Abstract

This paper examines the calculation of similarity between words in Indonesian by using the Word2Vec representation technique. Word2Vec is a model used to represent words into vector shapes. The model in this experiment was formed using the 4GB Indonesian Wikipedia corpus and then used the cosine similarity calculation method to determine its similarity value. This model is then tested with the gold standard set WordSim-353 and SimLex-999 which have been labeled with similarity values according to human ratings. To find out the accuracy of the correlation using the Pearson correlation. The results of the correlation of this study are 0.5663 for WordSim-353 test data using window size 14 and dimensions vector 150, and 0.3472 for SimLex-999 test data using window size 2 and dimensions vector 300. The results of the experiments show that the correlation between the gold standard and system techniques is still relatively weak.