ABSTRACT

Morphological segmentation is a subtask of natural language processing (NLP) that specializes in identifying the constituent morphemes of words in a language. As a subtask of morphological analysis, morphological segmentation is a crucial preprocessing step that improves the overall output of the NLP system. Low resource languages have recently received attention in research, with researchers aiming to improve NLP in these languages. However, the finer details within languages are often overlooked, which has led to low quality results among a variety of studies. Since morphological analysis is an important task in NLP, extracting the morphological syntax of verbs thus remains crucial. In this work, the researcher demonstrates a finite state transducer for the morphological segmentation of the Swahili verb. This research work set out to achieve four objectives, namely to analyze key parameters for the morphological segmentation techniques of Swahili verbs, to implement a web scraper to populate a dataset of Swahili verbs, to integrate morphological segmentation parameters into a finite state transducer for Swahili verb segmentation, and to validate the finite state transducer on the Swahili verb. The model performs morphological analysis of the Swahili verb by identifying morphological slots such as the subject, object, derivational suffixes, and any grammatical errors within the verb. It was implemented as a finite state network built out of regular expressions in an object-oriented programming (OOP) language. The same finite state transducer was also implemented in the Xerox Finite State Tools (XFST). Input verbs were extracted from an online dictionary using a web scrapper and separated into two datasets. Dataset A comprised 163 simple Swahili verbs while dataset B comprised 715 non-Arabic verbs. The OOP model outperformed its XFST counterpart, achieving a 98.77% accuracy on dataset A and 68.67% accuracy on dataset B. The results from the experiments prove that OOP rule-based techniques perform better than their XFST-based counterparts. The research work was quantitative, with the accuracy of the models evaluated using experiments. This work is beneficial in optimizing search engines that use Swahili, where verbal keywords need to be segmented to obtain their root. This work is also pivotal in assisting learners new to Swahili in understanding the structure of the verb and enabling them to explore possible combinations of morphemes that make up a correctly formed verb. Further, the work significantly contributes towards the development of a spell checker, a corpus and a syntax analyzer for Swahili verbs.