Towards Efficient Universal Neural Machine Translation

Biao Zhang[*]
School of Informatics
University of Edinburgh
b.zhang@ed.ac.uk

Humans benefit from communication but suffer from language barriers. Machine translation (MT) aims to overcome such barriers by automatically transforming information from one language to another. With the rapid development of deep neural networks, neural machine translation (NMT) – especially Transformer (Vaswani et al., 2017) – has achieved great success in recent years, delivering state-of-the-art and even near human performance on many bilingual text-based translation tasks (Akhbardeh et al., 2021). However, challenges remain particularly in 1) *efficiency* where a massive NMT model is a computational bottleneck for training and decoding, and 2) *universality* where extending NMT beyond bilingual and text-based scenarios (such as multilingual and speech-to-text translation) is still non-trivial. In this thesis, we investigate ways of developing simple and effective neural architectures to address these two challenges.

NMT is resource-hungry. Achieving high-quality translation demands complex network architectures and a large number of model parameters, which often takes hundreds or even thousands of training GPU hours and leads to slow inference. We tackle this computational inefficiency issue via three aspects: 1) simplifying model architectures, where we propose a lightweight recurrent network and root mean square layer normalization to enable higher model parallelization, as well as a merged attention network paired with depth-scaled initialization to improve deep Transformer; 2) exploring representation redundancy, where we demonstrate the feasibility of sparsifying encoder outputs in Transformer and propose a rectified linear attention to induce sparse attention weights efficiently; and 3) semi-autoregressive modeling, where we relax the independence assumption by allowing generation from the left-to-right and right-to-left directions simultaneously. Apart from benefiting efficiency, these techniques also lay the foundation for our research on universality, another topic of this thesis.

MT should be universal, i.e., being capable of transforming information between *any* languages in *any* modalities. Unfortunately, NMT still struggles with poor language coverage and cross-modality gap. As a step towards universal MT, we focus on (massively) multilingual NMT and direct speech-to-text translation (ST). Multilingual NMT suffers from capacity bottleneck and off-target translation; we thus study methods of increasing modeling capacity for multilingual Transformer, and propose random online backtranslation to bridge zero-short language pairs. We further explore when and where language-specific modeling matters via conditional language-specific routing, discovering the trade-off between shared and language-specific capacity. Unlike textual NMT, the modality gap between speech and text hinders ST. We narrow this gap by inventing adaptive feature selection, which automatically filters out uninformative speech features, improving translation as well as inference speed. Next, we extend our study to document-level speech translation to address the question whether and how context helps ST. We adopt contextual modeling for ST, and show its effectiveness on enhancing homophone and simultaneous translation.

Finally, we move forward to multilingual and multimodal modeling for translation by exploring multilingual ST, a critical path to universal NMT.

[*] Now at Google Deepmind.

We integrate the above methods into a single system and participate in the multilingual ST shared task in IWSLT2021. Our system achieves competitive performance in both supervised and zero-shot translation, where we observe the complementarity of different techniques in improving multilingual ST.

We believe that technologies nowadays are mature enough to pursue universal translation modeling. Along this path, challenges widely exist, but also opportunities. We released our source code to facilitate the development.[1]

Acknowledgements

The author would like to thank his PhD supervisors, Rico Sennrich and Ivan Titov, and his thesis examiners, Kenneth Heafield and Graham Neubig. The work presented in this thesis has been supported by Baidu Scholarship, the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreements 825460 (ELITR) and 825299 (GoURMET). The author also acknowledges the resources provided by the Cambridge Tier-2 system operated by the University of Cambridge Research Computing Service (http://www.hpc.cam.ac.uk) funded by EPSRC Tier-2 capital grant EP/P020259/1.

References

[Akhbardeh et al.2021] Akhbardeh, Farhad, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (wmt21). In *Proceedings of the Sixth Conference on Machine Translation*, Online, November. Association for Computational Linguistics.

[Vaswani et al.2017] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Guyon, I., U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, volume 30, pages 5998–6008. Curran Associates, Inc.

[1] `https://github.com/bzhangGo/zero`