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 Proccesing
Research in Computing Science 18, 2006, pp. 151-162