ABSTRACT :

An NMT framework that has been adapted to the domain of low-resourced and morphologically complex languages has been developed for translating from Amharic to English. Character-aware subword tokenization via SentencePiece and the Tanzil corpus allows the system to work around rare and compound words. A trained Transformer encoder-decoder model with multihead attention and feedforward subnets achieved 59.03 BLEU score on the Tanzil corpus compared to the 26.08 RNN with attention baseline. Integration of domain parallel data with underwent subword modeling and the design and development of a low-resource reproducible Transformer pipeline in addition to the consolidation of methodologically relevant parallel data for Amharic to English translation serves as primary and novel contributions of this research. Through the translation, domain adaptation and subword level segmentation pair and the results for the translated text confirms the Amharic to English translation underwent a boosted performance which speaks to the level of improvement of the translation model. This serves as a groundwork as an entry point for further research on new areas pertaining to the neural architectures of machine translations of Semitic and other languages of complicated morphology.