Suleiman D., Awajan A. 2019. Using Part of Speech Tagging for Improving Word2vec Model. Proceedings of the 2nd International Conference on new Trends in Computing Science (ICTCS'19). 9-11 October 2019, Amman- Jordan.

Abstract

Word2vec is an efficient word embedding model that convert words to vectors by considering syntax and semantic relationship between words. In this paper, an extension of the two approaches of word2vec model is proposed. The proposed model considers part of speech tagging of the words when exploring the probability of prediction output word given the input. Considering part of speech tagging provides deeper semantic meaning for words when training the model. Thus, the quality of the generated vectors becomes high and more representative. In addition, the proposed model equips the user with the ability to query about the words and their part of speech tagging. In this case, the words that have the same surface but different meaning will have different vector representations. This model can be used in all languages including English and Arabic languages. The focus of this paper is on Arabic language. The experiments are performed on OSAC datasets which consists of 22,429 documents. Several pre-processing stages include using Farasa stemmer take place. Moreover, part of speech tagging of the word is determined using Farasa toolkit.