Proceedings of the 5th WSEAS International Conference on Applied Computer Science, Hangzhou, China, April 16-18, 2006 (pp1010-1015)

Automatic Construction of Chinese Stop Word List

Feng Zou, Fu Lee Wang, Xiaotie Deng, Song Han, Lu Sheng Wang
Computer Science Department
City University of Hong Kong
Kowloon Tong, Hong Kong

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 a stop word list in Chinese language. Result analysis shows that our stop list is comparable with a general English stop word list, and our list is much more general than other Chinese stop lists as well. Our stop word extraction algorithm is a promising technique, which saves the time for manual generation and constructs a standard. It could be applied into other languages in the future.

Key-Words: stop word list, statistical modeling, information theory

1 Introduction

In information retrieval, a document is traditionally indexed by words [10, 11, 14]. 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, 18].

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 [17].

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 [18]. The English stop word lists available online [19][20] 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 stop word lists [1][2][3][9], 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 growth of online Chinese documents and the rapid increase of research interest in Chinese information retrieval, constructing a general Chinese stop word list 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.

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

information theory. Results from these two models are aggregated to generate the Chinese stop word list.

The rest of the paper is organized as following. Section 2 covers the methodology for the discovery of Chinese stop word list. Section 3 analyzes the result of the stop word list extraction experiments. Section 4 paints the conclusion.

2 Construction of 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 stop word list, 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 [15], 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 [16].

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.

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

If the characters are independent to one another, the *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.

2.2 Statistical Model (SM)

A statistical analysis was conduced on a corpus of 423 English articles in TIME magazine (total 245,412 occurrences of words), top 40 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.

Word	Freq.	Word	Freq.	Word	Freq.	Word	Freq.
The	15861	his	1815	U	955	Were	848
Of	7239	is	1810	had	940	Their	815
To	6331	he	1700	last	930	Are	812
A	5878	as	1581	Be	915	One	811
And	5614	on	1551	have	914	Week	793
In	5294	by	1467	who	894	They	697
That	2507	at	1333	not	882	Govern	687
For	2228	it	1290	has	880	All	672
Was	2149	from	1228	An	873	Year	672
With	1839	but	1138	S	865	Its	620

Table 1. Top 40 words with highest frequencies from 423 short TIME magazine articles (245,412 word occurrences, 1.6 MB)

Word	Var.	Word	Var.	Word	Var.
The	33.04	By	85.24	Wa	125.4
To	47.94	article	88.06	Also	127.6
Of	49.27	It	90.13	English	130.4
In	52.32	Be	93.87	He	143.6
A	57.15	As	95.92	Their	150.2
And	58.55	An	110.5	Government	155.5
On	70.84	Said	111.7	Been	156.2
For	75.37	At	111.7	But	156.5
That	76.67	Have	112.5	Other	157.5
Is	81.57	Which	116.9	All	162.9
Type	82.14	From	119.8	Country	163.4
Language	82.58	Not	120.1	Who	175.3
With	84.39	Will	120.5	Out	184.3

Table 2. Top 40 words with highest variances from TREC 5 English corpus

Traditional models extract stop words only based on the accumulated frequency without considering the distribution of words among documents. With the statistic results illustrated in Table 1 and Table 2, we purpose 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 two criteria.

First, we measure the mean of 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:

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

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:

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

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.

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

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.

In Table 1 and 2, words like “三” (three) and “积极” (active), with only high *MP* (Table 3) or lower *VP* (Table 4), will not show up at the top of Table *SAT* (Table 5). 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*.

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

Table 3. Top 10 words with highest *MP*

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

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
等	etc.	0.0491	74.86
上	above/on/up	0.0431	74.80
中国	China	0.0784	75.88

Table 5. Top 10 words with highest *SAT*

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 [13]. 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:

$$H(w_j) = \sum_{1 \le i \le N} P_{i,j} \times \log(\frac{1}{P_{i,j}}) \quad (5)$$

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.

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? The same problem was faced before social choice theory came into being. 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.

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

Table 6. Top 10 words with highest entropy H

Denote $\{S_1, S_2, \ldots, S_n\}$ as voters, and $\{t_1, t_2, \ldots, t_m\}$ as alternatives. Each voter gives out a list $\{t_{j,1}, t_{j,2}, \ldots, t_{j,m}\}$ of all alternatives in non-increasing order of his preference. For each individual voter S_i (i=1, …, n), we associate the number 1 with his most preferred alternative $t_{j,1}$ 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 voters assigned to it. The ranking of all the alternatives is proposed afterwards.

We apply this method to our final result generation to ensure the fulfillment of most of the ordered lists. The sorted lists from each model are treated as the voters' preferences and all the words are considered as alternatives. Each word will be associated with the sum of its rank in different lists. If word "的"(of) ranks at the top in the list generated by statistical model and ranks at the second in the list generated by information model, it will be associated with the weight 3. In the final ranking, a sorted list according to this weight is proposed.

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 corpura, which contain news reports from both Xinhua newspaper and People's daily newspaper. These corpura 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. Meanwhile, a comparison with other Chinese stop lists gave a better proof of its effectiveness and generality. The stop list generated with our methodology constructs a standard for the Chinese stop list in the future.

3.1 Stop Word List Generation

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 corpura, 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 linguistic, similar to English stop words, Chinese stop words are usually those words with part of speeches 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 English and Chinese Stop Word Lists Comparison

We give a comparison of our Chinese stop word list and 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 Chinese stop word list 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 list should be treated differently compared with English stop word list. As known, the meaning of Chinese word might change a lot according to the neighbors. This phenomenon changes the usage of Chinese stop word list a little bit. Recommended usage of Chinese stop word list 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.

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

Table 7. Partial of our Chinese Stop List and the general English Stop List

A detail comparison between Chinese stop word list generated in our algorithm and 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.

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

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

3.3 Chinese Stop Word Lists Comparison

To better prove the effectiveness of our Chinese stop word list, we compare our Chinese stop word list with other Chinese stop word lists available. For example, the stop list used in the application of Chinese term extraction from web pages [9], is listed in Table 9.

本月(this month), 乒乓(Ping Pang: onomatopoeia), 扑通(Pu Tong: onomatopoeia), 比较(Comparatively),毕竟 (after all), 必定 (absolutely),必然(absolutely),嘻嘻(Xi Xi: onomatopoeia),也不 (neither/nor),别看 (though), 别说 (though),何必 (no need to), 哎呀 (EiYa: exclamation), 我国 (my country), 起来 (begin to),来着(-ed), 啊 (A: a kind of exclamation),按(according to),吧 (Ba: a kind of exclamation), 被 (-ed), 比 (compared with/to),彼 (another),必 (definitely), 边 (while),便(then), 别 (don’t),并(and/with),不(no),到(unto), 等(else),点(a little bit),顶 (very/quite),都(both),对(to), 吨 (ton), 多(more),而 (while/ although), 尔(you)

Table 9. Partial of another Chinese Stop List

List in Table 9 does not contain some important tiny words like “的” (of), “和” (and) and “了” (-ed), which are considered to be quite helpful for Chinese segmentation [4]. Meanwhile, another difference of these two lists occurs at their applicability. Compare with our list, list in Table 9 lacks some generality. Words like “乒乓”(Ping Pang: onomatopoeia) and “扑通”(Pu Tong: onomatopoeia) could either be considered as stop words or content words in different situations. Besides, this list should come from corpora consisted of some local spoken languages. It includes many spoken phrases which are seldom used in formal Chinese language. Even though most of the words in Table 9 could be considered as candidates of stop words, this list could not be considered as a standard for further use. On the opposite, our list avoids this disadvantage. Since it is generated from large TREC news data instead of some local languages, its generality outperforms others.

4 Conclusion

Chinese stop word list is 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. Our stop list is comparable with a general English stop word list. Besides our list is much more general than other Chinese stop lists. Our stop word extraction algorithm is a promising technique, which saves the time for manual generation and constructs a standard. It could be applied into other languages in the future.

References:

[1] K.H. Chen, H.H. Chen, Cross Language Chinese Text Retrieval in NTCIR Workshop: towards Cross Language multilingual Text Retrieval, *ACM SIGIR Forum*, 35(2):12-19. 2001.

[2] L. Du, Y. Zhang, L. Sun, Y. Sun and J. Han, PM-Based Indexing for Chinese Text Retrieval, *Proceedings of the Fifth International Workshop on Information Retrieval with Asian Languages*.

[3] S. Foo, H. Li, Chinese word segmentation and its effect on information retrieval, *Information Processing and Management: an International Journal*, 40(1):161-190. 2004.

[4] X. Ge, W. Pratt, P. Smyth, Discovering Chinese words from unsegmented text, Proc. 22nd annual international ACM SIGIR conference on Research and development in information retrieval, August 15-19, 1999, Berkeley, CA USA, pp. 271-272.

[5] P.B. Kantor, J.J. 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*, 269-274, 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*, 4(4):314-355. 1990.

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

[9] H. Nakagawa, H. Kojima, A. Maeda, Chinese term extraction from web pages based on compound word productivity, *IJCNLP*, 269-279, *2005.*

[10] B.Y. Ricardo, R.N. Berthier, Modern Information Retrieval, *Addison Wesley Longman Publishing Co. Inc*.

[11] J. Rijsbergen, Information Retrieval, Second Edition, Department of Computer Science, *University of Glasgow, Butterworths, London*, 1979.

[12] R.B. Myerson, Fundamentals of social choice theory, Discussion Paper No.1162, 1996

[13] E. Shannon, mathematical theory of communication, *Bell System Technical Journal*, 27:, 379-423 and 623-656. 1948.

[14] G. Salton, C. Buckley, Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing and Management*, 24:513-523. 1988.

[15] Z. Wu, G. Tseng, Chinese text segmentation for text retrieval achievements and problems, *Journal of the American Society for Information Science*, 44(9):531-542. 1993.

[16] C.C. Yang, J.W.K. Luk, S.K. Yung, and J. Yen, Combination and Boundary Detection Approaches on Chinese Indexing, *Journal of the American Society for Information Science*, 51(4):340-351. 2000.

[17] Y. 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*.

[18] K. Zipf, Selective Studies and the Principle of Relative Frequency in Language, Cambridge, *MA; MIT Press*, 1932.

[19] DTIC-DROLS English Stop Word List http://dvl.dtic.mil/stop_list.html

[20] English Stop Word List in WordNet http://www.d.umn.edu/~tpederse/Group01/WordNet/words.txt