Computational Linguistics and Chinese Language Processing
vol. 3, no. 1, February 1998, pp. 27-44
©Computational Linguistics Society of R. O. C.

Unknown Word Detection for Chinese by a Corpus-based Learning Method

Keh-Jiann Chen*, Ming-Hong Bai*

Abstract

One of the most prominent problems in computer processing of the Chinese language is identification of the words in a sentence. Since there are no blanks to mark word boundaries, identifying words is difficult because of segmentation ambiguities and occurrences of out-of-vocabulary words (i.e., unknown words). In this paper, a corpus-based learning method is proposed which derives sets of syntactic rules that are applied to distinguish monosyllabic words from monosyllabic morphemes which may be parts of unknown words or typographical errors. The corpus-based learning approach has the advantages of: 1. automatic rule learning, 2. automatic evaluation of the performance of each rule, and 3. balancing of recall and precision rates through dynamic rule set selection. The experimental results show that the rule set derived using the proposed method outperformed hand-crafted rules produced by human experts in detecting unknown words.

1. Introduction

One of the most prominent problems in computer processing of Chinese language is the identification of the words in a sentence. There are no blanks to mark word boundaries in Chinese text. As a result, identifying words is difficult because of segmentation ambiguities and occurrences of out-of-vocabulary words (i.e., unknown words). For instance, in (1), the proper name 王英雄 'Wang, Ying-Xiong' is a typical example of an unknown word, and it has ambiguous segmentation of 王 'king' 英雄 'hero'. Another example in (1) 台灣大學生 'university student in Taiwan' also has ambiguous segmentations of 台灣 'Taiwan' 大學生 'university student' , 台灣大學 'National Taiwan University' 生 'give birth to' ,and 台灣 'Taiwan' 大學 'university' 生 'give birth to' etc.:

(1) 王英雄是一個典型的台灣大學生。
'Ying-Xiong Wang is a typical university student in Taiwan.'

* Institute of Information Sicence, Academia Sinica, Taipei, Taiwan, R. O. C. E-mail: {kchen, evan}@iis.sinica.edu.tw