Using Directed Graph based BDMM Algorithm for Chinese Word Segmentation

Yaodong Chen,[1] Ting Wang,[2] and Huowang Chen

National Laboratory for Parallel and Distributed Processing,
Changsha, Hunan, P.R.China 410073
[1] ydchen0104@yahoo.com.cn [2] wonderwang70@hotmail.com

Abstract. Word segmentation is a key problem for Chinese text analysis. In this paper, with the consideration of both word-coverage rate and sentence-coverage rate, based on the classic Bi-Directed Maximum Match (BDMM) segmentation method, a character Directed Graph with ambiguity mark is designed for searching multiple possible segmentation sequences. This method is compared with the classic Maximum Match algorithm and Omni-segmentation algorithm. The experiment result shows that Directed Graph based BDMM algorithm can achieve higher coverage rate and lower complexity.

1 Introduction

Word segmentation (WS) is a key problem in Chinese text processing. Many methods have been proposed, such as Forward Maximum Match (FMM), Backward Maximum Match (BMM) and Bi-Directed Maximum Match (BDMM), which are fast and simple, but deficiency in disambiguation. The accuracy of segmentation is vital to further processing, such as POS tagging and parsing. Because of the ambiguity in WS, multi-level linguistic knowledge should be considered [1]. It is rational to reserve multi-candidates of segmentation for further processing rather than only one result. On the other hand, too many redundant candidates produced by Omni-segmentation seem to cause low efficiency. Both FMM and BMM are viewed as an extreme in WS (the most simple way, producing only one candidate) and Omni-segmentation is another (the most complex one, producing all the possible candidates but most of which are incorrect), what we need is a tradeoff that gets over the segmentation blindness and the explosion of candidates [2]. This paper applies the BDMM with a character directed graph annotated with ambiguity mark, which aims to include all rational segmentation candidates and exclude the wrong ones for further processing.

2 Directed Graph based BDMM Algorithm (DGBS)

2.1 The Idea of DGBS

Given a sentence $S = c_1c_2\ldots c_n$ where c_i (i = 1, 2, ..., n) is Chinese character, a segmenting method M can produce a candidate set $T= \{T_1 ,..., T_l\}$ according to S. Let W