Abstract

Nahla A Belal

A Rule-Based Subject-Correlated Arabic Stemmer

Arabic is a derivational language that provides invaluable features. Arabic roots are basic forms that are used to formulate words. They are limited sets that encapsulate the word’s linguistic features. The knowledge of roots’ frequencies is a valuable additional feature, especially when it is bound to a specific topic. This paper utilizes collision resulting from the stemming process where two more words may have the same root. It minimizes the number of extracted roots within a specific subject using roots’ frequencies and explores its effect on multiple roots disambiguation.