Stemming Hausa text: using affix-stripping rules and reference look-up

Abstract

Stemming is a process of reducing a derivational or inflectional word to its root or stem by stripping all its affixes. It is been used in applications such as information retrieval, machine translation, and text summarization, as their pre-processing step to increase efficiency. Currently, there are a few stemming algorithms which have been developed for languages such as English, Arabic, Turkish, Malay and Amharic. Unfortunately, no algorithm has been used to stem text in Hausa, a Chadic language spoken in West Africa. To address this need, we propose stemming Hausa text using affix-stripping rules and reference lookup. We stemmed Hausa text, using 78 affix stripping rules applied in 4 steps and a reference look-up consisting of 1500 Hausa root words. The over-stemming index, under-stemming index, stemmer weight, word stemmed factor, correctly stemmed words factor and average words conflation factor were calculated to determine the effect of reference look-up on the strength and accuracy of the stemmer. It was observed that reference look-up aided in reducing both over-stemming and under-stemming errors, increased accuracy and has a tendency to reduce the strength of an affix stripping stemmer. The rationality behind the approach used is discussed and directions for future research are identified.