Word Frequency Approximation for Chinese without Using Manually Annotated Corpus

Sun Maosong[1] , Zhang Zhengcao[1] , Benjamin K Y T'sou[2], Lu Huaming[3]

[1] The State Key Laboratory of Intelligent Technology and Systems,
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
sms@tsinghua.edu.cn
[2] Language Information Sciences Research Center, City University of Hong Kong
rlbtsou@cityu.edu.hk
[3] School of Business, Beijing Institute of Machinery, Beijing 100085, China
huaminglu@sohu.com

Abstract. Word frequencies play important roles in a variety of NLP-related applications. Word frequency estimation for Chinese is a big challenge due to characteristics of Chinese, in particular word-formation and word segmentation. This paper concerns the issue of word frequency estimation in the condition that we only have a Chinese wordlist and a raw Chinese corpus with arbitrarily large size, and do not perform any manual annotation to the corpus. Several realistic schemes for approximating word frequencies under the framework of STR (frequency of string of characters as an approximation of word frequency) and MM (Maximal matching) are presented. Large-scale experiments indicate that the proposed scheme, MinMaxMM, can significantly benefit the estimation of word frequencies, though its performance is still not very satisfactory in some cases.

1 Introduction

Word frequencies play important roles in a variety of NLP-related applications, for example, TF in information retrieval. The estimation of word frequencies for English is very easy – it can be done by running a simple program to count word occurrences in a (in fact, any arbitrarily huge) corpus. In case of Chinese where no explicit word boundaries like spaces exist between words in texts, the task becomes very complex.

In general, a fully correct word-segmented Chinese corpus is a prerequisite for calculating word frequencies (Liu 1973). However, we face two difficulties in this respect. The first one is such a 'fully correct' corpus, or, a corpus with ideal segmentation consistency, is extremely difficult to obtain due to a main characteristic of Chinese word-formation: the borders between morphemes, words, and phrases of Chinese are fuzzy in nature (Dai 1992; Chen 1994), though the definition for 'word' from the linguistic perspective seems very clear (Zhu 1982;Tang 1992). A large number of linguistic constituents could be regarded as words by some linguists whereas be regarded as phrases by others (even for a specific linguist, his feeling to some constituents may change in between from time to time), resulting in serious