A Parsing Method for Identifying Words in Mandarin Chinese Sentences

*Liang-Jyh Wang **Tzusheng Pei *Wei-Chuan Li *Lih-Ching R. Huang

**Application Software Department
Computer and Communication Research Laboratories (CCL)
Industrial Technology Research Institute (ITRI)
W300, Blg. 14, 195 Chung Hsing Road, Section 4
Chutung, Hsinchu, Taiwan, R.O.C.
E-mail: x200hlc0@twnitril.bitnet

**Advanced Technology Center, CCL, ITRI
E000, Blg. 11, 195 Chung Hsing Road, Section 4
Chutung, Hsinchu, Taiwan, R.O.C.
E-mail: x200pts0@twnitril.bitnet

Abstract

This paper presents a parsing method for identifying words in mandarin Chinese sentences. The identification system is composed of a Tomita's parser augmented with tests originally a part of the English-Chinese machine translation system CCL-ECMT together with the associated augmented context-free grammar for word composition. The simple augmented grammar with the score function effectively captures the intuitive idea of longest possible composition of Chinese words in sentences and, at the same time, take into consideration the frequency counts of words. The identification rate of this system for the corpora taken from books and a newspaper is 99.6%. This identification system is simple, but the identification rate is relatively high. The minimum element for word-composition parsing is down to characters as opposed to sentence parsing down to Chinese words. It has the potential of incorporating phrase structures and semantic checking into the system. In this way, word identification, syntactic and even semantic analysis can be organized into a single phase. The results of testing the word identification on corpora taken from books and a Chinese newspaper are also presented.

1 Introduction

In processing Chinese sentences, the first phase is to identify words in sentences before doing further processing, such as syntactic and semantic analysis [Chen, 88]. Each Chinese sentence is composed of a sequence of Chinese characters. The character sequence is to be partitioned into segments with each segment corresponding to a Chinese word. Unfortunately, there is no boundary mark between any two consecutive words in a Chinese sentence. For many Chinese sentences, there are usually many ways to identify words in the sentences, i.e. the sentences are ambiguous in word composition. For an input sentence, the identification module first looks up possible words in the system dictionary. If there exists any ambiguity, the identification system should resolve it. The system is to identify the most favorable sequence of words for the input sentence. Most work done in resolving the ambiguity arising from identifying words is as follows. The approach uses a statistic method to group Chinese characters into two-character words making use of a measure of character association based on mutual information [Sproat and Shih, 90]. A statistic approach using frequency count of words [, 86][Liu et al., 75] is based on the statistical relaxation method widely used in image processing [Fan and Tsai, 87]. The structural method sets up heuristic rules for word-to-word relation to check the relationship among characters [Ho, 83][Liang, 87][Yeh and Lee, 88a][Zhang, 87]. In the unification-based approach [Yeh and Lee, 88b][Yeh and Lee, 90], the unification is a primitive operation. The ambiguous word strings are resolved by

ambiguity-resolution rules. Then the survived segmentations are ranked by the Markov process. Finally, an HPSG-based chart parser prunes results of identification of illegal syntactic and semantic construction.

This paper presents a simple but effective method for identifying Chinese words in sentences. It is based on the intuition of longest matching of Chinese words. The identification rate of testing on corpora of more than 16000 characters taken from a newspaper and books is about 99.6%. The word identification system is composed of a dictionary, a simple context-free grammar with augmented tests for word composition, a score function embedded in the tests of grammar to reflect longest matching of Chinese words, and a Tomita's parser [Tomita, 86] augmented with tests. The parser is originally a part of the English-Chinese machine translation system CCL-ECMT, formally ERSO-ECMT [Tang and Huang, 88]. Here, the minimum element for parsing is down to characters. The system usually outputs more than one result of word identification for an input Chinese sentence. The first output is the most favorable one which has highest score. The form of the Chinese word-composition grammar is the same as the one used for English sentence syntax in CCL-ECMT.

The structure of the Chinese word identification system is shown in Figure 1.

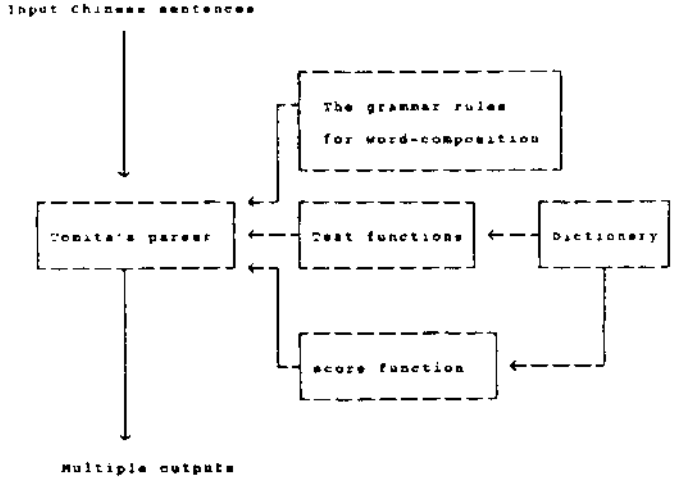


Figure 1. The architecture of the Chinese word identification system.

2 The Dictionary, the Grammar for Word Composition, and the Parser

The system Chinese dictionary contains a list of Chinese words sorted by character. Each word has at least one character. Each entry contains the word, its frequency count, and its part of speech. The total number of words is about 90000. The dictionary is taken from a source independent of the testing corpora.

The basic concept of constructing the context-free grammar for composition of Chinese words is as follows, but not necessarily appearing in this form.

```
S  --> WL
WL --> W WL
WL --> W
W  --> x
W  --> xx
W  --> xxx
W  --> xxxx
W  --> xxxxx
```

In above, S is a Chinese sentence, WL a list of words, W a Chinese word, and x a Chinese character. In general, Chinese words of more than five characters are very rare. Therefore, in the grammar for Chinese word composition, a Chinese word can only comprise one to five characters.

Word-composition tests can be augmented to grammar rules. The resulting grammar is called augmented context-free grammar. A rule is currently being used by the parser to construct parsing tree only under the condition that the tests should be met for the present parsing status. Mainly, the augmented tests include the following categories in the system. Some examples are shown in below.

1). Replication of text:

ABAB -- 討論討論.
AABB -- 高高興興.
ABAC -- 辦里糊塗.
AAB -- 吃吃飯.
A一A -- 看一看.
A不A -- 要不要.
AA -- 紅紅.

2). Numbers:

九萬八千四百二十.

3). Prefix: the first characters in the examples are the

prefixes.

初一, 第二, 老王.

4). Suffix: the last characters in the examples are the suffixes.

物理學, 科學家.

In general, unknown words not covered by the tests in above can not be identified.

This system has the capability of easily adding composition rules and any syntactic structures written in the same form of augmented context-free grammar, such as the determiner measure rules collected by the Chinese Lexicon Group, Academia Sinica, Taiwan. At present, the determiner measure has been added into the word-identification system. Syntactic and semantic checking for the current sentence segment under processing can also be added as tests under a grammar rule. The long-term goal would be developing syntactic and semantic analysis for complete sentences and the associated word identification module as one system. In fact, there is no clear boundary between word identification and further syntactic and semantic analysis of sentences. In general, the word identification can not be separated from the high level analysis, if we wish to do word identification well. In the identification system presented here, syntax of sentence segments may help identify words.

3 The Score Function

A score function is set up in the system. The formula for the score function is based on the intuition of longest Chinese words matching and taking into the consideration the frequency counts of words. The score function is implemented in the augmented tests under grammar rules. The score for an identification is accumulated until the end of parsing the whole sentence. The final accumulated score is the score of the result of identification. For a sentence, there are usually more than one parsing tree. Each tree has its own score. The parsing tree with highest score identifies the most favorable word composition of the input sentence. The score function for a sentence is as follows:

$$\text{SCORE} = \sum_{wordi} [\ (\text{length(word i)})^2 + \text{frequency-count(word i)/constant}\ ],$$

where length(word i) is the length of the ith word, frequency-count(word i) the frequency count of the ith word in the sentence, and constant set to 10,000,000,000 in the identification system. In fact, the constant can be any big number that makes

$$\sum_{wordi} [\text{frequency-count(word i)/constant}] < 1.$$

Since the summation in above is less than 1, the score from

$$\sum_{wordi} [\text{length(word i)}]^2$$

would dominate the total score. The frequency count dominates the score only when comparing words of the same length.

4 The Structure of the Determiner Measure

The Chinese Lexicon group, Computer Center, Academia Sinica, Taiwan, has developed a context-free grammar for determiner measure which is in the same form as the Chinese word composition grammar used here. This part of syntactic analysis has been merged into the word identification system.

For example, an expression with numbers is as follows.

```
NO = ( 一, 二, 兩, 三, 四, 五, 六, 七, 八, 九, 十,
       百, 千, 萬, 億, 兆, 零, 數, 幾, 好幾 )

           *
IN1 --> (NO)

Description = ( 大, 小 )

M = ( 個, 隻, 本, ..... )

DM --> ( IN1 ) (optional  Description )
       ( M )
```

In above, DM defines the determiner measure, IN1 specifies numbers, M is a collection of units, and Description is optional.

The determiner measure provides additional restrictions on syntactic structure of sentences with measure. It helps in identifying words of determiner measure.

5 An Example

Chinese sentence: 開發中國家人民生活.

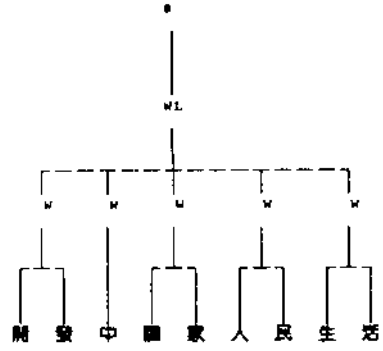
(People's life in developing country.)

The system outputs five trees. They corresponds to five results of identification. The first tree is the one with highest score. It should correspond to correct interpretation of human, if it is identified right. The trees are as follows.

The frequency counts of the possible words in the Chinese sentence are as follows.

開發	= 26	中	= 2200
國家	= 565	人民	= 1471
生活	= 656	中國	= 50
家	= 1802	家人	= 2
民	= 23	民生	= 1
活	= 298		

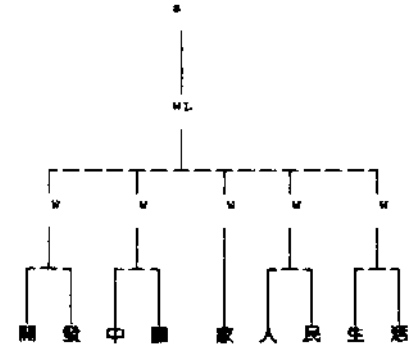
Tree 1:



The result is 開發 中 國家 人民 生活.

The score of tree 1 is 17.0000004918.

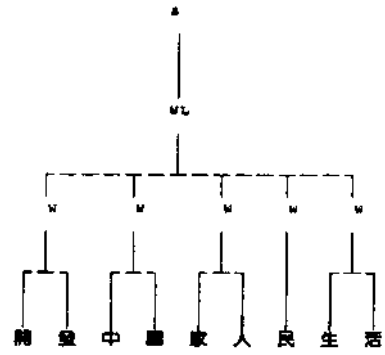
Tree 2:



The result is 開發 中國 家 人民 生活.

The score of tree 2 is 17.0000004005.

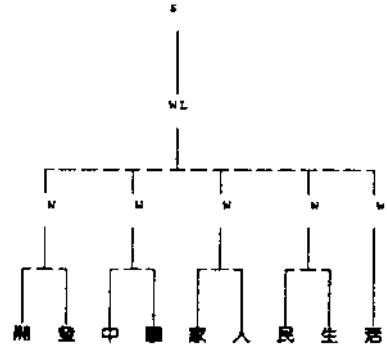
Tree 3:



The result is 開發 中國 家人 民 生活.

The score of tree 3 is 17.0000000757.

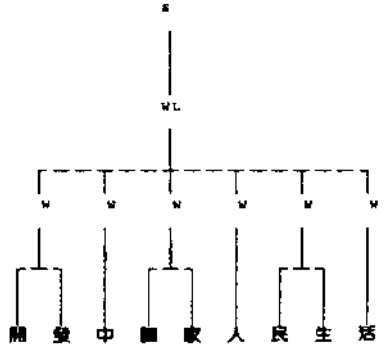
Tree 4:



The result is 開發 中國 家人 民生 活.

The score of tree 4 is 17.0000000377.

Tree 5:



The result is 開發 中 國家 人民生 活.

The score of tree 5 is 15.0000010405.

In the trees of this example, it clearly indicates that grouping of characters into words depends further on grouping of segments of sentence. The information of the system dictionary is apparently not enough to identify words in all sentences. To the extreme, the word identification process depends on syntactic and even semantic information of the sentence.

6 Experiment

The word identification system is implemented in LISP on TI microExplorer. The average speed of identification is about 13 characters per second. The testing corpora are taken from an article on optical character recognition, the preface of a book on QuickC and MS-DOS, and an article from a Chinese newspaper. There are more than 16,000 characters in the corpora. The identification rate is defined as the ratio of the number of characters correctly identified and the total number of characters in the testing corpus. The identification rate for our corpora as a whole is 99.6 %.

Without test functions under the grammar rules to prune unfavorable trees in advance, the parser usually generate so many parsing trees for some long Chinese sentences. That would take long parsing time. For some sentences, the parsing time may exceed 6 minutes, a time set for the word identification system. The parsing time would be considerably reduced if many trees can be pruned away in advance. But it is possible that, in some cases, favorable trees might be cut off. This is a tradeoff between identification rate and run time.

Since the system is based on longest matching, a shorter matching already covered by a longer matching may be ignored. This would considerably reduced run time for constructing parsing trees. In this way, for some long sentences, the number of trees which survives may be less than 10, compared with more than 1200 trees in the cases of some sentences without tests and actions added for eliminating shorter matching. This kind of tests and actions can be easily implemented under the grammar rules. At present, the grammar is small and there are not many tests in the system, so the system can be easily maintained.

There are still some infrequent unsolvable problems which reflect in the misidentified rate 0.4 % under this scheme. Most problems of this kind need syntactic and even semantic information to solve them.

A list of misidentified sentences taken from the corpora is collected in Appendix A.

7 Discussion and Conclusions

In processing Chinese sentence, identification of words is usually not done alone. In general, there is no clear boundary between identifying word composition and identifying the syntactic structure of sentences. The correctness of identification of some words even depends on the higher level syntactic and semantic structure.

Depending on the source of area, the identification rate may be different. The frequency count of words is collected from a corpus of a specific domain. It may not be applied to other domains with the same result.

Here a simple parsing method for Chinese word identification is presented. It has high identification rate and is flexible. Syntatic information for sentence segments and even semantic checking can be easily incorporated into the system.

8 Acknowledgments

The determiner measure presented in this paper is the work of Chinese Lexicon Knowledge Base Group, Computer Center, Academia Sinica, Nankang, Taiwan. Many thanks are due to all of them. We wish to thank Ming-Song Wang for his comments and support for testing the corpora. We also wish to thank United Informatics, Inc., for providing us with the database of Chinese text.

The paper is a partial result of the project No. 33H3100 conducted by ITRI under the sponsorship of the Ministry of Economic Affairs, ROC.

Appendix A

The following are typical misidentified Chinese sentences taken from the testing corpora. They can be roughly classified into two categories. In each example, the first one is the output of the word identification system, and the second one is what the sentence is supposed to be identified that way.

Category 1. In this category, longest matching is wrong. The long segments *of* Chinese characters should be further identified into shorter words to make sense. The syntactic word composition rules are apparently not enough for word identification. In general, semantic information is needed for this category. Fortunately enough, the method of longest matching can still identify words with identification rate of more than 99%.

1. 文 中將 限制 條件 分成 事量 順序
   文 中 將 限制 條件 分成 事量 順序

2. 需 經 實驗 才能 決定
   需 經 實驗 才 能 決定

3. 在 鍋牛 實驗 中用 的 交通 號誌 計時 碼
   在 鍋牛 實驗 中 用 的 交通 號誌 計時 碼

Category 2. In this category, the score contributed from length is the same for both the first result with highest score and the correct one. Here, the frequency

count dominates the difference. Unfortunately, the most frequently used words are not always the one to make sense. It still needs syntactic and semantic information to rank the outcomes of identification.

1. 印刷體 中 文字 識別 之 特徵 值
   印刷體 中文 字 識別 之 特徵 值

2. 必須 小於 某一 設 定值
   必須 小於 某一 設定 值

3. 是 為了 節省 字 庫存量
   是 為了 節省 字庫 存量

4. 除了 字 根號 以 2 byte 來 儲存 外
   除了 字根 號 以 2 byte 來 儲存 外

5. 最近 的 語法 把 他們 從 C 的 學習 曲線上 帶 跑 了
   最近 的 語法 把 他們 從 C 的 學習 曲線 上 帶 跑 了

6. 不 過日子 久 了
   不過 日子 久 了

7. 做 箇子 又 不 陷害 人
   做 箇子 又 不 陷 害人

References

[北京 , 86] 現代漢語頻率詞典 (XIANDAI HANYU PINLU CIDIAN), 北京語言學院語言教學研究所編. 北京語言學院出版社 1986年6月第一版.

[Chen, 88] Keh-jian Chen, "中文剖析的問題與對策 ", in Proceedings of R.O.C. Computational Linguistics Workshops I (ROCLING I), pp 19-28, 1988

[Fan and Tsai, 87] C. K. Fan and W. H. Tsai, "Automatic Word Identification in Chinese Sentences by the Relaxation Technique," in Proceedings of National Computer Symposium, pp. 423-431, National Taiwan University, Taipei, Taiwan, R.O.C, 1987.

[Ho, 83] W. H. Ho, "Automatic Recognition of Chinese Words," Master Thesis, National Taiwan Institute of Technology, Taipei, Taiwan, R.O.C. 1983.

[Liang, 87] Nanguan Liang, "On the Automatic Segmentation of Chinese Words and Related Theory," in Proceedings of the 1987 International Conference on Chinese Information Processing, pp.454-459, Beijing, 1987.

[Liu et al., 75] I. M. Liu, C. Z. Chuang, and S. C. Wang, "Frequency Count of Frequently Used Chinese Words," Luck Book Co., Taipei, Taiwan, R.O.C. 1975.

[Sproat and Shih, 90] Richard Sproat and Chilin Shih, "A Statistic Method for Finding Word Boundaries in Chinese Text," in Computer Processing of Chinese & Oriental Languages, Vol. 4, No. 4, March 1990.

[Tang and Huang, 88] An-Ching Tang and Tzu-Cheng Huang, "Tomita 增強型 LR 剖析器在機器翻譯系統的實驗應用 ", in Proceedings of ROCLING I, Taiwan, 1988.

[Tomita, 86] M. Tomita, "Efficient Parsing for Natural Language," Kluwer Academic Publishers, 1986.

[Yeh and Lee, 88a] Ching-Long Yeh and Hsi-Jian Lee, "Rule-Based Word Identification for Mandarin Chinese Sentences", in Proceedings of the International Conference on Computer Processing of Chinese and Oriental Languages, pp. 432-436, Toronto, Canada, 1988.

[Yeh and Lee, 88b] Ching-Long Yeh and Hsi-Jian Lee, "Unification-Based Word Identification for Mandarin Chinese Sentences," in Proceedings of the International Conference on Computer Processing of Chinese and Oriental Languages, pp. 27-32, Toronto, Canada, 1988.

[Yeh and Lee, 90] Ching-Long Yeh and Hsi-Jian Lee, "Unification-Based Word Identification for Mandarin Chinese Sentences-A Unification Approach," to appear in Computer Processing of Chinese & Oriental Languages.

[Zhang, 87] Chaosheng Zhang, "Adjacent Constraints and Automatic Segmentation of Chinese Words," in Proceedings of the 1987 International Conference on Chinese Information Processing, pp. 142-147, Beijing, 1987.