Abstract

Semantic similarity is the task of measuring relations between sentences or words to determine the degree of similarity or resemblance. Several applications of natural language processing require semantic similarity measurement to achieve good results; these applications include plagiarism detection, text entailment, text summarisation, paraphrasing identification, and information extraction. Many researchers have proposed new methods to measure the semantic similarity of Arabic and English texts. In this research, these methods are reviewed and compared. Results show that the precision of the corpus-based approach exceeds 0.70. The precision of the descriptive feature-based technique is between 0.670 and 0.86, with a Pearson correlation coefficient of over 0.70. Meanwhile, the word embedding technique has a correlation of 0.67, and its accuracy is in the range 0.76–0.80. The best results are achieved by the feature-based approach.