Automatic Identification of Chinese Stop Words

Feng Zou, Fu Lee Wang, Xiaotie Deng, Song Han

Computer Science Department, City University of Hong Kong
Kowloon Tong, Hong Kong
phenix@cs.cityu.edu.hk, {flwang, csdeng, shan00}@cityu.edu.hk

Abstract. In modern information retrieval systems, effective indexing can be achieved by removal of stop words. Till now many stop word lists have been developed for English language. However, no standard stop word list has been constructed for Chinese language yet. With the fast development of information retrieval in Chinese language, exploring Chinese stop word lists becomes critical. In this paper, to save the time and release the burden of manual stop word selection, we propose an automatic aggregated methodology based on statistical and information models for extraction of the stop word list in Chinese language. The novel algorithm balances various measures and removes the idiosyncrasy of particular statistical measures. Extensive experiments have been conducted on Chinese segmentation for illustration of its effectiveness. Results show that the generated stop word list can improve the accuracy of Chinese segmentation significantly.

1. Introduction

In information retrieval, a document is traditionally indexed by words [10, 11, 17]. Statistical analysis through documents showed that some words have quite low frequency, while some others act just the opposite. For example, words "and", "of", and "the" appear frequently in the documents. The common characteristic of these words is that they carry no significant information to the document. Instead, they are used just because of grammar. We usually refer to this set of words as stop words [10, 11, 21].

Stop words are widely used in many fields. In digital libraries, for instance, elimination of stop words could contribute to reduce the size of the indexing structure considerably and obtain a compression of more than 40% [10]. On the other hand, in information retrieval, removal of stop words could not only help to index effectively, but also help to speed up the calculation and increase the accuracy [20].

Lots of stop word lists have been developed for English language in the past, which are usually based on frequency statistics of a large corpus [21]. The English stop word lists available online [22, 23] are good examples. However, no commonly accepted stop word list has been constructed for Chinese language. Most current researches on Chinese information retrieval make use of manual or simple statistical stop word lists [1, 2, 3], some of which are picked up based on the authors experiences consuming a lot of time. The contents of these stop lists vary a lot from each other. With the fast

© *A. Gelbukh (Editor)*
Advances in Natural Language Processing
Research in Computing Science 18, 2006, pp. 151-162

growth of online Chinese documents and the rapid increase of research interest in Chinese information retrieval, constructing a general Chinese stop word list together with an applicable generating methodology becomes critical. In order to save the time and release the burden of manual stop word selection, an automatic aggregated methodology would be a better choice.

One of the difficulties for automatical identification of stop words in Chinese language is the absence of word boundaries. Different from texts in English and other western languages, which are segmented into words by using spaces and punctuations as word delimiters, Asian languages, such as Chinese, do not delimit words by space. Usually a Chinese word consists of more than one character and the number of characters contained varies. Meanwhile, Chinese characters carry a lot of different meanings. They could be interpreted differently when used together with different characters. The character "的", which is equivalent to the word "of" in English, is taken as an example. It could carry a different meaning in the combination with different characters, such as "的确"(certainly), "的士"(taxi), etc.

Although identification of stop words is quite a hard work without a correct segmentation of text [4], stop words play an important role during segmentation. In English, as an instance, texts are segmented into phrases with the help of stop words and punctuations. Researches show that using effective stop word list can improve the accuracy of Chinese segmentation as well [4].

In our paper, we propose an automatic aggregated methodology for construction of the stop word list in Chinese. Stop words are extracted from TREC 5 and 6 corpora which are widely accepted as standard corpora for Chinese processing. The stop word list is extracted based on statistical and information models. The statistical model extracts stop words based on the probability and distribution. The information model measures the significance of words by using information theory. Results from these two models are aggregated to generate the Chinese stop word list.

To demonstrate the effectiveness of the stop word list, we propose a novel Chinese segmentation algorithm based on it. Experiment has been conducted using the dataset of a recent competition on Chinese segmentation [15]. Results have shown that the stop word list can improve the accuracy of Chinese segmentation significantly. The outstanding performance illustrates the effectiveness of our Chinese stop word list.

The rest of the paper is organized as following. Section 2 covers the methodology for the discovery of the Chinese stop word list. Section 3 analyzes the result of the stop word list extraction experiments. To better prove the effectiveness of this methodology, in Section 3, we also propose an application of the stop word list in the field of segmentation, which is an important step for Chinese information retrieval. Section 4 paints the conclusion.

2. Construction of the Stop Word List in Chinese

Stop words, by definition, are those words that appear in the texts frequently but do not carry significant information. As a result, we propose an aggregated model to measure both the word frequency characteristic by statistical model and its information characteristic by information model. A proper segmentation of Chinese texts is required before construction of, because the word boundaries are not clear in Chinese texts. In this section, the texts in a large corpus of Chinese documents are first segmented, and then a standard Chinese stop word list is constructed based on the aggregated model.

2.1 Word segmentation

The difficulty of Chinese word segmentation is mainly due to the fact that no obvious delimiter or marker can be observed between Chinese words except for some punctuation marks. Segmentation methods existing for solving this problem of Chinese words include dictionary-based methods [18], statistical-based methods [8]. Other techniques that involve more linguistic information, such as syntactic and semantic knowledge [7] have been reported in the natural language processing literature. Although numerous approaches for word segmentation have been proposed over the years, none has been adopted as the standard. Since segmentation is not the main objective in our methodology, in our paper, we focus on a statistical approach using mutual information, called the boundary detection segmentation, which has been already proved to be effective [19].

Mutual information is to calculate the association of two events. In Chinese segmentation, mutual information of two characters shows how closely these characters associated with each another. Equation (1) shows the computation of mutual information of bi-grams "AB", where P(A,B) denotes the joint probability of two characters, and P(A), P(B) denote probabilities of character 'A' and 'B' respectively.

$$I(A,B) = \log_2\left(\frac{P(A,B)}{P(A)\times P(B)}\right) \qquad (1)$$

If the characters are mutually independent, P(A, B) equals to P(A)×P(B), so that I(A, B) equals 0. If 'A' and 'B' are highly correlated, I(A,B) will have a high value.

We first calculate the bi-grams and tri-grams mutual information of all the characters in documents. Based on these values and the change of values of the mutual information in one sentence, one can detect the segmentation points with the threshold value chosen manually.

2.2 Statistical model (SM)

A statistical analysis was conducted on a corpus of 423 English articles in TIME magazine (total 245,412 occurrences of words), top 50 words of which with their frequency are shown in Table 1. Stop words are ranked at the top with much larger frequency than the other words. On the other hand, stop words are also those words with quite a stable distribution in different documents. Statistics of the distribution of word frequencies in different documents (Table 2) offers a good demonstration. A combination of these two observations redefines the stop words as those words with stable and high frequency in documents.

Traditional models extract stop words only based on the accumulated frequency without considering the distribution of words among documents. With the statistical results illustrated in table 1 and table 2, we propose to extract the stop words according to the overall distribution of words. Since mean and variance are two important measurements of a distribution, we extract stop words based on these two criteria.

Table 1. Top 50 words with their frequencies from 423 short TIME magazine articles (245,412 word occurrences, 1.6 MB)

Word	Freq.	Word	Freq.	Word	Freq.	Word	Freq.	Word	Freq.
The	15861	his	1815	U	955	Were	848	Britain	589
Of	7239	Is	1810	Had	940	Their	815	when	579
To	6331	he	1700	Last	930	Are	812	out	577
A	5878	As	1581	Be	915	One	811	would	577
And	5614	on	1551	Have	914	Week	793	new	572
In	5294	by	1467	Who	894	They	697	up	559
That	2507	At	1333	Not	882	Govern	687	been	554
For	2228	It	1290	Has	880	All	672	more	540
Was	2149	from	1228	An	873	Year	672	which	539
With	1839	but	1138	S	865	Its	620	into	518

Table 2. Top 50 words with their variances from TREC 5 English corpus

Word	Var.	Word	Var.	Word	Var.	Word	Var.	Word	Var.
The	33.04	Type	82.14	At	111.7	Their	150.2	Hi	186.6
To	47.94	Language	82.58	Have	112.5	Government	155.5	Would	190.4
Of	49.27	With	84.39	Which	116.9	Been	156.2	About	190.7
In	52.32	By	85.24	From	119.8	But	156.5	More	195.4
A	57.15	article	88.06	Not	120.1	Other	157.5	After	196.0
And	58.55	It	90.13	Will	120.5	All	162.9	Between	196.5
On	70.84	Be	93.87	Was	125.4	Country	163.4	Up	197.6
For	75.37	As	95.92	Also	127.6	Who	175.3	There	198.2
That	76.67	An	110.5	English	130.4	Out	184.3	Or	198.7
Is	81.57	Said	111.7	He	143.6	Were	185.2	online	202.4

First, we measure the mean of the probability (*MP*) of each word in individual document. Suppose there are M distinct words and N documents all together. We denote each word as w_j (j=1, … , M) and each document as D_i (i=1, … , N). For each word w_j, we calculate its frequency in document D_i denoted as $f_{i,j}$. However, the document has different length. In order to normalize the document length, we calculate the probability $P_{i,j}$ of the word w_j in document D_i which is its frequency in the document D_i divided by the total number of words in document D_i. For each word w_j, the *MP* among different documents is summarized as following:

$$MP(w_j) = \frac{\sum_{1 \le i \le N} P_{i,j}}{N} \qquad (2)$$

An experiment has been conduced on a 153MB Chinese corpus consisting both of People's Daily news and Xinhua news from TREC 5 and 6. The top 10 words with highest mean of probability are shown in Table 3.

Table 3. Top 10 words with highest mean of *MP*.

Word	Equvilent Word in English	Mean	Variance
的	Of	0.5926	71.56
和	And	0.1324	72.78
在	In	0.1169	72.23
了	-ed	0.1075	72.98
中国	China	0.0784	75.88
一	One	0.0726	74.52
为	For	0.0670	74.01
有	Have	0.0661	79.79
三	Three	0.0535	80.46
等	etc.	0.0491	74.86

Since stop words should have high *MP* as well as stable distribution, the variance of probability (*VP*) of each word is calculated secondly. Based on the calculation of probability, the *VP* is defined by the standard formula:

$$VP(w_i) = \frac{\sum_{1\le i\le N}(P_{i,j} - \overline{P}_{i,j})^2}{N} \qquad (3)$$

The top 10 words with highest Variance of Probability are shown in Table 4.

Table 4. Top 10 words with lowest *VP*.

Word	Equvilent Word in English	Mean	Variance
的	Of	0.5926	71.56
在	In	0.1169	72.23
和	And	0.1324	72.78
了	-ed	0.1075	72.98
为	For	0.0670	74.01
一	One	0.0726	74.52
中	in/middle	0.0523	74.79
上	above/on/up	0.0431	74.80
等	etc.	0.0491	74.86
积极	Active	0.0117	75.04

Intuitively, the probability of a word to be a stop word is directly proportional to the mean of probability, but inversely to the variance of probability. A combination of these two criteria comes to the final formula. We call it the statistical value (*SAT*) of word w_j.

$$SAT(w_j) = \frac{MP(w_i)}{VP(w_i)} \qquad (4)$$

With all these values, a descending ordered lists will be generated. Those ranked in the top will have a larger chance to be considered as stop words in this model.

Table 5. Top 10 words with highest *SAT*.

Word	Equvilent Word in English	Mean	Variance
的	Of	0.5926	71.56
在	In	0.1169	72.23
和	And	0.1324	72.78
了	-ed	0.1075	72.98
为	For	0.0670	74.01
一	One	0.0726	74.52
中	in/middle	0.0523	74.79
等	etc.	0.0491	74.86
上	above/on/up	0.0431	74.80
中国	China	0.0784	75.88

In table 3 and 4, words like “三” (three) and “积极” (active), with only high *MP* or lower *VP*, will not show up at the top of table *SAT*. On the contrary, words like “的 ”(of) “和”(and) “在”(in) ranked at the top of all the tables, have both high *MP* and low *VP*.

2.3 Information model (IM)

From the viewpoint of information theory, stop words are also those words which carry little information. Entropy, one of the fundamental measurements of information [5], offers us another method for better describing stop word selection.

The basic concept of entropy in information theory is a measure to count that how much randomness is in a signal or in a random event. An alternative way to look at this is to talk about how much information is carried by the signal. As an example, consider some English text, encoded as a string of letters, spaces and punctuation (so our signal is a string of characters). Since some characters are not very likely (e.g. ‘z’) while others are very common (e.g. ‘e’) the string of characters is not really as random as it might be. On the other hand, since we cannot predict what the next character will be, it does have some ‘randomness’ and the randomness of each character will be different. Entropy is a measure of this randomness, suggested by Claude E. Shannon in his 1948 paper [14]. This could easily be applied to the Chinese text processing. Consider the distribution of each word over documents as an information channel. The high the entropy of this information channel is, the less random the character would be in all documents. Thus we measure the information value of the word w_j by its entropy.

The probability $P_{i,j}$ is its frequency in the document D_i divided by the total number of words in document D_i. We calculate the entropy value (H) for word w_j as following:

$$H(w_j) = \sum_{1\le i\le N} P_{i,j} * \log(\frac{1}{P_{i,j}}) \tag{6}$$

Similarly to statistical model, one ordered list is prepared for further aggregation. The higher entropy the word has, the lower information value of the word is. Therefore, the words with lower entropy are extracted as candidates of stop words.

Table 6. Top 10 words with highest *H*.

Word	Equvilent Word in English	Entropy
的	Of	0.0177
和	And	0.0059
在	In	0.0054
了	-ed	0.0050
中国	China	0.0039
一	One	0.0037
为	For	0.0034
有	Have	0.0034
三	Three	0.0028
中	In/middle	0.0028

In Table 6, words like "三" (three) and "中" (of), have lower *H* compared with those words such as "的"(of) "和"(and) "在" (in), which have high *H*. It is obviously that "的"(of) "和"(and) and "在" (in) would have better chances to be considered as stop word candidates.

2.4 Aggregation

The ordered lists generated according to two models reveal the features of stop words in different manners, which are all quite reasonable. How to get an aggregation of them? What kind of rules could assure the fairness of the final result? One of the popular solutions to it should be Borda's Rule [12], which covers all the binary relations even when many members of a population have a cyclic reference given a set of voters.

The sorted lists from each model $\{S_1, S_2\}$ are treated the voters' preferences and all the words $\{t_1, t_2, \ldots, t_m\}$ are considered as alternatives. Each model gives out a list $\{t_{j,1}, t_{j,2}, \ldots, t_{j,m}\}$ of all words in non-increasing order. We associate the number 1 with the most preferred word $t_{j,1}$ on the list, 2 with the second $t_{j,2}$ and so on. For all the words, we assign to each of them the number equal to the sum of the numbers all the models assigned to it. The ranking of all the words according to these weights is proposed finally.

3. Experiment and Analysis

To demonstrate the effectiveness of our methodology and to achieve a common Chinese stop word list, we experiment with TREC 5 and 6 Chinese corpora, which contain news reports from both Xinhua newspaper and People's daily newspaper. These corpora cover different aspects of our daily life which ensures the general applicability of our stop word list. The comparison of the list generated by our algorithm with an English stop word list shows that the intersection rate is very high. To clearly show the benefit of our stop word list to Chinese information retrieval, an experiment on segmentation using the Chinese stop word list we extracted is proposed. Experiment tells that our Chinese stop word list facilitates the process of word segmentation in Chinese information retrieval by increasing the accuracy of segmentation, which will result in a better performance in retrieval as well.

3.1 Generation of the Stop Word List

Experiments are conducted on a 153MB Chinese corpus consisting both of People's Daily news and Xinhua news from TREC 5 and 6. We eliminate all the non-Chinese symbols in the preprocessing step. Each uninterrupted Chinese character sequence is kept on one line in the transformed data. On the other hand, phrases like “新华社 ”(Xinhua News Agency), “人民日报”(people’s daily) and “完”(end), that are parts of the news’ format of Xinhua and People’s Daily corpora, are removed. We apply our methodology on these preprocessed documents and collect two ordered lists before aggregation, namely, statistical list and information list. These two lists are aggregated together to generate the final one.

From the viewpoint of linguistics, similar to English stop words, Chinese stop words are usually those words with parts of speech like adjectives, adverbs, prepositions, interjections, and auxiliaries. Adverb “的”(of), preposition “在”(in), conjunction “因为” (because of) and “所以”(so) are all examples. According to different domains, we could classify all stop words into two categories. One kind is called “generic stop words", which are stop words in the general domain.Another kind is document-dependent stop words. We call them “domain stop words". For example, words “Britain” and “govern” in the Zipf list (Table 1) are not included in most generic stop word list, because they are domain stop words of TIME magazine. That’s why in our preprocessing, we eliminated those words such as “新华社” (Xinhua News Agency), which are domain stop words in our news articles.

3.2 Comparison of English and Chinese Stop Word Lists

We compare of our Chinese stop word list with a general English stop word list in table 7. We find that most of the Chinese stop words have corresponding words in English stop word list. For example, word “的”(of) with “of”, “和”(and) with “and”. However, the specialty of sChinese stop word lists is that some words might have the same meaning, like “和”(and) and “与” (and), both of which means “and”. Another aspect worth mention is that Chinese stop word lists should be treated differently compared with English stop word lists. As known, the meaning of Chinese word might change a lot according to the neighbors. This phenomenon changes the usage of Chinese stop word lists a little bit. Recommended usage of Chinese stop word lists here in further task is to use a factor weakening the weight of these words instead of eliminating at one time. The advantage of this usage will be demonstrated in the segmentation application afterwards.

A detail comparison between the Chinese stop word list generated in our algorithm and the stop word list of Brown corpus [6], which is a well known and widely used corpus in English ,is done (Table 8). The result shows that the percentage of stop words intersected among two stop word lists is very high, which means that our stop word list in Chinese is comparable with English.

Table 7. Partial of theChinese Stop List and the general English Stop List

Chinese Stop Words	English Stop Words
的(of), 和(and), 在(in), 了(-ed), 一(one), 为(for), 有(have),	the, of, and, to, a, in,
中(in/middle), 等(etc.), 是(is), 上(above/on/up), 与(and),	that, is, was, he, for,
年(year), 对(to), 将(will/shall/would), 到(at/to), 从(from),	it, with, as, his, on, be,
不(not), 说(say), 目前(now/nowadays/present),	at, by, I, this, had, not,
百分之(percent), 还(also/and), 地(-ly), 并(also/else),	are, but, from, or,
使(cause/make), 他(he), 多(many/more/much), 进行(-ing),	have, an, they, which,
这些(these), 但是(but), 同(and/with), 一个(an/one),	you, were, her, all,
这个(the/this), 之后(after), 下(below/down), 有关(about),	she, there, would,
于是(so/therefore/thus), 而(moreover), 但是(but/however),	their, we, him, been,
也(also), 向(to), …	has, when, who, will,
	…

Table 8. Overlapping Comparison of the Chinese Stop List and the general English Stop List

No. of Stop Words at the Top of List	Overlapping of English and Chinese Stop List
100	81%
200	89%
300	92%

3.3 Stop Word Lists in Segmentation

The importance of word segmentation in Chinese text information retrieval has drawn attention of many researchers. Experiments prove that the effect of segmentation on retrieval performance is ineluctable. Better recognition of a higher number of words generally contributes to the improvement of Chinese information retrieval effectiveness.

As we mentioned, numerous methods have been proposed for segmentation, while none of them took into consideration of Chinese stop word lists. Either because of no standard list is available up till now, or nobody pays attention to these little words. Study on Chinese segmentation found that a large percentage of segmentation errors come from common single-character words, such as "的"(of), "是"(be), and "个"(ge) [4]. Therefore, we implemented a Chinese segmentation method using our stop word list to demonstrate its effectiveness.

We implemented the segmentation method by boundary detection, which is the same as the segmentation method we used in the construction section. The major modification is that while calculating the bi-grams and tri-grams mutual information, we will multiply the final values with factor 0.5, if any proper substring of exists in the stop word list. On the opposite, the values will be multiplied by a factor of 1.5 if the whole string is matched with entries in the stop word list. The reason for this modification is mainly because of the ambiguity of Chinese words or characters. Our motivation is to avid wrong elimination.

The mutual information equations will be modified as following:

$$\begin{cases} I'(a,b) = I(a,b) \times 0.5 & \text{if proper substring of } S \in \text{Stop List} \\ I'(a,b) = I(a,b) \times 1.5 & \text{if } S \in \text{Stop List} \\ I'(a,b) = I(a,b) & \text{Otherwise} \end{cases} \quad \textbf{(7)}$$

This approach purposes to detect the segmentation points. If any proper substring of a bi-gram or tri-gram appears to be a stop word means that it might be quite possible that it is not a word, so that the value of the point should be reduced and less than original.

中国 改革 和 发展的全局继续保持了稳定

(The progress of reformation and development of China keeps stable)

Without stop words:

[中国] [改革和发展的] [全局] [继续] [保持了] [稳定]

With stop words:

[中国] [改革] [和] [发展] [的] [全局] [继续] [保持] [了] [稳定]

Fig. 1. Comparison of segmentation results with and without using stop words.

In order to measure the accuracy of the segmentation result, we have measure the precision and recall of our segmented text against manually segmented texts. These training texts and test data are taken from a recent international Chinese segmentation competition. A sentence is taken as an example to show the difference of the segmentation results of our method and the original segmentation method (Figure1). Differences occur in the identification of words like "的" (of), "和" (and) and "了" (-ed). With the help of the stop word list, we could figure out those tiny words and segment correctly. In our experiment, the segmentation recall and precision is greatly improved from original 65.3% and 71.1% to 95.24% and 90.1% respectively. The competition has reported an average precision between 84.2% and 89%, and an average recall between 87.2% and 92.3%. The advantage of our segmentation, compared with other methodology [13], is that we make use of our stop word list in simple segmentation method and improve the performance quite a lot.

4. Conclusion

Chinese stop word lists are indispensable in the research of information retrieval. In the paper, we purpose an automatic algorithm for construction of stop word lists in Chinese and generate a generic Chinese stop word list as well. Comparison between our Chinese stop word list and a standard stop word list in English shows that the percentage of stop words intersection is very high, which verifies the effectiveness of our model. Extensive experiments have been conducted on Chinese segmentation to investigate the effectiveness of the stop word list extracted. The results showed that the stop word list can improve the accuracy of Chinese segmentation significantly. Our stop word extraction algorithm is a promising technique, which saves the time for manual generation. It could be applied into other languages in the future.

References

1. Kuang-hua Chen, Hsin-Hsi Chen, Cross Language Chinese Text Retrieval in NTCIR Workshop: towards Cross Language multilingual Text Retrieval, ACM SIGIR Forum, Volume 35 , Issue 2, 2001,pp.12-19.
2. Lin Du, Yibo Zhang, le Sun, yufang Sun and Jie Han, PM-Based Indexing for Chinese Text Retrieval, *Proceedings of the Fifth International Workshop on Information Retrieval with Asian Languages.*
3. Schubert Foo, Hui Li, Chinese word segmentation and its effect on information retrieval, *Information Processing and Management: an International Journal*, v.40 n.1, p.161-190, January 2004.
4. Xianping Ge,Wanda Pratt, Padhraic Smyth, Discovering Chinese words from unsegmented text, In *SIGIR-99* (Proceedings on the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, August 15-19, 1999, Berkeley, CA USA), pages 271-272.
5. Paul B. Kantor, Jung Jin Lee, The maximum entropy principle in information retrieval, Annual ACM Conference on Research and Development in Information Retrieval *Proceedings of the 9th annual international ACM SIGIR conference on Research and development in information retrieval*, page269-274, 1986.
6. H. Kucera, W. Francis, Computational analysis of present day American English, Providence, *RI: Brown University Press*, 1967.
7. I.M.Liu, Descriptive-unit analysis of sentences: Toward a model natural language processing, *Computer Processing of Chinese Oriental Languages*, Vol. 4, No.4, 1990, pp.314-355.
8. K.T.Lua, and G.W. Gan, An application of information theory in Chinese word segmentation, *Computer Processing of Chinese & Oriental Languages*, Vol. 8, No.1, 1994, pp.115-124.
9. Hiroshi Nakagawa, Hiroyuki Kojima and Akira Maeda, Automatic Term Extraction Based on Perplexity of Compound Words, *IJCNLP 2005*, pp. 269-279.
10. Baeza-Yates Ricardo and Ribeiro-Neto Berthier, Modern Information Retrieval, *Addison Wesley Longman Publishing Co. Inc.*
11. J.van Rijsbergen, Information Retrieval, Second Edition, Department of Computer Science, *University of Glasgow, Butterworths, London,* 1979.
12. Roger B. Myerson, Fundamentals of social choice theory, Discussion Paper No. 1162, September,1996

13. Wei-Yun Ma, Keh-Jiann Chen, Introduction to CKIP Chinese Word Segmentation System for the First International Chinese Word Segmentation Bakeoff, *Proceedings of ACL, Second SIGHAN Workshop on Chinese Language Processing*, pp. 31–38.
14. E.Shannon, mathematical theory of communication, *Bell System Technical Journal*, vol. 27, July and October, 1948, pp. 379--423 and 623--656.
15. Sproat, R., Emerson, T.: The First International Chinese Word Segmentation Bakeoff, *The Second SIGHAN Workshop on Chinese Language Processing*, Sapporo, Japan, July 2003.
16. Sproat, and C.L.Shih, A statistical method for finding word boundaries in Chinese text, *Computer Processing of Chinese & Oriental Languages*, vol.4, no. 4, 1990, pp. 336-351.
17. Salton, G., Buckley, C.: Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing and Management*, 24, (1988), 513-523.
18. Zimin Wu and Tseng Gwyneth, Chinese text segmentation for text retrieval achievements and problems, *Journal of the American Society for Information Science*, Vol. 44, No.9, 1993, pp.531-542.
19. Christopher C. Yang, JohnnyW.K. Luk, Stanley K. Yung, and Jerome Yen, Combination and Boundary Detection Approaches on Chinese Indexing, *Journal of the American Society for Information Science*, Vol. 51, No.4, 2000, pp.340-351.
20. Yiming Yang, Noise Reduction in a Statistical Approach to Text Categorization, Proceedings of SIGIR-95, *18th ACM International Conference on Research and Development in Information Retrieval.*
21. K. Zipf, Selective Studies and the Principle of Relative Frequency in Language, Cambridge, *MA: MIT Press*, 1932 .
22. DTIC-DROLS English Stop Word List http://dvl.dtic.mil/stop_list.html
23. An English stop word list in WordNet www.d.umn.edu/~tpederse/Group01/WordNet/words.txt