Marwah Alian, Arafat Awajan,Ahmad Al-Hasan, Raeda Akuzhia, Building Arabic paraphrasing benchmark based on Transformation Rules, ACM Transactions on Asian and Low-Resource Language Information Processing. 2021. ACM, (ISI,Scopus)

Abstract

Measuring semantic similarity between short texts is an important task in many applications of natural language processing, such as paraphrasing identification. This process requires a benchmark of sentence pairs that are labeled by Arab linguists and considered a standard that can be used by researchers when evaluating their results. This research describes an Arabic paraphrasing benchmark to be a good standard for evaluation algorithms that are developed to measure semantic similarity for Arabic sentences to detect paraphrasing in the same language. The transformed sentences are in accordance with a set of rules for Arabic paraphrasing. These sentences are constructed from the words in the Arabic word semantic similarity dataset and from different Arabic books, educational texts, and lexicons. The proposed benchmark consists of 1,010 sentence pairs wherein each pair is tagged with scores determining semantic similarity and paraphrasing. The quality of the data is assessed using statistical analysis for the distribution of the sentences over the Arabic transformation rules and exploration through hierarchical clustering (HCL). Our exploration using HCL shows that the sentences in the proposed benchmark are grouped into 27 clusters representing different subjects. The inter-annotator agreement measures show a moderate agreement for the annotations of the graduate students and a poor reliability for the annotations of the undergraduate students.