A55

Automated Corpora Creation using A Novel Arabic Stemming Algorithm

Eiman Tamah Al-Shammari, Jessica Lin
George Mason University

Text processing is a vital step in the information retrieval process, text mining, and natural language processing. Text processing includes several stages, such as normalization, stop words removal, and stemming. Stemming is the process of reducing the lexicon to its root. Stemming is language dependent approach to reduce a word to its root; this research is introducing a new Arabic stemming algorithm.

Arabic is one of the most complex languages, both spoken and written. However, it is also one of the most common languages in the world. It is also the base from which other languages are derived. Despite the wide usage of the language, technology has been slow in development for Arabic has been limited. The main reason relies on the formulation rules of Arabic. Arabic language exhibits a very complicated morphological structure.

The current Arabic leading stemmers are the root-based stemmer and light stemmer. Over-stemming and under-stemming are the main drawbacks of the root-based stemming and the light stemming algorithms respectively. Over-stemming, under-stemming and mis-stemming are all stemming errors that usually degrade the correctness of stemming algorithms.

Arabic stemmers blindly stem all the words and perform poorly especially with compound words, proper nouns and foreign Arabized words. The main cause of this problem is the stemmer lack of knowledge of the word lexical category (i.e. noun, verb, proposition ...etc.) This paper presents a new stemming Algorithm that relies on Arabic language morphology and Arabic language syntax.

The automated addition to the syntactic knowledge reduced both stemming error and stemming cost. Additionally, the new Algorithm automatically creates a global corpus of proper nouns and compound words based on the processed documents.