A STUDY ON MACHINE TRANSLATION FOR LOW-RESOURCE LANGUAGES

By TRIEU, LONG HAI

Written under the direction of Associate Professor Nguyen Minh Le

Abstract

Current state-of-the-art machine translation methods are neural machine translation and statistical machine translation, which based on translated texts (bilingual corpora) to learn translation rules automatically. Nevertheless, large bilingual corpora are unavailable for most languages in the world, called low-resource languages, that cause a bottleneck for machine translation (MT). Therefore, improving MT on low-resource languages becomes one of the essential tasks in MT currently.

In this dissertation, I present my proposed methods to improve MT on low-resource languages by two strategies: building bilingual corpora to enlarge training data for MT systems and exploiting existing bilingual corpora by using pivot methods. For the first strategy, I proposed a method to improve sentence alignment based on word similarity learnt from monolingual data to build bilingual corpora. Then, a multilingual parallel corpus was built using the proposed method to improve MT on several Southeast Asian low-resource languages. Experimental results showed the effectiveness of the proposed alignment method to improve sentence alignment and the contribution of the extracted corpus to improve MT performance. For the second strategy, I proposed two methods based on semantic similarity and using grammatical and morphological knowledge to improve conventional pivot methods, which generate source-target phrase translation using pivot language(s) as the bridge from source-pivot and pivot-target bilingual corpora. I conducted experiments on low-resource language pairs such as the translation from Japanese, Malay, Indonesian, and Filipino to Vietnamese and achieved promising results and improvement. Additionally, a hybrid model was introduced that combines the two strategies to further exploit additional data to improve MT performance. Experiments were conducted on several language pairs: Japanese-Vietnamese, Indonesian-Vietnamese, Malay-Vietnamese, and Turkish-English, and achieved a significant improvement. In addition, I utilized and investigated neural machine translation (NMT), the state-of-the-art method in machine translation that has been proposed currently, for low-resource languages. I compared NMT with phrase-based methods on low-resource settings, and investigated how the low-resource data affects the two methods. The results are useful for further development of NMT on low-resource languages. I conclude with how my work contributes to current MT research especially for low-resource languages and enhances the development of MT on such languages in the future.

Keywords: machine translation, phrase-based machine translation, neural-based machine translation, low-resource languages, bilingual corpora, pivot translation, sentence alignment