Chinese Base-Phrases Chunking

Yuqi Zhang and Qiang Zhou

State Key Laboratory of Intelligent Technology and Systems
Department of Computer Science and Technology
Tsinghua University, Beijing, 100084, P.R.China

{zyq, zhouq}@s1000e.cs.Tsinghua.edu.cn

Abstract

This paper introduces new definitions of Chinese base phrases and presents a hybrid model to combine Memory-Based Learning method and disambiguation proposal based on lexical information and grammar rules populated from a large corpus for 9 types of Chinese base phrases chunking. Our experiment achieves an accuracy (F-measure) of 93.4%. The significance of the research lies in the fact that it provides a solid foundation for the Chinese parser.

1 Introduction

Recognizing simple and non-recursive base phrases is an important subtask for many natural language processing applications, such as information retrieval. Gee and Grosjean (Gee and Grosjean, 1983) showed psychological evidence that chunks like base phrases play an important role in human language understanding. CoNLL-2000's shared task identified many kinds of English base phrases, which are syntactically related non-overlapping groups of words (Tjong and Buchholz, 2000). The shared task has significantly heightened the progress in the techniques of English partial parsing. For Chinese processing, Zhao (1998) put forward a definition of Chinese baseNP that is a combination of determinative modifier and head noun (Zhao, 1998). Based on that research, Zhao et al. (2000) extended the concept of baseNP to seven types of Chinese base phrases. These base phrases may consist of words or other base phrases, but its constituents, in turn, should not contain any base phrases.

In this paper, we put forward the new definition of Chinese base phrases, which are simple and non-recursive, similar to the CoNLL-2000's shared task. The definition enables us to resolve most local ambiguities and is very useful for NLP tasks such as name entity recognition and information extraction. We construct a hybrid model to recognize nine types of Chinese base phrases. Many researches in Chinese partial parsing (Zhou, 1996; Zhao, 1998; Sun, 2001) have shown that statistical learning is of great use for Chinese chunking, especially for large corpus. However, the lack of morphological hints in Chinese makes it necessary to use semantic and syntactic information such as context free grammar rules in Chinese processing. In our approach, viewing chunking as a tagging problem by encoding the chunk structure in new tags attached to each word, we use Memory-Based Learning (MBL) method to set a tag indicating type and position in a base phrase on each word. After which grammar rules are used to disambiguate the tags. Our test with a corpus of about 2 MB showed that the experiment achieves 94.4% in precision and 92.5% in recall.

2 Definitions of Chinese Base Phrases

The idea of parsing by chunks goes back to Abney (1991). In his definition of chunks in English, he assumed that a chunk has syntactic structure and he defined chunks in terms of major heads, which are all content words except those that appear between a function word f and the content word which f selects. A major head is the 'semantic' head (s-head) for the root of the chunk headed by it. However, s-heads can be defined in terms of syntactic heads. If the syntactic head h of a phrase P is a content word, h is also the s-head of P. If h is a function word, the s-head of P is the s-head of the phrase selected by h.

The research enlightens us about the definition of Chinese base phrases. In this paper, a Chinese base phrase consists of a single content word surrounded by a cluster of function words. The single content word is the semantic head of the base phrase. The forms of base phrases can be expressed as follows.

{Modifier} * + head + {complement}* or Coordinate structure

The components of 'modifier' and 'complement' are optional. A head could be a simple word as well as the structure of "modifier + head" or "head + complement", but not "modifier + head + complement". Coordinate structure could not consist of coordinate symbols such as comma and co-ordinating conjunction. The type of base phrases is congruent with its head's semantic information. In most cases, the type accords with the head's syntactical information, for example, when the head is a noun, the phrase is a noun phrase. However, when a head is a noun that denotes a place, the base phrase including that head is not a noun phrase, but a location phrase.

We consider 9 types of Chinese base phrases in our research: namely adjective phrase (ap), distinguisher phrase (bp), adverbial phrase (dp), noun phrase (np), temporal phrase (tp), location phrase (sp), verb phrase (vp), quantity phrase (mp), quasi quantity phrase (mbar). The inner grammar structures of every base phrase are very important too, but we will discuss that in another paper.

3 Overview

The frame of Chinese base phrase parsing is composed of two parts: one is the "Type and bracket tagging model", the other is the "Base phrases acquisition model" which consists of two modules which are "brackets matching "and "correct the types of base phrases". (See figure 1.) The input to the system is a sequence of POS. In the "Predict the phrase boundary" module, we predict the type, which each word belongs to, and the position of each word in a base phrase with Memory-Based Learning (MBL)(Using the software package provided by Tilburg University.). And the result is expressed as a pair formed by base phrase type and position information. Because our Chinese base phrases are non-recursive and non-overlapping, the left and right boundaries of base phrases must match with each other which means they should be a pair and alternative. However, the errors involving in the first part will lead to incorrect base phrases because the boundaries do not match, for example "[…[…]". In the second part, grammar rules that indicate the inner structures of base phrases are used to resolve the boundary ambiguities. Furthermore, it also takes lexical information into account to correct the type mistakes.

The corpus used in the experiment includes 7606 sentences. It comes from the Chinese Balance Corpus including about 2000 thousand words with four types: literature (44%), news (30%), academic article (20%) and spoken Chinese (6%). These 7606 sentences are split into 6846 training sentences and 760 held out for testing.

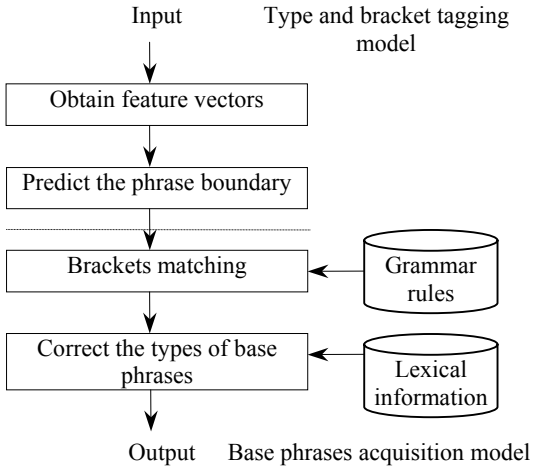

Figure 1: system overview

4 Predicting the phrase boundaries with MBL

Memory-Based Learning (MBL) is a classification based, supervised learning approach: a memory-based learning algorithm constructs a classifier for a task by storing a set of examples. Each example associates a finite number of classes. Given a new feature vector, the classifier extrapolates its class from those of the most similar feature vectors in memory (Daelemans et, al., 1999). The input to the "Predict the phrase boundary" module is some feature vectors, which compose of a sequence of POS. The solution of the module is to find $< r_i, c_i >$ (Wojciech and Thorsten, 1998), a duple formed by a type tag and a boundary tag for each word t_i. Here r_i indicates the boundary tag, while c_i denotes the type tag. $c_j \in \{np, vp, ap, dp, sp, tp, bp, mp, mbar, -\}$ ("-" denotes the word is not in any type of base phrases.) $r_i \in \{L, R, I, O, LR\}$ The r_i indicates the position of the word in a base phrase as shown below:

'L': the left boundary, 'R': the right boundary,

'I': the middle position, 'O': outside any base phrases, 'LR': the left and right boundary.

What information is used to represent data in feature vectors is an important aspect in MBL algorithms. We tried many feature vectors with various lengths. And it is interesting to note that the feature window is not the bigger the better. When the feature window is (-2, +2) in the context, the result is the best. So the feature vector in the experiment is: (POS-2, POS-1, POS0, POS+1, POS+2). The pattern describes the combination of feature vector and result duple $< r_n c_m >$ $0 \le n \le 4, 0 \le m \le 9$:

(POS-2, POS-1, POS0, POS+1, POS+2, $< r_n c_m >$).

For the experiment in the first step, we use [1] *TiMBL*, an MBL software package developed in the ILK-group (Daelemans et, al., 2001). The results of phrase boundary prediction with MBL shows in table 1.

Table1: The result of word boundary prediction

	Precision for $< r_n c_m >$	Recall For $< r_n c_m >$
np	92.27%	93.61%
vp	90.40%	89.65%
sp	75.15%	48.41%
tp	82.87%	71.62%
ap	93.52%	91.89%
bp	92.60%	76.38%
dp	97.56%	97.63%
mp	93.90%	92.38%
mbar	74.15%	72.26%
Total1	91.90%	91.65%
-	97.85%	98.41%
Total2	93.83%	93.83%

Table 1 shows that there is much difference between the results of various types of base phrases. The precisions and recalls of np, vp, mp, ap and dp are all almost over 90%. Comparatively, the results of sp, tp, bp and mbar are much lower, especially their recalls. This is due to some resemblances between sp, tp and np in Chinese syntactical grammars. Sp and tp may be considered as belong to NP, however, in the definition of Chinese base phrases, sp, tp and np are defined separately for the semantic difference. And the separation can also help in other tasks such as proper noun identification, information retrieval etc.

[1] *TiMBL* is a software bag about many MBL algorithms. It can be download free from http://ilk.kub.nl/

5 Obtaining Chinese base phrases

5.1 The errors in phrase boundary prediction

There are three types of errors in the results of first processing model.

(1) Boundary ambiguity: the r_i 's mistakes will cause the multiple choices regarding the boundaries. For example: "{np 这/rN } 一/m 创举/n }, /, 对/p {np 后世/t {np 针灸/n } 的/u {np 发展/vN 影响/vN } {ap 很/dD 大/a } 。/。". (Please pay attention to the '__' part.) There are altogether three modalities: "{ c_m …{ c_m …}", "{ c_m …}…}" and "{ c_m …{ c_m …}…}". These are caused by the redundancy and absence of boundaries.

(2) The type mistake of base phrases: For example: in the sentence of "{np 藏医/n } {dp 基本上/d } {vp 是/vC } {np 青藏高原/nS } 上/f {tp 藏族/nR 人民/n } 在/p…", the parser mistakes the type of "{ 藏族/nR 人民/n }" ,which is np, for tp. This error type commonly appears between sp, tp and np, as well as mbar and mp.

(3) Boundaries absence: For example, in the sentence of "{vp 包括/v } {np 内服/n } 、/、 {np 外用/n 药物/n } 以及/c {vp 放血/v }", "{np 外用/n 药物/n }" should be "{np 外用/n } {np 药物/n }". It is very difficult to correct this type of errors because the boundary distribution accords with the definition of Chinese base phrases. The left and right boundaries alternate with each other. Therefore, it is very difficult to find the errors in the sequence from the modalities.

5.2 Obtaining the whole base phrases with Grammar rules

With the bracket (boundary) representation, incorrect bracket will be generated but these will be eliminated in the bracket combination process. In the experiment, we attempt to apply grammar rules that represent the inner structures of Chinese base phrases to get rid of the boundary ambiguities. These grammar rules are derived from the corpus. On the other hand, boundary predictions can find many base phrases that do not accord with the limited grammar rules.

Figure 2 shows the main strategy of how to use the grammar rules. When if ()>1, there are more

Step 1: Finding the sequence where the errors appear. The sequences are three types:
"{…{…}", "{…}…}", "{…{…}…}".
Step 2: if (the number of sequences of POS in a pair of matched boundaries according with the grammar rules) >1
then {Select the boundaries that make the sequence longest}
Step 3: if (the number of sequences of POS in a pair of combined boundaries according with the grammar rules) =1
if (Only the sequence with the shortest length accords with the grammar rules).
then { Find partitions such as conjunctions, localizers, punctuations and some prepositions between the ambiguous boundaries in sequences;
if (The partitions exist)
then {Add boundaries to generate whole base phrases according to the partitions}
}

Figure 2: The Algorithm of Matching Boundaries

than one pair of combined brackets in which the sequences accord with the grammar rules. We are apt to choose the longest possible because the shorter sequences appear more in the corpus. The longer the sequence, the more weight it should carry. When there is only the shorter sequence according with grammar rules, it is more possible to be the correct one. In this case, one or more boundaries will be left. They often need some other boundaries to match, so we try to retrieve some missing boundaries through the partitions in the sentences that should not belong to any base phrases. These partitions are the marks of base phrase boundaries. If we find these partitions between two ambiguous boundaries, we will know where to place the new boundary.

5.3 Correct the type mistake with lexical information

In the Chinese language, some POS sequences may belong to different types. For example, "{vN n}" could be np, sp or tp. These sequences often appear in np, sp, tp, mp and mbar. It is difficult to know its right type even with the grammar rules, as we have done in section 5.2. In order to resolve this problem, we attempt to use lexical information because it implies semantic information to some extent.

The lexical information is distinctive between mp and mbar. mbar is often composed of numbers such as "1200" and numbers in Chinese such as "四". The lexical information between tp and np is also obvious, such as "时候", "时代" and "世纪" etc. For sp and np, the words are "地区", "流域" etc.

5.4 Experimental results

The simplest bracket combination algorithm is very strict: it only uses adjacent brackets if they appear next to each other in the correct order (first open and then close) without any intervening brackets. The result of the algorithm is shown in table 2, as the baseline of the boundary combination experiment.

Table 2: The base-line result

	Precision	Recall	F_M
Np	93.9%	86.1%	89.8%
Vp	90.6%	86.2%	88.4%
Sp	75.5%	47.7%	58.4%
Tp	85.4%	70.2%	77.0%
Ap	93.4%	83.4%	88.1%
Bp	93.4%	71.3%	80.9%
dp	97.7%	94.0%	95.8%
mp	92.0%	85.3%	88.5%
mbar	-------	0	-------
Total	92.9%	85.7%	89.2%

Table 3: The result of disambiguation with grammar rules

	Precision	Recall	F_M
np	94.3%	91.9%	93.1%
vp	95.0%	94.2%	94.6%
sp	73.6%	50.9%	60.2%
tp	84.9%	73.8%	79.0%
ap	93.5%	89.7%	91.5%
bp	91.6%	79.4%	85.0%
dp	97.6%	98.1%	97.8%
mp	86.7%	90.9%	88.7%
mbar	63.6%	12.6%	21.1%
Total	93.9%	92.0%	92.9%

From the table 2, we could see the recalls are commonly low. We change another strategy to obtain the whole base phrases as described in section 5.2. The result of using the grammar rules is shown in table 3.

With the help of grammar rules, all kinds of base phrases improved their f-measures though the precisions or recalls of some types decrease slightly. Comparing with the baseline results in table 2, all the recalls increase significantly. However, the recalls of sp, tp and mp still do not satisfy us. There are more than twenty structures of np which also belong to tp or sp. Except in the case where mp and mbar have the same structure {m}, they are easily distinguished in other structures. (Mbar is always composed of numerals and mp always ends with a quantifier.) In order to distinguish tp from np, sp from np and mbar from mp, we use lexical information for the type disambiguation. The results are shown in table 4.

Table 4: The result after using lexical information

	Precision	Recall	F_M
np	95.0%	91.9%	93.5%
vp	95.0%	94.3%	94.6%
sp	69.2%	71.3%	70.2%
tp	79.8%	84.1%	81.9%
ap	93.1%	90.0%	91.5%
bp	91.6%	79.4%	85.0%
dp	97.6%	98.1%	97.8%
mp	93.4%	90.9%	92.1%
mbar	67.6%	54.1%	60.1%
Total	94.4%	92.5%	93.4%

From the table 4, we could see improvement in all the results (precisions and recalls) of mp and mbar. It shows that the lexical information is effective for distinguishing between them. On the contrary, although the f-measures of np and sp increase, their precisions decline. Thus, those words marking tp and sp are not appropriate for disambiguation. We could see the effect of lexical information is limited because it is difficult to find the words that could distinguish different types of base phrases.

6 Conclusions

The experiment on identifying Chinese base phrases shows that the definition of Chinese base phrases is suitable for parsing. It shows good results and the efficiency of the proposed approach in simplifying sentence structures. Many tasks such as chunking on high level could benefit from this.

With the system described here, we get 9 types of Chinese base phrases, and acquire high precisions and recalls on most types of base phrases. The results of the experiment also show that the use of grammar rules is necessary. Grammar rules have effects on boundary disambiguation particularly. The lexical information is effective in distinguishing between mbar and mp.

Acknowledgements

This work was supported by the National Science Foundation of China (Grant No. 69903007), National 973 Foundation (Grant No. 1998030507) and National 863 Plan (Grant No. 2001AA114040).

References

Abney, Steven. (1991) Parsing by chunks. In Berwick, Abney, and Tenny, editors, *Principle-Based Parsing.* Kluwer Academic Publishers.

Erik F. Tjong Kim Sang and Sabine Buchholz. (2000). "Introduction to CoNLL-200 Shared Task: Chunking". *Proceedings of CoNLL-2000 and LLL-2000*. Lisbon, Portugal. 127-132.

J. P. Gee and F. Grosjean (1983) Performance structures: A psycholinguistic and linguistic appraisal. *Cognitive Psychology*, 15:411-458

Sun Honglin (2001) A Content Chunk Parser for Unrestricted Chinese Text, Dissertation for the degree of Doctor of Science, Peking University.

Walter Daelemans, Jakub Zavrel, Ko van der Sloot (2001) TiMBL:Tilburg Memory-Based Learner version 4.0 Reference Guide. http://ilk.kub.nl/downloads/pub/papers/ilk0104.ps.pz.

Wojciech Skut and Thorsten Brants (1998) Chunk Tagger, Statistical Recognition of Noun Phrase, In *ESSLLI-98 Workshop on Automated Acquisition of Syntax and Parsing*, Saarbrvcken.

Zhao Jun (1998) The research on Chinese BaseNP Recognition and Structure Analysis, Dissertation for the degree of Doctor of Engineering, Tsinghua University.

Zhao et al., (2000) Tie-jun ZHAO, et al. "Statistics Based Hybrid Approach to Chinese Base Phrase Identification", In *Proceedings of the Second Chinese Language Processing Workshop, ACL 2000*, 73-77.

Zhou, Qiang (1996). Phrase Bracketing and Annotating on Chinese Language Corpus, Ph.D. dissertation, Peking University.