

Title: Indonesian-English Neural Machine Translation

Author: Meisyarah Dwiastuti

Department: Institute of Formal and Applied Linguistics

Supervisor: Mgr. Martin Popel, Ph.D., Institute of Formal and Applied Linguistics

Abstract: In this thesis, we conduct a study on neural machine translation (NMT) for an under-studied language, Indonesian, specifically for English-Indonesian (EN-ID) and Indonesian-English (ID-EN) in a low-resource domain, TED talks. Our goal is to implement domain adaptation methods to improve the low-resource EN-ID and ID-EN NMT systems. First, we implement model fine-tuning method for EN-ID and ID-EN NMT systems by leveraging a large parallel corpus containing movie subtitles. Our analysis shows the benefit of this method for the improvement of both systems. Second, we improve our ID-EN NMT system by leveraging English monolingual corpora through back-translation. Our back-translation experiments focus on how to incorporate the back-translated monolingual corpora to the training set, in which we investigate various existing training regimes and introduce a novel *4-way-concat* training regime. We also analyze the effect of fine-tuning our back-translation models with different scenarios. Experimental results show that our method of implementing back-translation followed by model fine-tuning makes an improvement in our ID-EN NMT systems by around 1.5 BLEU point over a system without back-translation. Our ID-EN NMT systems show a comparable performance with Google Translate on WIT³ TED Talks tst2015-6 and tst2017plus test sets.

Keywords: neural machine translation, domain adaptation, back-translation