Kevin Duh
Assistant Research Professor, Computer Science
Senior Research Scientist, HLTCOE
Johns Hopkins University, Baltimore, USA

INSTITUTE OF FORMAL AND APPLIED
LINGUISTICS, CHARLES UNIVERSITY
PRAGUE, CZECHIA

February 1, 2022

Review of the doctoral dissertation "Non-Autoregressive Neural Machine Translation" by Jindřich Helcl

The thesis by Jindřich Helcl focuses on improving the decoding speed of neural machine translation (NMT) models. State-of-the-art neural translation models operate by encoding the input sentence with a neural network and then generating the output sentence one word at a time. This generation process is auto-regressive in the sense that output word prediction at time *t* is conditioned on the previous output word prediction at time *t-1*. Although this auto-regressive generation process is natural and effective, decoding speed may be slow. To address the problem of decoding speed, non-autoregressive models remove the dependency between output words and instead generate all of words in the sentence together in one step. This is currently a very active area of research, and the thesis makes important contributions.

Chapters 2 and 3 present background material on NMT and Non-Autoregressive (NAR) models. The discussion of the existing NAR models is admirably clear and well-structured. First, the fundamental difference between Autoregressive (AR) and NAR models are explained succinctly in math:

$$AR: p(y|x) = \prod_{t=1}^{T_y} p(y_t|y_{<t}, x, \theta) \qquad NAR: p(y|x) = \prod_{t=1}^{T_y} p(y_t|x, \theta)$$

Here, we seek to model the probability of output sequence y given input sequence x, where θ are the model parameters, T_y is the output length, y_t is the t-th individual output word, and $y_{<t}$ is the prefix of output words before time *t.* As seen here, the conditional independence assumption is responsible for both NAR's inference speed improvement as well as potential accuracy degradation. The thesis then continues by illustrating why accuracy may degrade and explains the multimodality problem and output length prediction problem.

Following the example in Gu et. al.'s 2017-2018 seminar work on NAR, the thesis explains the multimodality problem as follows: the English phrase "thank you" can be properly translated as both "danke schoen" and "vielen dank" in German (i.e., multiple modes in the *p(y|x)* distribution). An AR model, after generating "danke", will have high probability for "schoen" but not the other words. A NAR model, not conditioning on the prefix, would unfortunately give high probability to "danke dank", etc. The output length prediction problem is simply the inherent difficulty of pinpointing the length T_y given the input x. Most of the research in the field is focused on address these kinds of challenges, in one way or another.

The bulk of Chapter 3 provides a tour of several common solutions in the literature, divided into three broad categories: alignment-based methods, auxiliary training objectives, and iterative methods. The proposed work in Chapter 4 falls under the category of alignment-based methods. I have not seen the categorization before and find this part of the literature review to be extremely valuable, as it organizes the vast amount of existing work and provides a way to think about the NAR problem from different perspectives.

Chapter 4 introduces the main technical innovation of the work, which is a NAR model that employs the connectionist temporal classification (CTC) loss function. This is based on the EMNLP2018 paper, "End-to-End Non-Autoregressive Neural Machine Translation with Connectionist Temporal Classification", jointly authored with Jindřich Libovický. The CTC loss allows for inherent alignment between input and output and ameliorates the challenge of output length prediction in NAR. This is a novel idea, and it is great to see it tested in the context of machine translation. The implementation of this general idea, with techniques like source-to-target length expansion and LM rescoring, is reasonable and convincing.

While the application of CTC-NAR to machine translation is novel, note that there are work in the speech processing literature for combining CTC with NAR. I would suggest comparing the ideas in the following papers, in order to (a) to develop a better understanding of the various technical details, and (b) to learn whether speech and translation require different solutions.

- Tian, et. al., Spike-Triggered Non-Autoregressive Transformer for End-to-End Speech Recognition, Interspeech 2020
- Inaguma, et. al., Orthros: Non-autoregressive End-to-End Speech Translation with Dual-Decoder, ICASSP 2021
- Higuchi, et. al., Improved Mask-CTC for Non-autoregressive End-to-End ASR, ICASSP 2021
- Song, et. al., Non-autoregressive Transformer ASR with CTC-enhanced Decoder Input, ICASSP 2021

Chapter 5, titled "Experiments", pulls out all the stops and implements the current best practices for NAR, seeking to obtain the best-possible result. I learned a lot from reading this chapter. The best practices include data cleaning and back-translation for strong teacher models (for use in knowledge distillation of NAR models), lexical shortlist for speeding up inference, and comparison of various sizes of NAR models. Experiments are performed on both the WMT'14 News task and WMT'21 Efficiency Shared Task (German-English translation).

The comprehensive set of experiments will serve as a good reference for NAR researchers. In particular, the various methods are compared together in a fair way, making it easier to see what is working and when things are working. Further, evaluation on batched and non-batched GPU/CPU decoding are presented and contrasted. There are many tables and graphs in the chapter; they are extremely valuable data points for a future researcher looking to follow-up on this work.

Perhaps one missed opportunity in this chapter is the lack of Error Analysis. While the manual comparison of errors between AR and NAR models (in the previous chapter) is worthy and confirms some of our hypotheses about the challenges for NAR, I think it would be insightful to analyze the errors of different NAR variants in this chapter as well. The author is in a unique position of having so many NAR outputs, evaluated along the lines of speed and BLEU/ChrF/COMET. It would be interesting to compare the outputs of the various NAR models in, for example Table 5.6, to see if the BLEU differences reflect specific error patterns.

Finally, I appreciate the effort to demonstrate a very strong AR baseline. For example, Figure 5.3 shows that a strong AR baseline catches up in speed as batch size increases. The speed-accuracy improvements of the proposed NAR model are less impressive when seen in light of the strong AR baseline, but I agree with the author that it is far more important to provide an honest assessment of the tradeoff in practical deployment scenarios. This point perhaps needs to be emphasized more in our research field, which tends to favor work that show numerical gains. However, the nature of this multi-objective speed-accuracy problem implies that numerical gains (or lack thereof) are not easy interpret. So I very much appreciate the message presented in this chapter, such as the caveats discussed in Section 5.3.3.

In summary, I think this is a strong thesis in the field of machine translation. In terms of writing, I believe the thesis is succinct and provides the right level of detail. The descriptions are sufficiently clear for scientific reproducibility. The discussion of previous work is well-organized, and I would only suggest adding additional comparison to CTC-NAR work from the speech field.

In terms of technical contributions, the proposed CTC-NAR method is novel and good motivations; the comprehensive experiments on various setups will serve as a reference for future researchers. Overall, I think this thesis represents a valuable piece of work in the field of NAR for faster machine translation, and the author demonstrates the ability to contribute to the research community.

Best regards,

Kevin Duh
Johns Hopkins University