

A Comprehensive Roman (English)-to-Bangla Transliteration Scheme

**Naushad UzZaman, Arnab
Zaheen and Mumit Khan**

*Center for Research on Bangla
Language Processing*

**BRAC University
Bangladesh**

International Conference on
Computer Processing of Bangla
(ICCPB 2006)
17 February, 2006
Dhaka, Bangladesh

Outline

- ◆ Introduction on transliteration
- ◆ Discuss proposed transliteration scheme
- ◆ Performance of our proposed scheme

Introduction

◆ Transliteration

- In narrow sense: mapping of letters from one script to another

◆ Transcription

- Writing the sound of one language using the script of another language

◆ Transliteration

- In broader sense: both

Why transliteration?

- ◆ If you don't have a good text input system for your script
- ◆ Want to write the text phonetically in another script
 - e.g. chatting and smsing in English
 - Language is not too phonetic
 - ◆ dokkhin for দক্ষিণ (দ ক ষ ি ণ)

Previous work

- ◆ Roman (English) to Non-European languages
 - English - Japanese/Chinese/Arabic and many other languages
- ◆ Roman (English) to Bangla
 - ITRANS 1991
 - And few others
 - All of these are one-to-one mapping

Direct Mapping

◆ Trivial phonetic mapping that maps letters from one script to another.

- k - ক

- s - স

Why another direct mapping?

◆ Many-to-one mapping

- phul, phool, ful, fool - ফুল

◆ Using this direct mapping in phonetic mapping algorithm as well

Phonetic mapping

◆ Transcription + transliteration in narrow sense

- ottonto - অত্যন্ত (অ ত ্য ন ্ত)
- shondha - সন্ধ্যা (স ন ্ ধ ্য া)
- bebohar - ব্যবহার (ব ্য ব হ া র)

◆ Solution: Phonetic encoding

Phonetic encoding

◆ *Code a word based on its pronunciation.*

- অত্যন্ত - <ottnt>
- সন্ধ্যা- <shndha>
- ব্যবহার - <bebhar>

◆ Naushad UzZaman and Mumit Khan, *A Double Metaphone Encoding for Bangla and its Application in Spelling Checker*, Proc. IEEE NLP KE, Wuhan, China, 2005

Algorithm for Phonetic Mapping

- ◆ Generate phonetic codes for all Bangla words
- ◆ Input Bangla word using Roman (English) characters
- ◆ Generate phonetic code string of the input
- ◆ IF input's phonetic code
 - Matches only one word in the lexicon
 - ◆ Then convert input to that Bangla word; e.g. *ottonto* - অত্যন্ত
 - Matches to multiple words in the lexicon
 - ◆ Then produce suggestions of all relevant Bangla word and let the user select; e.g. *poddo* - পদ্ব/পদ্য
 - Does not match
 - ◆ Then use direct mapping; e.g. *ultapalta* - উলতাপালতা

Performance

◆ Problems:

- Input word does not exist in the lexicon
- Inflected form of the head word is missing
 - ◆ সরকার
 - ◆ সরকারের

◆ Solution

- Increase the lexicon size
- Use morphological generator to produce inflected forms

Performance on 2500 Newspaper words

- ◆ Words found in the lexicon: 68% (We had lexicon with more than 100,000 entries)
 - Given the word is in the lexicon, the instances it was handled properly by phonetic mapping with phonetic lexicon: 100%
- ◆ Words not found in the lexicon: 32%
 - direct mapping handles: 23%
 - absence of inflected words: **7%**
 - not handled properly by direct mapping: **2%**

Cross Language Information Retrieval Application

- ◆ User issues a query in one language to search a collection in a different language.
- ◆ Search সন্ধ্যা in a Bangla document querying shondha in English

Summary

- ◆ Transliteration
- ◆ Proposed direct mapping and phonetic mapping
- ◆ Prototype shows significant success
 - 91% accuracy for this sample set using lexicon of 100,000 entries
 - Can achieve 100% by increasing lexicon size
- ◆ Used in cross language information retrieval application

Acknowledgment

- ◆ Supported in part by the PAN Localization Project (www.panl10n.net), grant from the International Development Research Center, Ottawa, Canada and BRAC University.