinese BNP in terms of combination of determinative modifier and head noun. According to them a BNP in Chinese can be recursively defined as:

BaseNP ::= Determinative modifier + Noun | Nominalized verb(NV)

Determinative modifier ::= Adjective | Differentiable Adjective(DA) | Verb | Noun | Location | String | Numeral + Classifier

Inspired by these researches, we extend the concept of BNP to Base Phrase in Chinese. It is based on such knowledge that there are many structures, not only NP, in which the trivial components closely attach to their central words and constitute a basic phrase in a Chinese sentence. Obviously, resolving all these base phrases will greatly benefit Chinese parser by reliving it from some pre-processing (though non-trivial) and enable it focus on the most subtle syntactic structures.

Since the whole system of Chinese base phrase is still under discussing, this paper just presents some tentative research achievements on statistics based hybrid model to Chinese base phrase identification. For the 7 types we considered at present, our algorithm turns out promising results and smoothes the way for a better Chinese parser.

2 Statistics Based Hybrid Approach to Chinese Base Phrase Identification

2.1 Concepts and Definitions

In addition to BNP, constituents of many local structure in Chinese centers around a core word with certain fixed POS sequences. Therefore their identification is slightly different from parsing in that it bears relatively simple phenomenon. Like BNP identification, identification of these phenomena before parsing will provide a simpler sequence for parser, and thus deserves a separate research.

Currently, we are considering 7 Chinese base phrases in our research, namely base adjective phrase(BADJP), base adverbial phrase (BADVP), base noun phrase (BNP),

base temporal phrase (BTN), base location phrase (BNS), base verb phrase (BVP) and base quantity phrase (BMP) . Though theoretically definitions for these base phrases are still unavailable, Appendix I lists the preliminary illustrations for them in BNF format (necessary account for POS annotation can also be found)..

To frame the identification of Chinese base phrases, we further develop the following concepts:

Definition 1: Chinese based phrases are recognized as atomic parts of a sentence beyond words that posses certain functions and meanings. A base phrase may consist of words or other base phrases, but its constituents, in turn, should not contain any base phrases.

Definition 2: Base phrase tag is the token representing the syntactic function of the phrase. At present, base tag either falls in one of the 7 Chinese base phrases we are considering or not:

Phrase-Tag ::= BADJP | BADVP | BNP | BTN | BNS | BVP | BMP | NULL

Definition 3: Boundary tag denotes the possible relative position of a word to a base phrase. A boundary tag for a given word is either L(left boundary of a base phrase), R(right boundary of a), I(inside a base phrase) or O(outside the base phrase).

2.2 Duple Based HMM Parser

Based on above definitions, we could, in view of Wojciech's proposal [Wojciech and Thorsten, 1998], interpret the parsing of Chinese base phrases as the following:

Suppose the input as a sequence of POS annotations $T= (t_0, \ldots\ldots, t_n)$. The task is to find *RC*, a most possible sequence of duples formed by base phrase tags and boundary tags, among the POS sequence *T*.

$RC= (<r_0,c_0>, \ldots\ldots, <r_n,c_n>)$,

in which r_i $(1<i<=n)$ indicates the boundary tags, c_i represents the base phrase tags.

To go along with the POS tagger developed previously by us, we first think of preserving HMM (hidden Markov Model) for parsing Chinese base phrases. Thus the following formula is usually at hand:

$$\overline{RC} = \arg\max p(RC \mid T)$$

$$= \arg\max \frac{p(RC) * p(T \mid RC)}{p(T)}$$

For a given sequence of *T*, this formula can be transformed into:

$$\overline{RC} = \arg\max\ p(RC \mid T)$$

$$= \arg\max\ p(RC) * p(T \mid RC)$$

Essentially this model could be established through bigram or tri-gram statistical training by a annotated corpus. In practice, we just build our model from 10,000 manual annotated sentences with common bi-gram training:

$$p(RC) = \prod_{i=1}^{n} p(RC_i \mid RC_{i-1})$$

$$p(T \mid RC) = \prod_{i=1}^{n} p(T_i \mid RC_i)$$

In realization, a Viterbi algorithm is adopted to search the best path. An open test on additional 1000 sentences is performed to check its accuracy. Results are shown in Table1(note precision is calculated by word).

	Precision for R	Precision for C	Precision for Both R and C
Close Test	85.7%	87.5%	79.0%
Open Test	82.4%	85.1%	74.7%

Table 1. Results for Duple Based HMM

2.3 Triple Based MM Exploiting Linguistic Information

Although results shown in Table 1 is encouraging enough for research purposes, it is still lies a long way for practical Chinese parser we are aiming at. Reasons for errors may be account by too coarse-grained information provided by RC. Observing the fact that the Chinese base phrase occurs more frequently with some fixed patterns, i.e. some frozen POS chains, we decide to improved our previous model by emphasizing the contribution given by POS information.

Adding *t* denoting POS in the duple (r,

c), we develop a triple in the form of (t,r,c) for the calculation of a node. Naturally, the new model is changed into a MM (Markov model) as:

$$\overline{TRC} = \arg\max\ p(TRC)$$

$$= \arg\max \prod_{i=1}^{n} p(TRC_i \mid TRC_{i-1})$$

To train this model, we still using a bi-gram model. Applying the same corpus and tests described above, we got the performance of triple based MM identifier for Chinese base phrases (see Table 2).

	Precision for R	Precision for C	Precision for Both R and C
Close Test	89.2%	91.5%	84.6%
Open Test	88.4%	89.9%	83%

Table 2. Result for Triple Based MM

2.4 Further Improvement Through TBED Learning

Like other statistical models, the above model, whether duple based or triple based, both seem to reach an accuracy ceiling after enlarging training set to 12, 000 or so. To cover the remaining accuracy, we apply the transformation-based error driven (TBED) learning strategy described in [Brill, 1992] to acquired desired rules.

In our module, some initial rules are first designed as compensation of statistical model. Applying these rules will cause new mistakes as well as make correct identifications. Then the module will compare the processed texts with training sentences, generate new rules according to pre-defined actions and update its rule bank after evaluation (see Fig 1.).

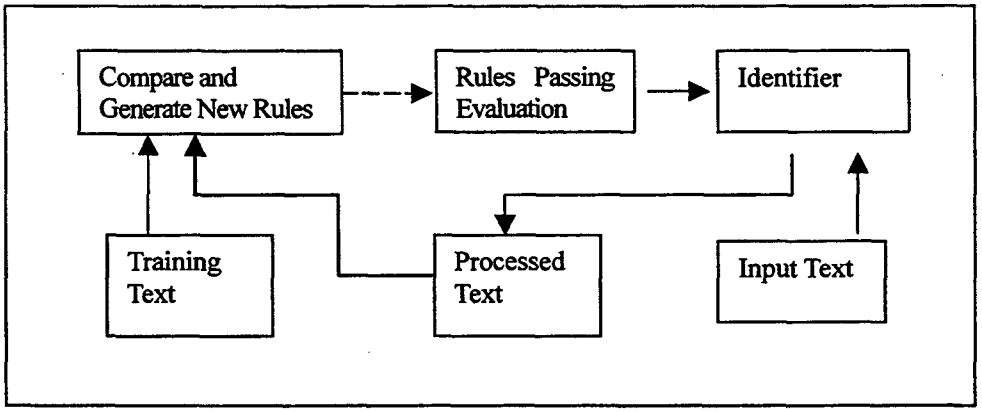

Figure 1. TBED Learning Module

The dotted line in fig 2. will stop functioning if pre-set accuracy is reached by the identifier for the Chinese base phrase. Evaluation of new rules is based on an greedy algorithm: only rule with max contribution (max correction and min error) will be added. Design of rule generation (pre-defined actions) is similar to those described in [Brill, 1992].

Table 3 shows a significant improvement after applying rules obtained through TBED learner. It is also the final performance of the proposed Chinese base phrase identification model.

	Precision for R	Precision for C	Precision for Both R and C
Close	91.2%	92.8%	89%
Open	90.4%	91.1%	87.1%

Table 3. Results after TBED Module

3 Conclusions and Discussions

We have accomplished preliminary experiments on identification of various types of base phrases defined in this paper. The data shown in last section prove that our method generates satisfactory results for

Chinese base phrase identification. The overall process of our method is outlined the following figure.

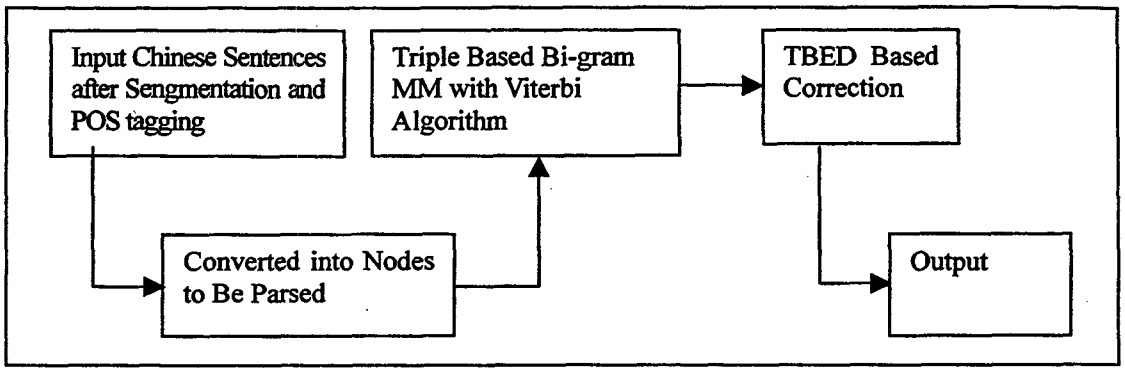

Fig 2. Processing of Chinese Based Phrase Identification

However, the 7 types Chinese base phrases we have proposed are far from perfection. Even what we have proposed for the 7 phrases is still under test. Further improvement will focus on two aspects: one is to discuss and add new base phrase for a broader coverage; the other is to define, theoretically or empirically, the Chinese base phrases with more strict constraints. Of course, new techniques to improved the accuracy of statistical model are the constant aim of our research.

To sum up, Chinese base phrase identification will reduce complexity of a Chinese parser. The successful identification of the 7 base phrases clearly simplifies the structure of the sentence. We expect that the research described in this paper will lay a solid foundation for a high-accuracy Chinese parser.

Reference

[Church, 1988] K. Church, A stochastic parts program and noun phrase parser for unrestricted text, In: Proc. of Second Conference on Applied Natural Language Processing, 1988

[Wojciech and Thorsten, 1998] Wojciech Skut and Thorsten Brants, Chunk Tagger, Statistical Recongnition of Noun Phrases, In ESSLLI-98 Workshop on Automated Acquisition of Syntax and Parsing, Saarbrvcken, 1998.

[Zhao & Huang, 1999] Zhao Jun and Huang Chang-Ning, The model for Chinese baseNP structure analysis, Chinese J. Computer, 22(2): pp141-146

[Zhou, et al, 1999] Zhou Qiang, Sun Mao-Song, Huang Chang-Ning, Chunk parsing scheme for Chinese sentences, Chinese J. Computer, 22(11): pp1159-1165

Appendix Illustration of 7 Chinese Base Phrases in BNF

The patterns listed here are far from complete (even for the 7 phrases themselves). Theoretical definition is beyond this paper and what we provide here is actually stage results of expert observation and linguistic abstraction.

BADJP ::= d^++a | d+BADJP | a^+ | a+BADJP | BADVP+a | BADVP+BADJP

BADVP ::= a+usdi(地) | d+usdi | vg+usdi | BADJP+usdi | BADVP+usdi | BMP+usdi

BMP ::= m^+ | m^++q^+ | m+q+m | d+m+q | f+m+q | r+m+q | BMP^+

BNP ::= a+n | a+usde(的)+n | a+usde+BNP | a+BNP | b+n | b+usde+n | b+usde+BNP | b+BNP | d+usde+n | f+n | f+usde+n | f+BNP | m+n | m+BNP | n^+ | n+usde+n | n+usde+BNP | n+usde+BMP | n+BNP | q+n | q+BNP | r+a+n | r+m+n | r+n | r+usde+n | r+usde+BNP | r+BNP | s+n | s+usde+n | s+usde+BNP | t+n | t+usde+n | t+usde+BNP | vg+usde+n | vg+usde+BNP | BADJP+n | BADJP+usde+n | BADJP+usde+BNP | BADJP+BNP | BMP+n | BMP+usde+n | BMP+usde+BNP | BMP+BNP | BNP+n | BNP+usde+n | BNP+usde+BNP | BNP+usde+BMP | BNP+BNP | BNS+usde+n | BNS+usde+BNP | BNS+BNP | BTN+usde+n | BTN+usde+BNP | BVP+usde+n | BVP+usde+BNP

BNS ::= a+nd | m+nd | n+s | r+nd | n+usde+f | n+usde+nd | n+usde+s | n+usde+BNS | nd^+ | r+usde+nd | r+usde+s | s+usde+nd | s+usde+BNS | BNP BNS | BNS^+

BTN ::= a+t | m+t | r+t | t^+ | t+usde+t | BMP+t | BTN+t | BNP+usde+t

BVP ::= a+vg | d+vg | vg+d+a | vg+d+vq | vg+d+vb | vg+usdf(得)+a | vg+usdf+d | vg+usdf+vq | vg+usdf+u | vg+usdf+BADJP | vg+ut | vg+vb | vg+ut+vq | vq+vg | vq+BVP | vz+vg | vz+BVP | BADJP+vg | BADVP+vg | BADVP+BVP | BVP+ut | BVP+vq | BVP+BVP

Symbol	Part-Of-Speech	Examples
a	Adjective	漂亮(beautiful), 浪漫(romantic)
d	Adverb	很(very), 依然(still)
f	Temporal/spacial position word	中(in), 上(on), 之间(between)
m	numeral	一(one), 二(two), 三(three)
n	noun	人民(people), 西红柿(tomato), 计算机(computer)
nd	Name of place	北京(Beijing), 哈尔滨(Harbin), 纽约(New York)
q	classifier	群(flock), 个(NULL)
r	pronoun	你(you), 我(I, me), 他(he, him)
s	location noun	附近(around), 室外(outside)
t	time noun	昨天(yesterday), 七月(July)
ut	tense auxiliary	着,了,过(NULL)
vb	Complemental verb	完,住(NULL)
vg	common verb	知道(know), 渴望(long for)
vq	directional verb	出,下来(NULL)
vz	modal verb	可以(can), 应该(should)

Table for POS symbols used in Appendix