A New Statistical Formula for Chinese Text Segmentation Incorporating Contextual Information

Yubin Dai
Christopher S.G. Khoo
Division of Information Studies
School of Applied Science
Nanyang Technological University
Singapore 639798
(65) 790-4602
dyb_lte@hotmail.com
assgkhoo@ntu.edu.sg

Teck Ee Loh
10 Kent Ridge Crescent
Data Storage Institute
Singapore 119260
(65) 874-8413
dsilohte@dsi.nus.edu.sg

ABSTRACT

A new statistical formula for identifying 2-character words in Chinese text, called the *contextual information formula*, was developed empirically by performing stepwise logistic regression using a sample of sentences that had been manually segmented. Contextual information in the form of the frequency of characters that are adjacent to the bigram being processed as well as the weighted document frequency of the overlapping bigrams were found to be significant factors for predicting the probablity that the bigram constitutes a word. Local information (the number of times the bigram occurs in the document being segmented) and the position of the bigram in the sentence were not found to be useful in determining words. The *contextual information formula* was found to be significantly and substantially better than the *mutual information formula* in identifying 2-character words. The method can also be used for identifying multi-word terms in English text.

Keywords

Chinese text segmentation, word boundary identification, logistic regression, multi-word terms

1. INTRODUCTION

Chinese text is different from English text in that there is no explicit word boundary. In English text, words are separated by spaces. Chinese text (as well as text of other Oriental languages) is made up of ideographic characters, and a word can comprise one, two or more such characters, without explicit indication where one word ends and another begins.

This has implications for natural language processing and information retrieval with Chinese text. Text processing techniques that have been developed for Western languages deal with words as meaningful text units and assume that words are easy to identify. These techniques may not work well for Chinese text without some adjustments. To apply these techniques to Chinese text, automatic methods for identifying word boundaries accurately have to be developed. The process of identifying word boundaries has been referred to as text segmentation or, more accurately, word segmentation.

Several techniques have been developed for Chinese text segmentation. They can be divided into:

1. *statistical methods*, based on statistical properties and frequencies of characters and character strings in a corpus (e.g. [13] and [16]).
2. *dictionary-based methods*, often complemented with grammar rules. This approach uses a dictionary of words to identify word boundaries. Grammar rules are often used to resolve conflicts (choose between alternative segmentations) and to improve the segmentation (e.g. [4], [8], [19] and [20]).
3. *syntax-based methods*, which integrate the word segmentation process with syntactic parsing or part-of-speech tagging (e.g. [1]).
4. *conceptual methods*, that make use of some kind of semantic processing to extract information and store it in a knowledge representation scheme. Domain knowledge is used for disambiguation (e.g. [9]).

Many researchers use a combination of methods (e.g. [14]).

The objective of this study was to empirically develop a statistical formula for Chinese text segmentation. Researchers have used different statistical methods in segmentation, most of which were based on theoretical considerations or adopted from other fields. In this study, we developed a statistical formula empirically by performing stepwise logistic regression using a sample of sentences that had been manually segmented. This paper reports the new formula developed for identifying 2-character words, and the effectiveness of this formula compared with the *mutual information formula*.

This study has the following novel aspects:

- The statistical formula was derived empirically using regression analysis.
- The manual segmentation was performed to identify "meaningful" words rather than simple words. "Meaningful" words include phrasal words and multi-word terms.
- In addition to the relative frequencies of bigrams and characters often used in other studies, our study also investigated the use of *document frequencies* and *weighted*

SIGIR '99 8/99 Berkley, CA USA

document frequencies. Weighted document frequencies are similar to document frequencies but each document is weighted by the square of the number of times the character or bigram occurs in the document.

- Contextual information was included in the study. To predict whether the bigram *BC* in the character string "… A B C D …" constitutes a word, we investigated whether the frequencies for *AB*, *CD*, *A* and *D* should be included in the formula.
- Local frequencies were included in the study. We investigated character and bigram frequencies *within the document* in which the sentence occurs (i.e. the number of times the character or bigram appears in the document being segmented).
- We investigated whether the position of the bigram (at the beginning of the sentence, before a punctuation mark, or after a punctuation mark) had a significant effect.
- We developed a segmentation algorithm to apply the statistical formula to segment sentences and resolve conflicts.

In this study, our objective was to segment text into "meaningful words" rather than "simple words". A simple word is the smallest independent unit of a sentence that has meaning on its own. A *meaningful word* can be a simple word or a compound word comprising 2 or more simple words – depending on the context. In many cases, the meaning of a compound word is more than just a combination of the meanings of the constituent simple words, i.e. some meaning is lost when the compound word is segmented into simple words. Furthermore, some phrases are used so often that native speakers perceive them and use them as a unit. Admittedly, there is some subjectivity in the manual segmentation of text. But the fact that statistical models can be developed to predict the manually segmented words substantially better than chance indicates some level of consistency in the manual segmentation.

The problem of identifying meaningful words is not limited to Chinese and oriental languages. Identifying multi-word terms is also a problem in text processing with English and other Western languages, and researchers have used the *mutual information formula* and other statistical approaches for identifying such terms (e.g. [3], [6] and [7]).

2. PREVIOUS STUDIES

There are few studies using a purely statistical approach to Chinese text segmentation. One statistical formula that has been used by other researchers (e.g. [11] and [16]) is the *mutual information formula*. Given a character string "… A B C D …", the mutual information for the bigram *BC* is given by the formula:

$$\mathrm{MI(BC)} = \log_2 \frac{freq(BC)}{freq(B) * freq(C)}$$

$$= \log_2 \mathrm{freq(BC)} - \log_2 \mathrm{freq(B)} - \log_2 \mathrm{freq(C)}$$

where *freq* refers to the relative frequency of the character or bigram in the corpus (i.e. the number of times the character or bigram occurs in the corpus divided by the number of characters in the corpus).

Mutual information is a measure of how strongly the two characters are associated, and can be used as a measure of how likely the pair of characters constitutes a word. Sproat & Shih [16] obtained recall and precision values of 94% using *mutual information* to identify words. This study probably segmented text into simple words rather than meaningful words. In our study, text was segmented into meaningful words and we obtained much poorer results for the *mutual information formula*.

Lua [12] and Lua & Gan [13] applied information theory to the problem of Chinese text segmentation. They calculated the information content of characters and words using the information entropy formula $I = - \log_2 P$, where P is the probability of occurrence of the character or word. If the information content of a character string is less than the sum of the information content of the constituent characters, then the character string is likely to constitute a word. The formula for calculating this "loss" of information content when a word is formed is identical to the mutual information formula. Lua & Gan [13] obtained an accuracy of 99% (measured in terms of the number of errors per 100 characters).

Tung & Lee [18] also used information entropy to identify unknown words in a corpus. However, instead of calculating the entropy value for the character string that is hypothesized to be a word (i.e. the candidate word), they identified all the characters that occurred to the left of the candidate word in the corpus. For each left character, they calculated the probability and entropy value for that character given that it occurs to the left of the candidate word. The same is done for the characters to the right of the candidate word. If the sum of the entropy values for the left characters and the sum of the entropy values for the right characters are both high, than the candidate word is considered likely to be a word. In other words, a character string is likely to be a word if it has several different characters to the left and to the right of it in the corpus, and none of the left and right characters predominate (i.e. not strongly associated with the character string).

Ogawa & Matsuda [15] developed a statistical method to segment Japanese text. Instead of attempting to identify words directly, they developed a formula to estimate the probability that a bigram straddles a word boundary. They referred to this as the segmentation probability. This was complemented with some syntactic information about which class of characters could be combined with which other class.

All the above mathematical formulas used for identifying words and word boundaries were developed based on theoretical considerations and not derived empirically.

Other researchers have developed statistical methods to find the best segmentation for the whole sentence rather than focusing on identifying individual words. Sproat et al. [17] developed a stochastic finite state model for segmenting text. In their model, a word dictionary is represented as a weighted finite state transducer. Each weight represents the estimated cost of the word (calculated using the negative log probability). Basically, the system selects the sentence segmentation that has the smallest total cost. Chang & Chen [1] developed a method for word segmentation and part-of-speech tagging based on a first-order hidden Markov model.

3. RESEARCH METHOD

The purpose of this study was to empirically develop a statistical formula for identifying 2-character words as well as to investigate the usefulness of various factors for identifying the words. A sample of 400 sentences was randomly selected from 2 months (August and September 1995) of news articles from the *Xin Hua News Agency*, comprising around 2.3 million characters. The sample sentences were manually segmented. The segmentation rules described in [10] were followed fairly closely. More details of the manual segmentation process, especially with regard to identifying meaningful words will be given in [5].

300 sentences were used for model building, i.e. using regression analysis to develop a statistical formula. 100 sentences were set aside for model validation to evaluate the formula developed in the regression analysis. The sample sentences were broken up into overlapping bigrams. In the regression analysis, the dependent variable was whether a bigram was a two-character word according to the manual segmentation. The independent variables were various corpus statistics derived from the corpus (2 months of news articles).

The types of frequency information investigated were:

1. *Relative frequency* of individual characters and bigrams (character pairs) in the corpus, i.e. the number of times the character or bigram occurs in the corpus divided by the total number of characters in the corpus.
2. *Document frequency* of characters and bigrams, i.e. the number of documents in the corpus containing the character or bigram divided by the total number of documents in the corpus.
3. *Weighted document frequency* of characters and bigrams. To calculate the weighted document frequency of a character string, each document containing the character string is assigned a score equal to the square of the number of times the character string occurs in the document. The scores for all the documents containing the character string are then summed and divided by the total number of documents in the corpus to obtain the weighted document frequency for the character string. The rationale is that if a character string occurs several times within the same document, this is stronger evidence that the character string constitutes a word, than if the character string occurs once in several documents. Two or more characters can occur together by chance in several different documents. It is less likely for two characters to occur together several times within the same document by chance.
4. *Local frequency* in the form of *within-document frequency* of characters and bigrams, i.e. the number of times the character or bigram occurs in the document being segmented.
5. *Contextual information*. Frequency information of characters adjacent to a bigram is used to help determine whether the bigram is a word. For the character string "… A B C D …" , to determine whether the bigram *BC* is a word, frequency information for the adjacent characters *A* and *D*, as well as the overlapping bigrams *AB* and *BC* were considered.
6. *Positional information*. We studied whether the position of a character string (at the beginning, middle or end of a sentence) gave some indication of whether the character string was a word.

The statistical model was developed using forward stepwise logistic regression, using the Proc Logistic function in the SAS v.6.12 statistical package for Windows. Logistic regression is an appropriate regression technique when the dependent variable is binary valued (takes the value 0 or 1). The formula developed using logistic regression predicts the probability (more accurately, the log of the odds) that a bigram is a meaningful word.

In the stepwise regression, the threshold for a variable to enter the model was set at the 0.001 significance level and the threshold for retaining a variable in the model was set at 0.01. In addition, preference was given to *relative frequencies* and *local frequencies* because they are easier to calculate than *document frequencies* and *weighted document frequencies*. Also, *relative frequencies* are commonly used in previous studies.

Furthermore, a variable was entered in a model only if it gave a noticeable improvement to the effectiveness of the model. During regression analysis, the effectiveness of the model was estimated using the measure of concordance that was automatically output by the SAS statistical program. A variable was accepted into the model only if the measure of concordance improved by at least 2% when the variable was entered into the model.

We evaluated the accuracy of the segmentation using measures of recall and precision. Recall and precision in this context are defined as follows:

$$\text{Recall} = \frac{\text{No. of 2-character words identified in the automatic segmentation that are correct}}{\text{No. of 2-character words identified in the manual segmentation}}$$

$$\text{Precision} = \frac{\text{No. of 2-character words identified in the automatic segmentation that are correct}}{\text{No. of 2-character words identified in the automatic segmentation}}$$

4. STATISTICAL FORMULAS DEVELOPED

4.1 The Contextual Information Formula

The formula that was developed for 2-character words is as follows. Given a character string "… A B C D …" , the association strength for bigram *BC* is:

$$\begin{aligned}\text{Assoc(BC)} = {} & 0.35 * \log_2 \text{freq(BC)} + 0.37 * \log_2 \text{freq(A)} + \\ & 0.32 \log_2 \text{freq(D)} - 0.36 * \log_2 \text{docfreq}_{wt}\text{(AB)} - \\ & 0.29 * \log_2 \text{docfreq}_{wt}\text{(CD)} + 5.91\end{aligned}$$

where *freq* refers to the relative frequency in the corpus and $docfreq_{wt}$ refers to the weighted document frequency. We refer to this formula as the *contextual information formula*. More details of the regression model are given in Table 1.

The formula indicates that contextual information is helpful in identifying word boundaries. *A* in the formula refers to the character preceding the bigram that is being processed, whereas *D* is the character following the bigram. The formula indicates that if the character preceding and the character following the bigram have high relative frequencies, then the bigram is more likely to be a word.

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate
INTERCPT	1	5.9144	0.1719	1184.0532	0.0001	.
Log freq(BC)	1	0.3502	0.0106	1088.7291	0.0001	0.638740
Log freq(A)	1	0.3730	0.0113	1092.1382	0.0001	0.709621
Log freq(D)	1	0.3171	0.0107	886.4446	0.0001	0.607326
Log $docfreq_{wt}$(AB)	1	-0.3580	0.0111	1034.0948	0.0001	-0.800520
Log $docfreq_{wt}$(CD)	1	-0.2867	0.0104	754.2276	0.0001	-0.635704

Note: *freq* refers to the relative frequency, and $docfreq_{wt}$ refers to the weighted document frequency.

Association of Predicted Probabilities and Observed Responses

Concordant = 90.1%	Somers' D = 0.803
Discordant = 9.8%	Gamma = 0.803
Tied = 0.1%	Tau-a = 0.295
(23875432 pairs)	c = 0.901

Table 1. Final regression model for 2-character words

Contextual information involving the *weighted document frequency* was also found to be significant. The formula indicates that if the overlapping bigrams *AB* and *CD* have high *weighted document frequencies*, then the bigram *BC* is less likely to be a word. We tried replacing the *weighted document frequencies* with the unweighted *document frequencies* as well as the *relative frequencies*. These were found to give a lower concordance score. Even with *docfreq (AB)* and *docfreq (CD)* in the model, $docfreq_{wt}$ *(AB)* and $docfreq_{wt}$ *(CD)* were found to improve the model significantly. However, local frequencies were surprisingly not found to be useful in predicting 2-character words.

We investigated whether the position of the bigram in the sentence was a significant factor. We included a variable to indicate whether the bigram occurred just after a punctuation mark or at the beginning of the sentence, and another variable to indicate whether the bigram occurred just before a punctuation mark or at the end of a sentence. The interaction between each of the "position" variables and the various relative frequencies were not significant. However, it was found that whether or not the bigram was at the end of a sentence or just before a punctuation mark was a significant factor. Bigrams at the end of a sentence or just before a punctuation mark tend to be words. However, since this factor did not improve the concordance score by 2%, the effect was deemed too small to be included in the model.

It should be noted that the contextual information used in the study already incorporates some positional information. The frequency of character A (the character preceding the bigram) was given the value 0 if the bigram was preceded by a punctuation mark or was at the beginning of a sentence. Similarly, the frequency of character D (the character following the bigram) was given the value 0 if the bigram preceded a punctuation mark.

We also investigated whether the model would be different for high and low frequency words. We included in the regression analysis the interaction between the relative frequency of the bigram and the other relative frequencies. The interaction terms were not found to be significant. Finally, it is noted that the coefficients for the various factors are nearly the same, hovering around 0.34.

4.2 Improved Mutual Information Formula

In this study, the *contextual information formula* (CIF) was evaluated by comparing it with the *mutual information formula* (MIF). We wanted to find out whether the segmentation results using the CIF was better than the segmentation results using the MIF.

In the CIF model, the coefficients of the variables were determined using regression analysis. If CIF was found to give better results than MIF, it could be because the coefficients for the variables in CIF had been determined empirically – and not because of the types of variables in the formula. To reject this explanation, regression analysis was used to determine the coefficients for the factors in the *mutual information formula*. We refer to this new version of the formula as the *improved mutual information formula.*

Given a character string "… A B C D …" , the *improved mutual information formula* is:

$$\text{Improved MI(BC)} = 0.39 * \log_2 \text{freq(BC)} - 0.28 * \log_2 \text{freq(B)} - 0.23 \log_2 \text{freq(C)} - 0.32$$

The coefficients are all close to 0.3. The formula is thus quite similar to the *mutual information formula*, except for a multiplier of 0.3.

5. SEGMENTATION ALGORITHMS

The automatic segmentation process has the following steps:

1. The statistical formula is used to calculate a score for each bigram to indicate its association strength (or how likely the bigram is a word).
2. A threshold value is then set and used to decide which bigram is a word. If a bigram obtains a score above the threshold value, then it is selected as a word. Different threshold values can be used, depending on whether the user prefers high recall or high precision.
3. A segmentation algorithm is used to resolve conflict. If two overlapping bigrams both have association scores above the

Recall	Precision		
	Comparative Forward Match	Forward Match	*Improvement*
Mutual Information			
90%	51%	-	-
80%	52%	47%	*5%*
70%	53%	51%	*2%*
60%	54%	52%	*2%*
Improved Mutual Information			
90%	51%	-	-
80%	53%	46%	*7%*
70%	54%	52%	*2%*
60%	55%	54%	*1%*
Contextual Information Formula			
90%	55%	54%	*1%*
80%	62%	62%	*0%*
70%	65%	65%	*0%*
60%	68%	68%	*0%*

Table 2. Recall and precision values for the *comparative forward match* segmentation algorithm vs. *forward match*

threshold value, then there is conflict or ambiguity. The frequency of such conflicts will rise as the threshold value is lowered. The segmentation algorithm resolves the conflict and selects one of the bigrams as a word.

One simple segmentation algorithm is the *forward match algorithm*. Consider the sentence "A B C D E …" . The segmentation process proceeds from the beginning of the sentence to the end. First the bigram *AB* is considered. If the association score is above the threshold, then *AB* is taken as a word, and the bigram *CD* is next considered. If the association score of *AB* is below the threshold, the character *A* is taken as a 1-character word. And the bigram *BC* is next considered. In effect, if the association score of both *AB* and *BC* are above threshold, the forward match algorithm selects *AB* as a word and not *BC*.

The *forward match* method for resolving ambiguity is somewhat arbitrary and not satisfactory. When overlapping bigrams exceed the threshold value, it simply decides in favour of the earlier bigram. Another segmentation algorithm was developed in this study which we refer to as the *comparative forward match algorithm*. This has an additional step:

If 2 overlapping bigrams *AB* and *BC* both have scores above the threshold value then their scores are compared. If *AB* has a higher value, then it is selected as a word, and the program next considers the bigrams *CD* and *DE*. On the other hand, if *AB* has a lower value, then character *A* is selected as a 1-character word, and the program next considers bigrams *BC* and *CD*.

The *comparative forward match* method (CFM) was compared with the *forward match* method (FM) by applying them to the 3 statistical formulas (the *contextual information formula*, the *mutual information formula* and the *improved mutual information formula*). One way to compare the effectiveness of the 2 segmentation algorithms is by comparing their precision figures at the same recall levels. The precision figures for selected recall levels are given in Table 2. The results are based on the sample of 300 sentences.

Recall	Precision		
	Mutual Information	*Improved Mutual Information*	*Contextual Information*
90%	57% (0.0)	57% (-2.5)	61% (-1.5)
80%	59% (3.7)	59% (-1.5)	66% (-0.8)
70%	59% (4.7)	60% (-1.0)	70% (-0.3)
60%	60% (5.6)	62% (-0.7)	74% (0.0)

* Threshold values are given in parenthesis.

Table 3. Recall and precision for three statistical formulas

The *comparative forward match* algorithm gave better results for the *mutual information* and *improved mutual information formulas* – especially at low threshold values when a large number of conflicts are likely. Furthermore, for the *forward match* method, the recall didn't go substantially higher than 80% even at low threshold values.

For the *contextual information formula*, the *comparative forward match* method did not perform better than *forward match*, except at very low threshold values when the recall was above 90%. This was expected because the *contextual information formula* already incorporates information about neighboring characters within the formula. The formula gave very few conflicting segmentations. There were very few cases of overlapping bigrams both having association scores above the threshold – except when threshold values were below –1.5.

6. EVALUATION

6.1 Comparing the Contextual Information Formula with the Mutual Information Formula

In this section we compare the effectiveness of the *contextual information formula* with the *mutual information formula* and the *improved mutual information formula* using the 100 sentences that had been set aside for evaluation purposes. For the *contextual information formula*, the *forward match* segmentation algorithm was used. The *comparative forward match* algorithm was used for the *mutual information* and the *improved mutual information* formulas.

The three statistical formulas were compared by comparing their precision figures at 4 recall levels – at 60%, 70%, 80% and 90%. For each of the three statistical formulas, we identified the threshold values that would give a recall of 60%, 70%, 80% and 90%. We then determined the precision values at these threshold values to find out whether the *contextual information formula* gave better precision than the other two formulas at 60%, 70%, 80% and 90% recall. These recall levels were selected because a recall of 50% or less is probably unacceptable for most applications.

The precision figures for the 4 recall levels are given in Table 3. The recall-precision graphs for the 3 formulas are given in Fig. 1. The *contextual information formula* substantially outperforms the *mutual information* and the *improved mutual information* formulas. At the 90% recall level, the *contextual information*

Avg Recall	Avg Precision		
	Mutual Information	*Improved Mutual Information*	*Contextual Information*
90%	57% (1.0)	58% (-2.3)	61% (-1.5)
80%	60% (3.8)	60% (-1.4)	67% (-0.7)
70%	59% (4.8)	60% (-1.0)	70% (-0.3)
60%	60% (5.6)	63% (-0.6)	73% (0.0)

* Threshold values are given in parenthesis.

Table 4. Average recall and average precision for the three statistical formulas

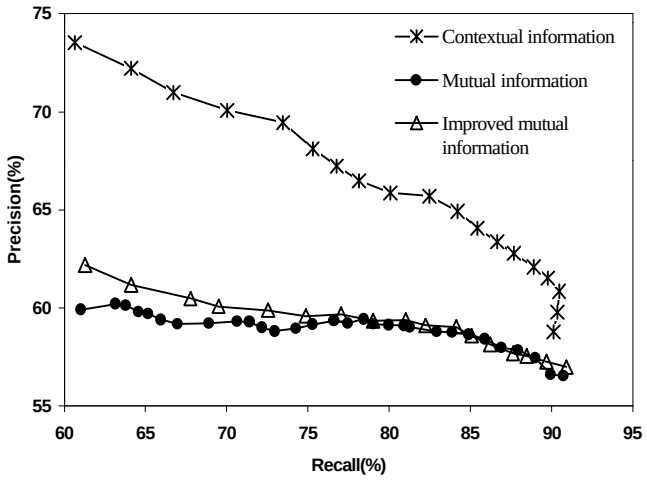

Fig. 1. Recall-precision graph for the three statistical

formula was better by about 4%. At the 60% recall level, it outperformed the *mutual information formula* by 14% (giving a relative improvement of 23%). The results also indicate that the *improved mutual information formula* does not perform better than the *mutual information formula*.

6.2 Statistical Test of Significance

In order to perform a statistical test, recall and precision figures were calculated for each of the 100 sentences used in the evaluation. The average recall and the average precision across the 100 sentences were then calculated for the three statistical formulas. In the previous section, recall and precision were calculated for all the 100 sentences combined. Here, recall and precision were obtained for individual sentences and then the average across the 100 sentences was calculated. The average precision for 60%, 70%, 80% and 90% average recall are given in Table 4.

For each recall level, an analysis of variance with repeated measures was carried out to find out whether the differences in precision were significant. Pairwise comparisons using Tukey's HSD test was also carried out. The *contextual information formula* was significantly better (α=0.001) than the *mutual information* and the *improved mutual information* formulas at all 4 recall levels. The *improved mutual information formula* was not found to be significantly better than *mutual information*.

Association Score>1.0 (definite errors)	
大学(农业大学)	university (agricultural university)
地质(地质时期)	geology (geologic age)
植物(陆地植物)	plant (upland plant)
主权(主权国家)	sovereignty (sovereign state)
Association Score Between –1.0 and 1.0 (borderline errors)	
统计(统计资料)	statistics (statistical data)
灾害(自然灾害)	calamity (natural calamity)
资源(人力资源)	resources (manpower resources)
教授(副教授)	professor (associate professor)
贫困(贫困化)	poor (pauperization)
十四(十四日)	fourteen (the 14th day)
二十(二十个)	twenty (twenty pieces)

Table 5. Simple words that are part of a longer meaningful word

Association Score >1.0 (definite errors)	
将由	will … through
日电	telegraph [on the] day [31 July]
Association Score Between –1.0 and 1.0 (borderline errors)	
还向	still … to
将是	will be
等人	people etc.
我要	I want
Person's name	
万文 (万文举)	Wan Wen Ju
Place name	
庄台(田庄台)	a village name in China
加拿(加拿大)	Canada
Name of an organization/institution	
华社 (新华社)	Xin Hua Agency
务院(国务院)	The State Department

Table 6. Bigrams incorrectly identified as words

7. ANALYSIS OF ERRORS

The errors that arose from using the *contextual information formula* were analyzed to gain insights into the weaknesses of the model and how the model can be improved. There are two types of errors: errors of commission and errors of omission. Errors of commission are bigrams that are identified by the automatic segmentation to be words when in fact they are not (according to the manual segmentation). Errors of omission are bigrams that are not identified by the automatic segmentation to be words but in fact they are.

The errors depend of course on the threshold values used. A high threshold (e.g. 1.0) emphasizes precision and a low threshold (e.g. –1.0) emphasizes recall. 50 sentences were selected from the 100 sample sentences to find the distribution of errors at different regions of threshold values.

Association Score between -1.0 and -2.0	
北段	the northern section of a construction project
残卷	fragments of ancient books
Association Score < -2.0	
九月	September
三日	3rd day
两浙	(name of a district in China)
南仓	(name of an institution)
周易	the Book of Changes

Table 7. 2-character words with association score below -1.0

We divide the errors of commission (bigrams that are incorrectly identified as words by the automatic segmentation) into 2 groups:

1. Definite errors: bigrams with association scores above 1.0 but are not words
2. Borderline errors: bigrams with association scores between –1.0 and 1.0 and are not words

We also divide the errors of omission (bigrams that are words but are not identified by the automatic segmentation) into 2 groups:

1. Definite errors: bigrams with association scores below –1.0 but are words
2. Borderline errors: bigrams with association scores between –1.0 and 1.0 and are words.

7.1 Errors of Commission

Errors of commission can be divided into 2 types:

1. The bigram is a simple word that is part of a longer meaningful word.
2. The bigram is not a word (neither simple word nor meaningful word).

Errors of the first type are illustrated in Table 5. The words within parenthesis are actually meaningful words but segmented as simple words (words on the left). The words lose part of the meaning when segmented as simple words. These errors occurred mostly with 3 or 4-character meaningful words.

Errors of the second type are illustrated in Table 6. Many of the errors are caused by incorrectly linking a character with a function word or pronoun. Some of the errors can easily be removed by using a list of function words and pronouns to identify these characters.

7.2 Errors of Omission

Examples of *definite errors* of omission (bigrams with association scores below –1.0 but are words) are given in Table 7. Most of the errors are rare words and time words. Some are ancient names, rare and unknown place names, as well as technical terms. Since our corpus comprises general news articles, these types of words are not frequent in the corpus. Time words like dates usually have low association values because they change everyday! These errors can be reduced by incorporating a separate algorithm for recognizing them.

The proportion of errors of the various types are given in Table 8.

8. CONCLUSION

A new statistical formula for identifying 2-character words in Chinese text, called the *contextual information formula*, was developed empirically using regression analysis. The focus was on identifying meaningful words (including multi-word terms and idioms) rather than simple words. The formula was found to give significantly and substantially better results than the *mutual information formula.*

Contextual information in the form of the frequency of characters that are adjacent to the bigram being processed as well as the weighted document frequency of the overlapping bigrams were found to be significant factors for predicting the probablity that the bigram constitutes a word. Local information (e.g. the number of times the bigram occurs in the document being segmented) and the position of the bigram in the sentence were not found to be useful in determining words.

Of the bigrams that the formula erroneously identified as words, about 80% of them were actually simple words. Of the rest, many involved incorrect linking with a function words. Of the words that the formula failed to identify as words, more than a third of them were rare words or time words. The proportion of rare words increased as the threshold value used was lowered. These rare words cannot be identified using statistical techniques.

This study investigated a purely statistical approach to text

Errors of Commission Association score > 1.0 (No. of errors=34)		**Borderline Cases** Association score: –1.0 to1.0 (No. of cases: 210)			**Errors of Omission** Association score < –1.0			
Simple words	Not words	Simple words	Not words	Meaning-ful words	Association score: –1.0 to –2.0 (No. of errors=43)		Association score < –2.0 (No. of errors=22)	
					Rare words & time words	Others	Rare words & time words	Others
82.3%	17.7%	55.2%	20.5%	24.3%	23.2%	76.8%	63.6%	36.4%

Table 8. Proportion of errors of different types

segmentation. The advantage of the statistical approach is that it can be applied to any domain, provided that the document collection is sufficiently large to provide frequency information. A domain-specific dictionary of words is not required. In fact, the statistical formula can be used to generate a shortlist of candidate words for such a dictionary. On the other hand, the statistical method cannot identify rare words and proper names. It is also fooled by combinations of function words that occur frequently and by function words that co-occur with other words.

It is well-known that a combination of methods is needed to give the best segmentation results. The segmentation quality in this study can be improved by using a list of function words and segmenting the function words as single character words. A dictionary of common and well-known names (including names of persons, places, institutions, government bodies and classic books) could be used by the system to identify proper names that occur infrequently in the corpus. Chang et al. [2] developed a method for recognizing proper nouns using a dictionary of family names in combination with a statistical method for identifying the end of the name. An algorithm for identifying time and dates would also be helpful. It is not clear whether syntactic processing can be used to improve the segmentation results substantially.

Our current work includes developing statistical formulas for identifying 3 and 4-character words, as well as investigating whether the statistical formula developed here can be used with other corpora. The approach adopted in this study can also be used to develop statistical models for identifying multi-word terms in English text. It would be interesting to see whether the regression model developed for English text is similar to the one developed in this study for Chinese text. Frantzi, Ananiadou & Tsujii [7], using a different statistical approach, found that contextual information could be used to improve the identification of multi-word terms in English text.

9. REFERENCES

[1] Chang, C.-H., and Chen, C.-D. A study of integrating Chinese word segmentation and part-of-speech tagging. Communications of COLIPS, 3, 1 (1993), 69-77.

[2] Chang, J.-S., Chen, S.-D., Ker, S.-J., Chen, Y., and Liu, J.S. A multiple-corpus approach to recognition of proper names in Chinese texts. Computer Processing of Chinese and Oriental Languages, 8, 1 (June 1994), 75-85.

[3] Church, K.W., and Hanks, P. Word association norms, mutual information and lexicography. In Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics (Vancouver, June 1989), 76-83.

[4] Dai, J.C., and Lee, H.J. A generalized unification-based LR parser for Chinese. Computer Processing of Chinese and Oriental Languages, 8, 1 (1994), 1-18.

[5] Dai, Y. Developing a new statistical method for Chinese text segmentation. (Master's thesis in preparation)

[6] Damerau, F.J. Generating and evaluating domain-oriented multi-word terms from texts. Information Processing & Management, 29, 4 (1993), 433-447.

[7] Frantzi, K.T., Ananiadou, S., and Tsujii, J. The C-value/NC-value method of automatic recognition for multi-word terms. In C. Nikolaou and C. Stephanidis (eds.), Research and Advanced Technology for Digital Libraries, 2nd European Conference, ECDL' 98 (Heraklion, Crete, September 1998), Springer-Verlag, 585-604.

[8] Liang, N.Y. The knowledge of Chinese words segmentation [in Chinese]. Journal of Chinese Information Processing, 4, 2 (1990), 42-49.

[9] Liu, I.M. Descriptive-unit analysis of sentences: Toward a model natural language processing. Computer Processing of Chinese & Oriental Languages, 4, 4 (1990), 314-355.

[10] Liu, Y., Tan, Q., and Shen, X.K. Xin xi chu li yong xian dai han yu fen ci gui fan ji zi dong fen ci fang fa ["Modern Chinese Word Segmentation Rules and Automatic Word Segmentation Methods for Information Processing"]. Qing Hua University Press, Beijing, 1994.

[11] Lua, K.T. Experiments on the use of bigram mutual information in Chinese natural language processing. Presented at the 1995 International Conference on Computer Processing of Oriental Languages (ICCPOL) (Hawaii, November 1995). Available: http://137.132.89.143/luakt/publication.html

[12] Lua, K.T. From character to word - An application of information theory. Computer Processing of Chinese & Oriental Languages, 4, 4 (1990), 304-312.

[13] Lua, K.T., and Gan, G.W. An application of information theory in Chinese word segmentation. Computer Processing of Chinese & Oriental Languages, 8, 1 (1994), 115-124.

[14] Nie, J.Y., Hannan, M.L., and Jin, W.Y. Unknown word detection and segmentation of Chinese using statistical and heuristic knowledge. Communications of COLIPS, 5, 1&2 (1995), 47-57.

[15] Ogawa, Y., and Matsuda, T. Overlapping statistical word indexing: A new indexing method for Japanese text. In Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (Philadelphia, July 1997), ACM, 226-234.

[16] Sproat, R., and Shih, C.L. A statistical method for finding word boundaries in Chinese text. Computer Processing of Chinese & Oriental Languages, 4, 4 (1990), 336-351.

[17] Sproat, R., Shih, C., Gale, W., and Chang, N. A stochastic finite-state word-segmentation algorithm for Chinese. Computational Lingustics, 22, 3 (1996), 377-404.

[18] Tung, C.-H., and Lee, H.-J. Identification of unknown words from a corpus. Computer Processing of Chinese and Oriental Languages, 8 (Supplement, Dec. 1994), 131-145.

[19] Wu, Z., and Tseng, G. ACTS: An automatic Chinese text segmentation system for full text retrieval. Journal of the American Society for Information Science, 46, 2 (1995), 83-96.

[20] Yeh, C.L., and Lee, H.J. Rule-based word identification for mandarin Chinese sentences: A unification approach. Computer Processing of Chinese and Oriental Languages, 5, 2 (1991), 97-118.