# A Chinese Word Segmentation System for Information Processing of Complex Sentences [1]

Shuangyun Yao[1], Dongtao Yu[2]
Center for Language and Language Education, Central China Normal University, Wuhan, China 430079
E-mail: ysy@mail.ccnu.edu.cn
2. Foreign Languages Department, Hunan Institute of Science and Technology, Yueyang, China 414006
Language Research Center, Shanghai Normal University, Shanghai, China 200234
E-mail: yudongtao@yahoo.cn

**Abstract**

*A large proportion of sentences in Chinese are complex ones, so it is significant to study them in Chinese information processing. There are marked and unmarked complex sentences. Connectives are the marks of marked complex sentences and the key to syntactic and semantic analysis of such sentences. Most connectives are conjunctions, while some of them are phrases which have been lexicalized and cannot be accurately extracted and tagged by the current Chinese word segmentation systems. The paper aims to set a Chinese word segmentation system that is capable of information processing for complex sentences by ameliorating the current ones. Statistics show that the new system is in a position to extract 87.2% of connectives and tag 85.8%, indicating its usefulness and potential of wide application.*

**Keywords**
*Complex sentences information processing, Chinese Word Segmentation System, connective, Corpus of Chinese Complex Sentences*

## 1. Introduction

A large proportion of sentences in Chinese are complex ones, and many of them are marked. According to the statistics collected by Yao(2006), 75% of Chinese sentences are complex sentences, among them 25% are marked ones. Therefore, information processing for complex sentences is important. Currently, natural language processing has made it possible to process sentences. To probe into the syntactic rules of Chinese, we should not stop at the level of simple sentences. Research on complex sentences is necessary and significant, and the key to this issue is the study of marked complex sentences, for the connectives in such sentences can serve as the mark of the syntactic and semantic relationships and make it easier for computer to recognize and process such sentences.

The premise of automatic analyzing and processing of complex sentences is the accurate recognition of the connectives. Concerning this issue, the current word

1 Research for this article was supported by a grant for a Youth Project of China Ministry of Education 'Cohesive Device and Textual Function Research on Connectives in Conversation Style' (07JC740018) and by 'Dangui Projects' provided by Central China Normal University (06DG027).

segmentation systems are insufficient, so we have ameliorated them and added the function of extracting and tagging connectives.

The paper first introduces some related information about complex sentences and word segmentation. Then, it probes into the deficiency in the connectives processing of the current Chinese word segmentation systems. Finally, it presents the procedures of our research in developing the word segmentation system.

## 2. Lexicalization of connective phrases and information processing

### 2.1 Distribution of connectives

Connectives are the linguistic units that link up clauses in complex sentences. Specific clausal relationship is tagged by specific connectives which are quite important in the information processing for complex sentences. According to their internal structures, connectives can be classified into the following four groups (Xing, 2002).

(i) Conjunction. Conjunctions are usually used to join clauses together, they are not any specific component in the clauses. For example, 因为(because)，所以(so)，虽然(although)，不但(not only)，而且 (but also) are of this type.

(ii) Relative adverb. Relative adverbs usually function both as an adverbial and a connective in a sentence. For example, 就(at once)，又(again)，也(also)，还(even, still) are of this type.

(iii) Auxiliary “的话” (if). “的话” is an auxiliary expressing subjunctiveness, which always appears at the end of a clause and marks the hypothetic relationship between the clauses.

(iv) Supra-lexical form. The supra-lexical form is not just one word, for example, 如果说 (if)，更不用说 (not to mention)， 要就是 (if it is)，etc.

In traditional grammar system, members of the first three groups are cataloged in words, while members of the fourth group are connective phrases. So *connective* is not a lexical concept, it covers both lexical and non-lexical expressions. Though supra-lexical connectives are different from those in other three groups, they are lexicalized units for their high frequency in use.

### 2.2 Advantages of processing lexicalized connectives as words in Chinese information processing

Word segmentation is one of the basic steps in natural language processing and the premise of other related researches. Consequently, accurate recognition of the connectives is the preliminary step in syntactic parsing and semantic interpretation of complex sentences. In traditional grammar system, supra-lexical connectives mentioned above are treated as phrases. Moreover, based on their internal structures, they are justified to be viewed as phrases. However, they are not only solidified in form but also functioning as a whole in grammar, so many people take them as conjunctions in practical use, and some dictionaries of function words even include certain supra-lexical connectives as conjunctions, such as “莫说是 (not to say)”, “简而言之 (in short)”, etc. Some word segmentation soft wares also tag such linguistic units as conjunctions.

As for Chinese information processing, it is necessary to process such connective phrases as conjunctions, because all the components in them realize the syntactic function and indicate the logic semantic relationship between clauses as a whole. Moreover, segmenting and tagging those phrases as words can provide explicit mark for syntactic parsing and semantic interpretation. Furthermore, this amelioration is also necessary for grammatical research. Adding such connective phrases to the enriched lexicon can further simplify the grammatical rule system and tally with the aim of establishing extensive corpus with comparatively simple grammar, which has been advocated by many contemporary linguists. (Xu, 2001) Moreover, it confirms to the reality that set expressions like “如果说 (if )” and “换言之 (in another word)” are treated as one lexeme. Similar treatment is made

to expressions like *in spite of*, *in order to* and *because of* in English, which are treated as one single word in morphological analysis in LOB. (Manning et al. 2005)

## 3. Deficiency of current word segmentation systems and the countermeasures

There are many tools for Chinese words segmentation, and two of them will be discussed here: ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System) developed by Institute of Computer Technology, Chinese Academy of Sciences, and CTSPT (Chinese Text Segmentation and POS Tagging, which can be tested on line at http://www.icl.pku.edu.cn/icl%5Fres/segtag98) by Institute of Computational Linguistics, Peking University. These two systems represent the highest level of Chinese word segmentation. ICTCLAS, which is of excellent performance and supports development of utilization under all types of circumstances, segments words with high speed and accuracy. CTSPT, which is based on a Chinese dictionary of very high quality, employs a method of both statistics-analysis and rules-application, and demonstrates both high accuracy and good performance. Even though these two systems have many advantages, some problems still exist in treating the connectives.

### 3.1 Problems for current word segmentation

As has been shown in the previous discussion, it is insufficient to study the complex sentences with the current word segmentation systems. The main problems are:

(i) Function words processing is not accurate enough.

Word-formation is a complicated matter in Chinese, because word is not a fixed grammatical unit here. For example, "只 (only)" belongs to three different word classes: adjective (for example, "只身 (alone)"), classifier (for example, "一只鸡 (a chicken)") and adverb (for example "他只会说英语 (He can speak English only.)"). "有 (have)" can be used only as a verb.

But when "只" and "有" are used together, the situation will become even more complicated. They can be combined together to form a word used as a conjunction, meaning "the only condition". For example:

(1) 只有 齐心协力， 才 能 把 事情 办 好。
Zhiyou qixinxieli cai neng ba shiqing ban hao
Only make concerted efforts can BA things do well
'Only by making concerted efforts, can we finish the tasks successfully. '

They can also be two independent words of a phrase, in which "只" is an adverb and "有" is a verb. For example:

(2) 张三 只 有 一 本 书。
Zhangsan zhi you yi ben shu
Zhangsan only have one classier book
'Zhangsan has only one book.'

The two examples above demonstrate clearly that some ambiguous word-groups may appear in Chinese word segmentation. Especially, when a function word appears together with other words, the possibility of ambiguity will increase greatly as shown in the following example.

(3) 我 只 有 一 支 铅笔。
Wo zhi you yi zhi qianbi
1S only have one classifier pencil
'I have only one pencil.'

Being extracted and tagged by ICTCLAS and CTSPT, the sentence turns out to be the same as the following.

(4) 我/r 只有/d 一/m 支/q 铅笔/n 。/w

In (3), “只有” are two independent words: adverb “只” and verb “有”, but in (4) they are missegmented as a connective.

Many similar examples can be found in Chinese. For example, “只要 (only if)” can be a connective of condition, “为此 (in this connection)” can be a connective of cause-effect, and “别说 (not to mention)” can be a connective of concession. But in many cases they are not used as words but a phrase, which are often missegmented.

Missegmentation will cause errors in word class tagging, especially in the tagging of multi-class words. For instance, “要” belongs to three different classes, which are in complementary distribution as shown in the following examples.

(5) 我 有 要 事 告诉 你。 (adjective)
Wo you yao shi gaosu ni
1S have important thing tell 2S
‘I have something important to tell you.’
(6) 他 要 了 一 台 电脑。 (verb)
Ta yao le yi tai diannao
3S take ASP one classifier computer
‘He took (asked for) a computer.’
(7) 他 要 不 来， 老师 一定 会 批评 他。 (connective)
Ta yao bu lai , laoshi yiding hui piping ta
3S if NEG come, teacher surely may criticize 3S
‘If he doesn’t come, the teacher will surely reproach him.’

Of the three part of speech of “要”, the frequency of the verb is much higher than that of the other two. So, “要” in (7) may easily be segmented and tagged as a verb. Moreover, the combination of “要” and “不” is an adversative connective, which are liable to be missegmented. The results of the tests with CTSPT and ICTCLAS are as follows.

(8) 他/r 要不/c 来/v ， /w 老师/n 就/d 会/v 批评/v 他/r 。/w (CTSPT)
(9) 他/r 要/v 不/d 来/v ， /w 老师/n 就/d 会/v 批评/v 他/r 。/w (ICTCLAS)

According to Yu (1998), 2.94% Chinese words are multi-class ones, which is one of the biggest problems in Chinese word segmentation.

(ii) The current word segmentation systems can not recognize relative adverbs (or connective adverbs).

The grammatical function of relative adverbs, the second group of connectives, is comparatively not definite in Chinese. Sometimes, they function as common adverbs, sometimes they don’t. For example, “就 (at once)”, “又 (again)”, and “也 (too)” are generally common adverbs, but they are connectives in the patterns like “如果……就 (if … then…)” and “即使…… 也 (even though)”. Because of the indefiniteness of their grammatical functions, the current word segmentation systems can not make an accurate division of them, as is shown in the following two examples.

(10) 你 向 前 走， 可 看见 一 条 小河。
Ni xiang qian zou, ke kanjian yi tiao xiaohe
2S towards front walk, can see one classifier small river
‘If you walk forward, you will see a small river.’
(11) 我 想 请 他 帮忙， 可 说 不 出 口。
Wo xiang qing ta bangmang，ke shuo bu chu kou.

1S want ask 3S help, can speak Neg out mouth
'I want him to help me, but I am too shy to ask .'

The results of tests with CTSPT and ICTCLAS are as follows:

(12) 你/r 向前/v 走/v，/w 可/d 看见/v 一/m 条/q 小河/n 。/w (CTSPT)
(13) 你/r 向前/v 走/v，/w 可/v 看见/v 一/m 条/q 小河/n 。/w (ICTCLAS)
(14) 我/r 想/v 请/v 他/r 帮忙/v ，/w 可/d 说/v 不/d 出/v 口/n 。/w (CTSPT)
(15) 我/r 想/v 请/v 他/r 帮忙/v，/w 可/v 说/v 不/d 出口/v 。/w (ICTCLAS)

Clearly, in(11),the adversative connective "可" is not tagged accurately.

(iii) The current word segmentation systems can not parse the lexicalized connectives.

Although connectives in the fourth group are functionally similar to other connectives, for they can grammatically link up clauses, they have not been widely viewed as words. The current word segmentation systems seldom segment and tag them as a whole. For example:

(16) 就算是 他 同意，我 也 不 同意。
Jiusuanshi ta tongyi, wo ye bu tongyi
Even though 3S agree, 1S also not agree.
'Even though he agrees, I won't.'

The results of tests with CTSPT and ICTCLAS are as follows:

(17) 就/d 算是/v 他/r 答应/v ，/w 我/r 也/d 不/d 同意/v 。/w (CTSPT)
(18) 就算/d 是/v 他/r 答应/v ，/w 我/r 也/d 不/d 同意/v 。/w (ICTCLAS)

In (16) "就算是" has been lexicalized as a connective of concession, which should be segmented and tagged as an individual word.

In Chinese there exist many lexicalized connectives, such as "要不是 (but for)", "以至于 (so…that)", "如果说 (if)", "假如说 (if)", "一方面 (on the one hand)", "另一方面 (on the other hand)", "与此相反 (contrary to this)", "换言之 (in another word)", "不单单 (not merely)", "尤其是 (especially)", "更何况 (let alone)", "甚至于 (even)", "要不然 (otherwise)", "以便于 (so that)", "为的是 (in order that)", "之所以 (so)", and so on. It will be very disadvantageous to the syntactic analysis of the sentences if these connectives are not segmented and tagged as a whole. The following is another example of a wrongly segmented and tagged sentence.

(19)如果/c 说/v 有/v 危机/n 存在/v 的话/u ，/w 那么/c 应该/v 面对/v ，/w 而/c 不/d 应该/v 逃避/v 。/w

In (19), not considering the holistic function of the supra-lexical form "如果说 (if)", the system segments and tags the lexicalized connective as a conjunction "如果 (if)" and a verb "说 (say)", which is disadvantageous to automatic parsing. According to the treatment of the system, there are three verbs in the first clause: "说 (say)", "有 (have)", and "存在 (exist)", and the coexistence of too many verbs increases the difficulty in analyzing the central verb, because computer cannot distinguish the different degrees of significance of those verbs. However, if "如果说" is parsed as a connective, the difficulty will be decreased.

### 3.2 Counter measures

The current word segmentation systems which tend to give priority to the strictness should not be blamed if they cannot segment and tag all connectives, because according to the previous discussion, connectives cover a wide range of linguistic units, from word to

lexicalized word groups. However, the significance of the connectives to the complex sentences processing makes it necessary and critical for a new function to be added to the current word segmentation systems, that is, the function of segmenting and tagging the connectives. Our plan is to recognize and tag the connective as a complete unit and process such information flexibly, without restructuring the current word segmentation system. To achieve the goal, we have adopted the following strategies: utilizing the statistics of the Corpus of the Chinese Compound and Complex Sentences to increase the accuracy of segmenting and tagging of the typical connectives; providing necessary instruction to increase the accuracy of automatic processing of relative adverbs; collecting the connectives lexicalized and being lexicalized and practically placing them on the word list to achieve holistic segmenting and tagging.

To make a detailed syntactic and semantic study of complex sentences, 12 types of connectives are classified, to which different symbols are assigned respectively: cyg (cause and effect)，ctd (inference), cjs (assumption)，ctj (condition)，cmd (purpose)，cbl (coordination)，clg (cohesion)，cdj (progression)，cxz (option)，czz (transition)，crb (concession)，cjz (assumption and transition).

A secondary division of the connectives on the basis of the current conjunctions is in accord with the trend of natural information processing, as well as the essence of Chinese word segmentation. Natural language processing is now relying more and more on the semantic information. The secondary tagging can provide richer syntactic and semantic information. For instance, noun can be secondarily tagged as nouns referring to time, place, people, organization, and so on. Taking the tagging of Chinese idioms as another example, Chinese idioms were tagged as “i”, but many scholars have found that this kind of tagging can only indicate the origin of the phrase but not the function, which is against the accurate processing of Chinese sentences. So some claim the need of the secondary tagging for idioms and it goes as follows: nominal idiom (in), verbal idiom (iv), adjectival idiom (ia) and connective idiom(ic). (Jin et al., 2003) Obviously, the adaptation is of great value.

## 4. Design of Our Word Segmentation System

Probing into the usage and matching of connectives, and distilling their contextual feature from the Corpus of Chinese Compound Sentences, we have set up a knowledge base that shows the usage of connectives, to guide the recognition of connectives and remove their ambiguity in word segmentation system.

### 4.1 Improvement of CLDC-LAC-2003-001 and related work

The word segmentation corpus used in our research is CLDC-LAC-2003-001, which was developed by the Artificial Intelligence Lab of Tsinghua University in 2003. CLDC-LAC-2003-001 lists more than 100,000 items. Each item includes word, phonetic symbol and frequency of use. We have made the following adaptations to the system: firstly, deleting the phonetic symbol, tagging the word class of each item and distributing the frequency according to the proportion of each word class; secondly, reordering the words with Delphi and memorizing the text file; and thirdly, setting a text file that covers all the connectives and their tags. As for research method, we have combined the automatic forward-backward matching with statistical stack search segmenting. Moreover, we have gauged the result with related knowledge and rules of connectives and secondarily tagged nouns like nouns referring to places, people, etc.

### 4.2 Definition

If the maxim results of forward-backward matching of the sentence are respectively SEG1 and SEG2, then

(i) the solo segmentation of SEG1 and SEG2 is defined as fragment (punctuations are exclusive);

(ii) the segment which can be extracted in different ways is defined as ambiguous

segment;

(iii) the sequential fragment words selected from the optimal result of forward and backward matching and words sideward are defined as unloaded words.

**4.3 Pretreatment of word segmentation method**

First, load the CLDC-LAC-2003-001 and initialize the related data. Second, parse the texts to be processed into sentences according to the punctuations. Third, select one sentence from the pretreated data and conduct primary extraction with the forward-backward matching with every possible word class tag and corresponding frequency attached. The non-inputted are tagged as noun and fragment as well.

If the results of forward parsing and backward parsing are identical, the results are defined as the correct segmentation. If not, to the cases with the same number of fragments, we maximized the probability of the ambiguous segment, that is, finding out all possible words of ambiguous segment, then finding out all the possible segmentation paths, and finally seeking the path of maximum probability to be the final output.

**4.4Processing of non-inputted words and connectives**

(i) Processing the unloaded words (including the nouns referring to persons, places and organizations) according to specific rules.

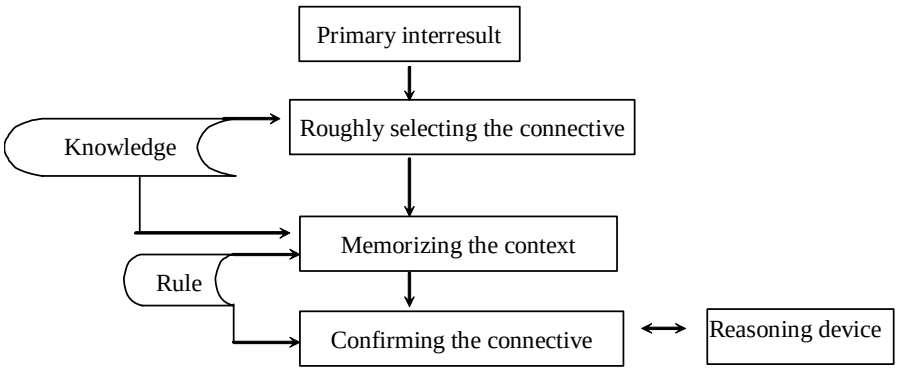
The set of rules includes two parts: positive rules and negative rules. Most of the unloaded are tagged as fragments after pretreatment. We seek the inconspicuous nouns that referring to the name of persons, places and organizations, and then confirm their categories with positive or negative rules. We have used the method of Charniak et al. (1993) in treating the unloaded words for reference.

(ii) Extracting the connectives according to the constituted connective rules.

The set of rules of connectives also includes two parts: extracting rules and tagging rules, each of which includes rewriting rule and triggering environment (Liu Ying,2002)

The Word Segmentation Model for connectives is shown in Diagram 1.

Diagram 1




Connective “所以 (so)” is taken as an example to illustrate the model here. The rewriting rule is to extract it into two words, and the triggering environment is: no connective “因为” exists in the previous clause, and “所以 (so)” is not at the beginning of the clause. Here is an example.

(20) 研究所 以 高昂 的 士气 投入 到 这个 项目 中 。
Yanjiusuo yi gaoang de shiqi touru dao zhege xiangmu zhong。
Research center with high Aux. morale put into this project in .

'The research center has put the project into operation with high morale.'

Sentence (20) has been tested with CTSPT and the output is as follows:

(21) 研究 所以 高昂 的 士气 投入 到 这个 项目 中 。

But with the adoption of the previous rules, the sentence is correctly parsed as following:

(22) 研究所 以 高昂 的 士气 投入 到 这个 项目 中。

**4.5Part-of-speech tagging**

The conversion matrix of parts of speech is the probability matrix of one part of speech converting to another one: P = (Pij) (* i and j are the indexes of two classes in the tagging system). The conversion matrix of part of speech is(cf. Liu,2000):

$$P = \begin{bmatrix} P_{00} & P_{01} & P_{02} & \cdots \\ P_{10} & P_{11} & P_{12} & \cdots \\ \cdots & \cdots & \cdots & \end{bmatrix}$$

$$P_{ij} = \frac{\text{times of co - occurrence of M i \& j}}{\text{times of co - occurrence of i}} \times 100\%$$

(1) Select the path of part-of-speech tagging by VOLSUNGA (cf. DeRose, 1998) algorithm and obtain the tag.

The conversion probability of part of speech here is a tiny positive. If the words are of a great number, the probability of every tag will be close to zero which cannot be figured on computer and comparison and analysis are not available. The solution is to figure the logarithmic summation of conversion probability. The logarithms of conversion probability are all negative and we get positive by reversing.

From the part of speech j of word i to the part of speech k of word i+1 is formulized as the following.

Fee (i, j, i+1, k)=-1log10(Pjk)-log10(Pk) (* Pk is the probability of word i+1 grammatically functioning as k. )

We can figure out the optimal tag, whose fee is the minimum in Linear Time, with the dynamic planning method.

(2) Scan and modify the tags according to the tagging rules of the connectives and make further complementation and modification to the rule base.

**5. Conclusion**

The ameliorated Word Segmentation System is to meet the need of complex sentences processing and improve the recognition and tagging of the connectives in such sentences in Chinese word segmentation system. statistics show the accuracies of extracting and tagging of connectives of the ameliorated one are respectively 87.2% and 85.8%, which indicates the bright future of its application. Our further attention will be given to the analysis of the errors in the extracting and tagging and the frame of further rules so as to increase the accuracy of Chinese Word Segmentation System.

**Acknowledgements**

For valuable enlightening comments and suggestions, we wish to thank the anonymous referees as well as Prof. Jie Xu. Our special thanks also go to Prof. Jinzhu Hu, Dr.Wei Shen and Dr. Chaohua Du for their contribution to the research and project.

**References**:

Charniak, Eugene. 1993. *Statiscal Language Learining*. Cambridge, MA: MIT Press.

DeRose, S. 1998. Grammatical Category Disambiguation by Statistical Optimization, *Computational Linguistics* .14.31-39.

Jin,Guangjin et al. 2003. On Standardization in Corpus Processing: a Discussion on *Standardized set of Chinese POS Markers for Computational Uses*, *Language Application*, 4.16-24.

Liu, Kaiying. 2000. *Chinese Word Automatic Parsing and Tagging*. Beijing: Commercial Press.

Liu, Ying. 2002. *Computational Linguistics*. Beijing: Tsinghua University Press.

Manning, Christopher D. and Hinrich Schütze. 2005. *Foundations of Statistical Natural Language Processing.* Trans. Yuanchun Fan et al. Beijing: Publishing House of Electronics Industry.

Yao, Shuangyun. 2006. *A Research on the Collocation of the Relation Markers of Chinese Compound Sentences and Some Relevant Explanation.* Doctoral Dissertation, Wuhan: Central China Normal University.

Yu, Shiwen et al. 1998. *The Grammatical Knowledge-base of Contemporary Chinese: A Complete Specification*. Beijing: Tsinghua University Press.

Xing, Fuyi. 2002. *A Study of Chinese Complex Sentences*. Beijing: Commercial Press.

Xu, Jie. 2001. *Grammatical Principles and Grammatical Phenomena*. Beijing: Peking University Press.