Chapter 14

A Linguistically Sortable Bengali Coding System and Its Application in Spell Checking: A Case Study of Multilingual Applications

M. Manzur Murshed
Monash University, Australia

Mahbubur Rahman Syed
Minnesota State University, USA

M. Kaykobad
Bangladesh University of Engineering and Technology, Bangladesh

INTRODUCTION

In the last decade, the use of Bengali scripts in daily computer usage has gained wide acceptance in Bangladesh. Although a wide range of commercial Bengali software have been developed so far to meet the ever-growing demand in the local market, a systematic and scientific efforts of integrating Bengali in modern computing systems remains in its infancy. One of the most important issues related to Bengali computing is to sort Bengali texts in order of linguistic order.

Sorting Bengali words is not same as sorting English words in ASCII. There are a number of coding schemes available in the market. Although Bangladesh Standards and Testing Institution (1996) has recently standardized a coding scheme to be used in all future Bengali text processing, the other existing coding schemes are likely to dominate for a few more years because of the availability of popular commercial software. Unlike English, Bengali letters can take more than one forms

Previously Published in Managing Information Technology in a Global Economy edited by Mehdi Khosrow-Pour, Copyright © 2001, Idea Group Publishing.
and two or more letters are sometimes combined into compound characters. Although a coding scheme should assist in both displaying characters as well as sorting texts efficiently, all of the available Bengali coding schemes, including the standard one, have placed emphasis more on displaying Bengali texts as close as possible to the actual forms.

It was surprising to learn that no Bengali coding scheme was available which could sort Bengali texts in complete linguistic order until we have recently proved in (Murshed and Kaykobad, 1998) that no completely linguistically sorted Bengali coding scheme exists. In (Murshed et al., 1998), we have further introduced an internal coding scheme, in addition to the primary coding scheme, to provide a linguistically sortable Bengali coding system. The internal coding scheme is made linguistically sortable by avoiding all the compound characters and introducing some artificial half-characters to compensate that. The compound letters of the primary scheme is supported by providing non-lossy transformation from the primary coding scheme to the internal scheme and vice versa.

Rahman and Iqbal (1998) have recently developed a very similar sorting algorithm using lossy transformation from the primary coding scheme to the internal scheme. The difference between lossy and non-lossy transformations lies on the fact that the later transformation is reversible, i.e., non-lossy transformation can also be used to convert a word from the internal coding scheme to the primary coding scheme.

By developing an inter-scheme text conversion utility, we have established in (Murshed et al., 1998) that use of non-lossy transformation instead of lossy transformation for sorting Bengali texts in linguistic order has some extra benefit. In this paper we discuss another very important application of non-lossy transformation by developing an efficient spell checking application for Bengali texts based on the internal coding scheme with non-lossy transformation. As usual, the handling of compound letters remains the key area where a Bengali text speller differs from its counterparts in other languages. Here we establish that using of the internal coding scheme in designing the dictionary and developing suggestion generating search engine not only provides a spell checking solution which is independent of any specific primary coding scheme but also assists in designing layered solution for efficient modularization and maintenance of coding.

This chapter is organized as follows. In the next section we present the basic properties of Bengali script. For the sake of completeness, some results and algorithms on sorting Bengali texts in linguistic order, developed in (Murshed et al., 1998), are given in the third section. In the fourth section, we discuss various issues of developing an efficient primary coding scheme independent spell checking application based on our solution to linguistically sorting Bengali texts. The final section concludes the paper.
A Component-Oriented Approach for Mixed Reality Applications
www.igi-global.com/chapter/component-oriented-approach-mixed-reality/8268?camid=4v1a