Graduate School of Science and Technology Master's Thesis Abstract

Laboratory name (Supervisor)	Natural Language Processing (WATANABE TARO (Professor))		
Student ID	2211402	Submission date	2024 / 7 / 19
Name	FREDERIKUS HUDI		
Thesis title	Disentangling Pretrained Representation to Incorporate Unseen Low-Resource Languages in Multilingual Machine Translation		

Abstract

Multilingual neural machine translation aims to encapsulate multiple languages into a single model. However, it requires an enormous dataset, leaving the low-resource language (LRL) underdeveloped. As LRLs may benefit from shared knowledge of multilingual representation, we aspire to find effective ways to integrate unseen languages in a pre-trained model. Nevertheless, the intricacy of shared representation among languages hinders its full utilisation. To resolve this problem, we employed an extra objective of target language prediction, a central language-aware layer, and a monolingual adapter layer to improve representation in integrating LRLs. Focusing on improving LRLs in the linguistically diverse country of Indonesia, we evaluated five languages using a parallel corpus of 1,000 instances each, with experimental results measured by BLEU showing zero-shot improvement of 8.4 from the baseline score of 7.1 to a score of 15.5 at best. Further analysis showed that the gains in performance are attributed more to the disentanglement of multilingual representation in the encoder with the shift of the target language-specific representation in the decoder.