ABSTRACT


This thesis aims to perform text feature weighting for summarization of document bahasa Indonesia using genetic algorithm. There are eleven text features, i.e., sentence position (f1), positive keywords in sentence (f2), negative keywords in sentence (f3), sentence centrality (f4), sentence resemblance to the title (f5), sentence inclusion of name entity (f6), sentence inclusion of numerical data (f7), sentence relative length (f8), bushy path of the node (f9), summation of similarities for each node (f10), and latent semantic feature (f11). We investigate the effect of the first ten sentence features on the summarization task. Then, we use latent semantic feature to increase the accuracy. All feature score functions are used to train a genetic algorithm model to obtain a suitable combination of feature weights. Evaluation of text summarization uses F-measure. The F-measure directly related to the compression rate. The results showed that adding f11 increases the F-measure by 3.26% and 1.55% for compression ratio of 10% and 30%, respectively. On the other hand, it decreases the F-measure by 0.58% for compression ratio of 20%. Analysis of text feature weight showed that only using f2, f4, f5, and f11 can deliver a similar performance using all eleven features.

Keywords: text summarization, genetic algorithm, latent semantic feature