様式

論　文　内　容　の　要　旨

氏　名　　YEMANE KELETA TEDLA

Tigrinya Morphological Segmentation with Bidirectional Long Short-Term Memory Neural Networks and its Effect on English-Tigrinya Machine Translation
(双方向長・短期記憶ニューラルネットワークを用いたティグリーニャ語の形態素分割と英語ティグリーニャ語機械翻訳への影響)

Natural language has a central role in the communication process of human beings. The field of Natural Language Processing (NLP) is the branch of Artificial Intelligence that enables machines to understand and process human language. NLP products have extensive applications in our day-to-day activities including grammar correction, spam filtering, personal digital assistance, language translation, recommendation systems, and so on. Significant NLP solutions have been reported for resourceful languages such as English. Unfortunately, the same is not true for the greater majority of the world languages. For instance, currently, Google Translate supports about 100 languages out of over 6000 languages in the world. While there are several reasons that contribute to this digital divide, the absence or scarcity of resources is a major bottleneck impeding NLP advances in low-resource languages.

Tigrinya is one of the languages with very limited support of language resources. It is a morphologically rich Semitic language spoken by over seven million people in Eritrea and Ethiopia. We aim to initiate Tigrinya language processing from the foundation by constructing essential annotated and unannotated text corpora and building fundamental NLP components. In resource building, we constructed a news text corpus comprising over 15 million tokens, with a lexicon of over 593,000 unique tokens. We processed this corpus to generate important word lists including Tigrinya stop words and affix lists. The corpus is [illegible]ransliteration, clitics normalization, and text cleaning. In an Earlier research, a part of the corpus containing 72,080 tokens was manually annotated for parts-of-speech (POS). In this research, we constructed another new resource that consists of over 45,000 morphologically segmented tokens extracted from the POS tagged corpus. Moreover, we compiled and properly aligned an English-to-Tigrinya parallel corpus for machine translation research. These resources are employed in the following research.

First, we approached POS tagging as a classification as well as a sequence labeling problem employing support vector machines (SVM) and conditional

random fields (CRF) respectively. We utilized the unique morphological patterns of Tigrinya to boost performance of particularly unknown words. As a result, our method doubled the accuracy of unknown words from around 39% to 80%. Hence, the overall accuracy also improved to about 90.89%. Furthermore, we obtained 91.6% accuracy (state-of-the-art) approaching POS tagging as a sequence-to-sequence labeling using bidirectional long short-term memory (BiLSTM) networks with word embeddings forgoing feature engineering.

Second, we presented the first research of morphological segmentation for Tigrinya. We explored language-independent character and substring features based on CRF. In addition, we obtained state-of-the-art F1 score of 95.07% with BiLSTM networks using concatenated character and word embeddings. This approach does not require feature engineering to extract linguistic information, which is useful for languages lacking sufficient resources. In this research, we explored several Begin-Inside-Outside (B-I-O) tagging schemes to discover the recommended strategy in Tigrinya morphological segmentation.

Finally, we explored English-to-Tigrinya statistical machine translation. Translation from English to the morphologically rich language Tigrinya has several challenges including out-of-vocabulary problem, language model perplexity, and poor word alignment. We introduced shallow and fine-grained morphological segmentation to mitigate these problems and improve convergence of the two languages. Generally, the results showed that translation using the morphologically segmented models can improve translation quality.