International Conference on Computing, Mechanical and Electronics Engineering (ICCMEE'2015), July 9-10, 2015 Singapore

Statistical Analysis of Sinhala Spelling Error Patterns for Spelling Error Correction

Himesha Wijekoon, and Gamini Wijayarathne

***Abstract*—** Spelling error detection & correction techniques are used extensively 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 difficult 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. Initially spelling errors are categorized based on the literature. A word dictionary is used to identify the errors and a special software tool is developed in order to record statistical data regarding the spelling errors of Sinhala documents. This tool is used by a Sinhala language expert to record statistical data related to spelling errors in a selected sample of documents. A total of 18795 words are analyzed with the help of this software tool. 94.5% of total misspelled words are single error words while very small occurrences are found for two or more error words, first character misspelled words, run-on words and split words. Spelling errors are analyzed for insertion, deletion, substitution and transposition error types as well. Then these errors are further analyzed up to the character level in order to find the language specific patterns. Finally the validity of these statistics are tested for the significance. The possible reasons of language specific error patterns are discussed and a weight based decision tree format is derived as an outcome which can be used to find the best correction from a word dictionary to replace a misspelled word.

***Keywords*—**Non-word errors, Sinhala, spell checking, spelling error patterns.

Himesha Wijekoon is with the Department of Industrial Management, University of Kelaniya, Sri Lanka (corresponding author's phone: 094718325763; e-mail: himesha@kln.ac.lk).

Gamini Wijayarathne is with the Department of Industrial Management, University of Kelaniya, Sri Lanka (e-mail: gamini@kln.ac.lk).