Abstract

This paper presents the first part-of-speech (POS) tagging research for Tigrinya (Semitic language) from the newly constructed Nagaoka Tigrinya Corpus. The raw text was extracted from a newspaper published in Eritrea in the Tigrinya language. This initial corpus was cleaned and formatted in plaintext and the Text Encoding Initiative (TEI) XML format. A tagset of 73 tags was designed, and the corpus for POS was manually annotated. This tagset encompasses three levels of grammatical information, which are the main POS categories, subcategories, and POS clitics. The POS tagged corpus contains 72,080 tokens. Tigrinya has a unique pattern of root-template morphology that can be utilized to infer POS categories. Subsequently, a supervised learning approach based on conditional random fields (CRFs) and support vector machines (SVMs) was applied, trained over contextual features of words and POS tags, morphological patterns, and affixes. A rigorous parameter optimization was performed and different combinations of features, data size, and tagsets were experimented upon to boost the overall accuracy, and particularly the prediction of POS for unknown words. For a reduced
tagset of 20 tags, an overall accuracy of 90.89% was obtained on a stratified 10-fold cross validation. Enriching contextual features with morphological and affix features improved performance up to 41.01 percentage point, which is significant.

References


Index Terms

Computer Science Information Sciences

Keywords

Semitic languages, Tigrinya corpus, Tigrinya part-of-speech tagging, morphological patterns