Abstract

Indonesian grapheme-to-phoneme (G2P) conversion represents a task of mapping each grapheme / spelling symbol in any Indonesian word to its phonemic representation / pronounciation symbol

A selection for the best method is in this final project results in determining a model called IG-tree + best-guess strategy as the chosen model to solve G2P conversion problem. The model is basically in decision-tree structure built based on a trainingset, constructed using concept of information gain (IG) in weighing the relative importance of attributes, and equipped with the best-guess strategy in classifying new instances. However, the system is in this final project leveraged with new features added to its pre-existing structure to improve its performance. A pruning mechanism is proposed for the model for two objectives: (1) improving its generalization ability, and (2) minimizing its dimension. Another new feature, the homograph case handler using a text-categorization method, is proposed for the system to handle its special case of a few sets of words which are exactly the same in graphemic representations but are different each other in phonemic representations.

It is shown in this final project that the model in general performs well while the additional features really give additional benefits as expected.

Keywords: grapheme-to-phoneme conversion, Indonesian, IG-tree, best-guess strategy, pruning, homograph-case handler