Automatic Spell Checker for Malay Blog

Abstrak

Spell checker is a system that is used to detect and correct misspelled word. Misspelled word is a word that exists in the existing lexicon that is not correctly spelled or in shortened form. These misspelled words often result in ineffective results of the Information Retrieval (IR) application such as document retrieval. This is because IR application should be able to recognize all words in a particular language in order to be more robust. The current spell checker for the Malay language uses a dictionary that contains pair of commonly misspelled word and its correctly spelled word in detecting and correcting misspelled word. However, this type of spell checker can only correct misspelled words that exist in the existing dictionary; otherwise it requires user interaction to correct it manually. This approach works well if the spell checker is a standalone system but it is not really an effective system when the spell checker is part of another IR application such as document retrieval for weblog. This is because there will be always new misspelled words created along with the increasing number of weblog pages. Thus, the number of misspelled words will also grow extremely. In this paper, we propose a new spell checker that detects and automatically corrects misspelled words in Malay without any interaction from the user. The proposed approach automatically replaces the misspelled word if it exists in the reSpellWord dictionary. Otherwise, it will go through the process of Selangor Slang Identification or Repetitive word Identification or Opposite Word Identification whichever is suitable. If the word cannot be identified as a misspelled word, 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 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 dictionary in order to update the dictionary. The proposed approach is evaluated by using texts that are selected randomly from the popular Malay blog. Based on the

experimental results obtained, the proposed approach is found to be effective in detecting and correcting the Malay misspelled word automatically.