Abstract

The need to capture intra-word information in natural language processing (NLP) tasks has inspired research in learning various word representations at word, character, or morpheme levels, but little attention has been given to syllables from a syllabic alphabet. Motivated by the success of compositional models in morphological languages, we present a Convolutional-long short term memory (Conv-LSTM) model for constructing Swahili word representation vectors from syllables. The unified architecture addresses the word agglutination and polysemous nature of Swahili by extracting high-level syllable features using a convolutional neural network (CNN) and then composes quality word embeddings with a long short term memory (LSTM). The word embeddings are then validated using a syllable-aware language model (31.267) and a part-of-speech (POS) tagging task (98.78), both yielding very competitive results to the state-of-art models in their respective domains. We further validate the language model using Xhosa and Shona, which are syllabic-based languages. The novelty of the study is in its capability to construct quality word embeddings from syllables using a hybrid model that does not use max-over-pool common in CNN and then the exploitation of these embeddings in POS tagging. Therefore, the study plays a crucial role in the processing of agglutinative and syllabic-based languages by contributing quality word embeddings from syllable embeddings, a robust Conv–LSTM model that learns syllables for not only language modeling and POS tagging, but also for other downstream NLP tasks.