Alian M. Awajan A. 2020. Evaluating Factors Affecting Sentences Similarity And Paraphrasing Identification Using K-Means Clustering. Proceeding of the 35th IBIMA conference. 1-2 April, Seville, Spain.

Abstract

Sentence similarity determines whether two sentences are close in their structure and meaning. The detection of sentence similarity can be affected by several factors such as sentence representation, similarity measure, and words weighting function. In this study, the impact of three factors that influence similarity detection and paraphrasing identification is evaluated using clustering algorithms. In the evaluation of the impact of these factors, we tried different word embedding models, clustering algorithms, and weighting methods for the context words. The clustering algorithms are applied to an Arabic paraphrasing benchmark that consists of 1010 pairs of Arabic sentences constructed on the basis of Arabic transformation rules and labeled for similarity and paraphrasing. Experimental results show that pre-trained embedding, weighting context words with part of speech, and labeling sentence pairs by the majority of experts provides better recall and precision.