Sedki E., Alzaqah A., Awajan A. 2015. Arabic Text Dimensionality Reduction Using Semantic Analysis. WSEAS Transactions on Information Science and Applications. Volume 12, 2015, Art. #21. Pages 209-218.

Abstract

An efficient method to compress and reduce the dimensionality of Arabic texts using semantic model-based representation of text is introduced. The proposed system creates equivalence classes, where similar words, generated according to the rich productive morphology of the language and based on the stem root-pattern paradigm, are grouped together and represented by a class identifier. In addition, synonyms and similarly named entities are regrouped in order to improve the representation and reduce its size. The reduced representation of the text is accessible to most machine learning algorithms and natural language processing applications that require heavy computational complexity. Distributional similarity measures were used to create equivalence classes of similar words. These measures were applied to the word-context matrix associated with the document in order to identify similar words based on a text's context. The results confirmed that the proposed method shows that incorporation of sematic information in vector representation is superior to classical bag-of-words representation, in terms of size reduction and results quality of applications. The best results are achieved for the clustering of words that are semantically similar, based on their stems. In addition, regrouping differently named entities representing the same concepts improved the reduction amount by 5%.