Abstract

Paraphrase identification is an important process within natural language processing. The idea is to automatically recognize phrases that have different forms but contain same meanings. For example if we input query “causing fire hazard”, then the computer has to recognize this query that this query has same meaning as “the cause of fire hazard”. On the other hand, paraphrase is expressing the meaning of (the writer or speaker or something written or spoken) using different words or forms, especially to achieve greater clarity. In this research we will focus on classifying two Indonesian sentence whether it is a paraphrase to each other or not. There are three step in this research, first is preprocessing, second is classifier training, and third is performance evaluation.

Preprocessing consists of tokenization, non-alphanumerical removal, and stemming. After preprocessing we will conduct feature extraction in order to build new features from given dataset. First feature is syntactic which is the result from computation of distance between two sentences using Normalized Levensthein Distance method. The second feature is semantic that is obtained by calculate similarity of pair sentence based on semantic trees. Semantic distance computation using Wu and Palmer method. After feature extraction data will be splitted into two parts, training data and testing data. The purpose of training data is to train/build classifier, while testing data will be used when testing the performance of trained classifier. After the dataset was splitted, then we discretize the features by clustering them using K-Means method. We use Bayesian Networks as the method of training the classifier. Parameter estimation that we use is called MAP(Maximum A Posteriori) and Multinomsial Distribution Probability. The average result that we get from testing the clasifier as follows: Precision 61.2%, Recall 84.8%, Acuration 66.2%, and F1-Measure 71.5%.

Keywords: paraphrase identification, preprocessing, bayesian networks, MAP