ABSTRACT

Textual Semantic Similarity is one of the tasks that fall within the Natural Language Processing. The tasks performed included the calculation of similar meanings similarity score of a given text, in this study is word pairs from three different types of dataset which are Simlex-999, WordSim-353 and Rubenstein & Goodenough. From the calculation of these scores, this score of similarity is a prefix of the implementation of the development of further NLP research.

Score calculation is done by using the help of library JWNL, WordNet lexical database, using the method of Resnik, Lin and Jiang Conrath which are the several methods provide by Information Content based measurements. Several condition tested to the environment, such as finding IC value by its frequency and using the latest development of finding IC value, hyponym count. The effect using of with or without sense tagged, and analyze word characteristics by its POS, NOUN and VERB. With the method chosen, the result value shows that Lin method reach the highest correlation among the others with 85.5% on R&G dataset. With using frequency to calculate IC value, shows better result when it sense tagged, an vice versa for hyponym. Word pair that has POS.NOUN shows better value correlation with 59.8% than wordpairs that has POS.VERB.

Keywords: Semantic Textual Similarity, Information Content, Lin, WordNet, SimLex-999, hyponym count