A Hybrid Method for Abstracting Newspaper Articles

James Liu*
Department of Computing, Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong.
E-mail: csnkliu@comp.polyu.edu.hk

Yan Wu
Department of Computer, Harbin Institute of Technology, Harbin, People's Republic of China 150001.
E-mail: wuy@insun.hit.edu.cn

Lina Zhou
Institute of Computational Linguistics, Peking University, Beijing, People's Republic of China 100086.
E-mail: lzhou@cmi.arizona.edu

This paper introduces a hybrid method for abstracting Chinese text. It integrates the statistical approach with language understanding. Some linguistics heuristics and segmentation are also incorporated into the abstracting process. The prototype system is of a multipurpose type catering for various users with different requirements. Initial responses show that the proposed method contributes much to the flexibility and accuracy of the automatic Chinese abstracting system. In practice, the present work provides a path to developing an intelligent Chinese system for automating the information.

Introduction

Electronic information access has become quite popular in recent years. As the variety of retrievable materials continues to increase, the demands on information systems are growing, but are at the same time hard to meet, particularly when users are confronted with information volumes that increasingly exceed individual processing capabilities (Stanfill & Waltz, 1991). Consequently, there is an urgent need for user aids such as abstracting techniques, i.e., automatic methods of abstracting text, which will become more critical.

Generally, an abstract can be considered as a concise, short summary of an article. It carries more information than the title, but less than that of the whole text (article). The abstract is often written by someone other than the author, and it is very difficult to achieve satisfactory automatic abstraction. There has been a lot of research into issues related to information abstraction (Gokcay & Gokcay, 1995; Rama & Srinivasan, 1993; Stanfill & Thau, 1991; Stanfill & Waltz, 1991; Tsou, Ho, Lin, Liu, Lun, & Heung, 1990; Watanabe, 1996). Typically, abstraction systems select suitable methods by first taking user's requirements into consideration. We group users of this type of system into four categories: (a) they have a clear topic in mind and are satisfied with the result of something such as keyword searching; (b) they want to glance at the content of the text and expect the system to find the most important sentences from it; (c) they hope the extraction can be done automatically by the system while matching their defined criterion at the same time; and (d) they require the extracted sentences to connect with each other to provide a meaningful and fluent content of the original text.

It is found that all these users are becoming more and more demanding.

To identify the text content, a number of approaches have been suggested: the structural method, and conceptual, statistical, and syntactic methods. Broadly speaking, they can be divided into two classes: the understanding-based approach, and the extraction-based approach. The former employs the content features of the article in abstraction. It can provide better concept identification and representation, but such systems seem inherently to involve extensive human effort in knowledge engineering (Jacobs, 1992). Knowledge bases have been constructed for suitably narrow domains, and it is not an easy task to transfer such a knowledge base to a different domain.

The latter method employs the style feature and matching techniques to extract an abstract. Recently, research work by the IBM group and others (Church & Hanks, 1989; Cocke, Pietra, Jelinek, Mercer, & Roosin, 1988) has renewed interest in statistical methods for processing texts. The extraction-based method provides a simple and powerful perspective for text analysis. However, the success and applicability of statistical methods in literature is still an issue of controversy.

* To whom correspondence should be addressed.

Received January 21, 1997; revised March 17, 1999; accepted April 1, 1999.

In this article, a hybrid method for abstraction of Chinese text is proposed and tested in a window-based Chinese Extraction System. This is to enhance the abstraction process with better flexibility. The system is of a multipurpose type catering to various users with different requirements. It provides options allowing for the automatic selection of keywords and the most important sentences using relevant statistical information, and also for the analysis of the surface structure and features for each sentence in a given text. An abstract of the article can then be generated recursively according to the required structure and the user defined criterion. In the next section, the components and processes of the method are introduced in detail. Then, the abstraction system incorporated with the features mentioned above is presented, and the result is evaluated with the questionnaire. The limitation of the prototype system and improvement on this method is discussed in the last section. Finally, the method is applied to Chinese newspaper articles, and the results of this are given.

Chinese Automatic Abstraction

Research into Chinese automatic abstraction started in the earlier 1990s. Basically, methods similar to those for non-Chinese systems have been adopted. However, characteristics that are specific to Chinese include:

1. The Chinese sentence is a continuous string of characters. Compared with Indo-European languages, it has to go through a segmentation stage before the word can be identified. On the other hand, apart from some particles, Chinese has no morphology change, meaning that it can easily make out every occurrence of the same word without morphology analysis.
2. The Chinese language has a history of more than 3,000 years, and is currently used by one-quarter of the world's population. Chinese characters are ideograph in nature. According to the Kang Shi 康熙 Dictionary published in 1716 AD, there were 49,030 Chinese characters at that time. In the 1986 Mainland China Dictionary, more than 57,000 characters are listed. Some statistics are helpful here to help us understand the frequency occurrence of Chinese characters. If we take newspapers as a sampling target, about 2,500 characters are commonly used to cover roughly 98% of the paper stories. Among the most frequently used characters is 的, which represents around 3 to 5% of the total usage. The top 120 to 130 most frequently used Chinese characters cover roughly 50% of the total usage.
3. In the Chinese language, there is no strict correspondence between a word's classification and its distribution properties, nor are there obvious inflections to designate part of speech. This reflects the complexity of homograph identification, which involves some words that can take on several different meanings or functions without changing the written form. Homograph disambiguation for Chinese is done by using information nearby (part of speech, syntactic/semantic properties of neighboring words), and also by scanning the sentence to locate words with syntactic and semantic properties that may shed light on the probable part of speech of the homographic word.
4. There is a close set of conjuncts in Chinese, which is helpful in analyzing the rhetorical structure of a Chinese text. The rhetorical relations in Chinese context are analyzed as inferential, contrastive, resultive, listing, summative, apposition, and transitional in the method. Some conjuncts of the resultive and summative types are usually followed by important sentences.
5. The sentence position is an effective factor for abstraction evaluation. Usually, important sentences are located in the first and last paragraphs, and behind Cue Words such as "the conclusion is," "From the above," etc. The first and last sentences in every paragraph are also comparatively important. The values of the sentences at these positions should be high. It is also a good way of composing the abstract by extracting these sentences of high values.

A Hybrid Extraction Method

In literature, abstraction methods can range from the relatively mechanical extraction of keywords or more complex linguistic elements from the original text to the production of a pragmatically and stylistically well-formed summary based on text understanding. Generally speaking, existing automated text abstraction systems use one of the following approaches to perform abstraction on the original text:

Abstraction based on text extraction—The mechanical method is intended to extract important and representative sentence fragments, sentences, or even paragraphs from the original text of any domain by pattern matching, word frequency statistics, heuristic functions, and special term dictionaries. The abstract generated by a text extraction system using this method generally consists of keywords or disconnected sentences rather than well-connected prose. This was promising when it first emerged, but it was later discovered that this method was restricted by the data structure of the dictionary and experimental heuristic function (Jacobs, 1992). Therefore, automatic abstraction tends to be performed on a restricted domain and appended with a natural language understanding mechanism, which signifies the second phase.

Abstraction based on text understanding—This attempts to remedy the text extraction approach by including a knowledge base (KB) related to the area of discourse to assist in the analysis of the sentence structures, and to identify the focus and theme as well as other aspects of the content of the input text. A relevant, static, and complete knowledge base with syntactic and semantic rules and inference mechanism must be constructed before the text abstraction process can actually proceed. In addition, knowledge representation methods (Zhou, Shen, Yu, & Liu, 1996) were developed extensively to formalize the content of the text. This brings a higher quality to the output abstraction. However, its application is limited, and the KB would generally be domain dependent.

Our hybrid method takes advantage of both extraction-based and understanding-based approaches. It is based on

the extraction method, and incorporates some surface linguistic heuristics without involving much sophisticated language understanding work.

Extraction-Based Method

The objective behind automatic abstraction and document retrieval is the same: identifying the main theme and related topics in a specific document set. A number of different techniques for extraction are employed in our method, for example, pattern matching, statistical analysis of word frequency, and conceptual methods that rely on the use of knowledge bases.

Keyword approach

Keywords are the words that carry the major concepts of the article. They appear in the form of nouns/noun phrases. This approach is intended to facilitate the retrieval of potentially relevant items from a collection of Chinese natural language text. According to the user-defined keywords, the approach attempts to extract the matching sentences from the original text to form an abstract. Figure 1 shows the logic of keyword matching. Taking the user-input information as a pattern, the efficiency of the process depends very much on the pattern-matching base.

Several pattern-matching rules (Liu, Goss, & Murray, 1994) are compiled in the system as follows.

Given a sequence of characters,

$$P = p_1, p_2, \ldots, p_m$$

e.g., P = BBBSSFSBBSOFSBBSOFFOFBSSSBBFFB

For simplicity reason, these characters can be regarded as patterns (symbols) representing different Chinese characters:

Exact matching. The process compares every character pair of the keyword,

$$Q = q_1, q_2, \ldots, q_n$$

From P. If exact matching is found in all m pairs, (q_1, p_r), (q_2, p_{r+1}), $\ldots$, (q_n, p_{r+n-1}) where $1 \leq r \leq n \leq m$, then the matching is successful, for example, an exact match of Q below is found in P,

P = BBBSSFSBBSOFSBBSOFFOFBSSSBBFFB

Q = BBS

Note that the pattern Q occurs three times in P at different positions in this typical example. It could indicate that pattern P of the text is statistically important.

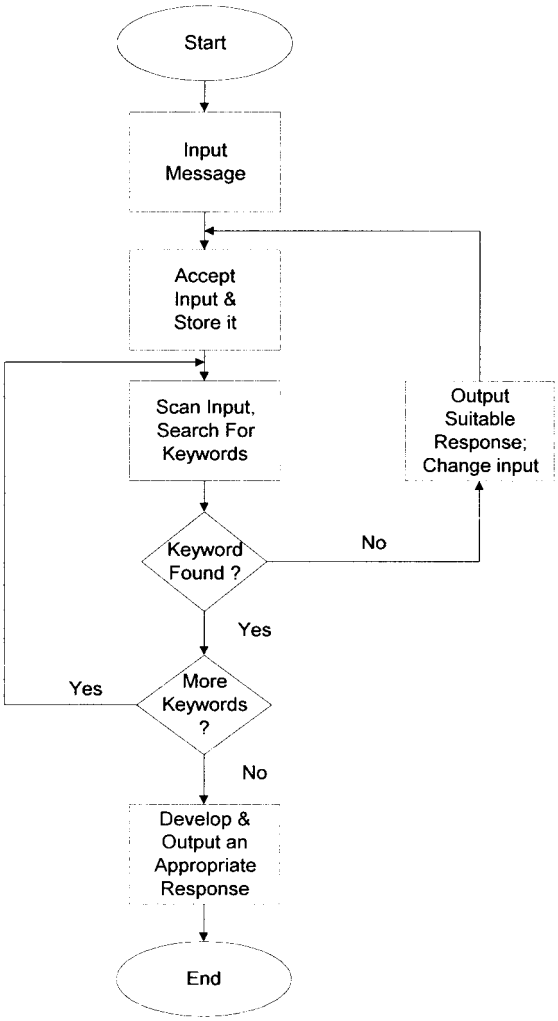


FIG. 1. A flow chart of keyword approach.

Partial matching. The process performs something like wild-card matching against a sequence, ignoring other subelements of the pattern, for example, a partial match of Q below is found in P,

P = BBBSSFSBBSOFSBBSOFFOFBSSSBBFFB

Q = B - -F

Such matching is useful when noise occurs in phrases or sentences. For example, "in that paper" and "in that abbreviated paper" can be considered as the matched pattern having the same meaning.

Variable matching. This matching allows a word in P to be a variable. For example, "聰 (clever)" is a variable in which a sample pattern is " . . . 聰 . . . ." It can be adopted to detect the synonyms at the corresponding location on Q. The character sequences

" . . . 聰明 (wise) . . . "

" . . . 聰敏 (smart) . . . "

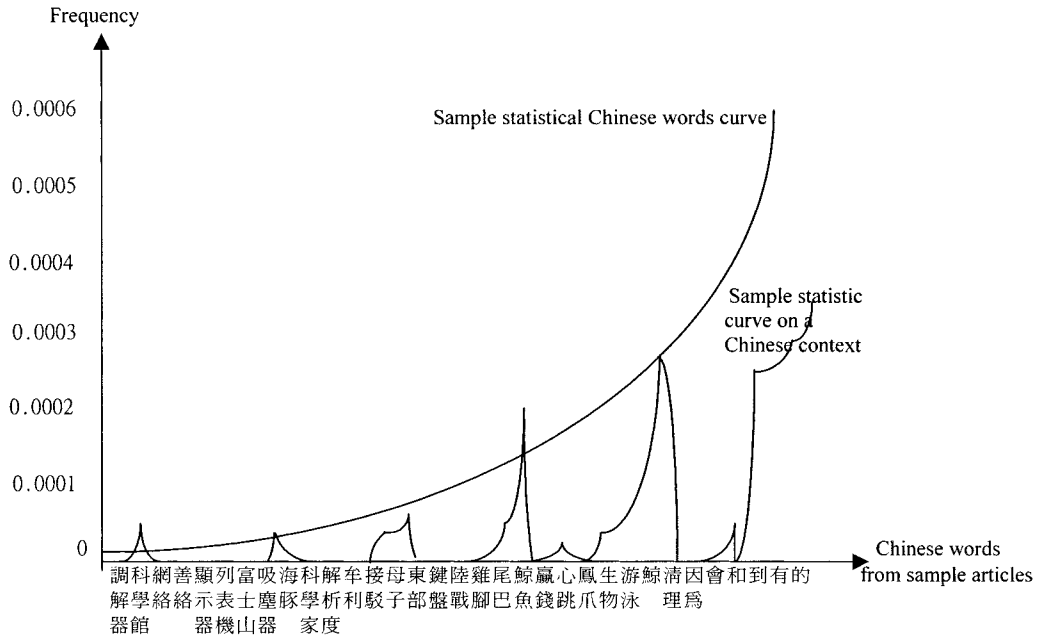


FIG. 2. Sample frequency graph of one of the Chinese texts.

" . . . 天聰 (talent) . . . "
can all be matched to Q = 聰敏, for example.

Flexible matching. This is a combination of the above three techniques to achieve keyword approach under different user criteria.

Statistical approach

Word frequency is calculated in the texts and compared with the standard word frequency statistics of 520,000 Chinese words. There are two types of word: closed word (pronoun, demonstrative, etc.); and open-word (noun, verb, etc.). Given an article, by ordering the open words according to their frequency statistics and removing those frequencies that are lower than a preset threshold, a set of effective words can be obtained. This information is used to weigh sentences. The value of a sentence can be represented by

$$\omega(S) = \text{Max}|\alpha(E_i)/\beta(\psi_i)| \tag{1}$$

where $\alpha(E_i)$ is the count of effective word E_i, and $\beta(\psi_i)$ is the total word count between effective words ψ_i in sentence S.

The word frequency statistics can also be used to extract keywords automatically and assign weights to those words. Keywords are the substantives (mostly open words) carrying the major concepts of the article. According to our corpus, Chinese function words (or Chinese empty words) usually have more frequency than Chinese substantives. The selection strategy is that those words, which have the least variance between their actual distribution and the standard ones, tend to be the important keywords Kw_i. The actual distribution is represented by current frequency Cfw_i of the word W_i in the current context; while the standard distribution uses the standard frequency information Sfw_i. The formula for calculating the word frequency fw_i is:

$$fw_i = \frac{O(W_i)}{\sum_{i=1}^{N} O(W_i)} \tag{2}$$

$$Kw_i = w_i$$

where

$$\underset{wi \in W}{\text{Min}} (|Cfw_i - Sfw_i|) \tag{3}$$

Here, $O(W_i)$ is the function for counting the occurrence of word W_i. N is the total number of different words in the article. W is the whole set of words in the text. To produce a consistent result for the information content of the characters and words, we have collected standard frequency data from a well-recognized resource, i.e., *Statistics and Analysis of Chinese Lexicon* 漢語詞匯的統計與分析, which is one of the major efforts of the Beijing Institute of Linguistics. It covers the primary, secondary, and tertiary Chinese textbooks, published between 1978–1985. Among the total number of 520,000 characters that we studied, there are 18,177 words. We applied these frequency statistics as a standard to make a comparison with the given text input into the system. A sample distribution is shown in Figure 2. The entries on the x-axis represent different Chinese words found in a sample article.

Keywords with different degrees of importance can then be extracted from the text automatically. For instance, a

word "我們 (*we*)" appears 10% of the time, and another word "網路 (*network*)" appears 0.02% in the actual frequency distribution of words, but it does not mean that "我們" is the most important word in the context. In the standard frequency distribution, "我們" appears 15% and "網路" appears 0.0015%. From the variance in distribution of these two words, we can see that "網路" has less variance value. According to Equations (2) and (3), "網路" will probably be selected as the keyword than the other word "我們".

After the most important word has been identified, a multiregression process is performed to guide the sentence output. This means that, if the most important word used in extracting the sentences cannot fulfill the required percentage of the text, the algorithm will take the second most important word as the extraction keyword. If the extraction still cannot satisfy the requirement, more keywords will be considered and the matching process regressed. The algorithm presented here performs much better than that based on the actual frequency analysis (Zhou et al., 1996).

Conceptual Approach

This is useful for those people who want to scan only a portion of a Chinese document for their favorite topics. In fact, this approach is a combination of the first two approaches. However, the difference is that the system will work on the keyword matching part first. If the extraction satisfies the user's requirements, no regression is needed. Otherwise, an extra statistical procedure will be executed. Figure 3 illustrates the logic of this method.

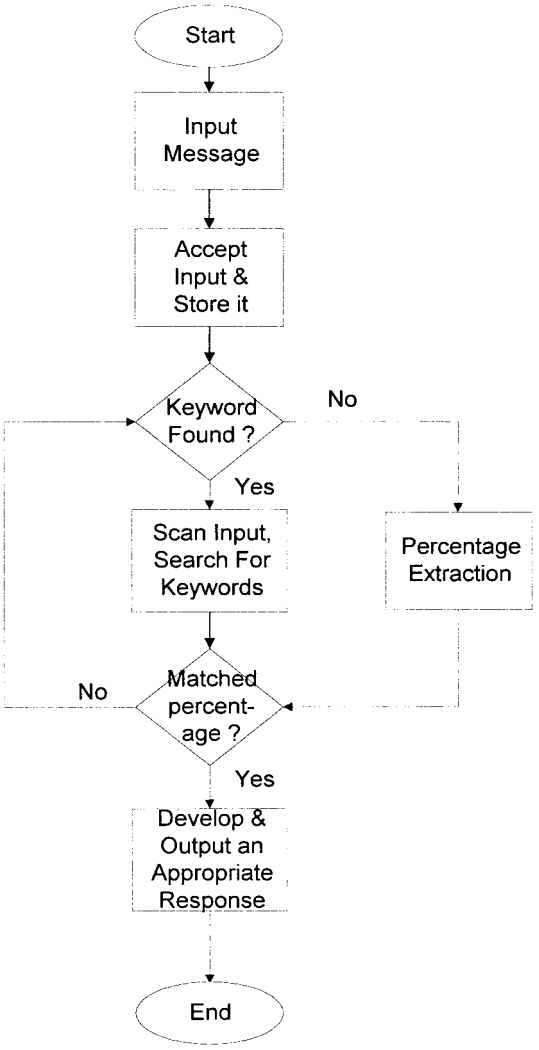


FIG. 3. A flow chart of the conceptual method.

Understanding-based method. In addition to the extraction-based method, we tried to use the rhetorical structure analysis to capture the main theme of the text. In a strict sense, this method is somewhere between extraction-based and understanding-based methods. It uses features in the article content. The system needs to understand the content on at least four levels:

First level: this is the word level. The system not only recognizes the words, but also has to know the functions of every word in a sentence. Basically, it classifies words into smaller classes for word identification and specification of the part of speech of the word, for example, 大衛 (David), 我 (I), 我們 (We), 你 (You), 你們 (You), 他 (He), 他們 (They).

Second level: this is concerned with the missing elements and the usage of demonstrative pronouns (dp), i.e., knowledge about what information has been left out and what is referred to by a particular demonstrative pronoun. For example, 這/dp 就/f 是/v 我/p 的/u 目標/n (this is my goal), 這 (this) is a demonstrative pronoun. The system should find the "目標 (goal)" before this sentence in the text. (f—adverb, v—verb, n—noun, u—auxiliary, p—pronoun)

Third level: the system recognizes a sentence according to its structure, and the functions and role of each word and phrase in the sentence. For example, the positions of subject and object in a sentence may not be exchanged, for example, (a) 我/p 是/v 中國人/n (I am Chinese). The relation between 我 and 中國人 is "Is-A," i.e., the first noun is a small concept and the second noun phrase is a large concept. It is a decision sentence of an implication relation, so the positions of subject and object cannot be interchanged. (b) 廚房/n 滿/f 是/v 廢料/n (Rubbish fills the kitchen). Since 廚房 (kitchen) is a location noun or direction phrase and 廢料 (rubbish) is some existing object, it is a decision sentence of an existential relation. Therefore, their positions cannot be interchanged.

Also, a simple description (meaning) can possibly be represented by different sentences. For example, 我/p 送/v 她/p—份禮物/n (I gave her a present), 我/p 送/v—份禮物/n給/v她/p (I gave a present to her),—份禮物/n 我/p **送給**/v 了/u 她/p (A present that I gave to her). In tackling this kind of problem, case grammar (Pun & Lum, 1989) is used to identify the case role and the number of cases. We shall treat, for example in 我送—份禮物給她, 給 (to) as an explicit tag, 我 (I) and 送 (gave) as implicit tags. The context constraints of a sentence have to be employed to determine these tags.

Fourth level: this is the relationship between clauses, i.e., the coherence relationship, causal relationship, etc. Each

sentence describes an event, and several events join together to describe a context. This is related to the understanding of sentence group or a paragraph in an article. Our approach adopts the rhetorical structure theory, which was developed by Mann and Thompson to describe the linguistic structure extraction (Tsou, Lin, Ho, Lai, & Chan, 1995). An abstract is worked out via the following four consecutive processing steps:

Chinese Automatic Word Segmentation

The first step in Chinese natural language processing is automatic word segmentation (Kwok, 1997). This segments a Chinese character string into a Chinese word string. That is,

$$s_1, s_2 \ldots s_n = w_1 w_2 \ldots w_j \quad j \leqq n$$

Here s_i is a Chinese character; w_i is a Chinese word.

The general methods of Chinese word segmentation consist of FMM (Forward Maximum Matching), RMM (Reverse Maximum Matching), WSM (Word Segmentation with Marks), and so on. The difficulties in automatic segmentation are ambiguity and new words.

Ambiguity resolution: segmenting ambiguities contain three types: ambiguity of combination type, overlap type, and hybrid type.

Definition 1. Suppose AB is a Chinese character string, if all of AB, A, and B are Chinese words, then AB is called an ambiguous phrase of combination type.

For example, the Chinese character string "他將來上海工作" (He will be coming to Shanghai for work) is an ambiguous phrase of combination type, because all of "將來 (will be coming)," "將 (will be)," and "來 (coming)" are Chinese words. With FMM, we get the error result "他/將來/上海/工作"; With RMM, we also get the error result "他/將來/上海/工作." The correct segmentation is "他/將來/上海/工作."

Definition 2. Suppose ABC is a Chinese character string; if all of AB and BC are Chinese words, ABC is called ambiguous phrase of overlap type.

For example, the Chinese character string "他的確切菜了(He is actually chopping the vegetable now)" is an ambiguous phrase of overlap type, because all of "的確 (actually)" and "確切 (really)" are Chinese words. With FMM, we can get the correct result "他/的確/切/菜/了"; With RMM, we get the error result "他/的/確切/菜/了."

Definition 3. Suppose ABC is a Chinese character string; if all of AB, BC, A, and B are Chinese words, ABC is called ambiguous phrase of hybrid type. This ambiguous type contains the features of ambiguity of combination type and ambiguity of overlap type.

For example, see the following Chinese character strings "這篇文章寫得太平淡了(This article was written too plainly)" and "牆抹得太平了(The wall was dubbed too flat)," here "太平淡(too plainly)" is an ambiguous phrase of hybrid type, because all of "太平淡 (too plainly)," "太平 (too flat)," "太 (too)," and "平 (flat or level)" are Chinese words. With FMM, we can get the error result "這/篇/文章/寫/得/太平/淡/了" and "牆/抹/得/太平/了"; with RMM, we can get the correct result "這/篇/文章/寫/得/太/平淡/了" and the error result "牆/抹/得/太平/了." The correct result is "牆/抹/ 得/太/平/了."

Suppose $S = c_1 c_2 \cdots c_n$ is a Chinese character string, and there are m kinds of segmentation results, they are $S_1 = w_{11} w_{12} \cdots w_{1n_1}$, or $S_2 = w_{21} w_{22} \cdots w_{2n_2}$, . . . , or $S_3 = w_{m1} w_{m2} \ldots w_{mn_m}$. We design the following mathematical model to select the segmenting result:

$$f(w_{i1} w_{i2} \cdots w_{in_i}) = \max_{i=1,m} \prod_{j=1}^{n_i} f(w_{ij})$$

Here, $f(w_{ij})$ is the frequency of word w_{ij}.

We have built a dictionary that contains a frequency of words collected from a standard Chinese dictionary. Using the equation above, we can effectively solve the ambiguities. For example, the segmenting result of the phrase "結合成分子時 (at the time of forming to a particle)" is "結合/成/分子/時"; the segmenting result of "他將來上海工作" is "他/將/來/上海/工作."

New word resolution. The electronic dictionary does not include all Chinese words, for example, names, addresses, and professional terms. These new words could present a problem for automatic segmentation.

We solve new word problems with some local corpus. Local corpus comprises of current text and other background knowledge. This method can solve the following new word cases.

If a Chinese character string appears two or more times in the local corpus, we consider it a Chinese word. According to statistics, we can find the following condition: the more times the Chinese characters conjointly appear, the more likely they are a Chinese word. In other words, the more times the Chinese word appears, the more times every Chinese character that comprises the word conjointly appears. We called the conjoint appearance concomitance. The synchronic frequency can perfectly reflect the reliability that the Chinese characters make a Chinese word.

Suppose ABC is a Chinese character string, A and C are all Chinese words, and B is a new word, then we can consider B as a Chinese word.

This method can solve more than 95% of new word problems in automatic segmentation (Wang, Li, & Wu, 1995).

Our segmentation system has been applied to Chinese natural language. The experimental results indicate that its accuracy is up to 98% (Wang, Li, & Wu, 1995).

Rhetorical Structure Analysis

In a Chinese context, especially argumentative discourse, the chains of reasoning are commonly indicated by explicit

syntactic markers, which express the rhetorical relationships between the constituent propositions. We then cut out the less important parts in the extracted structure to generate an abstract of the desired length. In Chinese text, these syntactic markers, called "[關聯詞] (conjunct)," include a relatively stable set of words with a few hundred entries.

In this step, it is not necessary to understand the meaning of each individual sentence. The only requirement is the ability to deconstruct the source text into constituent simple propositions separated by punctuation symbols and syntactic markers. Take the following paragraph as an example:

這個部分的程式負責的工作是在啓動中文系統的時候,把原先儲存在磁片上的中文文字字型資料讀到電腦主機的記憶体里, 因爲 (Because) 常用的中文字合起來約有一万三仟多字, 所以 (Therefore) 光這些字型資料就有已經很多, 如果 (Since) 全部擺到記憶体里, 就沒有空間執行其它的應用程式, 那麼 (consequently) 只能較常用的几仟個字型讀進來, 那麼要用到不在記憶体里的字怎麼辦?

$S1$ = 常用的中文字合起來約有一万三仟多字 (the commonly used Chinese characters amount to more than 13,000).

$S2$ = 光這些字型資料就有已經很多 (these characters are already associated with lots of information).

$S3$ = 全部擺到記憶体里, 就沒有空間執行其它的應用程式 (all of this information is resident in the memory, so there will not be enough RAM space for the application program).

$S4$ = 只能較常用的几仟個字型讀進來 (we can only read in those commonly used characters a few thousand at a time).

Q1(S4)—P(S3(Q2))//*inferential relation*
Q2(S2)—P(S1)//resultive relation
Q1, Q2, and P stand for the clause.

In this example, we can use the rhetorical structure to define the relations in the paragraph.

Rhetorical Structure Transformation

After understanding the relations between the sentences, the proper Chinese conjuncts will be used for the corresponding rhetorical relations. Based on the coarse analysis of the structure of the sentences, the system evaluates those sentences to decide which one should be discarded in the final abstract. In the mean time, some referential anaphors can be recognized and resolved.

The rhetorical relations in the above example state that it should be *cause and effect* 因果關係 and *probable premise and condition relations* 或然推理關係. By applying the rules that have been compiled for the associated type of relations, the above clause examples can be changed into:

$Q2(S2)$—$P(S1)$//*cause and effect*
becomes
$Q2$ = subject_rule($S2$)
(光這些字型資料就有已經很多)
- - - - - -
中文 (Chinese)
$Q1(S4)$—$P(S3(Q2))$//*probable premise and condition*
becomes
$Q1$ = probable_rule($S4$)
(只能較常用的几仟個字型讀進來)
或 (perhaps)

Hence, the final extracted result is $Q2 + Q1$:

光中文這些字型資料就有已經很多,或者只能較常用的几仟個字型讀進來] (these Chinese characters are already associated with lots of information, and perhaps we can only read in those commonly used characters a few thousand at a time).

Postprocessing

According to the statistical figures listed in *Statistics on Chinese Word Frequencies: Fast-Learning Word Selections* 漢字頻度統計—速成識讀优選法, Chinese words can be divided into five levels as follows. The first level has 500 words, the second level also takes up 500 words and so on until the fifth level, which contains 2991 Chinese words. It is guaranteed that the words in the first level have the highest frequency and those in the fifth level have the lowest frequency. Their distributions of usage differ sharply, which show that: (a) the first level caters for 77.419% of the total sample words; (b) the second level + first level accounts for 90.819%; and (c) the third level + second + first occupies 95.898%.

Inspired by the above statistics, the postprocessing scans the Chinese texts and the top 30% of words are chosen for extraction. A sentence, which does not contain any words from the 30% of keywords, will be removed, because their inclusion is regarded as nonimportant.

Text Structure Understanding

Because the length of abstract limits the abstracting result, the statistical abstraction only provides the main part of the subject, but discarding the other discussed ones in the text. Wu, Liu, and Wang (1998) have designed an algorithm to automatically divide the semantic paragraph of text. The semantic paragraph is a set of natural paragraphs that describe the same subject.

A text may be about a number of subjects. In a text, the words or phrases describing the subject may appear in a paragraph or consecutive paragraphs. If a paragraph is a start paragraph about a subject, it may contain a number of words or phrases that appear for the first time in the text.

According to the idea discussed above, we introduce rules for dividing the semantic paragraph as follows.

RULE: suppose $p_1, p_2, \ldots, p_n$ represents n paragraphs in the text T, we call $p_i - p_j$ semantic paragraphs if they satisfy the following conditions:

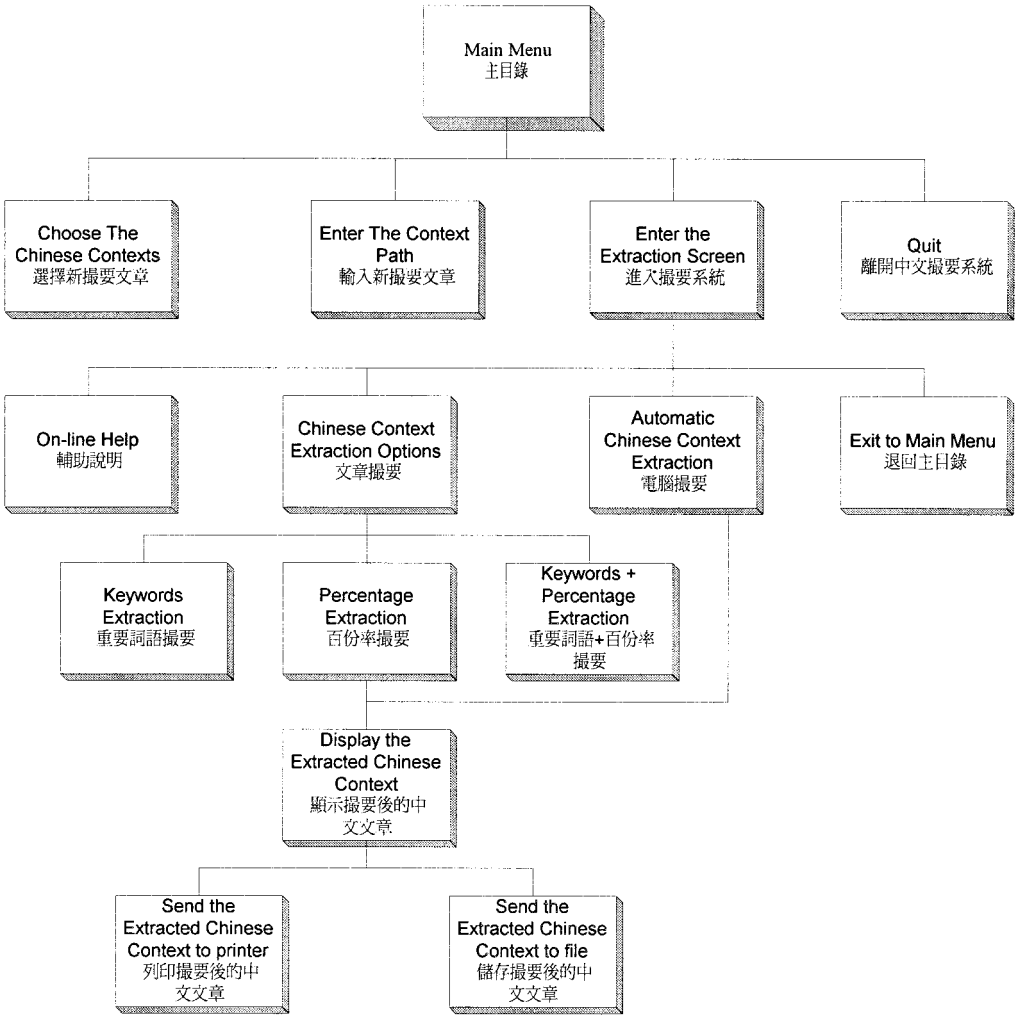


FIG. 4. An architecture of the Chinese Extraction System.

$$\alpha_{i-1,i} < \alpha_{k,k+1},\ k = i, \ldots, j-1; \tag{1}$$

$$\alpha_{j,j+1} < \alpha_{k,k+1},\ k = i, \ldots, j-1; \tag{2}$$

$$\left| C_i - \bigcup_{i_1=1}^{i-1} C_{i_1} \right| \div |C_i| \geq K_1 \tag{3}$$

$K_1 \leq 1$ is a constant.

$$\left| \bigcap_{i_1=1}^{i-1} C_{i_1} \right| \div \left| \bigcup_{i_1=1}^{j} C_{i_1} \right| \geq K_2 \tag{4}$$

$K_2 \leq 1$ is a constant.
Here, C_i is a set of noun words in the ith paragraph,

$$\alpha_{ij} = |C_i \cap C_j|$$

Conditions 1 and 2 ensure more coherence of every paragraph of the same semantic paragraph than other one; condition 3 ensures that a semantic paragraph must contain new concepts; condition 4 ensures that every paragraph in the same semantic paragraph must contain the same subject.

Text Generation

There are three types of task for a generator: text planning, sentences planning, and surface realization (Hovy, Noord, Neumann, & Bateman, 1995). Text planning selects from a knowledge pool that information is to be included in the output, and out of this will come a text structure to ensure coherence. Sentence planning organizes the content of each sentence and order of its parts. Surface realization converts sentence-sized chunks of representation into grammatically correct sentences.

After collecting information from the paragraph, sentence and clause levels using the above steps, an abstract is formed.

System Implementation and Evaluation

Structure of the Abstracting System

The architecture of the implemented system is listed in Figure 4. The system supports various kinds of keywords, or

TABLE 1. The sample questionnaire.

1. Do you think the use of the percentage extraction can help you understand the content of Chinese articles quickly?
   Disagree Partially disagree Neutral
   Partially agree Agree
2. Do you think the use of the Keyword extraction can help you find the summary of the Chinese context easily?
   Disagree Partially disagree Neutral
   Partially agree Agree
3. Do you think this system can help you search a specific topic in a large pool of documents?
   Disagree Partially disagree Neutral
   Partially agree Agree
4. Can the system be useful in helping you to prepare a summary on an article by applying the Percentage or Keyword extraction?
   Disagree Partially disagree Neutral
   Partially agree Agree
5. Do you think it is more useful to use Keyword + Percentage extraction than just Keyword or Percentage extraction?
   Disagree Partially disagree Neutral
   Partially agree Agree
6. Can the Chinese Automatic Extraction function give you an idea about Chinese document summarization?
   Disagree Partially disagree Neutral
   Partially agree Agree
7. Which option do you think is the most useful in helping you to scan all documents?
   Keywords Percentage
   Automatic Extraction
8. Which option do you think is the most useful in helping you to generate a summary of a Chinese article?
   Keywords Keyword + Percentage
   Percentage Automatic Extraction

percentage extraction, or both. It can generate keywords automatically, and perform rhetorical structural analysis of the article to produce an abstract of the desired length. Some typical examples are given in Appendix.

Performance Evaluation

Table 1 is a sample questionnaire given to people to record their response during the system implementation and testing phase. A group of people was (a sample of 35 computer users of Chinese processing systems) invited to test the prototype system and provide feedback for evaluation.

An analysis of user responses based on the questionnaire is given in Table 2. It shows that a greater proportion of users (>65%) tended to agree that the percentage requirement for extraction can help (1) generate a summary of a Chinese article, and (2) search for some specific topic. More than 60% of users thought that keyword and percentage extraction applied together could be helpful in producing the summary.

However, it was found that most people hesitated to use the system for text abstraction. They preferred using the keyword + percentage method to do the extraction. This indicates that most users would like to use the system to extract certain articles of specific interest (given some keywords), and perhaps would like to see the system being able to accomplish the abstraction with those understanding-based methods. It appears that the function, which provides automatic extraction, is in need of improvement to strengthen the quality of abstraction and its adoption.

Limitation of Our Method and Future Direction

The extraction methods provided in this article expose some problems in developing an abstracting system.

Overfiltering

If a Chinese article unexpectedly contains words such as "的 (*of*)," "他 (*he*)," "我 (*I*)," which appear with high frequency in an article, the system will count it as usual and it may not be taken as keyword, because the probability of these keywords being the most important keywords is quite low. The incorporation of understanding-based methods is expected to improve keyword extraction.

Redundancy

It is unfavorable when an article contains many duplicated sentences. The system will extract the same sentences with the same keywords many times.

Larger Percentage

If an article contains 1,000 words, but the user only wants to extract 5% of them, it may be that the most important sentences unexpectedly contain a total of 300 words. In this case, the system will extract 30% of the whole article.

Insufficient Data

In the prototype system, the sample testing is not exhaustive and comprehensive enough to reflect most linguistic phenomena.

Due to the above limitations and drawbacks, we proposed several possible improvements in this kind of system:

TABLE 2. Respondents by attitude toward the Chinese Extraction System (figures in %).

Question number	Disagree	Partially disagree	Neutral	Partially agree	Agree
*Q*1	0	0	15	25	60
*Q*2	5	15	50	30	0
*Q*3	0	5	25	50	20
*Q*4	5	5	45	35	10
*Q*5	5	15	20	30	30
*Q*6	15	30	35	20	0
*Q*7	Keyword 30	Percentage 60	Automatic extraction 10		
*Q*8	Keyword 15	Percentage 35	Keyword + Percentage 45		Automatic extraction 5

1. Natural language understanding should look more deeply for keyword identification and important sentence generation. The analysis of the text can go further in two dimensions: subparagraph, and superparagraph level.

In the first of these, the intermediate representation provided in Zhou, Shen, Yu, and Liu (1996) is a good alternative, using three-level generation technology to obtain more concise abstracts. In the second, the document structure (Sumita, Ono, & Miike, 1993), which represents logical chunks of sentences in each section and the rhetorical relationships between them, is identified to provide a basis for automatically extracting and summarizing the important parts of the document.

1. It is suggested that the automatic extraction system can incorporate a user interactive component. The system and the users have to communicate to get more information about the Chinese article. Using the database plus the inference rule, a letter and more intelligent Chinese extraction system can be achieved a popular topic in Chinese Extraction.
2. To increase the volume and expand the type of texts processed, inclusion of all Chinese words is needed to make the system more powerful and complete. Intelligent text-based systems will vary as to the degree of difficulty of the texts they deal with. Many systems will have to work on difficult texts. Usually, it is the complexity of the text that makes the system desirable in the first place (Jacobs, 1992). It is for the construction of such systems that we need to think about deeper methods of natural language understanding, those that are more robust and suitable for processing long texts without human help.

Conclusion and Discussion

In this study, we have discussed a hybrid method for Chinese extraction in great detail. Our system was developed using Windows to secure better human–computer interaction and a user-friendlier interface. Also, the Pen Power 蒙恬筆式環境 system was also used as a Chinese input tool during implementation and testing.

In application, the system will be developed along two dimensions. To provide a system for fast information searching in Chinese contexts, and to serve as a tool for people who want to extract the relevant content from a Chinese document.

We explored the integration of two types of method to address the problem of text extraction. The first was the extraction-based method, where the system extracted text according to user criteria, i.e., keywords and percentage. The other obtained more accurate results and produced a more concise abstract through understanding text structure.

Future methods need to tackle the deeper syntactic and semantic meaning of the texts, improving their performance in information processing applications.

Supplied with similar resources used for other languages, we believe that the proposed method can be extended to languages other than Chinese.

Acknowledgments

The authors appreciate the partial support of Hong Kong Polytechnic University Research Grants G-S134 and G-YB11.

References

Church, K.W., & Hanks, P. (1989). Word association norms, mutual information and lexicography. Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics (pp. 76–83). Vancouver, BC.

Cocke, B.J., Pietra, S.D., Pietra, V.D., Jelinek, F., Mercer, R., & Roosin, P. (1988). A statistical approach to language translation. Proceedings of the International Congress on Computational Linguistics (pp. 71–75). Budapest, Hungary.

Gokcay, D., & Gokcay, E. (1995). Generating titles for paragraphs using statistically extracted keywords and phrases. 1995 IEEE International Conference on System, Man and Cybernetics (pp. 3174–3179).

Hovy, E., Noord, V.G., Neumann, G., & Bateman, J. (1995). Language generation. In R.A. Cole, J. Mariani, H. Uszkoreit, A. Zaenen, V. Zue, G. Varile, & A. Zampolli (Eds), Survey of the State of the Art in Human Language Technology, 161–187. On-line.

Jacobs, P.S. (Ed.). (1992). Text-based intelligent systems: Current research and practice in information extraction and retrieval. Hillsdale, NJ: Lawrence Erlbaum Associates Press.

Kwok, K.L. (1997). Comparing representations in Chinese information retrieval, Proceedings of ACM SIGIR International Conference on Research and Development in Information Retrieval (pp. 34–41).

Lin, H.L., Tsou, K.B., Ho, H.C., Lai, T.B.Y., Lun, S.C., Choi, C.Y., & Kit, C.Y. (1991). Automatic Chinese text generation based on inference trees. Proceedings of ROCLING IV. Taipei (pp. 215–236).

Liu, J., Goss, S., & Murray, G. (1994). Similarity comparison and analysis of sequential data. IEEE Proceedings of International Conference on Expert Systems for Development (pp. 138–143). Bangkok, Thailand.

Pun, K.H., & Lum, B. (1989). Resolving ambiguities of complex noun phrases in a Chinese sentence by case grammar. Computer Processing of Chinese & Oriental Languages, 4(2–3), 185–236.

Rama, D.V., & Srinivasan, P. (1993). An investigation of content representation using text grammars. ACM Transactions on Information Systems, 11, 51–75.

Stanfill, C., & Thau, R. (1991). Information Retrieval Thau. Information Retrieval on the Connection Machine: 1-8192 gigabytes. Information Processing and Management, 27, 285–310.

Stanfill, C., & Waltz, L.D. (1991). Statistical methods, artificia intelligence, and information retrieval. In P.S. Jacobs (Ed.). Text-based intelligent systems (pp. 215–226). Hillsdale, NJ: Lawrence Erlbaum Associates Press.

Sumita, K., Ono, K., & Miike, S. (1993). Document structure extraction for interactive document retrieval systems. Proceedings of SIGDOC'93. Waterloo, Ont., Canada.

Tsou, B.K., Lin, H.L., Ho, H.C., Lai, T.B.Y., & Chan, T.Y.W. (1995). Automated Chinese full-text abstraction based on rhetorical structure analysis. Proceedings of International Conference on Computer Processing of Oriental Languages (pp. 259–266). Honolulu, HI.

Tsou, K.B., Ho, H.C., Lin, H.L., Liu, G.K.F., Lun, C.S., & Heung, A.Y.L. (1990). Automated Chinese text abstraction: A human-machine cooperative approach. Proceedings International Conference on Computer Processing of Chinese and Oriental Languages (pp. 33–39). Changsha, China.

Tsou, K.B., Lin, H.L., Ho, H.C., & Lai, B.Y. (1992). From argumentative discourse to inference trees: Using syntactic markers as cues in Chinese text abstraction. Proceedings of 3rd International Conference on Chinese Information Processing (pp. 76–93). Beijing, China.

Wang, K.Z., Li, J.J., & Wu, Y. (1995). Study of non dictionary Chinese segmentation. Proceedings of the 3rd National Conference on Chinese computing linguistics (pp. 359–359).

Watanabe, H. (1996). A method for abstracting newspaper articles by using surface clues. Proceedings of COLING'96 (pp. 974–977). Copenhagen, Denmark.

Wu, Y. Liu, J.N.K., & Wang, K.Z. (1998). A method of automatic analysis of text structure. Proceedings of the 1998 International Conference on Chinese Information Processing (pp. 304–309), Beijing, China.

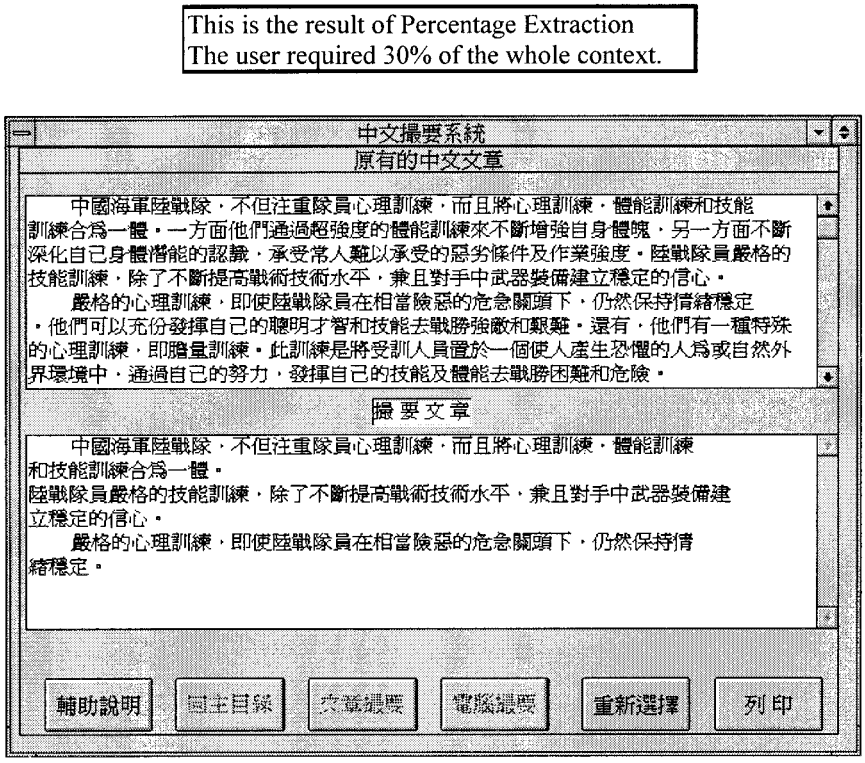
Zhou, L.N., Shen, G., Yu, S.W., & Liu, J. (1996). Uniword-oriented natural language generation technology. Proceedings of ICCC'96 (pp. 257–261). Singapore.

Appendix

Sample Chinese Article (in English)

Chinese marine troops not only emphasize the training of their minds, but also integrates that with physical and skill training. On the one hand, they have to go through a very strict training to build up good physical strength. On the other, they have to learn and understand their skill, and are able to adapt to work under harsh and difficult conditions hardly tolerable by normal people. The strict skill training not only increases their skill level in battle, but also builds up confidence with the acquired weapons.

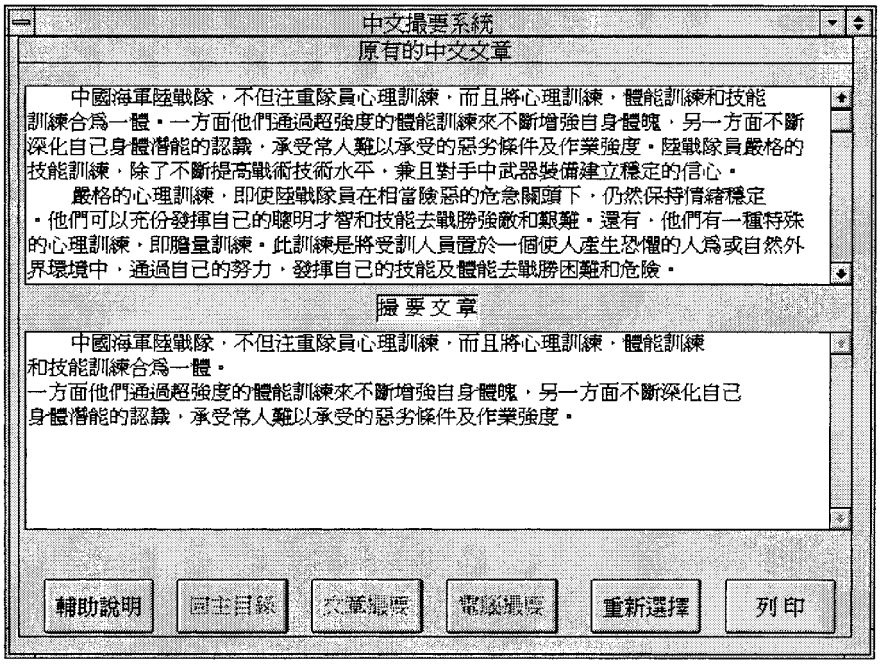
With intensive mind training, even when the troops are in quite dangerous conditions, they can still maintain stable behavior. They can overcome hardship, and beat their enemies using their intelligence and skill. They also have to undergo very special psychological training. This is the bravery training. Under horrifying conditions, is it natural or artificial, the trainee can make use of his skill and physical strength to overcome those difficulties and dangers.



Sample abstract (in English)

Chinese marine troops not only emphasize the training of their minds, but also integrate that with physical and skill training. The strict skill training not only increases their skill level in battle, but also builds up confidence with the acquired weapons. With intensive mind training, even the troops are in quite dangerous conditions; they can still maintain a stable behavior.

This is the result of using Keyword+Percentage method.
The user required 30% of the whole context and
keywords 訓練 #　技能 (training #　　skill).



Sample abstract (in English)

Chinese marine troop not only emphasizes the *training* of their minds, but also integrates that with physical and *skill* training.

On the one hand, they have to go through a very strict training to build up good physical strength. On the other, they have to learn and understand their skill, and are able to adapt to work under harsh and difficult conditions hardly tolerable by normal people.