A Corpus-Based Statistical Approach to Automatic Book Indexing

Jyun-Sheng Chang*, **Tsung-Yih Tseng**, **Ying Cheng**, **Huey-Chyun Chen**, **Shun-Der Cheng**
Department of Computer Science
National Tsing Hua University
Hsinchu, Taiwan 30043, ROC
jschang@cs.nthu.edu.tw

Sur-Jin Ker
Department of Computer Science
SooChow University

John S. Liu
Software Research Office
Sampo Research Institute

Abstract

The paper reports on a new approach to automatic generation of back-of-book indexes for Chinese books. Parsing on the level of complete sentential analysis is avoided because of the inefficiency and unavailability of a Chinese Grammar with enough coverage. Instead, fundamental analysis particular to Chinese text called word segmentation is performed to break up characters into a sequence of lexical units equivalent to words in English. The sequence of words then goes through part-of-speech tagging and noun phrase analysis. All these analyses are done using a corpus-based statistical algorithm. Experimental results have shown satisfactory results.

1. Introduction

Preparing back-of-book indexes is of vital importance to the publishing industry but is a very labor intensive task. Attempts have been made over the years to automate this procedure for the apparent benefits of cost saving, shorter preparation time, and possibility of producing more complete and consistent indexes. Early work involves using occurrence characteristics of contents words [Borko, 1970]. Later people came to realize that indexes are often multi-word terms and their generation might involve more elaborated syntactic analysis on phrasal or sentential level [Salton, 1988; Dillon and McDonald, 1983]. However, a full syntactical approach [Salton, 1988] to this task has real problem with efficiency and coverage for unrestricted text. No viable automatic solution is currently in use.

Indexing Chinese books involves another severe obstacle, namely the word segmentation problem. Chinese text consists of a sequence of characters which roughly correspond to letters in English. However, there are no spaces to mark the beginning and end of a word as in English. Until recently, this problem has been considered difficult to solve without elaborated syntactical and semantic analyses [Chen, 1988].

Recent research advances may lead to the development of viable book indexing methods for Chinese books. These include the availability of efficient and high precision word segmentation methods for Chinese text [Chang et al., 1991; Sproat and Shih, 1990; Wang et al., 1990], the availability of statistical analysis of a Chinese corpus [Liu et al., 1975] and large-scale electronic Chinese dictionaries with part-of-speech information [Chang et al., 1988; BDC, 1992], the corpus-based statistical part-of-speech tagger [Church, 1988; DeRose, 1988; Beale, 1988], as well as phrasal and clausal analyzers [Church 1988; Ejerhed 1990]

2. Problem description

As being pointed out in [Salton, 1988], back-of-book indexes may consist of more than one word that are derived from a noun phrase. Given the text of a book, an indexing system, must perform some kind of phrasal and statistical analysis in order to produce a list of candidate indexes and their occurrence statistics in order to generate indexes as shown in Figure 1 which is an excerpt from the reconstruction of indexes of a book on transformational grammar for Mandarin Chinese [Tang, 1977].

Before phrasal analysis can be performed, the text must go through the more fundamental morphological and part-of-speech analysis. The morphological analysis for Chinese text is mainly a so-called *word segmentation* process, which segments a sequence of Chinese character into a sequence of words. See Figure 2 for illustration.

The noun phrase generation process described in this paper is based on a corpus-based statistical analysis and does not use an explicit syntactical representation. Examples of noun phrases found are underlined as shown in Figure 2.

* This research was supported by ROC National Science Council under Contract NSC 81-0408-E-007-529.

Figure 1. Indexes

當/兩/個/以上/的/邏輯/述辭/
[dan/lian/ge/yishang/de/luogi/shuci]
P/Q/CL/LOC/CTM/NC/NC/
When two or more logical predicates

在/同/一/個/句子/裡面/前/後/出現/的/時候/
[zai/tong/yi/ge/juzi/limian/qian/hou/cuxian/de/shihou]
P/D/Q/CL/NC/LOC/LOC/LOC/V/CTM/NC/
apear at the same sentence,

我們/就/說/後面/的/述辭/
[wuomen/jiou/shuo/houmian/de/shuci]
NP/ADV/V/NC/CTM/NC/
we then say that the predicate

在/前面/的/述辭/的/邏輯/範圍/
[zai/qianmian/de/shuci/de/luoji/fangwei/]
P/NC/CTM/NC/CTM/NC/NC/
is within the logical scope of predicates before it.

Figure 2. Segmentation, tagging, and noun phrase finding

3. Generating Indexes

3.1. Word Segmentation

Segmentation through Constraint Satisfaction

The word segmentation problem for Chinese can be simply stated as follows: Given a Chinese sentence, segment the sentence into words. For example, given

把 劉 顯 仲 的 確 實 行 動 作 了 分 析

we are supposed to segment it into

把/劉 顯 仲/的/確 實/行 動/作/了/分 析
[ba /liuxianzhong/de/queshi/xiendong/cuo/le/fenxi]
Xian-Zhong Liu 's exact action was given an analysis.

where 劉 (Liu) is a surname and 顯仲 (Xian-Zhong) is a last name. In the following, we will describe a method that extends our previous work on segmentation [Chang et al., 1991a] to handle surname-names [Chang et al., 1991b]. Segmentation is solved as a constraint satisfaction problem.

1. V	Verbe (Predicative)
2. NC	Nouns
3. NP	Proper Names or Pronouns
4. A	Adjectives (Non-Predicative)
5. P	Prepositions
6. ADV	Adverbs
7. CJ	Conjunctions
8. D	Determiners
9. Q	Quantifiers
10. CL	Classifers
11. LOC	Locatives
12. ASP	Aspect Markers
13. CTS	Sentential Clitics
14. CTN	Noun Clitics
15. CTM	Modifiers Clitics
16. INT	Interrogatives
17. S	Sentences
18. PP	Prepositional Phrases
19. PREF	Prefixes
20. SUF	Suffixes

Figure 3. List of part-of-speeches

The constraint satisfaction problem

The constraint satisfaction problem involves the assignment of values to variables subject to a set of constraining relations. Examples of CSPs include map coloring, understanding line drawing, and scheduling [Detcher and Pear, 1988]. The CSP with binary constraints can be defined as follows: Given a set of n variables X_1, X_2, ..., X_n and a set of binary constraints K_{ij}, find all possible n-tuples (x_1, x_2, ..., x_n) such that each n-tuple is an instantiation of the n variables satisfying

$$(x_i, x_j) \text{ in } K_{ij}, \text{ for all } K_{ij}$$

Segmentation as a Constraint Satisfaction Problem

The word segmentation problem can be cast as a CSP as follows: Suppose that we are given a sequence of Chinese character (C_1, C_2, ..., C_n) and are to segment the sequence into subsequences of characters that are either words in the dictionary or surname-names. We can think of a solution to this segmentation problem as an assignment of *break/continue* (denoted by the symbols '>' and '=' respectively) to each place X_i between two adjacent characters C_i and C_{i+1}:

$$\begin{array}{ccccccccc} | & C_1 & | & C_2 & | & C_3 & \ldots & | & C_n & | \\ X_0 & & X_1 & & X_2 & & \ldots & X_{n-1} & & X_n \end{array}$$

subject to the constraint that the characters between two closest breaks correspond to either a Chinese word in the dictionary or surname-names. (For convenience, we add two more places; one at the beginning, the other at the end.) So the set of constraints can be constructed as follows:

For each sequence of characters C_i, ..., C_j, (j >= i) which are a Chinese word in the dictionary or a surname-name,

if j = i, then put (>,>) in $K_{i-1,i}$.

if j > i, then put (>,=) in $K_{i-1,i}$, (=,=) in $K_{i,i+1}$, ..., and (=,>) in $K_{j-1,j}$.

For example, consider again the following:

把 劉 顯 仲 的 確 實 行 動 作 了 分 析

The corresponding CSP is

$K_{0,1}$ = {(>,>)},
$K_{1,2}$ = {(>,=)},
$K_{2,3}$ = {(=,>),(=,=)},
$K_{3,4}$ = {(=,>)},
$K_{4,5}$ = {(>,>)},
$K_{5,6}$ = {(>,>)},
$K_{6,7}$ = {(>,>),(>,=)},
$K_{7,8}$ = {(>,>),(=,>),(>,=)},
$K_{8,9}$ = {(>,>),(=,>),(>,=)},
$K_{9,10}$ = {(>,>),(=,>),(>,=)},
$K_{10,11}$ = {(>,>),(=,>),(>,=)},
$K_{11,12}$ = {(>,>)},
$K_{12,13}$ = {(>,>),(>,=)},
$K_{13,14}$ = {(=,>)},

since

把/劉顯/劉顯仲/的/的確/確實/實行/
行動/動作/作/了/分/分析/

are either words in the dictionary or probable surname-names (hypothesized words).

Typically, there will be more than one solution to this CSP. So the most probable one with highest product of probability of hypothesized words is chosen to be the solution. Ordinary words are listed in the dictionary along with this kind of probability estimated from a general corpus [Liu et al., 1975]. As for proper names such as Chinese surname-names not listed in the dictionary, their probability are approximated by using another corpus containing more than 18,000 names as described in the following subsection.

The Problem with Proper Names in Chinese Text

Proper nouns account for only about 2% of average Chinese text. However, according to a recent study on word segmentation [Chang et al., 1991a], they account for at least 50% of errors made by a typical segmentation system. Moreover, proper names are oftentimes indexes. Therefore their correct segmentation is crucial to automatic generation of back-of-book indexes.

The difficulties involved in handling proper names are due to the following: (1) No apparent punctuation marking is given like capitalization in English. (2). Most of characters in proper names have different usage. So this problem has been held impossible to solve in the segmentation process. And it was suggested that proper names are best left untouched in the segmentation process and rely on syntactical and semantic analysis to solve the problem when nothing can be made out of the characters representing them [Chen, 1988]. Using the corpus-based statistical approach, we have shown that it is possible to identify most Chinese surname-names (姓名) without using explicit syntactical or semantic representation.

Most surnames are single character and some rare ones are of two characters (*single-surnames* and *double-surnames*). Names can be either one or two characters (*single-names* and *double-names*). Some characters are more often used for names than others. Currently, there are more double-names than single-name in Taiwan.

The formation of hypothesized surname-names is triggered by the recognition of a surname. In the example above, 劉 (Liu) is one of some 300 surnames. Subsequently, we will take one character and two characters after the surname as probable last names, in this case 顯 (Xian) and 顯仲 (Xian-Zhong). A general corpus, G and a surname-name corpus N are used to evaluate the probability of a surname-name. For instance, the probability of a most common kind of 3-character name (single-surname/double-name) such as 劉顯仲 is :

P(劉顯仲) = P(single-surname/double-names in G) x
P(劉 being a surname in N) x
P(顯 being 1st character in names in N) x
P(仲 being 2nd character in names in N)

Names of other combinations can be handled similarly.

The Algorithm

To sum up, the whole process of word segmentation with surname-name identification is as follows:

1. Scan from left to right across the sentence
2. Check to see if the prefix of what is being scanned is a hypothesize word, by
2.1. dictionary lookup of an ordinary word and its probability
2.2. checking for the existence of a surname
2.2.1. forming possible combinations of the surname-name
2.2.2. evaluating the probability of each combination
3. Post the constraints of the CSP and probability for each hypothesized word
4. Solve the CSP
5. Find the most probable solution to CSP through dynamic programming

3.2. Part-of-speech Tagging

As far as we know, there has been only scarce research done on part-of-speech tagging for Chinese [Chang et al., 1988; Chen, 1991; Bai and Xia, 1991; BDC, 1992]. As for English, there are at least three independently developed

taggers [Church 1988; DeRose 1988; Beale 1988]. We started out using an electronic dictionary [Chen; 1991; Chang et al., 1988] with a very elaborated part-of-speech system based on Chao's work [Chao, 1968]. Because it is difficult to get sufficient manualy tagged data for a large tag set, we have since switched to another electronic dictionary with some 90,000 entries and a much smaller tag set. The dictionary is actually a bilingual one (Chinese-English) developed by Behavior Design Corporation [BDC, 1992]. The list of part-of-speeches is shown in Figure 3. The algorithm is essentially the same as [DeRose, 1988]. The BDC Chinese-English Dictionary is used to obtain the list of possible part-of-speeches for each segmented word. Currently, the collocation probabilities of part-of-speech are estimated from a manually tagged text of about 4,000 words.

3.3. Finding Noun Phrases

Instead of using a full-blown parser to find noun phrases, we first mark the noun phrases in the same text of about 4,000 words and compute the statistical characteristics of categoric patterns of noun phrase and then use the statistics in a stochastic algorithm for finding noun phrases in a manner similar to [Church 1988; Ejerhed 1990].

Extracting keywords from a noun phrase is somewhat heuristic unlike the rigorous approach of using the syntactical structure within the noun phrase in [Salton, 1988].

4. The Experimental Results

The algorithm described in Section 3 is currently under development and the programs are written in C and ProFox, and run on an IBM PC compatible machine. The segmentation, tagging, and NP identification parts are completed, while the statistical analysis of the occurrence of NPs is being implemented now. The statistics used in the system consists of four parts:

(*S1*) Appearance counts of 40,032 distinct words from a corpus of 1,000,000 words of Chinese text [Liu el al., 1975].

(*S2*) The BDC Chinese-English Dictionary [BDC, 1992].

(*S3*) A general corpus of 300,000 words. Some 4,000 words of text from this corpus is tagged and marked with NP.

(*S4*) A name corpus of some 18,000 surname-names.

The performance of the completed parts of the system is as follows: The hit rate of word segmentation is about 97% on the average. For the surname-names alone, we get 90% average hit rate which eliminate about 40% of errors produced by our previous segmentation system. About 98% of part-of-speeches are tagged correctly. And about 95% of the noun phrases are found successfully.

5. Concluding Remarks

The preliminary results that we have obtained seem very promising. The approach presented here does not rely on a fully developed Chinese grammar for syntactical analysis on the sentential level. Thus the efficiency in system development and generation of indexes is reasonable and cost of building and maintaining such a system is acceptable. Currently, we are working on (1) handling translated names, (2) improving the hit rate of tagging and NP identification by using a larger and more correctly tagged and marked training corpus, and (3) completion of the statistical analysis of occurrence of noun phrases.

Acknowledgements

Thanks are due to for Dr. Keh-Yih Su for making the BDC dictionary available to us. Preliminary work in segmentation has been done using the electronic dictionary developed by the Chinese Dictionary Group, Academia Sinica and acquired from Computer and Communication Research Laboratories through the Technology Diffusion Program of ITRI.

References

Shuanhu Bai and Ying Xia. A Scheme for Tagging Chinese Running Text. In *Proceedings of Natural Language Processing Pacific Rim Symposium*, pages 345-350, Singapore, 1991.

Andrew David Beale. Lexicon and Grammar in Probabilistic Tagging of written English, In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 211-216, Buffalo, 1988.

Behavior Design Corporation. *BDC Electronic Chinese-English Dictionary*, Hsinchu, Taiwan, 1992.

H. Borko. Experiments in Book Indexing by Computer, *Information Storage and Retrieval*, 6(1):5-16, 1970.

Jyun-Sheng Chang, Chi-Dah Chen, and Shun-Der Chen. Chinese Word Segmentation through Constraint Satisfaction and Statistical Optimization, In *Proceedings of ROC Computational Linguistics Conference*, pages 147-165, Kenting, Taiwan, 1991, (in Chinese).

Jyun-Sheng Chang, Shun-Der Chen, Ying Chen, John S. Liu, and Sue-Jin Ker. A Multiple-corpus Approach to Identification of Chinese Surname-Names, In *Proceedings of Natural Language Processing Pacific Rim Symposium*, pages 87-91, Singapore, 1991.

Li-Li Chang et al. *Part-of-Speech Analysis for Mandarin*

Chinese, Technical Rep. T0002, Computation Center, Academia Sinica, Taiwan, 1975, (in Chinese).

Yuen Ren Chao, *A Grammar for Spoken Chinese*, University of California Press, California, 1968.

Chih-Dah Chen. *Segmentation and Part-of-speech Tagging for Chinese,* master thesis, National Tsing-Hua University, Hsinchu, Taiwan, 1991.

Keh-Jiann Chen and Chu-Ren Huang, Word Classifications and Grammatical Representation in Chinese, *manuscript*, 1991.

Keh-Jiann Chen. Problems and Strategies in Parsing Chinese Sentences - A Tutorial, In *Proceedings of ROC Computational Linguistics Workshop*, Sitou, Taiwan, September, 1988, pp. 19-24, (in Chinese).

Kenneth Ward Church. A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. In *Proceedings of Second Conference on Applied Natural Language Processing*, pages 136-143, Austin, 1988.

Steven J. DeRose. Grammatical Category Disambiguation by Statistical Optimization, *Computational Linguistics*, 14(1):31-39, Winter 1988.

Rina Dechter and Judea Pearl, 1988, Network-Based Heuristics for Constraint-Satisfaction Problems, J. of Artificial Intelligence 34(1):1-38, 1988.

M. Dillon and L.K. McDonald. Fully Automatic Book Indexing, *Journal of Documentation*, 39(3):135-154, 1983.

Eva I. Ejerhed. Finding Clauses in Unrestricted Text by Finitary and Stochastic Methods, In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 219-227, Austin, 1988.

In-mao Liu et al. *Frequency Counts of Chinese Words*, Lucky Book Co., Taipei, Taiwan, 1975.

Gerard Salton. Syntactical Approaches to Automatic Book Indexing, In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 204-210, 1988.

Richard Sproat and Chilin Shih, A Statistical Method for Finding Word Boundaries in Chinese Text, *Journal of Computer Processing of Chinese and Oriental Languages*, 4(4):336-351, March, 1990.

Ting-chi Tang. *Studies in Transformational Grammar of Chinese, Volume I: Movement Transformations*, Taipei, Student Book Co., 1977, (in Chinese).

Lian-Jyh Wang, Tzusheng Pei, Wei-Chuan Li, and Lih-Ching R. Huang. A Parsing Method for Identifying Words in Mandarin Chinese Sentences, Identification of Chinese Name, In *Proceedings of International Joint Conference on Artificial Intelligence*, pages 1018-1023, Sidney, 1991.