

Accessor Variety Criteria for Chinese Word Extraction

Haodi Feng*
Shandong University
City University of Hong Kong

Kang Chen†
Tsinghua University

Xiaotie Deng‡
City University of Hong Kong

Weimin Zheng†
Tsinghua University

We are interested in the problem of word extraction from Chinese text collections. We define a word to be a meaningful string composed of several Chinese characters. For example, 百分之, 'percent', and 越来越, 'more and more', are not recognized as traditional Chinese words from the viewpoint of some people. However, in our work, they are words because they are very widely used and have specific meanings. We start with the viewpoint that a word is a distinguished linguistic entity that can be used in many different language environments. We consider the characters that are directly before a string (predecessors) and the characters that are directly after a string (successors) as important factors for determining the independence of the string. We call such characters accessors of the string, consider the number of distinct predecessors and successors of a string in a large corpus (TREC 5 and TREC 6 documents), and use them as the measurement of the context independency of a string from the rest of the sentences in the document. Our experiments confirm our hypothesis and show that this simple rule gives quite good results for Chinese word extraction and is comparable to, and for long words outperforms, other iterative methods.

1. Introduction

Words are the basic linguistic units of natural language processing. The importance of word extraction is stressed in many papers. According to Huang, Chen, and Tsou (1996), the word is the basic unit in natural language processing (NLP), as it is at the lexical level where all modules interface. Possible modules involved are the lexicon, speech recognition, syntactic parsing, speech synthesis, semantic interpretation, and so on. Thus, the identification of lexical words and/or the delimitation of words in running texts is a prerequisite of NLP. Teahan et al. (2000) state that interpreting a text as a sequence of words is beneficial for some information retrieval and storage tasks: for example, full-text searches, word-based compression, and key-phrase extraction. According to Guo (1997), words and tokens are the primary building blocks in almost all linguistic theories and language-processing systems, including Japanese (Kobayasi, Tokumaga, and Tanaka 1994), Korean (Yun, Lee, and Rim 1995), German (Pachunke et al. 1992), and English (Garside, Leech, and Sampson 1987), in various media, such

* School of Computer Science and Technology, Jinan, PRC; Department of Computer Science, Tat Chee Avenue, Kowloon, Hong Kong. E-mail: fenghd@cs.cityu.edu.hk or fenghaodi@hotmail.com.

† Department of Computer Science and Technology, Peking, PR China. E-mail: {ck99,zwm-dcs}@mails.tsinghua.edu.cn.

‡ Department of Computer Science, Tat Chee Avenue, Kowloon, Hong Kong. E-mail: csdeng@cityu.edu.hk.

as continuous speech and cursive handwriting, and in numerous applications, such as translation, recognition, indexing, and proofreading. The identification of words in natural language is nontrivial since, as observed by Chao (1968), linguistic words often represent a different set than do sociological words.

Chinese texts are character based, not word based. Each Chinese character stands for one phonological syllable and in most cases represents a morpheme. This presents a problem, as only less than 10% of the word types (and less than 50% of the tokens in a text) in Chinese are composed of a single character (Chen et al. 1993). However, Chinese texts, and texts in some other Oriental languages such as Japanese, do not have delimiters such as spaces to mark the boundaries of meaningful words. Even for English text, some phrases consist of several words. However, the problem in English is not as dominant a factor as in Chinese. How to extract words from Chinese texts is still an interesting problem. Note that word extraction is different from the very closely related problem of sentence segmentation. Word extraction aims to collect all of the meaningful strings in a text. Sentence segmentation partitions a sentence into several consecutive meaningful segments. Word extraction should be easier than sentence segmentation, and the problems involved in it can be solved using simpler methods.

Some Chinese information-retrieval systems operate at the character level instead of the word level, for example, the Csmart system (Chien 1995). However, to further improve the efficiency of natural Chinese processing, it is commonly thought to be important to apply studies from linguistics (Kwok 1997). Lexicon construction is considered to be one of the most important tasks. Single Chinese characters can quite often carry different meanings. This ambiguity can be resolved when the characters are combined with other characters to form a word. Chinese words can be unigrams, bigrams, trigrams, or n -grams, where $n > 3$. According to the *Frequency Dictionary of Modern Chinese* (Beijing Language Institute 1986), among the 9,000 most frequent Chinese words, 26.7% are unigrams, 69.8% are bigrams, 2.7% are trigrams, 0.007% are four-grams, and 0.002% are five-grams. There are lexicons for identifying some (and probably most of the frequent) words. However, sometimes less-frequent words are more effective. Weeber, Vos, and Baayen (2000) recently extracted side-effect-related terms in a medical-information extraction system and found that many of the terms had a frequency of less than five. This indicates that low-frequency words may also carry very important information. Our experiments show that we can extract low-frequency words using a simple method without overly degrading the precision.

There are generally two directions in which words can be formed (Huang, Chen, and Tsou 1996). One is the deductive strategy, whereby words are identified through the segmentation of running texts. The other is the inductive strategy, which identifies words through the compositional process of morpho-lexical rules. This strategy represents words with common characteristics (e.g., numeric compounds) by rules. In Chinese text segmentation there are three basic approaches (Sproat et al. 1996): pure heuristic, pure statistical, and a hybrid of the two. The heuristic approach identifies words by applying prior knowledge or morpho-lexical rules governing the derivation of new words. The statistical approach identifies words based on the distribution of their components in a large corpus. Sproat and Shih (1990) develop a purely statistical method that utilizes the mutual information between two characters: $I(x, y) = \log \frac{p(x, y)}{p(x)p(y)}$; the limitation of the method is that it can deal only with words of length two characters. Ge, Pratt, and Smyth (1999) introduce a simple probabilistic model based on the occurrence probability of the words that constitute a set of predefined assumptions. Chien (1997) develops a PAT-tree-based method that extracts significant words by observing mutual information of two overlapped patterns with the significance function

$SE_c = \frac{Pr(c)}{Pr(a)+Pr(b)-Pr(c)}$, where a and b are the two biggest substrings of string c . Zhang, Gao, and Zhou (2000) propose the application of a statistical method that is based on context dependence and mutual information. Yamamoto and Church (2001) experiment with both mutual information and residual inverse document frequency (RIDF)¹ as criteria for deciding Japanese words, and their main contribution is in affording a reduced method for computing term and document frequency. In almost all of the work cited to this point, the dimension that is used to compute mutual information is term frequency. Chen and Bai (1998) propose a corpus-based learning approach that learns grammatical rules and automatically evaluates them. Chang and Su (1997) use an unsupervised Viterbi training process to select potential unknown words and iteratively truncate unlikely unknown words in the augmented dictionary. Teahan et al. (2000) propose a compression-based algorithm for Chinese text segmentation. Paola and Stevenson (2001) demonstrate an effective combination of deeper linguistic knowledge with the robustness and scalability of a statistical technique to derive knowledge about thematic relations for verb classification. Mo et al. (1996) deal with the identification of the determinative-measure compounds in parsing Mandarin Chinese by developing grammatical rules to combine determinators and measures.

We introduce another concept, accessor variety (AV) (for a detailed definition, refer to subsection 3.1), to describe the extent to which a string is likely to be a meaningful word. Actually, Harris (1970) uses similar criteria to determine English morpheme boundaries, and our work is partially motivated by his success. We first discard those strings with accessor varieties that are smaller than a certain number (called the threshold; see subsequent discussion). The remaining strings are considered to be potentially meaningful words. In addition, we apply rules to remove strings that consist of a word and **adhesive characters** (clarified in subsection 3.2). Our experiment shows that even for small thresholds, quite good results can be obtained.

In Section 2, we introduce examples of unknown words, the identification of which is the task of our work. In Section 3, we discuss our method. In Section 4, we present our experimental results. We conclude our work with a discussion and a comparison to previous results in Section 5. In Section 6, we list some future work that can be pursued following the concept of AV. We note that although our method is quite simple, it is marginally better than previous comparable results. This method distinguishes itself from statistically based approaches and grammatical rules. Because of its simplicity, it can be used easily in computer-based applications. Moreover, innovative variations of our method and its combination with statistical methods and grammatical methods are worthy of further exploration.

2. Unknown Words

As defined by Chen and Bai (1998), unknown words are words that are not listed in an ordinary dictionary, and word extraction seeks to identify such words. To give readers an intuitive view of these words, we list the types of unknown words that most frequently appear (Chen and Bai [1998] list 14 different types). What we should point out here is that except for numeric-type compounds, which are extracted separately, we extract all the other types of words together.

1. **Proper names.** These include acronyms, Chinese names, and those words that have been borrowed from other languages: for example, 中银, ‘Bank of China’; 冯好娣,

¹ $RIDF = \text{observed } IDF - \text{predicted } IDF = -\log \frac{df}{D} + \log(1 - e^{-\frac{tf}{D}})$, where tf , df , and D are term frequency, document frequency, and number of documents, respectively.

'Feng Haodi' (Chinese girl's name); 爱德华王子, 'Prince Edward'; 微软, 'Microsoft'; and 大不列颠及北爱尔兰联合王国, 'the United Kingdom of Britain and Northern Ireland'. To recognize proper names is the first task for Chinese word extraction, because their meanings cannot be obtained through the combination of smaller words, as in the compound words that are described next. Therefore, a reasonable way to approach them is to deduce them from Chinese text collections.

2. **Compound words.** These are strings with specified meanings that are composed of shorter meaningful words: for example, 中国工商银行, 'Industry and Commerce Bank of China', is composed of 中国, 'China', 工商, 'industry and commerce', and 银行, 'bank'; and 外商投资企业, 'foreign businessmen invested company', is composed of 外商, 'foreign businessmen', 投资, 'invest', and 企业, 'company'. Compound words account for a large proportion of Chinese words because it is very easy to compose a new compound word out of smaller known words. There are about 5,000 commonly used Chinese characters, but the number of compound Chinese words is unpredictable. We want to extract those compounds that are accepted as *words* by most people.

3. **Derived words.** These are words that have affix morphemes: for example, 现代化, 'modernization', and 电脑化, 'computerization', both of which contain affix morpheme 化.

4. **Numeric-type compounds.** Some examples of numeric-type compounds would be 1999 年, '1999'; 第一届, 'the first session'; 两千年, 'year 2000'; and 十一条街, '11 streets'. Although these words have specific meanings and are used frequently, most dictionaries do not contain them. It is not very difficult to identify them, since there are morphological rules (Mo et al. 1996) for generating these words. Such numeric-type compounds contain numbers as the main components, and measure characters or words are used nearby.

3. Proposed Approach

One of the important parameters that is employed in statistical methods for automatic Chinese word extraction is word or character frequency. Equivalent frequencies, such as document frequency and term frequency, are used analogously. Algorithms that are based on these frequencies are used to measure how likely it is that a particular string of characters is a meaningful word, according to the belief that "when a string is repeated many times, it must carry a meaning." However, in this article, we use not frequency, but accessor variety. This can be explained as "when a string appears under different linguistic environments, it may carry a meaning." We introduce the concept *accessor variety* as a new criterion for identifying meaningful Chinese words.

3.1 Accessor Variety

In Chinese text, each substring of a whole sentence can potentially form a word, but only some substrings carry clear meanings and thus form a correct word. For example, the sentence 门把手弄坏了 has 21 substrings, but only four substrings, 门把, 把手, 弄坏, and 门把手, can be considered words (we do not consider single-character words here). In some implementations, the segmentation method is used to extract those words (recent reviews on Chinese word segmentation include Wang, Su, and Mo [1990] and Wu and Tseng [1993]). There are several commonly used segmentation methods such as **forward maximum matching** and **backward maximum matching** (Teahan et al. 2000; Dai, Loh, and Khoo 1999; Sproat et al. 1996). If the dictionary includes the words 门把手, 把手, and 弄坏, then forward maximum matching will extract two words, 门把手 and 弄坏, after segmenting the sentence. If 门把手 is deleted

from the dictionary, then the sentence will be segmented into 门, 把手, 弄坏, and 了, and two words, 把手 and 弄坏, are obtained. Furthermore, if 把手 is removed from the dictionary, then another, different segmentation pattern will be achieved. Therefore, the dictionary is an important factor in these methods. In fact, this sentence has ambiguities ('The door handle is broken' or 'the door hurts the hand'), and the segmentation methods try to find a reasonable way to solve this problem. We do not segment the sentence but extract those substrings that might possibly form words. The accessor variety criterion is used to decide whether a substring should be retained or discarded. Sentences (1)–(4) can be used to illustrate the meaning of accessor variety:

- (1) 门把手弄坏了, 'The door hurts the hand' or 'The door handle is broken'.
- (2) 小明修好了门把手, 'Xiao Ming fixed the door handle'.
- (3) 这个门把手很漂亮, 'This door handle is very beautiful'.
- (4) 这个门把手坏了, 'This door handle is broken'.

Consider how to extract the word 门把手 from these four sentences. In fact, the three-character string 门把手 has *three* distinct prefixes, "S", 了, 个 ("S" denotes the start of a sentence), and *four* distinct suffixes, 弄, "E", 很, 坏 ("E" denotes the termination of a sentence). This means that the string can be used in at least three different environments and might carry meanings that are independent of those of the other characters in these four sentences. In this case $three = \min\{three, four\}$ is called the **accessor variety** of string 门把手.

We use the criterion accessor variety to evaluate how independently a string is used, and thus how likely it is that the string can be a word. The accessor variety of a string s of more than one character is defined as

$$AV(s) = \min\{L_{av}(s), R_{av}(s)\}.$$

Here $L_{av}(s)$ is called the **left accessor variety** and is defined as the number of distinct characters (predecessors) except "S" that precede s plus the number of distinct sentences of which s appears at the beginning. Similarly, the **right accessor variety** $R_{av}(s)$ is defined as the number of distinct characters (successors) except "E" that succeed s plus the number of distinct sentences in which s appears at the end. In other words, characters "S" and "E" are repeatedly counted. The reason for doing this is that some words usually appear at the beginning or the end of sentences. For example, 突然, 'suddenly', is often used separately as a short sentence. Therefore, "S" and "E" will be counted multiple times, and we regard 突然 as a meaningful word, although there are probably very rarely other characters preceding or succeeding it.

The extracted words should ensure an AV value of no less than a predefined threshold, which means that such strings should appear in enough different environments and therefore be considered meaningful. Our experiments show that even with a small threshold, the result is quite precise.

3.2 Adhesive Characters

There are some characters, such as auxiliary characters 的 (a mark following an adjective) and 地 (a mark following an adverb), that often adhere to other words as heads or tails to compose a string with a high AV value that is not an actual linguistic word; we call these characters **adhesive characters**. For example, 的人们, 'of people', has a very high AV value because many adjectives (and hence many predecessors) precede 的人们, 'people' (e.g., 这里的人们, 'here people', and 勤劳的人们, 'diligent people'). Moreover, many words (and hence many successors) can succeed it to describe the behavior of the people (e.g., 这里的人们以从事商业谋生, 'people here make a living

out of commerce', and 勤劳的人们在工作, 'diligent people are working'). It seems that 的 combines with 人们 very firmly, but 的人们 cannot be accepted as a word by most people. There are also some nonauxiliary characters that very frequently adhere to other, shorter words. In our method, we ignore the difference between the auxiliary and nonauxiliary adhesive characters and extract them under the same criteria. Recalling the discussion about the AV value, we divide the adhesive characters into two groups. The **head-adhesive characters** often stick at the heads of other words and have high R_{av} values, and the **tail-adhesive characters** often stick at the tails of other words and have high L_{av} values. How adhesive characters are found will be discussed in the article.

The adhesive characters should be stripped from the string for constructing a well-formed word. According to the places in which these characters appear, three cases are considered. If the leftmost consecutive characters of a string are all head-adhesive characters, then we say that it is in the $h+core$ style. If the rightmost consecutive characters of a string are all tail-adhesive characters, then we say that it is in the $core+t$ style. A string that is in both $h+core$ and $core+t$ styles is said to be in the $h+core+t$ style, where the *core* is the inner part of the string found by removing the left consecutive head-adhesive characters and right consecutive tail-adhesive characters. For example, 的我, 'of I', is in the $h+core$ style and 我, 'I', is the core, 我的, 'my', is in the $core+t$ style and 我, 'I', is the core; and 的过程是, 'of procedure is', is in the $h+core+t$ style and 过程, 'procedure', is the core. In other words, none of the strings matching these three cases should be considered words, that is, they should be discarded.

With the help of adhesive characters, we can introduce the *ADHESIVE-JUDGE* rules to discard all those strings that have a high AV score but are unlikely to be real words:

1. A string that is composed of two characters in any of the $h+core$, $core+t$, and $h+core+t$ (no core in the case of two characters) styles should be discarded if it does not appear in a specified electronic dictionary. For example, strings such as 的我, 'of I', and 一的, 'one of', will be discarded, whereas strings such as 的确, 'surely', 了解, 'comprehend', and 珍妮, 'Jane' (a girl's name), will remain. Under this rule, most meaningful two-character strings that are unknown to the dictionary will be recognized as meaningful words because they rarely contain adhesive characters.
2. A string that is made up of more than two characters in any of the three styles ($h+core$, $core+t$, and $h+core+t$) should be discarded if the core is a meaningful multi-character word.
3. The most frequently used auxiliary words, such as 的, 'of', 有, 'have', 了 (a mark indicating completion), and 在, 'at', must be used to delimit the original string. If any token is found to be an identified multicharacter word (a word in the specified dictionary or extracted by this algorithm before processing the string under consideration), then the original string is abandoned.

All of the strings will be kept as meaningful words if they survive elimination according to these rules. According to these rules, strings such as 的过程是, 'of procedure is', 一的, 'one of', and 的我, 'of I', should be abandoned, whereas 的确的确, 'actually', and 实事求是, 'seek truth from facts', should remain even though they all contain auxiliary words.

3.3 Numeric-Type Compounds

We define **numeric-type compounds** to be strings of numbers or strings that contain substrings of numbers followed by measures. For example, 千千万万, ‘thousands upon thousands’, 第一届奥运会, ‘the first Olympic games’, and 五十公斤, ‘50 kilograms’, will be considered as potential numeric-type compounds, whereas 一心一意, ‘wholehearted’, will not be, because 心 and 意 are not measures. Numbers include Arabic numbers of SBC case and DBC case and Chinese numbers in both simplified form and traditional form. Special words such as 几, ‘several’, 约, ‘about’, and 左右, ‘or so’, are treated as numbers too. Measures include both Chinese measures and foreign measures (e.g., 亩, ‘mu’, 尺, ‘chi’, 盎司, ‘ounce’, and 加仑, ‘gallon’). Because of the special nature of words of this type, some lexicons do not include them, which is why we extract them separately. In our method, a numeric-type compound must be first a maximal numeric-type string, which means that the string cannot be preceded or succeeded by other numbers or measures in the sentence under consideration. For example, when processing the sentence 一九七七年十月二日是他的生日, ‘October 2nd, 1977, is his birthday’, strings 一, ‘one’, 一九, ‘nineteen’, 一九七七年, ‘1977’, 十月, ‘October’, and 二日, ‘second day’, are not extracted. The only numeric-type compound that is extracted from this sentence is 一九七七年十月二日, ‘October 2nd, 1977’. The numeric-type compound candidates are then further examined according to the *ADHESIVE_JUDGE* rules, and the survivors of that examination are eventually accepted as numeric-type compounds. Notice that for strings of only numbers or strings of numbers followed by measures, we set the threshold to one.

As we process numeric-type compounds separately, we ignore strings that contain numeric-type compounds when we extract the ordinary words.

4. Experimental Results

4.1 Setup of the Experimental Environment

The corpus-based word extraction method described in the previous section was tested on a 153 MB corpus consisting of *People’s Daily* news and Xinhua news from TREC5 and TREC6 (Linguistic Data Consortium, n.d.). We also conducted experiments on a small corpus that has approximately 1.7 MB of data and is a part of the former corpus. Neither corpus was annotated. The system dictionary that we used in each experiment was downloaded from <http://www.mandarintools.com/segmenter.html> and contains 119,538 terms from two to seven characters long. In our method, a preprocessing step was performed on the corpus in which we eliminated all of the non-Chinese symbols. Each uninterrupted Chinese character sequence was kept as one line in the transformed data. For each line in the data file, all possible substrings were extracted, along with their predecessors and successors. Those predecessors and successors were finally merged, and the AV , L_{av} , and R_{av} values were calculated. Different thresholds were used for discarding those strings with low AV values and checking how the threshold affects the results. Moreover, the *ADHESIVE_JUDGE* rules were used for the further discarding of those strings that seemed unlikely to be words.

A list of adhesive characters is needed when we apply the *ADHESIVE_JUDGE* rules. We constructed the adhesive character list based on the accessor variety information of single characters. Characters with high L_{av} values were considered to be tail-adhesive characters. Characters with high R_{av} values were considered to be head-adhesive characters. Characters with very high AV values were considered to be the **delimiters** that are used in rule (3) of the *ADHESIVE_JUDGE* rules. In the end, we placed 68 tail-adhesive characters, 66 head-adhesive characters, and 16 delimiters on our list.

In our experiments, we performed only one step of each of the *ADHESIVE_JUDGE* rules (in either direction) for discarding meaningless multicharacter strings. That is, in any of the three styles (*h+core*, *core+t*, or *h+core+t*), only the leftmost or rightmost character was considered among all of the head- or tail-adhesive characters. If the first character of a string was a head-adhesive character and the remaining substring (after stripping the first character) was found in the system dictionary or the preextracted shorter word lists (and thus a core was found), such a string was considered to be in the *h+core* form and thrown away. The same judgment process was used in the *core+t* and *h+core+t* styles. In other words, only the first or last character, or both, of a string were used in rule (2) of the *ADHESIVE_JUDGE* rules. Such simplification does not hurt the results too much.

The *AV* value threshold is another important factor in this method. We tested different thresholds to evaluate how they influenced the performance. One might imagine that a higher threshold will result in higher precision while causing the loss of some recall. This phenomenon was certainly observed in our experiments. Word length has a relationship with the threshold: that is, longer words required a smaller threshold to reach the same precision, or higher precision could be obtained on longer words with the same threshold, because longer words have more specific usage and appear in fewer environments.

Our first experiment was carried out on the small corpus of Xinhua news. Strings with lengths varying from two to ten characters were examined. In the following, we tested our method on the large corpus and all strings with lengths from two to seven characters. In the end, we extracted the numeric-type compounds from each corpus.

In addition, there is no commonly accepted standard for evaluating the performance of word extraction methods, and it is very hard to decide whether a word is meaningful or not (Sproat et al. 1996). We define **precision** as the number of extracted words that would be meaningful in a Chinese native speaker's opinion, divided by the total number of extracted compounds. As it is very hard to find all of the words in the original corpus that would be found meaningful by a Chinese person, it is very hard to count recall in the traditional way, that is, the number of meaningful words extracted divided by the number of all meaningful words in the original data. On the other hand, it is also impossible to approach traditional precision and traditional recall by comparing the hand-segmented sample sentences and the automatically segmented sentences, as people usually do, because our method does not touch upon segmentation. The reason that we do not consider segmentation is that we aim only to investigate the performance of *AV* itself, whereas the involvement of a segmentation module would inevitably influence our judgment on the performance of *AV*. Therefore, we substitute **partial recall** for traditional recall. We define *partial recall* as the number of extracted meaningful words (from the whole corpus) that appear in a sample corpus divided by the total number of meaningful words in the sample corpus. Evidently, the partial recall value will be no smaller, and usually greater, than the recall value calculated in the traditional way. This point will be clearly reflected by the following experimental results. What should be pointed out here is that some people use the **F-measure** as an evaluation metric (Ricardo and Berthier 1999; Chang and Su 1997). However, this is difficult to interpret according to Beeferman, Berger, and Lafferty (1999). In our opinion, as the F-measure or precision-recall curves are based on two parameters, recall and precision, it is enough for us only to list the partial recall and precision.

4.2 Experiments on the Small Corpus

As noted previously, the small corpus contained approximately 1.7 MB data of Xinhua news. We processed all of the strings in the corpus with lengths from one to ten

Table 1
Some of the words extracted from the small corpus.

粤港经济	Economy of GuangDong and Hong Kong
浦东新区	new region of PuDong
西哈努克	Sihanouk (name)
意大利队	Italian Team
自然保护区	nature protection region
巴解组织执委会	Administration Committee of PLO
联合国教科文组织	UNESCO
海峡两岸关系协会	Association of Relations Across the Taiwan Straits (ARATS)
光大国际信托投资公司	GuangDa International Trust Investment Company
仪征化纤工业联合公司	YiZheng Chemical Fibre United Company
中国海洋石油总公司	Parent Ocean Petroleum Company of China
小浪底水利枢纽工程	XiaoLangDi Irrigation Hinge Project
法国网球公开赛	French Open Tennis
香港特别行政区	Hong Kong Special Administration Region
联合国安理会	United Nations Security Council
亚洲开发银行	Asia Development Bank
经济体制改革	Innovation of the Economy System
最惠国待遇	most-favoured-nation clause
克里斯托弗	Christopher (name)
预委会	Preparing Committee
曼德拉	Mandela (name)
尤伯杯	UBA Championship Cup

characters. Table 1 shows some of the extracted correct words that were not contained in the system dictionary.

We can see that almost all of these words are compound words, proper names, or derived words. It would be almost impossible to list all of them in a general-purpose dictionary. Furthermore, some of them occur only a few times. For example, **亚洲开发银行** only occurs three times in this corpus. The method we used has the ability to extract low-frequency words.

Table 2 shows the overall precision performance when the word length is not specified. We set the threshold from two to nine and observed that with a larger threshold we could obtain more precise results. As the number of words extracted was very large (approximately 30,000 words), we randomly chose a portion (often approximately 1,000 words) of the total set of extracted words as the test set to calculate the precision; that is, we listed all of the extracted words, and then for each word, we generated a random number between zero and one. If the number was smaller than the number of test words divided by the number of all extracted words (here 1,000/30,000), then the corresponding word was chosen. Human judgment was then used to check whether an extracted word was a correct or spurious word.

In the evaluation phase we found that the method performed differently on strings of different lengths. Hence, we also checked the precision performance with specified word lengths. We set the threshold to three and obtained the data in Table 3. Again we used the sampling method just described to test the overall precision.

From Table 3 we can see that the method worked almost equally well on all word lengths except length three. After checking the results, we found that three-character strings are often constructed from a two-character legitimate word together with a single character. It is difficult to judge with such a simple method whether such three-character strings are legitimate words.

Beyond precision, another concern is partial recall. In other words, how many words will be missed using such a method? The corpus contained 55,788 sentences.

Table 2
Experiments on the threshold–precision relationship of the small corpus.

Threshold	Precision	Number of words
2	64.4%	37,093
3	83.8%	14,468
4	89.6%	8,648
5	94.1%	6,147
6	96.8%	4,757
7	97.4%	3,800
8	97.3%	3,162
9	97.7%	2,734

Table 3
Experiments on the word length–precision relationship of the small corpus.

Word length	Precision	Number of words
2	90.2%	6,962
3	56.6%	2,532
4	91.4%	3,417
5	85.1%	712
6	90.4%	493
7	89.4%	180
8	90.1%	111
9	80.3%	61

Table 4
Experiments on the threshold–partial recall relationship of the small corpus.

Threshold	Partial recall	Number of words
2	76.7%	37,093
3	66.5%	14,468
4	59.0%	8,648
5	54.3%	6,147
6	50.3%	4,757
7	47.1%	3,800
8	44.0%	3,162
9	41.5%	2,734

We checked only a small portion (a random sample of approximately 2,000 sentences) of the total corpus. We used this sample to find meaningful words by hand. The result of automatic extraction from the whole corpus was then compared with that of hand extraction of the sample sentences. The partial recall was computed as the number of words in both sets divided by the number of words in the human extraction set. We list the experimental partial-recall values in Table 4.

We analyzed the instance with the threshold of two. Some of the words were missed because they occurred only once, which was less than the threshold. Some of the words were missed because they occurred only in very restricted environments. This means that although they appeared more than once in the corpus, their accessor variety value was only one. In the latter case, we could extract the strings that contained such strings as substrings. The details are discussed in the section on error analysis.

Table 5
Some words extracted from the large corpus.

荒路	desolate road
黄衍平	Huang Yanping (Chinese name)
鹅毛扇	goose feather fan
鼻通灵	Bi Tong Ling (name of a Chinese medicine)
鸿雁传情	send love by swan goose
黄金热土	beloved hometown
麻姑献寿	MaGu offers birthday present
经营自主权	right of independent management
北京博物馆	the Peking Museum
华东工学院	the Technology Institute of East China
假冒伪劣商品	fake and bad merchandise
北京书法庙会	the Peking penmanship temple fair
星期日图片报	Sunday photo newspaper
社会主义现代化	socialistic modernization

4.3 Experiments on the Large Corpus

The corpus that was used for these experiments was the TREC Chinese corpus (Linguistic Data Consortium, n.d.), which contains 160,000 articles, including articles that were published in the *People's Daily* from 1991 to 1993 and a portion of news released by the Xinhua News Agency in 1994 and 1995. In the experiment, we extracted words with lengths of two to seven characters. The data contained approximately 7,000,000 sentences. We first eliminated the non-Chinese characters. All of the experiments that were carried out on the small corpus were also conducted on the large corpus. In Table 5 we first show some legitimate words that were extracted from the large corpus. Notice that these words cannot be found in the word list that was extracted from the small corpus or in the system dictionary.

In Table 6, we show the overall precision performance. The performance trends that were observed in Table 2 can be also observed here. However, as this corpus is much larger than the previous one, many characters have the chance to occur together to form spurious words. That is why the precision is much lower than that for the small corpus. Nevertheless, as the corpus is much larger now, a correct word can occur in many more environments than in the small corpus, which suggests that we can improve the precision by using a large threshold for the accessor variety value without overly degrading the partial recall. For example, when the threshold is set to nine, the precision is as high as 73.4% and the partial recall remains as high as 80.4%.

Table 6
Experiments on the threshold–precision relationship of the large corpus.

Threshold	Precision	Number of words
2	51.2%	2,854,700
3	58.3%	1,269,378
4	69.0%	788,964
5	70.3%	562,407
6	70.4%	432,830
7	73.8%	349,511
8	74.2%	291,688
9	73.4%	249,904

Table 7
Experiments on the word length–precision relationship of the large corpus with threshold three.

Word length	Precision	Partial recall	Number of words
2	37.8%	92.3%	266,027
3	22.9%	83.5%	335,557
4	68.9%	80.9%	360,413
5	67.0%	83.3%	141,153
6	76.0%	81.6%	123,392
7	70.7%	64.3%	42,836

Table 8
Experiments on the word length–precision relationship of the large corpus with threshold nine.

Word length	Precision	Partial recall	Number of words
2	71.7%	90.0%	77,200
3	52.7%	73.0%	55,015
4	74.6%	70.2%	78,868
5	75.0%	63.9%	18,775
6	86.9%	63.2%	15,663
7	89.4%	42.9%	4,383

The precision and partial-recall performance in respect to the word length was also tested on the large corpus. The same sample method was used, and the results for thresholds three and nine are shown in Tables 7 and 8, respectively.

Note that there is a great jump in the precision for word lengths two and three after we change the threshold from three to nine, but the partial recall does not change much. For longer words, the method even performs well with threshold three.

The next experiment was intended to test the partial-recall performance for all of the words with lengths from two to seven. The result is shown in Table 9, which indicates that the partial-recall value is satisfactory even with a large threshold. This means that we can extract most of the words in the corpus.

4.4 Experiments on Numeric-Type Compounds

In this section, we consider numeric-type compounds. Some of the compounds of this type that were extracted from the large corpus are listed in Table 10.

Table 9
Experiments on the threshold–partial recall relationship of the large corpus.

Threshold	Partial recall	Number of words
2	89.2%	2,854,700
3	87.2%	1,269,378
4	85.6%	788,964
5	84.2%	562,407
6	83.0%	432,830
7	82.0%	349,511
8	81.2%	291,688
9	80.4%	249,904

Table 10
Numeric-type compounds extracted.

3 月 2 日	March 2nd
第一次	first time
一九九二年五月四日	May the Fourth, 1992
海峡两岸	two sides of the Strait
两国关系	relationship between two countries
三十公斤左右	thirty Kilograms or so
100 港元	one hundred Hong Kong dollars
200 盎司	two hundred ounces
四万亩	forty thousand mu

4.5 Error Analysis

Two kinds of errors occurred: the extraction of meaningless strings as meaningful words and the neglect of meaningful words. Some errors of the two types are listed below.

4.5.1 Meaningless Strings Extracted. A number of meaningless strings were extracted: for example, 解决波黑, ‘solve the Republic of Bosnia and Hercegovina’; 会议今天, ‘meeting today’; 采用国际, ‘employ international’; 锦标赛第, ‘title match order’; 目前中国, ‘today China’; 有关部, ‘related part’; 赛今天, ‘game today’; 国际经, ‘international pass’; 市人民, ‘city people’; 将于明, ‘will next’; 界人, ‘field people’; 成国, ‘become country’; 省第, ‘province order’; 地指, ‘point to’; 市首, ‘city first’; and 人注, ‘people attention’.

Most of these errors occurred because the strings are made up of one shorter meaningful word and one character that has a high accessor variety value but is absent from the adhesive-character list. For example, 市人民 is composed of 市 (accessor variety value 133 in the large corpus) and 人民, but 市 is not on the adhesive-character list. Therefore, 市人民 was extracted as a word. However, if we list too many characters as adhesive characters, the partial recall will be degraded. To give another example, 中国银行发, ‘bank of China deliver’, was extracted as a meaningful word even though its meaning is very unclear. In the string 中国银行发, we considered 中 and 发 to be adhesive characters and regarded 中国银行发 as being in the *h+core+t* style. However, the core 国银行 is not in the system dictionary or the shorter word list that we extracted previously. Hence, it passed the *ADHESIVE_JUDGE* rules and was retained as a word. It is hard to discard strings such as 采用国际 and 解决波黑, even though their meanings are not at all clear.

4.5.2 Meaningful Words Missed. A number of meaningful words were missed: for example, 抹掉, ‘clear’; 气态, ‘gaseous state’; 游兴未尽, ‘sight-seeing interest is not fulfilled’; 荒凉, ‘barren’; 弘扬, ‘carry forward’; 海基会, ‘Straits Exchange Foundation’; 近日, ‘recently’; and 非国大, ‘African National Congress’.

The main reason for these errors is that the strings occur only once in the corpus or their accessor varieties are smaller than the threshold. One way to solve this problem is to use a larger corpus to improve the partial recall. Another reason for these errors is that the string is composed of a shorter word plus an adhesive character, in which case it was discarded according to the *ADHESIVE_JUDGE* rules. For example, 非国大 is composed of 非国 and 大, where 非国 is a word in the system dictionary and 大 is an adhesive character. To solve this problem, we can use fewer adhesive characters at the cost of some precision. To give another example, 长江三角洲, ‘Chang Jiang triangle

region', is a meaningful word that appeared in the corpus but was not extracted, because it contains a substring 三角 that can be interpreted as a numeric compound 'three jiao' (which means 0.3 Chinese RMB), and therefore we discarded it. However, we can extract this string as a numeric-type compound.

5. Conclusion

We have described a hybrid method for extracting Chinese words from the Chinese text corpus using accessor variety and adhesive characters. We tested the method on the performance of different thresholds and word lengths and different corpus sizes.

We conclude that the method based on accessor variety and adhesive characters performs efficiently in fulfilling word extraction tasks. The precision with the small corpus we used was much larger than that with the large corpus, but the situation was opposite for partial recall. For example, when the threshold was set to three, the precision and partial recall with the small corpus were 83.8% and 66.5%, respectively, whereas with the large corpus they were 58.3% and 87.2%, respectively. When the threshold was set to nine, the corresponding numbers were 97.7% and 41.5% versus 73.4% and 80.4%. As even human judges differ when facing the task of segmenting a text into words and test corpora differ from system to system (Sproat et al. 1996), it is very difficult to compare two methods.

To convincingly illustrate the efficiency of our method, we chose one of the most direct ways: We implemented Chang and Su's (1997) method and our own method on a corpus, the size of which was similar to the one that was used in their paper. We chose Chang and Su's paper as reference for two reasons: Their approach was unsupervised, just like ours, and it was a complicated iterative method that integrated several commonly used word-filtering techniques (including Viterbi training, mutual information, entropy, and joint Gaussian mixture density function) to improve their result. Their segmentation system contains two modules: One is the segmentation module, which is used to segment words and calculate the frequencies of the words; the other is the filtering module, which is used to rank the likelihood ratios of the words, and further to filter out those words with low likelihood ratios from the augmented dictionary and add those words with high likelihood ratios into the augmented dictionary. The system iteratively repeats these two modules until a predefined condition is fulfilled. We will show that even compared to such a deliberate approach, our simple method is marginally better. For simplicity, we will use IT to refer to Chang and Su's method and AV to refer to our method, where the symbol IT implies iterative and AV implies accessor variety.

We combined PD9208.SGML and PD9209.SGML (files of *People's Daily* as published in August and September 1992, which is a proportion of the TREC Chinese corpus [Linguistic Data Consortium, n.d.]) to form a file of 376,053 sentences after the clearing step (notice that in Chang and Su's [1997] paper, 311,591 sentences were used).

We conducted two comparison experiments, one for extracting words with lengths of two to four characters and the other for extracting words with lengths of two to seven characters. The reason for selecting these two sets of word lengths is that Chang and Su considered only words with lengths of two to four characters, whereas in our method we consider words with lengths of two to seven characters. In both experiments, the number of iterations for IT was 21 (because Chang and Su also conducted 21 iterations), and the AV value threshold (when the AV value of a string is greater than or equal to this threshold, it is considered to be a word) for our method is three.

Because we do not segment the file in AV, it is impossible to count the precision and recall by comparing the hand-segmented sample sentences with the automatically

segmented sample sentences. (In this case, sample sentences are first obtained, then they are segmented both by hand and automatically by the method under examination. The precision is equal to the number of words that are extracted both by hand and automatically, divided by the total number of words that are extracted automatically. The recall is equal to the number of words that are extracted both by hand and automatically, divided by the total number of words that are extracted by hand.) This evaluation method was applied in Chang and Su's (1997) original work. Instead, we evaluated both IT and AV with the method that we described in the previous sections. We randomly chose 1,000 words of each word length (in the first experiment, word length varied from two to four, and in the second experiment, word length varied from two to seven) from the output dictionary that was generated by each method. The precision of each word length was then defined as the proportion of correct words among the 1,000 sample words of the same word length. Regarding partial recall (we used partial recall as a substitute for traditional recall, as discussed previously), we first randomly chose sentences from the unsegmented file, and then segmented them by hand. Then we extracted words with different lengths from this set of sentences. The partial recall of each word length was then defined as the number of words of that length that were extracted both from the hand-segmented sample sentences and from the automatically generated output dictionary, divided by the total number of words of that length that were extracted from the hand-segmented sample sentences.

The system dictionaries that we used in each experiment were derived from the large dictionary described before (i.e., a dictionary downloaded from <http://www.mandarin-tools.com/segmenter.html> that contains 119,538 terms from two to seven characters long). In each experiment, the size of the system dictionary and the size of the applied corpus were chosen to approach those of the system dictionary and the corpus that were mentioned in Chang and Su's (1997) original work.

In each experiment, all of the values of precision and partial recall of both IT and AV were counted by the same person. Therefore, the evaluation results should be reasonably credible.

In the experiment of extracting words of lengths two to four, the system dictionary contained 24,705 bigrams, 4,355 trigrams, and 4,252 four-grams, that is, a total of 33,312 entities. We randomly chose 979 sentences and segmented them by hand. Suppose that the word set obtained was S . We then removed from S those segments that occurred in the system dictionary and those segments that appeared less than five times in the original corpus (the 376,053 sentences). The latter removal was undertaken because Chang and Su (1997) did not consider segments with frequency of less than five. Hence, from S , we obtained 580 bigrams, 156 trigrams, and 135 four-grams. These words were considered to be new words extracted by hand from the sample sentences and were used to test the partial recall for each method, IT and AV. In Table 11, we list the precision and partial-recall value for each word length from two to four for both IT and AV.

We can see from the table that IT outperforms AV for word length two, but the situation is just the opposite for word length four. With word length three, the two methods perform comparatively, and AV's performance is slightly worse. Considering that our method, AV, is much simpler than IT, we conclude that it is quite promising.

Because we observed from this experiment that the performance of our method improves with increased word length, we conducted another experiment to further examine this phenomenon. In this experiment, we extracted words with lengths from two to seven characters. The system dictionary that we used contained 38,097 entries, with 27,986 bigrams, 4,906 trigrams, 4,834 four-grams, 238 five-grams, 89 six-grams and 44 seven-grams. We randomly chose 1,989 sentences and segmented them by hand.

Table 11
Precision and partial recall of word lengths two to four of the first experiment on IT and AV.

	Bigram	Trigram	Four-gram
Precision			
IT	57.69%	26.18%	56.93%
AV	47.04%	25.75%	68.76%
Partial recall			
IT	85.69%	84.62%	81.48%
AV	75.34%	81.41%	87.41%

Table 12
Precision and partial recall of word lengths two to seven of the second experiment on IT and AV.

	Bigram	Trigram	Four-gram	Five-gram	Six-gram	Seven-gram
Precision						
IT	49.85%	25.38%	59.12%	32.71%	56.60%	32.62%
AV	42.70%	28.28%	68.86%	54.66%	73.77%	70.23%
Partial recall						
IT	84.84%	71.59%	78.05%	70.37%	80.65%	84.62%
AV	80.83%	81.06%	88.35%	83.33%	90.32%	76.92%

After filtering out the segments that appeared in the system dictionary and those with frequencies less than five, the numbers of new words that were extracted by hand from the sample sentences of word lengths two to seven were 699, 264, 369, 54, 31, and 13, respectively. These words were used to test the partial recall. In Table 12, we list the results of the second experiment. The precision and partial-recall values were computed in the same way as were the values in Table 11.

This table strongly indicates that AV outperforms IT for all word lengths except for word length two. Two characters have greater chances of occurring together in different environments than larger numbers of characters. This degrades the precision of our method in the case of bigrams, as the threshold that we used for AV value was three, i.e., when the AV value of a bigram was greater than or equal to three, we regarded it as a word. The reason for the lower partial recall of AV with word length two is that we filtered out all of the bigrams that were both absent from the system dictionary and had adhesive characters. For larger word lengths, only those grams with specific meanings had chances of occurring together in different environments; that is, they had higher AV values, which resulted in a higher precision value in our method. The reason for higher partial recall values of AV with longer grams is that even when a longer gram with higher AV value both was absent from the system dictionary and had adhesive characters, we did not dogmatically filter it out. Alternatively, we furthered examined whether it was in one of the three styles *h+core*, *core+t*, or *h+core+t* (as discussed in Section 3). If it was in one of these styles, then we filtered it out.

There are several ways to explain the performance of IT being better than that of AV with bigrams but worse with longer grams. First, IT does not consider adhesive characters, which helps improve the partial recall while degrading the precision, as many grams contain adhesive characters and they are hard to inspect (note that in our method, we filtered out some of the grams with adhesive characters). Second, IT uses several techniques to filter out the bad candidates for real words, which is intended to help improve the precision. But there are several deficiencies in this design. In the IT

segmentation module, a longer segment is preferred. (For each sentence, IT tries to find the segmentation with the highest likelihood, where the likelihood is defined as the multiplication of the relative frequencies of all the segments, and the relative frequency of one segment is defined as the frequency of that segment divided by the sum of the frequency of all grams [Chang and Su 1997]. Therefore, if a segmentation has more segments, then its likelihood value is smaller.) This will inevitably degrade the partial recall of shorter grams. On the other hand, because system dictionaries usually contain very limited numbers of longer terms, IT's filter module (i.e., a likelihood-ranking model) has inadequate information to correctly describe the feature functions of word class or nonword class for longer grams. This will inevitably degrade both the precision and the partial recall for longer grams, as real words might be considered nonwords, and nonwords might be considered real words. Finally, although the combination of several features seems more comprehensive, it also generates more noise than using only one feature.

We think that all of the factors that we described above can roughly explain the phenomenon that is presented in Table 12. Comparing Table 12 to Table 11, we find that the results are slightly different even for the same word lengths. One reason for this is that in different experiments, we used different system dictionaries. Note that all of the results were obtained only on new words. Therefore, the size of the system dictionary will affect the result of the experiment. Usually, the larger the system dictionary is, the poorer are the precision and recall that are obtained. In the dictionary that we used in the latter experiment, there were more bigrams than in the dictionary that we used in the first experiment. That is why the precision value and the partial-recall value for bigrams are smaller than those in the first experiment. As there are similar numbers of trigrams and four-grams in both dictionaries, the results for these grams are very close in both experiments. Another factor that may lead to these differences is the use of different sample sentences and different methods to segment them by hand for testing partial recall. In the former experiment, we considered only terms with lengths from two to four characters, and hence only segmented the sample sentences to terms of lengths from two to four. In the latter experiment, we considered all terms with lengths from two to seven characters.

6. Discussion of Future Work

In this work, we have proposed accessor variety as an alternative to the commonly used frequency criterion. Our approach may give rise to new research directions in Chinese text processing. Our promising results for word extraction make it a potential useful method for other problems as well.

In addition, word extraction is the basic step for many text-processing tasks. It is related to but different from word segmentation. Extracted words can be used as the fundamental elements for related application problems, such as creating a text summary for a bundle of articles and text clustering.

Furthermore, words as sequences of letters occur not only in language processing, but also in other application areas. Our method may be of some heuristic value to other related problems, such as those involving substring processing (Deng, Li, and Wang 2002; Thijs et al. 2002; Narasimhan et al. 2002), and biomedical concepts identification (Majoros, Subramanian, and Yandell 2003).

Finally, in our simple method, we process the data only once, and no iterative refinement is applied. The result is comparable even to that of very comprehensive systems and shows some improvement with longer grams. The simplicity of our method makes it especially suitable for processing large corpora.

Acknowledgments

We would like to thank Jing-Shin Chang for help in our implementation of their method *IT*. Many thanks also go to Chun-yu Kit for his suggestions. And we also thank the anonymous reviewers for their constructive advice in revising the original manuscript. The work described in this article was fully supported by a grant of NSFC/RGC joint research scheme (N_CityU 102/01, NSFC60131160743).

References

- Beeferman, Doug, Adam Berger, and John D. Lafferty. 1999. Statistical models for text segmentation. *Machine Learning*, 34(1-3):177-210.
- Beijing Language Institute. 1986. *Xian dai han yu pin lu ci dian* [Word Frequency Dictionary of Modern Chinese]. Beijing Language Institute Press.
- Chang, Jing-Shin and Keh-Yih Su. 1997. An unsupervised iterative method for Chinese new lexicon extraction. *International Journal of Computational Linguistics and Chinese Language Processing*, 2(2):97-148.
- Chao, Yuen-Ren. 1968. *A Grammar of Spoken Chinese*. University of California Press, Berkeley and Los Angeles.
- Chen, Ching-Yu, Shu-Fen Tseng, Chu-Ren Huang, and Keh-Jiann Chen. 1993. Some distributional properties of Mandarin Chinese—A study based on the Academia Sinica corpus. *Proceedings of Pacific Asia Conference on Formal and Computational Linguistics I*, pages 81-95, Taipei.
- Chen, Keh-Jiann and Ming-Hong Bai. 1998. Unknown word detection for Chinese by a corpus-based learning method. *International Journal of Computational Linguistics and Chinese Language Processing*, 3(1):27-44.
- Chien, Lee-Feng. 1995. Csmart—A high-performance Chinese document retrieval system. In *Proceedings of the 1995 International Conference of Computer Processing of Oriental Languages (ICCPOL)*, pages 176-183, Hawaii.
- Chien, Lee-Feng. 1997. PAT-tree-based keyword extraction for Chinese information retrieval. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 50-58, Philadelphia.
- Dai, Yubin, Teck Ee Loh, and Christopher Khoo. 1999. A new statistical formula for Chinese text segmentation incorporating contextual information. In *SIGIR '99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 82-89, Berkely, CA.
- Deng, Xiaotie, Guojun Li, and Lusheng Wang. 2002. Center and distinguisher for strings with unbounded alphabet. *Journal of Combinatorial Optimization*, 6(4):383-400.
- Garside, Roger, Geoffrey Leech, and Geoffrey Sampson. 1987. *The Computational Analysis of English: A Corpus-Based Approach*. London, Longman.
- Ge, Xian-Ping, Wanda Pratt, and Padhraic Smyth. 1999. Discovering Chinese words from unsegmented text. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 271-272, Berkeley, CA.
- Guo, Jin. 1997. Critical tokenization and its properties. *Computational Linguistics*, 23(4):569-596.
- Harris, Zellig S. 1970. Morpheme boundaries within words. In *Papers in Structural and Transformational Linguistics*. D. Reidel, Dordrecht, pages 68-77.
- Huang, Chu-Ren, Keh-Jiann Chen, and Ben-Jamin K. Tsou. 1996. Readings in Chinese Natural Language Processing. In *Journal of Chinese Linguistics Monograph Series no. 9*, edited by Chu-ren Huang et al., pages 1-22.
- Kobayasi, Yosiyuki, Takenobu Tokumaga, and Hozumi Tanaka. 1994. Analysis of Japanese compound nouns using collocational information. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING'94)*, vol. 2, pages 865-869, Kyoto, Japan.
- Kwok, Kui-Lam. 1997. Comparing representations in Chinese information retrieval. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 34-41, Philadelphia.
- Linguistic Data Consortium. n.d. *TREC Mandarin-Text Retrieval Conference Mandarin Newswire*, LDC 2000T52.
- Majoros, William H., G. Mani Subramanian, and Mark Yandell. 2003. Identification of key concepts in biomedical literature using a modified Markov heuristic. *Bioinformatics*, 19(3):402-407.
- Mo, Ruo-Ping J., Yao-Jung Yang, Keh-Jiann Chen, and Chu-Ren Huang. 1996. Determinative-measure compounds in Mandarin Chinese: Formation rules and parser implementation. In *Readings in*

- Chinese Natural Language Processing, Journal of Chinese Linguistics Monograph Series* no. 9, edited by Chu-ren Huang et al., pages 123–146.
- Narasimhan, Giri, Changsong Bu, Yuan Gao, Xuning Wang, Ning Xu, and Kalai Mathee. 2002. Mining protein sequences for motifs. *Journal of Computational Biology*, 9(5):707–720.
- Pachunke, Thomas, Oliver Mertineit, Klaus Wothke, and Rudolf Schmidt. 1992. Broad coverage automatic morphological segmentation of German words. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING'92)*, vol. 4, pages 1218–1222, Nantes, France.
- Paola, Merlo and Suzanne Stevenson. 2001. Automatic verb classification based on statistical distribution of argument structure. *Computational Linguistics*, 27(3):373–408.
- Ricardo, Baeza-Yates and Ribeiro-Neto Berthier. 1999. *Modern Information Retrieval*. ACM Press, Addison Wesley Longman.
- Sproat, Richard and Chilin Shih. 1990. A statistical method for finding word boundaries in Chinese text. *Computer Processing of Chinese and Oriental Languages*, 4:336–351.
- Sproat, Richard, Chilin Shih, William Gale, and Nancy Chang. 1996. A stochastic finite-state word-segmentation algorithm for Chinese. *Computational Linguistics*, 22(3):377–404.
- Teahan, William J., Yingying Wen, Rodger J. McNab, and Ian H. Witten. 2000. A compression-based algorithm for Chinese word segmentation. *Computational Linguistics*, 26(3):375–393.
- Thijs, Gert, Kathleen Marchal, Magali Lescot, Stephane Rombauts, Bart De Moor, Pierre Rouze, and Yves Moreau. 2002. A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. *Journal of Computational Biology*, 9(2):447–464.
- Wang, Yong-Heng, Hai-Ju Su, and Yan Mo. 1990. Automatic processing of Chinese words. *Journal of Chinese Information Processing*, 4(4):1–11.
- Weeber, Marc, Rein Vos, and R. Harald Baayen. 2000. Extracting the lowest frequency words: Pitfalls and possibilities. *Computational Linguistics*, 26(3):301–317.
- Wu, Zimin and Gwyneth Tseng. 1993. Chinese text segmentation for text retrieval: Achievements and problems. *Journal of the American Society for Information Science*, 44(9):532–542.
- Yamamoto, Mikio and Kenneth W. Church. 2001. Using suffix arrays to compute term frequency and document frequency for all substrings in a corpus. *Computational Linguistics*, 27(1):1–30.
- Yun, Bo-Hyun, Ho Lee, and Hae-Chang Rim. 1995. Analysis of Korean compound nouns using statistical information. In *Proceedings of the 1995 International Conference on Computer Processing of Oriental Languages (ICCPOL-95)*, Hawaii.
- Zhang, Jian, Jianfeng Gao, and Ming Zhou. 2000. Extraction of Chinese compound words—An experimental study on a very large corpus. *Proceedings of the Second Chinese Language Processing Workshop*, pages 132–139, Hong Kong.