Improved Automatic Spell Checker for Malay Blog

Abstract

Automatic Spell Checker for Malay Blog is a spell checker approach that is designed to detect and correct the Malay misspelled words automatically. The approach is able to detect and automatically correct misspelled words in Malay with minimal interactions from the user. The approach automatically replaces the misspelled word if it exists in the reSpellWord.txt dictionary. Otherwise, it will go through either one of these processes entitled "Selangor" slang identification process, repetitive word identification process or the opposite word identification process. If the word still cannot be identified by applying those processes, the misspelled word will go through the process called a "spelling embodiment," where a few alternative words will be suggested and they are ranked using the Levenshtein Distance in order to choose the most likelihood word for the misspelled word. The correctly-spelled alternative word that has the highest ranking will be chosen as a replacement for the misspelled word. This misspelled word and its correctly-spelled word are then added automatically into the predefined dictionary (e.g., reSpellWord.txt dictionary) in order to update the dictionary. In short, this paper proposes an improved approach to automatically identify and correct misspelled Malay blog texts. Processes that are embedded into the proposed approach includes filtering of the proper name and the alphanumeric words, adding rules into the stemmer and adding the n-gram formula in order to improve the spelling embodiment result. Based on the experimental results obtained, the proposed approach is found to be effective in reducing the percentages of error in detecting and correcting the Malay misspelled word automatically.