

Two-Level Shallow Parser for Unrestricted Chinese Text

Sun Honglin^{1,3}

sunhl@bku.edu.cn

Lu Qin²

csluqin@comp.polyu.edu.hk

Yu Shiwen¹

yusw@pku.edu.cn

¹Institute of Computational Linguistics
Peking University, Beijing

²Department of Computing

Hong Kong Polytechnic University

³Center for Language Information Processing

Beijing Language & Culture University

Abstract: Shallow parsing provides partial solution for syntactic parsing of sentences. Some previous work applied statistical methods to assign phrase boundaries based on the POS sequences. Because many Chinese words have several different grammatical functions without any change of forms, the conventional POS sequences can not give enough information for successful deduction of phrase boundaries. This paper presents a method applying two-level analysis for the shallow parsing of Chinese text. At the first step, during POS tagging, verbs and adjectives are assigned based on tags representing their grammatical functions in the context. The second step analysis works on the output of the first step and assigns non-recursive phrase boundaries between tag pairs. Both levels are based on Hidden Markov Models. The experiment shows that the results of this method are satisfactory.

Keywords: shallow parsing, POS tagging, Hidden Markov Models

1. Introduction

Shallow parsing, also referred to as partial parsing which aims at the partial parses of sentences, has caught more and more attention in recent years. Both rule-based and statistics-based approaches have been taken in

this field [Abney 1997]. The statistical methods have the advantages of simplicity and robustness. [Church 1988] proposed a simple stochastic technique for recognizing the non-recursive base noun phrase (base NP) in English. He utilized Hidden Markov Models (HMMs) to insert the open brackets and close brackets representing the beginning and the end of non-recursive noun phrases (base NPs) respectively and received very satisfactory results. [Li Wenjie et. al. 1995] applied a similar method for recognizing the maximal-length noun phrase in Chinese. But, the results were unsatisfactory. The unsatisfactory results are due to two reasons. Firstly, the maximal-length noun phrases have more recursive structures within themselves and this will make it more difficult to correctly find their boundaries than the simple non-recursive base NP. Secondly, some of the Chinese words have several different grammatical functions without any change of forms. The sequences of conventional POS tags cannot give enough information to deduce the phrase boundaries. For example, a Chinese verb can function as subject, object, modifier or head in noun phrases, complement, etc., without any change of forms. Thus when a verb is encountered in the text, there are too many boundary candidates

around it to make a choice with high accuracy. This makes it difficult to get distinguishable statistical information on the correlation between the POS tags and phrase boundaries. Zhou (1996) assigned brackets between tag pairs as the first step towards the full parse of Chinese sentences and he reported that the assigning of brackets could get an accuracy over 91%. However, his corpus mainly came from the elementary school Chinese textbooks of Singapore and the sentences are short and simple. Therefore, it cannot represent the difficulties in natural text.

In this paper, we present a new method applying two-level analysis for the shallow parsing of unrestricted Chinese text. At the first level, in the process of POS tagging, verbs and adjectives which are most ambiguous in syntactic functions are assigned with tags representing their grammatical functions in the context. The second level analysis works on the output of the first level and assigns non-recursive unlabeled brackets between tag pairs. At both levels, HMMs based statistical method is used.

The remainder of the paper is organized as follows. Section 2 describes the tagging of words with syntactic functions. Section 3 presents the shallow parsing. Section 4 gives concluding remarks.

2. Tagging words with syntactic functions

2.1 Extension of tagset

Automatic part-of-speech (POS) tagging has achieved quite significant results in the last decade. A number of POS taggers are readily available and widely used in major languages

in the world. However, all these systems only take into account information about the static categories of words. As we pointed out in Section 1, many words can appear in significantly different grammatical positions without any change of forms because there is no inflection in Chinese. This makes it difficult to predict the phrase boundaries based on the POS tags.

To deal with the serious structural ambiguities produced by the POS sequences, we extend the conventional POS tagset to attach the information about the syntactic functions of words. The most serious ambiguities exist in the classes of verb and adjective. At first, we divide the class of verb into 4 subclasses: (1) auxiliary verb, e.g. “能, 会, 可以”; (2) copular verb, e.g. “是, 像”; (3) formal verb, e.g. “进行, 加以”; (4) general verb, e.g. “吃, 发展”. This classification is based on the static syntactic properties of the words. There is almost no overlapping between the subclasses. Furthermore, the general is divided 9 categories based on syntactic functions: (1) verb as modifier in noun phrase, e.g. “研究目的, 生产成本”; (2) verb as head in noun phrase, e.g. “理论探讨, 经济的发展”; (3) verb as complement, e.g. “送走客人, 做完手术”; (4) verb with noun phrase as object, e.g. “发展农业, 挖掘潜力”; (5) verb with verb phrase as object, e.g. “表示感谢, 推进改革”; (6) verb with adjective as object, e.g. “解除痛苦, 提供方便”; (7) verb with clause as object, e.g. “祝愿老人身体健康, 保持交通畅通”; (8) verb with pivot object, e.g. “帮助民工及时返乡, 供家庭和个人收看”; (9) verb in other positions including:

a) the first part of “DE construction”, e.g.

“采取的措施要及时, 得到的教益很深”;

b) the second part of “SUO construction”, e.g. “宪法所规定的权利, 我们所反对的”;

c) preceding direction/position word, e.g. “会谈前, 退休后, 在试验中, 谈判中”;

d) as the head of predicate but without object, e.g. “依法经营, 生活能自理”;

e) as subject or object, e.g. “建设无法进行, 得到帮助, 发起挑战”.

The syntactic function tags for adjectives contain: (1) adjective as adverbial, e.g. “认真学习, 突然降临”; (2) adjective as subject or object, e.g. “稳定是首要的, 感到满意.”; (3) adjective as the head of noun phrase, e.g. “最大的苦恼, 社会的温暖”; (4) adjective in other position, including as predicate, modifier and complement.

2.2. Experimental results

Because many verbs and adjectives can have multiple grammatical functions, the additional functional tags will make these words have more tag candidates to be chosen when tagging. To test the feasibility of tagging words with syntactic functions automatically, we conducted a tagging experiment, applying two tagsets to the same corpus. (1)TAGSET1, containing 20 tags; (2) TAGSET2, contains 34 tags. TAGSET1 is the tagset used in the large syntactic lexicon described in Yu(1998) and is commonly used in the community of Chinese information processing. TAGSET2 is the extension of TAGSET1. The classes of verb and adjective in TAGSET1 are extended as described in Section 2.1. The tags in TAGSET1 and the extensions in TAGSET2 are list in Appendices 1 and 2, respectively.

In the experiment, we used a corpus of 203,499 words, comprising 157 articles from

People’s Daily. The whole corpus was divided into two parts: (1) the training set containing 137 files and 183,050 words; and (2) the testing set containing 20 files and 20,449 words. Two tagging algorithms were applied separately. The first algorithm, referred to Algorithm1, chooses the most likely tag for every word in the lexicon . The second algorithm, referred to as Algorithm2, is a HMMs based Viterbi algorithm which uses both tag transition probabilities ($P(T_i | T_{i-1})$) and the emit probabilities ($W_j | T_i$) and utilizes dynamic programming to choose the path with greatest possibility from all possible paths in one sentence (DeRose 1988). The experimental results are shown in Table 1 (unknown words were not taken into account).

Table 1 Accuracy Rate of Tagging

| | TAGSET1 | TAGSET2 |
|----------------|---------|---------|
| Algorithm1 (%) | 95.50 | 89.26 |
| Algorithm2 (%) | 96.60 | 92.85 |

The above experiment shows that although the accuracy of TAGSET2 is 3.75 percent lower than that of TAGSET1. The whole accuracy is up to 93%. This meets the requirements of many NLP tasks. On the other hand, comparing the two algorithms, we find that TAGSET2 has greater improvement of accuracy than TAGSET1 when applying Algorithm2 as compared to applying Algorithm1. The improvement is 1.1 percent for TAGSET1 and 3.59 percent for TAGSET2. The difference between the two algorithms can be explained as that transition probabilities in the tag sequences have more effect for Algorithm2 than Algorithm1. The experiment shows that it is feasible to extend

the conventional POS tagset to contain dynamic information about the words' grammatical functions in a framework of simple statistical POS tagging.

3. Shallow parsing

3.1. Parsing Paradigm

[Church 1988] and [Li Wenjie et.al.1995] used two boundary marks: the beginning of NP and the end of NP. When reviewing Church's work, [Abney 1997] proposed five types of brackets: (1) “[” marking the beginning of a base NP; (2) “]” marking the end of a base NP; (3) “[” marking the end of one base NP and the beginning of another one; (4) “[” marking the inside of a base NP; and (5) “[” marking the outside of a base NP.[Zhou 1996] applied three types of phrase boundaries, i.e. open bracket marking the beginning of one phrase, close bracket marking the end of one phrase and no bracket. [Taylor & Black 1998] assigned one of the two possible breaks, break or no break, to any pair of tags.

Abney's five types of brackets enumerate all the possibilities of brackets between tag pairs when determining the phrase boundaries of base NP. But, his fifth type of brackets is unnecessary if we want to mark the boundaries of any phrases instead of any specific type. We add “[” to Zhou's paradigm, since this will make all types of brackets to form a whole set of finite states. The four types of phrase boundaries between any word pair W_i W_{i+1} are defined as follows:

(1) [: indicates that W_i and W_{i+1} are not immediate constituents of a construction and W_{i+1} combines with one word or phrase on its

right side;

(2)] : indicates that W_i and W_{i+1} are not immediate constituents of a construction. W_i combines with one word or phrase on its left side;

(3)][: indicates that W_i and W_{i+1} are not immediate constituents of a construction. W_i combines with one word or phrase on its left side and W_{i+1} combines with one word or phrase on its right side;

(4) * : indicates that W_i and W_{i+1} are immediate constituents of a construction.

We do not consider the recursiveness of phrases and assume that only one of the four types of phrase boundaries can appear between any pair of tags. Such boundaries as “[[“, “[”]”, “[”]”, etc. are not permitted. This assures that the shallow parser assigning brackets between tag pairs can be abstracted as a finite state automata, so the HMMs can be applied.

For example, in the Chinese sentence “国家积极改善投资的法律环境”, after POS tagging and bracketing, we will get the following result:

[国家/n [积极/az [改善/vgn [投资/vg * 的/u] [法律/n * 环境/n]

In the above sentence, after bracketing, we can determine two phrases in it immediately, i.e., “投资的” and “法律环境”.

In parsing schema, we choose to use binary tree to represent the syntactic structure for any constructions, even for multi-branching structures like conjunction. For example, the phrase “选育、推广、普及农业、林木、畜禽和水产的优良品种” will be analyzed as below if multi-branching is permitted:

[选育、推广、普及] [农业、林木、畜禽和水产] 的] [优良品种]

But in our notation, it will be analyzed as:

[选育/vgn \ /、] 推广/vgn \ /、 普及/vgn][农业/n \ /、] 林木/n \ /、] 畜禽/n] 和/c] 水产/n] 的/u][优良/a 品种/n]

Because of the use of binary-tree paradigm, out of the sixteen possible boundary bigrams only half are valid. The boundary bigrams are listed below (the two adjacent boundaries are delimited by hyphens):

(1) valid bigrams: [-, [-*,]-],]-[,][-,][-*, *-], *-][.

(2) invalid bigrams: [-,]-[,]-[,]-*,][-],][-], ***, *-[-.

This paradigm has two advantages: (1) It can prevent some of the bracketing errors, even for the errors^{*} produced by the human annotators; and (2) It can help to alleviate the sparse data problem in statistics.

3.2. Algorithm for bracketing

The input to the shallow parser is a sequence of word/tag pairs and the output is a sequence of phrase boundaries. The HMMs are used in bracketing, where hidden states are phrase boundaries and the transitions between states represent the likelihood of particular sequences of phrase boundaries. Each state has an observation probability distribution giving how likely that state is to have produced a sequence of POS tags. The state observation probabilities are called the POS sequence model and the set of transition probabilities is called the phrase boundary model. Bayes equation is used to relate the two and the most likely phrase boundary sequence for a given tag sequence can be found by searching through the model and selecting the most likely path. In searching

process, dynamic programming algorithm is used to reduce the search space (DeRose 1988).

The POS sequence model is trained by the statistics of the occurrences of distinct sequences of POS tags preceding and following the boundary when a boundary type is given. The POS sequence is a window of N tags around a boundary b_i with M tags preceding b_i and N-M tags following b_i . The probability of a POS sequence given a phrase boundary can be estimated by maximum likelihood estimation as:

$$P(T|b_i) = \frac{\text{Count}(T|b_i)}{\text{Count}(b_i)} \quad (1)$$

where T indicates the tags preceding and following the phrase boundary b_i . When N is 2, M has the value of 1, which means one tag preceding b_i and one tag following b_i are chosen in the window. Thus Equation (1) can be rewritten as:

$$P(T|b_i) = \frac{\text{Count}(T_i, T_{i+1}|b_i)}{\text{Count}(b_i)} \quad (2)$$

where T_i indicates the tag for the i-th word in a sentence or word string.

The phrase boundary model is trained by the statistics of the probability of N-gram of phrase boundaries in the corpus. It is assumed that the occurrence of the N-th boundary is only dependent on the previous N-1 boundaries. Thus the estimation is:

$$P(b_i | B_{i-1}^N) = P(b_i | b_{i-1}, b_{i-2}, \dots, b_{i-N+1}) \quad (3)$$

$$= \frac{\text{Count}(b_i, b_{i-1}, b_{i-2}, \dots, b_{i-N+1})}{\text{Count}(b_{i-1}, b_{i-2}, \dots, b_{i-N+1})} \quad (4)$$

When N=3, only the previous two boundaries

are taken into account, Equation (4) can be rewritten as:

$$P(b_i|b_{i-1}, b_{i-2}) = \frac{\text{Count}(b_{i-2}b_{i-1}b_i)}{\text{Count}(b_{i-2}b_{i-1})} \quad (5)$$

This is the trigram model of the phrase boundaries.

3.3. Experimental Results

In the task of assigning phrase boundaries, we selected 4,051 sentences in the tagged corpus described in Section 2.2. The total word tokens is 131,516 and the average sentence length is 32.5 words per sentence, ranging from 2 words per sentence to 278 words/sentence¹. At first, the 4,051 sentences were manually bracketed. Then, they were divided into two parts: 3,051 sentences for training and 1,000 sentences for open testing. In the bracketing experiment, we tried three algorithms. The first algorithm chooses the most likely bracket between two tags. The second is the Viterbi algorithm and the bigram model is applied when estimating the transition probabilities of phrase boundaries. The third is also the Viterbi algorithm but trigram model is used (i.e. Equation (5)). In the test, we took 1,000 sentences from the training set for close testing compared with the open testing. The results are shown in Table 2.

Table 2 Experimental results of shallow parsing

| | accuracy of close testing | accuracy of open testing |
|-------------|---------------------------|--------------------------|
| Algorithm 1 | 85.4% | 83.6% |
| Algorithm 2 | 86.2% | 84.5% |
| Algorithm 3 | 87.8% | 86.3% |

¹ We delimited the sentences by one of the three punctuation marks: full stop, question mark and exclamation mark, except for the titles, which were followed by a paragraph mark in our corpus. The shortest sentence comprising 2 words is a title.

The experiment shows that the third algorithm works best within the three algorithms.

4. Conclusion

In this paper, we proposed a new method of shallow parsing for unrestricted Chinese text. In view of the fact that there is no inflection in Chinese, we divided the task into two steps. During POS tagging, the conventional POS tagset is first extended to contain the dynamic information about grammatical functions of words in the context. Then, one of the four types of phrase boundaries is assigned between each pair of tags. The experiment shows that it is feasible to add the syntactic functions to the POS tag, and based on the output of the first step, the shallow parser can obtain satisfactory results in processing unrestricted Chinese text. In both steps, the simple stochastic method based on HMM model is applied, so the system is very easy to implement.

In the future, we will continue our work in two directions. First, we will extend the scale of the experiment and improve the algorithm further. Second, we will conduct research on identifying linguistic structures, such as noun phrases formed by consecutive content words, V-O phrases and other collocations, from very large Chinese corpus with the help of the shallow parser described here.

References

- [1] Abney, S. 1997. Part-of-Speech Tagging and Partial Parsing, In *Corpus-Based Methods in*

- Language and Speech Processing*, edited by Young S. and Bloothoof G, Kluwer Academic Publishers, pp.118-136.
- [2] Church, K. 1988. A Stochastic stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the Second Conference on Applied Natural Language Processing*, pp136-143.
- [3] DeRose, S. 1988. Grammatical Category Disambiguation by Statistical Optimization, *Computational Linguistics*, Vol. 14 (1), pp. 31-39.
- [4] Li, Wenjie et al. 1995. Corpus-based Maximal-length Chinese Noun Phrase Extraction, In *Proceedings of NLPRS'95*, pp. 246-251.
- [5] Taylor P., Black A. 1998. Assigning Phrase Breaks from part-of-speech sequences, *Computer Speech and Language*, Vol.12, pp. 99-117.
- [6] Yu, Shiwen, et al. 1998. Grammatical Knowledge Base of Contemporary Chinese: A Complete Specification. Tsinghua University Press, Beijing.
- [7] Zhou, Qiang 1996. A Model for Automatic Prediction of Chinese Phrase Boundary Location, *Journal of Software*, Vol. 7 Supplement, pp. 315-322.

Appendix 1 List of tags in TAGSET1

| TAG | NOTE | TAG | NOTE | TAG | NOTE |
|-----|----------------------|-----|-------------|-----|----------------|
| n | noun | v | verb | u | auxiliary word |
| t | time word | m | numeral | y | modal particle |
| s | location word | q | classifier | o | onomatopoe |
| f | direction & position | r | pronoun | e | exclamation |
| a | adjective | p | preposition | k | affix |
| b | distinguishing word | d | adverb | i | idiom |
| z | static word | c | conjunction | | |

Appendix 2 Extensions of verb and adjective from TAGSET1 to TAGSET2

| TAG | NOTE | TAG | NOTE | TAG | NOTE |
|-----|------------------------|-----|-----------------------------|-----|-----------------------------|
| va | auxiliary verb | vgc | verb as complement | az | adjective as adverbial |
| vi | copula verb | vgn | verb with NP as object | as | adjective as subj. or obj. |
| vf | Formal verb | vgv | verb with VP as object | ax | adjective as head in NP |
| vg | verb without object | vga | verb with adjective as obj. | a | adjective in other position |
| vgp | verb as modifier in NP | vgs | verb with clause as object | | |
| vgx | verb as head in NP | vgj | verb with pivot object | | |