Abstract

Paraphrase identification is one of the process in Natural Languange Processing, in purpose to teach the machine if one query is having same meaning to another. For example, query ”when rain falls” should be recognized to have same meaning with ”when it rains”. The meaning of paraphrase itself is restatement of text or passage using different words. In this research we focusing on classifying two Indonesian sentence whether it is a paraphrase or not. The step itself is preprocessing, classifier building, and performance evaluation.

Preprocessing consist of tokenization, normalization, stepword removal, and stemming. After preprocessing comes feature extraction in order to build new features from dataset. There are two kinds of feature and method to use, first is syntactic feature using Levhenstein Distance, second is semantic feature using Wu and Palmer. After extraction, dataset will be splitted into two kinds of data, data training and data testing. Data training will be used to train classifier, while data testing will be used when testing the performance of classifer. After dataset splitted, then we will making classifier using Naive Bayes as the method of classifer. The result from classifier testing as follows: Accuracy 0.713, Precision 0.688, Recall 0.798, and F1-Measure 0.735.

Keyword : *Text Mining, Naive Bayes, Paraphrase Identification.*