

Awajan A. 2014. Unsupervised Approach for Automatic Keyword Extraction from Arabic Documents, Proceedings of the 26th International Conference on Computational Linguistics and Speech Processing ROCLING 2014, Pages 175-184.

Abstract

In this paper, we present an unsupervised two-phase approach to extract keywords from Arabic documents that combines statistical analysis and linguistic information. The first phase detects all the N-grams that may be considered keywords. In the second phase, the N-grams are analyzed using a morphological analyzer to replace the words of the N-grams with their base forms that are the roots for the derived words and the stems for the non-derivative words. The N-grams that have the same base forms are regrouped and their counts accumulated. The ones that appear more frequently are then selected as keywords. An experiment is conducted to evaluate the proposed approach by comparing the extracted keywords with those manually selected. The results show that the proposed approach achieved an average precision of 0.51.