Awajan A. 2016. Semantic Similarity Based Approach for Reducing Arabic Texts Dimensionality. International Journal of Speech Technology (Springer). June 2016, Volume 19, Issue 2, pp 191–201. DOI. 10.1007/s10772-015-9284-6. 13 Pages

Abstract

In this paper, we introduce an efficient method to represent Arabic texts in comparatively smaller sizes without losing significant information. The proposed method uses the linguistic features of the Arabic language, mainly its very productive morphology and its richness in synonyms, to reduce the dimension of the document vector and to improve its vector space model representation. We have incorporated semantic information from word thesauri like WordNet to create clusters of similar words extracted from the same root and regrouped along with their synonyms. Distributional similarity measures are applied on the word-context matrix associated with the document in order to identify similar words based on a text's context. The experimental results have confirmed that the proposed method significantly reduces the size of text representation by about 20% compared with the stem-based vector space model and by about 40% compared with the traditional bag of words model.