Abstract

Nahla A Belal

CBAS: Context Based Arabic Stemmer

Arabic morphology encapsulates many valuable features such as word's root. Arabic roots are being utilized for many tasks the process of extracting a word's root is referred to as stemming. Stemming is an essential part of most Natural Language Processing tasks, especially for derivative languages such as Arabic. However, stemming is faced with the problem of ambiguity, where two more roots could be extracted from the same word. On the other hand, distributional semantics is a powerful co-occurrence model. It captures the meaning of a word based on its context. In this paper, a distributional semantics model utilizing Smoothed Pointwise Mutual Information (SPMI) is constructed to investigate its effectiveness on the stemming analysis task. It showed an accuracy of 81.5%, with a at least 9.4% improvement over other stemmers.