Non-Autoregressive Neural Machine Translation

Jindřich Helcl
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
V Holešovičkách 2, 180 00, Prague, Czech Republic
helcl@ufal.mff.cuni.cz, +420 951 552 955

Supervisor: **Jan Hajič**
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 118 00, Prague, Czech Republic
hajic@ufal.mff.cuni.cz, +420 951 554 257

Thesis Overview

In the presented thesis, we explore approaches to non-autoregressive neural machine translation (NAR NMT, or NAT) and discuss common flaws in evaluation methodology in the current literature on the topic. We make a point that if these flaws are not properly addressed, the field of non-autoregressive text generation research might lead the progress in a wrong direction.

An emerging research area within NMT, non-autoregressive translation models promise high decoding speed, which is an attractive feature for users without access to powerful hardware on computational clusters. However, this speed-up is gained by assuming conditional independence between the output tokens (and therefore enabling parallel decoding) which is a major modeling constraint.

The thesis consists of 6 chapters, including the introduction and the conclusions. The second chapter gives self-contained overview of NMT in general; along with "historical" approaches such as using LSTM networks, we describe in detail the current state-of-the-art architecture used in wide spectrum of not-only-NLP applications, the Transofrmer model.

The third chapter of the thesis gives a thorough overview of the published research on NAR models. After the general description of what NAR models are and which challenges researchers face when dealing with these models, we

attempt to categorize the existing methods in four non-exclusive groups: those based on relaxing the one-to-one alignment between output positions and tokens in the reference sentence, methods using auxiliary objectives during training, iterative methods, and "miscellaneous" which did not fit into any of the three categories above. For each group, we list the known approaches and outline their ideas. In the end of the chapter, we summarize the research as a whole and point out issues most papers have in commmon, namely weak autoregressive baseline models (which leads to overestimating the performance of NAR models, both in terms of translation quality and speed) and unsound evaluation methodology (inconsistent speed evaluation, only reporting BLEU scores on the WMT 14 test set).

In Chapter 4, we present our starting point: Non-autoregressive NMT using connectionist temporal classification (CTC). The chapter is based on our paper that introduced this technique, which was among the first publications on this topic and is still highly cited. It explains the technique in detail and presents the results, which are in the thesis considered preliminary, since the original paper did not use techniques crucial to obtain results comparable to the rest of the research. We also include section about an interesting idea of using n-gram language models to refine the decoding process.

As the number of publications on this topic increased over time, we identified the systematic issues with the evaluation and further improved on our previous work. Chapter 5 describes our main experiments with CTC-based NAR models. In these experiments, we focus on fair comparison to the current research on NAR methods, as well as to the current state-of-the-art in NMT, including the comparision to efficient autoregressive models, an aspect which is commonly being ignored. We find that despite achieving results on the same level as the leading approaches in non-autoregressive NMT, when compared to an efficient autoregressive models, there is a substantial gap *both* in terms of speed and translation quality. This suggest that there is still a long way to go before non-autoregressive research "closes this gap", as it is very often said in NAR literature to already have happened.

This submission to the EAMT Best Thesis Award is structured as follows. After this overview, we include a copy of the author's CV with list of publications relevant to the thesis, followed by the copy of the thesis itself. At the end, we attach reviews by two reviewers.