Abstract

In recent years, a number of mehtods for improving the decoding speed of neural machine translation systems have emerged. One of the approaches that proposes fundamental changes to the model architecture are non-autoregressive models. In standard autoregressive models, the output token distributions are conditioned on the previously decoded outputs. The conditional dependence allows the model to keep track of the state of the decoding process, which improves the fluency of the output. On the other hand, it requires the neural network computation to be run sequentially, and thus it cannot be parallelized. Non-autoregressive models impose conditional independence on the output distributions, which means that the decoding process is parallelizable and hence the decoding speed improves. A major drawback of this approach is lower translation quality compared to the autoregressive models. The goal of the non-autoregressive translation research is to find methods that improve the translation quality, while retaining high decoding speed. In this thesis, we explore the research progress so far and identify flaws in the generally accepted evaluation methodology. We experiement with non-autoregressive models trained with connectionist temporal classification. We find that even though our models achieve state-of-the-art performance on the standard WMT 14 benchmark, there is a large room for improvement when we compare non-autoregressive methods to highly optimized autoregressive models.