BLENDING SEGMENTATION WITH TAGGING IN CHINESE LANGUAGE CORPUS PROCESSING [1]

Zhou Qiang , Yu Shiwen

Institute of Computation Linguistics
Peking University, Beijing 100871, P.R.China

ABSTRACT

This paper proposes a new method for Chinese language corpus processing. Unlike the past researches, our approach has following charactericstics : it blends segmentation with tagging and integrates rule-based approach with statistics-based one in grammatical disambiguation. The principal ideas presented in the paper are incorporated in the development of a Chinese corpus processing system. Experimental results prove that the overall accuracy for segmentation is 97.68% and that for tagging is 94.55% in about 400,000 Chinese characters.

1. Introduction

Processing a Mandarin Chinese corpus needs to go through several stages. From initial text corpus, through word segmentating, grammatic category tagging, syntactic analysis (bracketing) , semantic and pragmatic analysis, one can get corpora with different tags, such as segment-ational tags, word categories, phrase categories and so on. In current paper, we will focus on the first two stages, i.e. word segmentation and category (i.e. part of speech) tagging.

Word segmentation is essential in Chinese information processing because there are no obvious delimiting markers between Chinese words except for some punctuation marks. Matching input characters against the lexical entries in a large dictionary is helpful in identifying the embedded words. However some ambiguous segmentation strings(ASSs) and unregistered words (i.e. the word that is not registered in the dictionary) in the text will lower the segmentation accuracy. To resolve these problems, various knowledge sources might have to be consulted.

In the past decade, two different methodologies were used for word segmentation: some approaches are rule-based([1--5]), while others are statistics-based([6--8]). Many automatic word segmentation systems adopting the above models have been developed and significant results have been achieved. But these systems were developed only on word level. They did not take large-scale corpus category tagging into account and were short of a objective evaluaton for segmentation accuracy from category level. So the development of these automatic segmentation systems is restricted.

Grammatical category tagging for Chinese language is very difficult, because Chinese words are frequently ambiguous. One Chinese word can represent lexical items of different categories. Apart from this, unlike English and other Indo-European languages, Chinese has no inflexions and therefore there are not obvious morphological variations in Chinese text which are helpful to distinguish one grammatic category from others.

In some Chinese category tagging systems, statistics-based algorithms were used([10--12]). The basic processing procedure of these systems is: First, a tagged corpus was made through editing. Then, a dictionary containing category tagging entries and a matrix of category collocational probabilities were derived from the tagged corpus. Using these arguments, a probability model was built and category tagging was completed automatically. Up to now, there are not any reports about rule-based approach to Chinese language category tagging.

Comparing with the above researches on segmentation and tagging, our method has the following new characteristics:

First, it blends segmentation with tagging. We use a segmentation dictionary, in which every word is marked with its word category, to complete segmentation and initial tagging simultaneously. The category becomes a bridge linking segmentation and tagging.

Second, it integrates rule-based approach with statistics-based approach in category tagging. Therefore it inherits the advantages of the two approaches and overcome their respective disadvantages.

The following sections will discuss this method in detail.

[1] The project is support by National Science Fundation of China

2. Corpus processing blending segmentation with tagging

In practice of segmenting many Chinese sentences, we find that it is helpful to make use of word category in automatic segmentation processing. In general, there are three advantages:

1). Using category collocational relation of different words in ASSs and the contextual word categories, one can resolve most segmentation ambiguities.

As we know, there are two types of ASS : intersecting ASS (IASS) and combining ASS(CASS).

An IASS S=ABC has two possible segment-ation : AB+C and A+BC. Thus it results in two category combinations : $C_{AB} + C_C$ and $C_A + C_{BC}$. But the probility for them to appear in a given context is not the same. Depending on their context and the difference between two category collocational probabilities ($P(C_{AB}|C_C)$ and $P(C_A|C_{BC})$), we can select a correct segmentation.

Sometimes a CASS S=AB can be segmented into two words: A+B, but occasionally it is only one word S. Since the CASS itself can not provide the special information for correct segmentation, it is necessary to take the relation between it and its forward word or its backward word into consideration. In this sense, the categories of the words in the CASS and those one beside the CASS play a very important role.

2). Help to compound new words by using Chinese word-formation theory

In Chinese, a word is composed of morphemes. The combination of morphemes has its special rules. These rules tell us which and what kind of morphemes can be combined into a word. Using these rules, we can find out some unregistered words and segment them correctly from a sentence. For example, typical word-compounding cases of nouns are :

A). mono-syllablic noun + mono-syllablic noun
ma(horse) + *che*(car) --> *mache*(carriage)
B). mono-syllablic noun + bi-syllablic noun
shou(hand) + *zhijia*(nail) --> *shouzhijia*(finger nail)
C). bi-syllablic noun + mono-syllablic noun
dianliu(current) + *biao*(table) --> *diaoliubiao*(galvanometer)
D). bi-syllablic verb + mono-syllablic noun
zhengming(prove) + *xin*(letter) --> *zhengmingxin*(testimonial)

From such word-compounding cases, we can sum up many useful word formation rules that are based on category combination. Therefore, we will achieve a better segmentation effect in spite of using a smaller segmentation dictionary.

3). Be helpful to discover some segmentation errors

In Chinese sentence, the frequency of some category collocations is very low, such as d+n+$, v+u+d+$ and so on, where d is adverb, n is noun, v is verb, u is auxiliary, $ is the ending mark of a sentence. Therefore, if there is such a category combination in the segmented sentence, we will almost be certain that this segmentation may be wrong. In the following examples, there are such errors :

i). *mai*/v *le*/u *yitou*/d *niu*/n ./w
(buy -ed head cow
correct result : bought a cow)
ii). *ta*/r *qiu*/n *da*/v *de*/u *zuihao*/d ./w
(he ball play Prt had better
correct result : He plays basketball best.)

Here, we can see that the category information provides a powerful means to check segmentation errors automatically.

Based on all the above understandings, we proposed a method combining segmentation with tagging and used it in the practice of segmentation and category tagging on a large-scale Chinese language corpus. The basic processing procedures are :

First, complete automatic segmentation by using a segmentation dictionary with word categories. On the meantime, assign an initial tag(all possible categories for a word) to every segmentation unit.

Second, carry out some basic word-compounding words, such as combining stems with affixes , combining overlapping morphemes, integrating Chinese numberal words and so on.

Third, implement automatic category tagging through grammatic category disambiguation and assign a single category tag to every word.

Fourth, find and combine unregistered words which accord with Chinese word formation rules and assign a suitable category to the combined new words.

Fifth, check the category combination in segmented sentences, find some possible errors and then go back to the segmentation process.

3. The designing strategy of category tagging

Comparing with many past automatic category tagging systems([10--12]), our current processing has some new properties. The basic idea can be briefly summarized as following:

1). Be based on a dictionary with word categories

In current process, the initial category tagging was made by looking up the segmentation dictionary with word categories during segmentation. The category is derived from the "Grammar Knowledge Base for Chinese Words" (GKBCW), which has been developed by the Institute of Computational Linguistics of Peking University in the past five years[13]. Since the information in the dictionary was provided by linguists who refer themselves to the standard of classification based on the distribution of grammatical functions[14], it is of high accuracy. Therefore, applying this information to initial category tagging, the coherence and reliability of the tagging results can be guaranteed. This has laid good foundation for the following disambiguation processing.

2). Use a small tag set

In our current system, category tagging is restricted to the basic category descriptions, i.e. 26 categories. Meanwhile, in order to keep the new information that was found during manually proofreading, such as proper names, proper addresses, and so on, we add up several subcategories: ng(proper noun), ngp(proper noun for a person), and Ng(noun morpheme), Ag (adjective morpheme) and Vg(verb morpheme). All these categories and subcategories form a tag set of 31 tags.

A small tag set can help us concentrate on the ambiguous words that appear the most frequently in a sentence. Therefore, the processing complex can be reduced and tagging accuracy will be improved.

3). Form a stereo knowledeg base by combining tagged results with the information in the dictionary

Although our tag set is small, we can easily expand the tag set for the different application by linking with the GKBCW. Because in our GKBCW, each category has many features, which were proposed by liguists. These features help to describe the grammatic functions and distributions of every category completely. For example, verb category has about forty features, and noun category has twenty-five features([13]). In general, these grammatic features are also one kind of information for classification.

If we use the word and its basic category in tagged corpus as a keyword to look up GKBCW, we can get the detailed grammatic features of each word. Therefore, taking all tagged words as a plane, and the grammatic features of every word as a depth, we will give a stereo knowledge base. According to different needs, we can tag different grammatic categories or subcategories to the words in corpus by using the grammatic features in knowledge base. In addition, using the stereo knowledge base, we can also analyse the phrase structure of sentence in corpus.

4). Integrate rule-based approach with statistics-based approach in disambiguation

Because rule-based approach and statistics-based approach have their respective advantages, we tried to integrate them in our category tagging system. Our method is: First, through statistical analysis (manually or automatically) in a large-scale corpus, find the the most frequent ambiguous phenomona, study their context, and extract some contextual frame rules to eliminate those most frequently appearing and comparatively simpler ambiguities. Then, using the arguments trained by correctly tagged corpus, make a probability model to disambiguate some ambiguous category combination of lower frequence and deduce the category of the unregistered words.

But during actually processing, we lay different particular emphasis on these two approaches at different stages. At first, because there was not a large-scale corpus tagged with correct category, a small-scale corpus had to be tagged using rule-based approach and its remaining ambiguities and some tagging errors were corrected manually. After statistic analysis on the correctly tagged corpus, the rule base was adjusted and some trained arguments were given. Then some new sentences were added to the old corpus to form a new middle-scale corpus. Using the new adjusted rules and trained arguments, the new corpus was tagged through both rule-based approach and statistics-based approach. In this way, the scale of the corpus was increased gradually like a snowball. Due to the increase in corpus scale, the descriptions of rule became more and more accurate and the statistic information became more and more comprehensive. Therefore the manual proofreading work will decrease drastically. As a result, a best integration of these two approaches was achieved.

4. Disambiguation in automatic category tagging

4.1. rule-based approach

The basic strategy of rule-based approach is to determine one category for a categorically ambiguous word based on its syntatic or semantic context. In our system, in order to highten the tagging effect, the task is divided into three stages:

1). disambiguate against special word

In Chinese running text, some multi-tag words appear frequently, especially the mono-syllablic words, such as, "yi", "zhe", "le", "guo", "ba", "lai", "hao", "jiu", and so on. For these words, we set some special disambiguation rules, which describe the different context for these words with different category. Therefore, the category of words in one sentence can be determinated easily. This is a word-oriented disambiguation.

2). disambiguation against special multi-tag

According to statistic analysis, some multi-tag combinations, such as v-q, p-v,v-n,q-n,v-d,a-v and so on, appear frequently in corpus. In order to construct the disambiguation rules for these multi-tag combinations, the probability that one special tag is selected from a multi-tag set in the different context is counted. At the same time, the grammatic function features of category, especially the distribution inforamtion which distinguishes one category from the others are summed up and extracted. Then the ambiguities can be eliminated by these rules. This is a multitag-oriented disambiguation.

3). disambiguate by context constraint

The approach applies a set of context frame rules. Each rule, when its context is satisfied, has the effect of deleting one or more candidates from the list of possible tags for one word. If the number of the candidates is reduced to one, disambiguation is considered successful. This is a frame-oriented disambiguation.

4.2. statistics-based approach

Formally, the statistic scheme can be described as following:

Let $W=W_1...W_n$ be a span of ambiguous words in sentence and W_1,W_n are unambiguous, $C=C_1...C_n$ be a possible tag sequence for the span, where C_i is a category of W_i . $P(C|W)$ is conditional probability from W to C. Therefore, the goal of disambiguation is equivalent to find a list of category sequence C' with the largest score $P(C'|W)$, i.e.

$$P(C'|W)=\max_{C'\in C} P(C|W)$$

Computing the above formula with bi-gram model, we get:

$$P(C|W)=\prod_{i=2}^{n} P(C_i|C_{i-1})\, P(W_i|C_i)$$

where $P(C_i|C_{i-1})$ are the contextual probabilities and $P(W_i|C_i)$ are the lexical probabilities. The approximation of two probablitics can be calculated from the trained arguments.

During actual process, the category of the unregistered word is deduced firstly. Let C_u is a possible tag set for unregistered word, C_l is the tag of its left word and the C_r is the tag of its right word. T is the set of total tags in corpus. Therefore, $Cu=\{C_1, C_2\}$, where :

$$C1 = \operatorname*{argmax}_{Ci\in T} P(C_i|C_l)$$

$$C2 = \operatorname*{argmax}_{Cj\in T} P(C_r|C_j)$$

So the unregistered word phenomenon is changed into categorically ambiguous problem.

For a span of ambiguous words (bounded by unambiguous words), if we arrange the different tags of every word vertically and the different words horizontally, we will form a direct chart whose nodes are tagged with $P(W_i|C_i)$ and whose arcs are tagged with $P(C_i|C_j)$. Using VOLSUNGA algorithm ([9]) to get the best path in direct chart, we will complete the automatic category disambiguation.

5. Experimental results and future work

A segmentation and tagging system was built based on the above mentioned. The programs of the system are written by C language. Using this system, a verb usage corpus with about 400,000 Chinese characters or 300,000 Chinese words was segmented and tagged. The test results are: segmentation accuracy --- 97.68%, tagging accuracy --- 94.55% .

Some better processing results of previous segmentation systems and tagging systerms are : about 99% segmentation accuracy on 150,000 Chinese characters ([5]) and 94.82% tagging accuracy by close test on 150,000 words of tagged corpus ([12]). Compared with these systems, the result of our system is promising.

In our future research, we try to make further improvement on our method and add some new funtions to our segmentation and tagging system, such as, unregistered word deduce during segmentation, identity management in knowledge base, analysis belief degree on tagging results. Then we will extend our corpus' scale to about five million words.

In addition, we will persue research on Chinese phrase structure analysis and try to tag phrase category in corpus. We hope the work will be helpful for the study on Mandarin Chinese grammar.

References

[1]. Liang N.Y., (1987). An Automatic Word Segmentation System of Written Chinese---CDWS. *Journal of Chinese Information Processing (JCIP)*, Vol 2.

[2]. Li G.C., Liu K.Y. & Zhang Y.K. , (1988). Segmenting Chinese Word and Processing Different Meaning structures. *JCIP*, Vol 3.

[3]. Huang Y.X., (1989). A 'Produce-Test' Approach to Automatic Segmentation of Written Chinese. *JCIP*, Vol 4.

[4]. Yao T.S. , Zhang G.P. & Wu Y., (1990). A Rule-Based Chinese Word Segmentation System. *JCIP*, Vol 1.

[5]. He K.K. , Xu H. & Sun B. , (1991). The Implement of Automatic Segmentation Expert System of Written Chinese. *JCIP*, Vol 3.

[6]. Li B.Y. & i.e. , (1992). A MM Automatic Segmentation Algorithm using Corpus tag to Disambiguation. *Proc of ROCLING IV*, P147-165.

[7]. Zhang J.S. , Chen Z.D. , Shen S.D. , (1992). A method of word Identification for Chinese by Constraint Satisfaction and Statistical Optimization Techniques. *Proc. of ROCLING IV*, P147-165.

[8]. Sun M.S. , Lai T.B.Y. , Lun S.C. & Sun C.F., (1992). Some Issues on the Statistical Approach to Chinese Word Identification. *ICCIP 92*, Vol 1, P246-253

[9]. Steven J. DeRose, (1989). Grammatical category disambiguation by statistical Optimization. *Computional Linguistics*, Vol 14, P31-39

[10]. Liu K.Y. , Zhen J.H. & Zhao J. , (1992). A Research on Several Algorithms for the Assignment Parts of Speech to Words in Corpus. *Advances on Research of Machine Translation*, P378-386.

[11]. Bai S.H. , Xia Y. & Huang C.N. , (1992). The Methordic Research of Grammatical tagging Chines Corpus. *Advances on Research of Machine Translation*, P408-418

[12]. Bai Shuanhu & Xia Ying, (1991) . A Scheme For Tagging Chinese Runing Text. *NLPRS'91*, P345-349

[13]. Yu S.W. , Zhu X.F. , Guo L., (1992). Outline of the Grammar Knowledge Base for Chinese Words and its Developing Approachs. *ICCIP 92*, P186-191

[14]. Zhu Dexi , (1979). *Xufa Jianxi (Lectures of Grammar).* Business Press.