

Abstract

Alquran is a way of life for Muslims and have many interpretations to make it easy to understand. Alquran have many verses in it, some verses in Alquran have similar meanings. To facilitate the search for similarity of the verse then conducted research on the text of the translation of the Quran. Many methods in the field of text mining and NLP can be used to measure semantic similarity between sentences, some of it is use alignment and semantic vectors. Alignment is one of method that can calculate semantic similarity by doing alignment of words. Features of alignment used in this research are identical word, PPDB, word sequence and named entity. While semantic vector is the methods used to calculate the similarity of a word from the distribution of adjacent words. For the use of the semantic vector it needs corpus text8 taken from Matt Mahoney. In order to obtain better results, a combination of alignment method and semantic vector using ridge regression was used. In addition, the manufacture of gold standard was made as a benchmark to determine the value of the correlation and become contribution of the authors to add data that already exist in research last year. The combination of alignment and semantic vector methods to calculate semantic similarity in Alquran translation data in 2016 resulted in a correlation value of 0,90467 higher by 0,002029 from research conducted by Dwi Jayanti, while for the translation of Alquran data collected in 2017 resulted in a correlation value of 0,79756 lower by 0,00593 compared to the system built by Md. Sultan.

Keywords: *text mining, nlp, word alignment, semantic, semantic vector, quran, moslem, regression.*