Word Extraction Based on Semantic Constraints in Chinese Word-formation

Maosong Sun [1], Shengfen Luo,[1] and Benjamin K T'sou [2]

[1] National Lab. of Intelligent Tech. & Systems, Tsinghua University, Beijing 100084, China
sms@mail.tsinghua.edu.cn

[2] Language Information Sciences Research Centre, City University of Hong Kong
rlbtsou@cityu.edu.hk

Abstract. This paper presents a novel approach to Chinese word extraction based on semantic information of characters. A thesaurus of Chinese characters is conducted. A Chinese lexicon with 63,738 two-character words, together with the thesaurus of characters, are explored to learn semantic constraints between characters in Chinese word-formation, forming a semantic-tag-based HMM. The Baum-Welch re-estimation scheme is then chosen to train parameters of the HMM in the way of unsupervised learning. Various statistical measures for estimating the likelihood of a character string being a word are further tested. Large-scale experiments show that the results are promising: the F-score of this word extraction method can reach 68.5% whereas its counterpart, the character-based mutual information method, can only reach 47.5%.

1 Introduction

Processing of unknown words is important for Chinese word identification in running texts. New words are generated quite often with the rapid development of Chinese society. In experience, the accuracy of a word identification system will decrease about 10% if unknown words are not treated properly [12].

Chinese is an isolating language. Methods for processing of unknown words in inflective languages, like, for example [5], may not be appropriate for Chinese because of its different morphological structure. A Chinese word is composed of either single or multiple Chinese characters. In most cases, a Chinese character has at least one sense, and can stand independently at the morphological level. The task of extracting Chinese words with multi-characters from texts is quite similar to that of extracting phrases (e.g., compound nouns) in English, if we regard Chinese characters as English words.

Researches in this field have been done extensively. Generally, there are two kinds of methods for word/phrase extraction, i.e., rule-based and statistic-based. The latter has become the mainstream of the state-of-the-art. In statistic-based approaches, the soundness of an extracted item being a word/phrase is usually estimated by the associative strength between constituents of it. Two widely used statistical measures for quantifying the associative strength are frequency and mutual information [1, 2, 7,