World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:18, No:02, 2024

Named Entity Recognition System for Tigrinya Language

Authors : Sham Kidane, Fitsum Gaim, Ibrahim Abdella, Sirak Asmerom, Yoel Ghebrihiwot, Simon Mulugeta, Natnael Ambassager

Abstract : The lack of annotated datasets is a bottleneck to the progress of NLP in low-resourced languages. The work presented here consists of large-scale annotated datasets and models for the named entity recognition (NER) system for the Tigrinya language. Our manually constructed corpus comprises over 340K words tagged for NER, with over 118K of the tokens also having parts-of-speech (POS) tags, annotated with 12 distinct classes of entities, represented using several types of tagging schemes. We conducted extensive experiments covering convolutional neural networks and transformer models; the highest performance achieved is 88.8% weighted F1-score. These results are especially noteworthy given the unique challenges posed by Tigrinya's distinct grammatical structure and complex word morphologies. The system can be an essential building block for the advancement of NLP systems in Tigrinya and other related low-resourced languages and serve as a bridge for cross-referencing against higher-resourced languages.

Keywords : Tigrinya NER corpus, TiBERT, TiRoBERTa, BiLSTM-CRF
Conference Title : ICNLPMSA 2024 : International Conference on Natural Language Processing, Morphological and Semantic Analysis
Conference Location : Jeddah, Saudi Arabia
Conference Dates : February 19-20, 2024