Abstract

Natural Languange Processing is one of science disciplines which focusing to generate knowledge from human written text which is not structured. The measuring of Semantic Similarity to word pair is one of the task in Natural Language Processing that the main idea is to find a semantic similarity score to word pair. This score is representing how similar the word pair is. One of methods for measuring semantic similarity is Pointwise Mutual Information$_{max}$ (PMI_{max}). PMImax estimate the maximum correlation to word pair and the closest sense between those two words because a word often has multiple sense or can be called with Word Polysemy.

In this final project is implemented semantic similarity measure to word pair using PMI_{max} with estimates of word polysemy. The context of word sourced from Brown Corpus and Gutenberg dataset. The result of the score compared to gold standard dataset WordSim-353, Miller Charles, and Simlex-999.

Research results obtained by using PMImax shows that the best correlation is 66,5% with WordSim-353 semantic similarity dataset using Pearson correlation and the value of sense on the analysis of variables p and q. The semantic similarity score for each word pair is depend on Co-Occurence value, high Co-Occurence value will produce high semantic similarity score.

Keywords : *Semantic similarity, Pointwise Mutual Information, Word Polysemy*