Annotated Guidelines and Building Reference Corpus for Myanmar-English Word Alignment

Nway Nway Han
nwaynwayhan@ucsm.edu.mm
University of Computer Studies, Mandalay, UCSM

Aye Thida
ayethida@ucsm.edu.mm
University of Computer Studies, Mandalay, UCSM

Reference corpus for word alignment is an important resource for developing and evaluating word alignment methods. For Myanmar-English language pairs, there is no reference corpus to evaluate the word alignment tasks. Therefore, we created the guidelines for Myanmar-English word alignment annotation between two languages over contrastive learning and built the Myanmar-English reference corpus consisting of verified alignments from Myanmar ALT of the Asian Language Treebank (ALT). This reference corpus contains confident labels sure (S) and possible (P) for word alignments which are used to test for the purpose of evaluation of the word alignments tasks. We discuss the most linking ambiguities to define consistent and systematic instructions to align manual words. We evaluated the results of annotators agreement using our reference corpus in terms of alignment error rate (AER) in word alignment tasks and discuss the words relationships in terms of BLEU scores.