Abstract

Neural machine translation (NMT) is currently considered the state-of-the-art for language pairs with vast amounts of parallel data. In this thesis project, we utilize such systems to provide translations between four languages in the psychology domain, where the biggest challenge is posed by in-domain data scarcity. Therefore, the emphasis of the research is laid on exploring domain adaptation methods in this scenario. We first propose a system for automatically building in-domain adaptation corpora by extracting parallel sentence pairs from comparable articles of *Wikipedia*. To this end, we use supervised classification and regression methods trained on NMT context vector similarities and complementary textual similarity features. We find that the best method for our purposes is a regression model trained on continuous similarity labels. We rerank the extracted candidates by their similarity feature averages and use the top-N partitions as adaptation corpora. In the second part of the thesis we thoroughly examine multilingual domain adaptation by transfer learning with respect to the adaptation data quality, size, and domain. With clean parallel in-domain adaptation data we achieve significant improvements for most translation directions, including ones with no adaptation data, while the automatically extracted corpora prove beneficial mostly for language pairs with no clean in-domain adaptation set. Particularly in these latter cases, the combination of the two adaptation corpora yields further improvements. We also explore the possibilities of reranking N-best translation lists with in-domain language models and similarity features. We conclude that adapted systems produce candidates that can result in a higher improvement in translation performance than the ones of unadapted models, and that remarkable improvements can be achieved by similarity-based reranking methods.