A Study of Sinhala Spelling Error Patterns for Spelling Error Correction

Himesha Wijekoon

Abstract:

Spelling error detection & correction techniques are used widely by word processing, machine translation, information retrieval and natural language processing systems. Even though it is straight forward to verify a misspelled word by looking up in a word dictionary, it is very hard to suggest the best correction. For a morphologically rich and a complex Indic language like Sinhala a probabilistic method is the best approach for qualifying the best correction for a detected misspelled word over the other existing methods. This research intends to identify & analyze non-word spelling error patterns in Sinhala. A word dictionary will be used to identify the errors and a special software tool will be developed in order to record statistical data regarding the spelling errors of Sinhala documents. This tool will be used by a Sinhala language expert to record data related to spelling errors in a selected sample of documents. Errors will be categorized into different types along with statistical results and will be analyzed. The reasons of language specific error patterns will be discussed and a weight based decision tree format will be derived as an outcome which can be used to find the best correction from a word dictionary to replace a misspelled word.

Keywords: Sinhala, spelling error patterns, spell checking, non-word errors