A Framework for Computational Processing of Spelling Variation

Anil Kumar Singh
Language Technologies Research Centre
International Institute of Information Technology, Hyderabad, India

Abstract

Many languages like Hindi, especially those which have been used as (spoken) link languages and don't have a long history of standardization, allow variant spellings of the same words. One of the reasons for this is the influence of the writer's first language or dialect. In this paper we present a computational framework for predicting spelling variations based on the speaker's dialect. This method is mainly based on using a phonetic model of scripts. Such a method is especially suited for Brahmi origin scripts which are used for most major Indian languages. The same method can also be used for studying regional variation given some written corpora in the concerned dialects as well as for determining the standard or the normalized form given a variant.

The phonetic model used here differs from others such models in that it has been defined in terms of mainly phonetic features (vowel, voiced, etc.) augmented with some orthographic features. Other models mostly try to map graphemes to phonemes directly.

Our focus in the current work is on the framework for computational processing of phonological and spelling variation rather than on specific (statistical or rule based) techniques like hidden markov modelling (HMM). The idea was to come up with a framework under which we can experiment with various such techniques. Such a framework is needed for languages like Hindi because a lot of errors while trying to solve natural language processing (NLP) problems occur because of the non-standard spellings and dialectal or regional variations.

The framework that we propose would consists of the following components:

- A phonetic model of Indian language scripts, which includes a function for measuring the phonetic distance (or similarity) between two letters or words
- A 'trie' based compilation tool for lexicon to enable fast and easy search
- A syllable or *akshar* forming component, as the Indian language scripts we are considering are syllabic in nature, in addition to having phonetic properties
- A decoder which uses the lexicon 'trie' and the distance function to determine the standard or normalized form corresponding to the spelling variations
- A generator to generate possible spelling variations given a standard or normalized form
- A tool for studying spelling (and indirectly, phonological) variation given annotated corpora

The decoder can use techniques like Gaussian mixture based HMM combined with dynamic time warping algorithm, as is done in speech recognition systems. The generator can be rule based or could use some statistical approach.

As far as the corpus is concerned, we have manually marked up two Hindi novels written by native speakers of two different 'dialects of Hindi' and are considered major 'regional' novels in Hindi. These novels contain a lot of 'Hindi words' with spellings which correspond to their phonological variants in those dialects. We have also marked up two major English novels by Indians which contain many Hindi words. These will be used for studying the spelling variations of Hindi words or Indian names in written English.