FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

ABSTRACT OF DOCTORAL THESIS

Jindřich Helcl

Non-Autoregressive Neural Machine Translation

Institute of Formal and Applied Linguistics

Supervisor:	prof. RNDr. Jan Hajič, Dr.
Study programme:	Computer Science
Study branch:	Mathematical Linguistics

Prague 2021

The results of this thesis were achieved in the period of a doctoral study at the Faculty of Mathematics and Physics, Charles University in years 2013–2019.

Student:	Mgr. Jindřich Helcl
Supervisor:	prof. RNDr. Jan Hajič, Dr. Institute of Formal and Applied Linguistics Faculty of Mathematics and Physics Charles University Malostranské nám. 25, 118 00 Prague 1
Department:	Institute of Formal and Applied Linguistics Faculty of Mathematics and Physics Charles University Malostranské nám. 25, 118 00 Prague 1
Opponents:	prof. Lucia Specia, Ph.D. Department of Computing Faculty of Engineering Imperial College London 180 Queen's Gate, South Kensington, London, United Kingdom
	Ing. Jan Čech, Ph.D. Department of Cybernetics Faculty of Electrical Engineering Czech Technical University Karlovo naměstí 13, 121 35 Prague 2, Czech Republic

The thesis defence will take place on June 13, 2019 at 10:10 a.m. in front of a committee for thesis defences in the branch Mathematical Linguistics at the Faculty of Mathematics and Physics, Charles University, Malostranské nám. 25, Prague 1, room S1.

Chairman of Academic Council:	doc. Ing. Zdeněk Žabokrtský, Ph.D. Institute of Formal and Applied Linguistics Faculty of Mathematics and Physics Charles University Malostranské nám. 25, 118 00 Prague 1

The thesis can be viewed at the Study Department of Doctoral Studies of the Faculty of Mathematics and Physics, Charles University, Ke Karlovu 3, Prague 2.

This abstract was distributed on May 23, 2019.

MATEMATICKO-FYZIKÁLNÍ
FAKULTA
Univerzita Karlova

AUTOREFERÁT DISERTAČNÍ PRÁCE

Jindřich Helcl

Neautoregresivní neuronový strojový překlad

Ústav formální a aplikované lingvistiky

Školitel:	prof. RNDr. Jan Hajič, Dr.
Studijní program:	Informatika
Studijní obor:	Matematická lingvistika

Praha 2021

Disertační práce byla vypracována na základě výsledků získaných během doktorského studia na Matematicko-fyzikální fakultě Univerzity Karlovy v letech 2013–2019.

Doktorand: Mgr. Jindřich Helcl

Školitel: prof. RNDr. Jan Hajič, Dr.
Ústav formální a aplikované lingvistiky
Matematicko-fyzikální fakulta
Univerzita Karlova
Malostranské nám. 25, 118 00 Praha 1

Školicí pracoviště: Ústav formální a aplikované lingvistiky
Matematicko-fyzikální fakulta
Univerzita Karlova
Malostranské nám. 25, 118 00 Praha 1

Oponenti: prof. Lucia Specia, Ph.D.
Department of Computing
Faculty of Engineering
Imperial College London
180 Queen's Gate, South Kensington, Londýn, Spojené království

Ing. Jan Čech, Ph.D.
Katedra kybernetiky
Fakulta elektrotechnická
České vysoké učení technické
Karlovo naměstí 13, 121 35 Praha 2

Obhajoba disertační práce se koná dne 13. června 2019 v 10:10 před komisí pro obhajoby disertačních prací v oboru Matematická lingvistika na Matematicko-fyzikální fakultě UK, Malostranské nám. 25, Praha 1, v místnosti S1.

Předseda RDSO: doc. Ing. Zdeněk Žabokrtský, Ph.D.
Ústav formální a aplikované lingvistiky
Matematicko-fyzikální fakulta
Univerzita Karlova
Malostranské nám. 25, 118 00 Praha 1

S disertační prací je možno se seznámit na studijním oddělení Matematicko-fyzikální fakulty UK, Ke Karlovu 3, Praha 2.

Autoreferát byl rozeslán dne 23. května 2018.

Contents

1. Introduction

In real-world applications of machine translation (MT), efficiency is often crucial. Most commercial neural machine translation (NMT) models are available through a cloud-based service, such as Microsoft Translator[1] or Google Translate.[2] Scaling cloud-based solutions for large user bases is simple but costly. Even with a large pool of computational resources, it is worthwhile to implement optimizations that decrease model latency and improve user experience.

Locally deployed NMT models provide a number of advantages over cloud-based solutions. First, the service does not rely on the internet connection. Second, the data is not being sent to a third party server and, therefore, it is suitable for translating private or confidential data. However, without optimization, running state-of-the-art translation models locally often requires specialized hardware, such as one or more GPUs. Otherwise, the time to translate a single sentence can easily exceed one second on a standard CPU.

Higher decoding speeds can be achieved by model optimization. In their 2019 submission to the Workshop on Neural Generation and Translation (WNGT) Efficiency Shared Task, Kim et al. (2019) successfully employed knowledge distillation, quantization, shortlisting (Jean et al., 2015) and a simpler recurrent unit design to bring the throughput of a translation model up to 3,600 words per second on a CPU, with a modest drop in the translation quality. Following this work, Bogoychev et al. (2020) reported further improvements with attention head pruning (Voita et al., 2019). Their work has been part of the Bergamot Research Project, which aims to bring offline translation models to a browser.[3]

Non-autoregressive (NAR) models present an alternative approach to model optimization, using different architecture and a different decoding algorithm which has lower time complexity. In NMT, a non-autoregressive decoding algorithm does not access previously decoded outputs, imposing conditional independence assumption on the output token probability distributions. This assumption allows for parallelization of the decoding, which can significantly reduce the latency of the translation system. On the other hand, it also presents a challenge to the language model, which usually leads to poorer translation quality.

[1] https://microsoft.com/translator/
[2] https://translate.google.com/
[3] https://browser.mt/

2. Non-Autoregressive Neural Machine Translation

The defining feature of a non-autoregressive (NAR) model is the assumption of conditional independence between the output distributions across time steps. The output distribution in autoregressive models is defined as follows:

$$p(y|x) = \prod_{t=1}^{T_y} p(y_t|y_{<t}, x, \theta) \tag{2.1}$$

Unlike Equation 2.1, NAR models do not condition the output token probabilities on previously decoded outputs $y_{<t}$. The probability of an output sentence y given an input sequence x can then be modeled as:

$$p(y|x) = \prod_{t=1}^{T_y} p(y_t|x, \theta) \tag{2.2}$$

Although technically possible, making the outputs in RNN-based models conditionally independent does not reduce the time complexity because in RNNs, the value of each hidden state depends on the value of the preceding state. However, in the Transformer model, hidden states in each layer depend only on the states from the previous layer. This allows for parallel computation at the layer level

In the following paragraphs, we discuss the necessary alterations to the Transformer architecture. Since the outputs are conditionally independent, we cannot feed the previously decoded outputs into the Transformer decoder. We need to provide the input to the decoder and estimate the target length. The causal mask over decoder self-attention is now unnecessary. We also address the main issue and the reason autoregressive (AR) models are still superior in modeling language.

Multimodality Problem. In one of the first applications of a non-autoregressive model to neural machine translation (NMT), Gu et al. (2018) describe the *multimodality problem* which arises when the outputs are conditionally independent.

When estimating the probability of a word on a given position, there may be multiple words which get a high probability. These words are the so-called *modes* of the distribution. In autoregressive models, once a word is selected, other modes are ignored in the following time steps. However, a non-autoregressive model does not base its decision for a given position on the preceding ones, so when multiple positions have multiple modes, the model has no means of coordinating the selection of modes across different time steps.

A well-known example of the multimodality problem is the translation of the sentence "thank you" into German, which has two equally likely translations: "vielen dank" and "danke schön." In this case, the pair of German tokens "danke" and "vielen" create the two modes in the first position, and the tokens "dank" and "schön" are the modes in the second position. If an autoregressive model chooses to generate "danke" in the first position, the token "dank" in the second position will no longer receive high probability from the model. However, when a non-autoregressive model assigns high probabilities to the correct translations, it also has to assign high probabilities to the other (incorrect) two combinations, "danke dank" and "vielen schön" (Gu et al., 2018).

Decoder Inputs. A NAR Transformer decoder cannot receive the previously decoded tokens on the input. A solution proposed by Gu et al. (2018) is to use a simple fertility model, which also serves as the explicit target length estimator.

Compared to the autoregressive Transformer, the model has the following modifications. First, the inputs to the decoder are made up of the sequence of encoder inputs, either uniformly stretched to the predicted target sentence length, or copied using a fertility model. Second, the decoder self-attention does not use the causal mask, since all states can now attend to all other states in both directions. Third, a *positional attention* sub-layer is added to every decoder layer, where the positional encoding (see Equation **??** in Section **??**) is used as queries and keys, and the decoder states as values. Gu et al. (2018) argue that providing positional information directly to the decoder layers could improve the potential of the decoder to model local reodering.

In Gu et al. (2018), the multimodality problem (and length estimation) is addressed by introducing latent fertility variables $F = f_1, \ldots, f_{T_x}$ sampled from a prior distribution. Each $f_i \in \mathbb{N}_0$ denotes the number of times x_i is copied to the decoder input (summing up to the target length T_y). The output probability is then conditioned on the latent vector F, which is marginalized out:

$$p(y|x,\theta) = \sum_{F \in \mathcal{F}} p(F|x,\theta) \cdot p(y|x,F,\theta) \tag{2.3}$$

where the fertility model $p(F|x,\theta)$ and the translation model $p(y|x,F,\theta)$ can be trained jointly using a variational lower bound with a candidate distribution q:

$$\begin{aligned} \mathcal{L}(\theta) &= \log p(y|x,\theta) = \log \sum_{F \in \mathcal{F}} p(F|x,\theta) \cdot p(y|x,F,\theta) \\ &\geq \mathbb{E}_{F \sim q} \left(\sum_{t=1}^{T_y} \log p(y_t|x,F,\theta) + \sum_{t=1}^{T_x} \log p(f_t|x,\theta) \right) + \mathcal{H}(q) \end{aligned} \tag{2.4}$$

where q is an external deterministic fertility model (and, therefore, $\mathcal{H}$ is a constant), and the expectation is also deterministic. The fertility model depends on an external module which is not trained together with the model. The authors fine-tune the trained translation model using reinforcement learning (Williams, 1992) to estimate the gradients of the fertility model.

During decoding, marginalizing over all possible fertility values is intractable. Therefore, Gu et al. (2018) experiment with three approximation methods – argmax, average decoding, and noisy parallel decoding (NPD). In argmax decoding, the fertility with the highest probability is chosen in each step, similarly to greedy decoding. The average method chooses the expected fertility given the distribution in each position. NPD is based on sampling and rescoring with an autoregressive model, as explained below.

3. Connectionist Temporal Classification

4. Experiments

5. Conclusions

Bibliography

Bogoychev, N. – Grundkiewicz, R. – Aji, A. F. – Behnke, M. – Heafield, K. – Kashyap, S. – Farsarakis, E.-I. – Chudyk, M. Edinburgh's Submissions to the 2020 Machine Translation Efficiency Task. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, p. 218–224, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.ngt-1.26. Available at: `https://aclanthology.org/2020.ngt-1.26`.

Chen, P. – Helcl, J. – Germann, U. – Burchell, L. – Bogoychev, N. – Barone, A. V. M. – Waldendorf, J. – Birch, A. – Heafield, K. The University of Edinburgh's English-German and English-Hausa Submissions to the WMT21 News Translation Task. In *Proceedings of the Sixth Conference on Machine Translation*. Association for Computational Linguistics, 2021.

Gu, J. – Bradbury, J. – Xiong, C. – Li, V. O. K. – Socher, R. Non-Autoregressive Neural Machine Translation. In *6th International Conference on Learning Representations, ICLR 2018*, Vancouver, BC, Canada, April 2018. Available at: `https://openreview.net/forum?id=B1l8BtlCb`.

Helcl, J. – Libovický, J. – Kocmi, T. – Musil, T. – Cífka, O. – Variš, D. – Bojar, O. Neural Monkey: The Current State and Beyond. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, p. 168–176, Boston, MA, March 2018. Association for Machine Translation in the Americas. Available at: `https://aclanthology.org/W18-1816`.

Helcl, J. – Libovický, J. Neural Monkey: An Open-source Tool for Sequence Learning. *The Prague Bulletin of Mathematical Linguistics*. Apr 2017, 107, 1, p. 5–17. ISSN 0032-6585.

Jean, S. – Cho, K. – Memisevic, R. – Bengio, Y. On Using Very Large Target Vocabulary for Neural Machine Translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, p. 1–10, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1001. Available at: `https://aclanthology.org/P15-1001`.

Kim, Y. J. – Junczys-Dowmunt, M. – Hassan, H. – Fikri Aji, A. – Heafield, K. – Grundkiewicz, R. – Bogoychev, N. From Research to Production and Back: Ludicrously Fast Neural Machine Translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, p. 280–288, Hong Kong, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5632. Available at: `https://aclanthology.org/D19-5632`.

Libovický, J. – Helcl, J. Attention Strategies for Multi-Source Sequence-to-Sequence Learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, p. 196–202, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-2031. Available at: `https://aclanthology.org/P17-2031`.

Libovický, J. – Helcl, J. End-to-End Non-Autoregressive Neural Machine Translation with Connectionist Temporal Classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, p. 3016–3021, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1336. Available at: `https://aclanthology.org/D18-1336`.

Libovický, J. – Helcl, J. – Mareček, D. Input Combination Strategies for Multi-Source Transformer Decoder. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, p. 253–260, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6326. Available at: https://aclanthology.org/W18-6326.

Miceli Barone, A. V. – Helcl, J. – Sennrich, R. – Haddow, B. – Birch, A. Deep architectures for Neural Machine Translation. In *Proceedings of the Second Conference on Machine Translation*, p. 99–107, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4710. Available at: https://aclanthology.org/W17-4710.

Voita, E. – Talbot, D. – Moiseev, F. – Sennrich, R. – Titov, I. Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 5797–5808, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1580. Available at: https://aclanthology.org/P19-1580.

Williams, R. J. Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning. *Machine learning*. 1992, 8, 3-4, p. 229–256.

List of Publications

Helcl, J. – Libovický, J. Neural Monkey: An Open-source Tool for Sequence Learning. *The Prague Bulletin of Mathematical Linguistics*. Apr 2017, 107, 1, p. 5–17. ISSN 0032-6585. 54 citations.

Libovický, J. – Helcl, J. Attention Strategies for Multi-Source Sequence-to-Sequence Learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, p. 196–202, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-2031. Available at: `https://aclanthology.org/P17-2031`. 140 citations.

Miceli Barone, A. V. – Helcl, J. – Sennrich, R. – Haddow, B. – Birch, A. Deep architectures for Neural Machine Translation. In *Proceedings of the Second Conference on Machine Translation*, p. 99–107, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4710. Available at: `https://aclanthology.org/W17-4710`. 80 citations.

Helcl, J. – Libovický, J. – Kocmi, T. – Musil, T. – Cífka, O. – Variš, D. – Bojar, O. Neural Monkey: The Current State and Beyond. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, p. 168–176, Boston, MA, March 2018. Association for Machine Translation in the Americas. Available at: `https://aclanthology.org/W18-1816` 7 citations.

Libovický, J. – Helcl, J. – Mareček, D. Input Combination Strategies for Multi-Source Transformer Decoder. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, p. 253–260, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6326. Available at: `https://aclanthology.org/W18-6326`. 43 citations.

Libovický, J. – Helcl, J. End-to-End Non-Autoregressive Neural Machine Translation with Connectionist Temporal Classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, p. 3016–3021, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1336. Available at: `https://aclanthology.org/D18-1336`. 67 citations.

Chen, P. – Helcl, J. – Germann, U. – Burchell, L. – Bogoychev, N. – Barone, A. V. M. – Waldendorf, J. – Birch, A. – Heafield, K. The University of Edinburgh's English-German and English-Hausa Submissions to the WMT21 News Translation Task. In *Proceedings of the Sixth Conference on Machine Translation*. Association for Computational Linguistics, 2021. 0 citations.

Only publications relevant to this thesis are included. The number of citations was computed using Google Scholar. Total number of citations of publications related to the topic of the thesis (without self-citations): 391