Evaluation of Stop Word Lists in Chinese Language

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

Department of Computer Science, 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 the evaluation of Chinese stop word lists becomes critical. In this paper, to save the time and release the burden of manual comparison, we propose a novel stop word list evaluation method with a mutual information-based Chinese segmentation methodology. Experiments have been conducted on training texts taken from a recent international Chinese segmentation competition. Results show that effective stop word lists can improve the accuracy of Chinese segmentation significantly.

1. Introduction

In information retrieval, a document is traditionally indexed by frequency of words in the documents (Ricardo & Berthier, 1999; Rijsbergen, 1975; Salton & Buckley, 1988). Statistical analysis through documents showed that some words have quite low frequency, while some others act just the opposite (Zipf, 1932). 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 (Raghavan & Wong, 1986; Ricardo & Berthier, 1999; Zipf, 1932; Zou, Wang, Deng & Han, 2006).

The application of stop words has already been explored in many fields. In digital libraries, for instance, elimination of stop words could significantly reduce the size of the indexing structure and obtain a compression ratio of more than 40% (Ricardo & Berthier, 1999). On the other hand, a stop word list provides a good resource for information retrieval. It can speed up the calculation and increase the accuracy at the same time (Salton & Buckley, 1988; Yang, 1995).

Up to now, a lot of stop word lists have been developed for English language. These stop word lists are traditionally extracted by frequency analysis of all the words in a large corpus (Yang, 1995). Results from different corpora are usually quite similar to each other and they are commonly used as standards (Zipf, 1932). Different from English language, no commonly accepted stop word list has been constructed yet for Chinese language. Some research work on Chinese information retrieval makes use of manual stop word lists (Chen & Chen, 2001; Du, Zhang, Sun, Sun & Han, 2000; Nakagawa & Hojima, 2005), others might automatically generate stop word list. These Chinese stop word lists vary a lot to each other. None of them has been accepted as a standard.

In order to produce a stop word list which is widely accepted as a standard, it is extremely important to compare the performance of different stop word list. On the other hand, it is also essential to investigate how a stop word list can affect the completion of related tasks in language processing. With the fast growth of online Chinese documents, the evaluation of these stop word lists becomes quite an essential topic. In this paper, we propose a novel stop word list evaluation method with a Chinese segmentation methodology called boundary detection (Yang, Luk, Yung, Yen, 2000; Yang & Li, 2005) based on mutual information.

We aim at comparing the performances of segmentation with and without using stop word lists to evaluate the effectiveness of these lists. As known to all, many research works in Chinese segmentation have made use of mutual information (Sproat & Shih, 1990), which is used to calculate how strongly these characters are associated with one another. We make some modifications on the mutual information-based boundary detection methodology. A new factor is added into the segmentation process, which purposes to help detect the segmentation points and demonstrate the effectiveness of the Chinese stop word list used.

The rest of the paper is organized as following. Section 2 covers the methodology of the Chinese segmentation algorithm and our modification to the algorithm. Section 3 presents the experimental results. Section 4 gives a short conclusion.

2. Chinese Word Segmentation

The importance of Chinese 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.

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 (Wu & Tseng, 1993), statistical-based methods (Yang, Luk, Yung & Yen, 2000; Lua & Gan, 1994). Other techniques that involve more linguistic information, such as syntactic and semantic knowledge (Liu, 1990) 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 (Yang, Luk, Yung & Yen, 2000).

2.1 Boundary Detection Segmentation

As known to all, many research works in Chinese segmentation have made use of mutual information (Sproat & Emerson, 2003), which 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. 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.

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

If the characters are independent to one another, *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. A threshold value is chosen to extract all the bi-grams from the text.

The computation of mutual information of a tri-gram "*ABC*", either a combination of a bi-gram "*AB*" with a unigram "*C*" or a combination of a unigram "*A*" with a bi-gram "*BC*", is shown in equation (2) and (3):

$$I(A,BC) = \log_2\left(\frac{P(A,BC)}{P(A)\times P(BC)}\right) \quad (2)$$

$$I(AB,C) = \log_2\left(\frac{P(AB,C)}{P(AB)\times P(C)}\right) \quad (3)$$

Similarly, the computation of mutual information of a quad-gram "*ABCD*", is calculated as in equation (4), (5) and (6):

$$I(A,BCD) = \log_2\left(\frac{P(A,BCD)}{P(A)\times P(BCD)}\right) \quad (4)$$

$$I(AB,CD) = \log_2\left(\frac{P(AB,CD)}{P(AB)\times P(CD)}\right) \quad (5)$$

$$I(ABC,D) = \log_2\left(\frac{P(ABC,D)}{P(ABC)\times P(D)}\right) \quad (6)$$

Algorithm was proposed to segment Chinese text based on the mutual information, (Yang, Luk, Yung & Yen, 2000). Texts are divided into trunks with *n* consecutive characters that are called *n*-grams. We calculate the mutual information of adjacent characters in the *n*-grams to determine the segmentation points.

The detailed algorithm is given below:

Segmentation Algorithm:

1. *Counting occurrence frequencies*
   Obtain occurrence frequencies for all possible *n*-gram, for *n* = 1 to 4.

2. *Extracting bi-grams*
   Compute mutual information for all *n*-grams. Determine the bi-grams with the highest mutual information value and remove it from the sentence. Repeat the removal of bi-grams until no more bi-gram existing in the sentence or the mutual information values are less than a threshold, T_1. (Sproat & Shih, 1990).

3. *Combining the extracted bi-grams and uni-tri-grams to form tri-grams*
   Compute the mutual information for all combinations of uni-gram and bi-gram to form tri-grams, i.e., (bi-gram. uni-gram) or (uni-gram, bi-gram). Combine the bi-grams and uni-grams with the highest mutual information value until no such patterns exist in the sentence or the mutual information values are less than a threshold, T_2.

4. *Combining the extracted tri-grams, bi-grams and uni-grams to form quadra-grams*
   Compute the mutual information for all combinations of uni-grams, bi-grams and tri-grams to form quad-grams, i.e., (uni-gram, tri-gram), (bi-gram, bi-gram), or (tri-gram, uni-gram). Combine the uni-grams, bi-grams, or trigrams with the highest mutual information value until no such patterns exist in the sentence or the mutual information values are less than a threshold, T_3.

The boundary detection segmentation methodology first calculates 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.

2.2 Modified Boundary Detection Segmentation

In order to demonstrate the effectiveness of stop word list, we make some modifications on this mutual information-based boundary detection methodology.

A new factor is added into the segmentation process. While calculating the bi-grams, tri-gram and quad-grams mutual information, we will multiply the original values of

mutual information with a factor of 0.5, if any proper substring exists in the stop word list. On the opposite, a factor of 1.5 will be multiplied to these values if the whole string is matched with some entry in 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 of an *n*-gram X will be modified as following:

$$\begin{cases} I'(X) = I(X) \times 0.5 & \text{if proper substring of } X \in \text{Stop List} \\ I'(X) = I(X) \times 1.5 & \text{if } X \in \text{Stop List} \\ I'(X) = I(X) & \text{otherwise} \end{cases} \quad (7)$$

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

3. Experiments

We have done experiments on training texts taken from a recent international Chinese segmentation competition (Sproat & Emerson, 2003). The benefit of this test data is that all the texts have already been segmented, which offers us convenience to evaluate experiment results.

We compare the performances of segmentation with and without using stop word lists. Experiments show that differences occur in the identification of words like "的", "和" and "了", which have already been demonstrated to be quite essential words in Chinese processing (Ge, Pratt & Smyth, 1999). Stop word lists with these tiny words outperform those without them. With the help of effective stop word lists, we could figure out these words and segment all the texts correctly.

Here is an example. In one of the experiments, we make use of the stop word list in (Zou, Wang, Deng & Han; 2006), a part of which is listed in figure 1. The segmentation of a sentence with and without this stop word list is illustrated in figure 2.

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), 从(from),
不(not), 将(will/shall/would), 到(at/to), 说(say),
地(-ly), 使(cause/make),
目前(now/nowadays/present), 他(he),
百分之(percent), 也(also), 还(also/and), 向(to),
并(also/else), 多(many/more/much),进行(-ing),
这些(these), 但是(but), 之后(after),同(and/with),
一个(an/one), 这个(the/this), 下(below/down),
而(moreover), 于是(so/therefore/thus),
但是(but/however)

Figure 1: part of a Chinese stop word list

In our experiments, by using this effective stop word list, the average segmentation recall and precision is greatly improved from original 65.3% and 71.1% to 95.24% and 90.1% respectively. Compared to the average precision between 84.2% and 89%, and the average recall between 87.2% and 92.3% reported in that competition, it is a great improvement. Experiments show that segmentation using an effective stop word list outperforms segmentation without a stop word list significantly. This evaluation methodology illustrates the effectiveness of Chinese stop word lists, which can improve the accuracy of Chinese segmentation.

4. Conclusion

Chinese stop word lists is indispensable in the research of information retrieval of Chinese language. In the paper, we propose a novel segmentation algorithm for the evaluation of stop word lists in Chinese language. Experiments have been conducted on a large corpus to investigate the effectiveness of the stop word lists. The results show that an effective stop word list can improve the accuracy of Chinese segmentation significantly. Our stop word evaluation algorithm is a promising technique, which gives out the effective proof of different kinds of Chinese stop word lists. It could be applied into other languages in the future.

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 is still keeping stable)

Segmentation without stop words:

[中国] [改革和发展的] [全局] [继续] [保持了] [稳定]
[China] [reformation and development of] [progress] [still] [keeping] [stable]

Segmentation with stop words:

[中国] [改革] [和] [发展] [的][全局] [继续] [保持][了] [稳定]
[China][reformation][and][development][of][progress] [still] [keep][-ing] [stable]

Figure 2: Comparison of segmentation results with and without using stop words.

References

Chen K.H., and Chen H.H., (2001), Cross-Language Chinese Text Retrieval in NTCIR Workshop: towards Cross-Language multilingual Text Retrieval. *ACM SIGIR Forum*, Vol. 35, No. 2, pp. 12-19.

Du L., Zhang Y.B., Sun L., Sun Y.F., and Han J., (2000), PM-Based Indexing for Chinese Text Retrieval, *In Proceedings of Fifth International Workshop on Information Retrieval with Asian Languages*, pp. 55-59, Hong Kong.

Ge X.P., Wanda P., Padhraic S. (1999), Discovering Chinese Words from Unsegmented Text. *In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR 1999*, Berkeley, pp. 271-272.

Liu I.M., (1990), Descriptive-unit analysis of sentences: Toward a model natural language processing. *Computer Processing of Chinese Oriental Languages*, Vol. 4, No. 4, pp. 314-355.

Lua K.T., and Gan G.W., (1994), An application of information theory in Chinese word segmentation. *Computer Processing of Chinese & Oriental Languages*, Vol. 8, No.1, pp. 115-124.

Nakagawa H., Kojima H., and Maeda A., (2005), Chinese term extraction from web pages based on compound word productivity, *IJCNLP* 2005, pp. 269-279.

Raghavan V.V., and Wong S.K.M, (1986) A critical analysis of vector space model for information retrieval. *Journal of the American Society for Information Science*, Vol. 37, No. 5, pp. 279-287.

Ricardo B.Y., Berthier R.N., (1999), Modern Information Retrieval. Addison Wesley Longman Publishing, Boston.

Rijsbergen C.V., (1979), Information Retrieval. Butterworths, London, 1975.

Sproat R., and Emerson T. (2003). The First International Chinese Word Segmentation Bakeoff. *In Proceedings of The Second SIGHAN Workshop on Chinese Language Processing*, Sapporo, Japan, pp. 133-143.

Sproat R., and Shih C. (1990). A statistical method for finding word boundaries in Chinese text. *Computer Processing of Chinese and Oriental Languages*, Vol. 4, pp. 336-351.

Salton G., and Buckley C., (1988) Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing and Management*, Vol. 24, pp. 513-523.

Wu Z., Tseng G. (1993). Chinese text segmentation for text retrieval: Achievements and problems. *Journal of the American Society for Information Science*, Vol. 44, No.9, pp.531-542.

Yang C.C., Luk J.W.K., Yung S.K., and Yen J., (2000), Combination and Boundary Detection Approaches on Chinese Indexing. *Journal of the American Society for Information Science*, Vol. 51, No. 4, pp. 340-351.

Yang C.C., and Li K.W. (2005). A Heuristic Method Based on a Statistical Approach for Chinese Text Segmentation. *Journal of the American Society for Information Science and Technology*, Vol. 56, No. 13, pp. 1438-1447.

Yang Y.M., (1995), Noise Reduction in a Statistical Approach to Text Categorization. *In Proceedings of the 18th ACM International Conference on Research and Development in Information Retrieval (SIGIR'95).*

Zipf G.K. (1949). Human Behavior and the Principle of Least Effort. *Addison-Wesley*.

Zou F., Wang F.L., Deng X., Han S.,. (2006), Automatic Identification of Chinese Stop Words. *A special issue on Advances in Natural Language Processing of the journal Research on Computing Science*, ISSN 1665-9899.