Pre-training Methods for Neural Machine Translation

Mingxuan Wang
ByteDance AI Lab
wangmingxuan.89@bytedance.com

Lei Li
ByteDance AI Lab
lileilab@bytedance.com

1 Tutorial Introduction

Pre-training is a dominant paradigm in Nature Language Processing (NLP) (Radford et al., 2019; Devlin et al., 2019; Liu et al., 2019a), Computer Vision (CV) (He et al., 2019; Xie et al., 2020) and Auto Speech Recognition (ASR) (Bansal et al., 2019; Chuang et al., 2020; Park et al., 2019). Typically, the models are first pre-trained on large amount of unlabeled data to capture rich representations of the input, and then applied to the downstream tasks by either providing context-aware representation of the input, or initializing the parameters of the downstream model for fine-tuning. Recently, the trend of self-supervised pre-training and task-specific fine-tuning finally fully hits neural machine translation (NMT) (Zhu et al., 2020; Yang et al., 2020; Chen et al., 2020).

Despite its success, introducing a universal pre-trained model to NMT is non-trivial and not necessarily yields promising results, especially for the resource-rich setup. Unique challenges remain in several aspects. First, the objective of most pre-training methods are different from the downstream NMT tasks. For example, BERT (Devlin et al., 2019), a popular pre-trained model, is designed for language understanding with only a transformer encoder, while an NMT model usually consists of an encoder and a decoder to perform cross-lingual generation. This gap makes it not feasible enough to apply pre-training for NMT (Song et al., 2019). Besides, machine translation is naturally a multi-lingual problem, but general pre-training methods for NLP mainly focus on English corpus, such as BERT and GPT. Given the success of transfer learning in multi-lingual machine translation, it is very appealing to introduce multi-lingual pre-training for NMT (Conneau and Lample, 2019). Finally, speech translation has attracted much attention recently, while most pre-training methods are focused on text representation. How to leverage the pre-training methods to improve the speech translation becomes a new challenge.

This tutorial provides a comprehensive guide to make the most of pre-training for neural machine translation. Firstly, we will briefly introduce the background of NMT, pre-training methodology, and point out the main challenges when applying pre-training for NMT. Then we will focus on analysing the role of pre-training in enhancing the performance of NMT, how to design a better pre-training model for executing specific NMT tasks and how to better integrate the pre-trained model into NMT system. In each part, we will provide examples, discuss training techniques and analyse what is transferred when applying pre-training.

The first topic is the *monolingual pre-training for NMT*, which is one of the most well-studied field. Monolingual text representations like ELMo, GPT, MASS and BERT have superiorities, which significantly boost the performances of various natural language processing tasks (Peters et al., 2018; Devlin et al., 2019; Radford et al., 2019; Song et al., 2019). However, NMT has several distinct characteristics, such as the availability of large training data (10 million or larger) and the high capacity of baseline NMT models, which requires carefully design of pre-training. In this part, we will introduce different pre-training methods and analyse the best practice when applying them to different machine translation scenarios, such as unsupervised NMT, low-resource NMT and rich-source NMT (Zhu et al., 2020; Yang et al., 2020). We will cover techniques to finetune the pre-trained models with various strategies, such as knowledge distillation and adapter (Bapna and Firat, 2019; Liang et al., 2021).

The next topic is *multi-lingual pre-training for NMT*. In this context, we aims at mitigating the English-centric bias and suggest that it is possible

Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Tutorial Abstracts, pages 21–25, August 1st, 2021.

to build universal representation for different language to improve massive multi-lingual NMT. In this part, we will discuss the general representation of different languages and analyse how knowledge transfers across languages. These will allow a better design for multi-lingual pre-training, in particular for zero-shot transfer to non-English language pairs (Johnson et al., 2017; Qi et al., 2018; Conneau and Lample, 2019; Pires et al., 2019; Huang et al., 2019; Lin et al., 2020; Liu et al., 2020; Pan et al., 2021; Lin et al., 2021).

The last technical part of this tutorial deals with the *Pre-training for speech NMT*. In particular, we focus on leverage weakly supervised or unsupervised training data to improve speech translation. In this part, we will discuss the possibilities of building a general representations across speech and text. And shows how text or audio pre-training can guild the text generation of NMT (Wang et al., 2019; Liu et al., 2019b; Bansal et al., 2019; Wang et al., 2020; Baevski et al., 2020a,b; Huang et al., 2021; Long et al., 2021; Dong et al., 2021b,a; Han et al., 2021; Ye et al., 2021).

We conclude the tutorial by pointing out the best practice when applying pre-training for NMT. The topics cover various of pre-training methods for different NMT scenarios. After this tutorial, the audience will understand why pre-training for NMT is different from other tasks and how to make the most of pre-training for NMT. Importantly, we will give deep analyze about how and why pre-training works in NMT, which will inspire future work on designing pre-training paradigm specific for NMT.

2 Tutorial Outline

PART I: Introduction (15 min)

- Background of NMT
- General pre-training paradigm
- Unique Challenges
  - Objective difference
  - Multi-lingual generation
  - Modality disparity

PART II: Monolingual Pre-training for NMT (60 min)

- The early stage
  - NMT initialized with word2vec
  - NMT initialized with language model
- BERT fusion in NMT
  - BERT Incorporating methods
  - BERT Tuning methods
- Unified sequence-to-sequence pre-training
  - MASS, Bart, etc.

PART III: Multi-lingual Pre-training for NMT (45 min)

- Multilingual fused pre-training
  - Cross-lingual Language Model Pre-training
  - Alternating Language Modeling Pre-training
  - XLM-T: Cross-lingual Transformer Encoders
- Multilingual sequence to sequence pre-training
  - mBART
  - CSP
  - mRASP

PART IV: Pre-training for Speech Translation (45 min)

- MT pre-training
- ASR pre-training
- Audio pre-training
- Raw text pre-training
- Bi-modal pre-training

PART V: Conclusion and Future Directions (15 min)

3 Type of Tutorial

Cutting-edge. In this tutorial, we will discuss the most advanced techniques of pre-training for neural machine translation. The instructors will also present their own practical experiences in enhancing a machine translation service as a product, which are usually not found in papers.

4 Tutorial Breadth

Based on the representative set of papers listed in the selected bibliography, we anticipate that 70%-80% of the tutorial will cover other researchers' work, while the rest concerns the work where at least one of the presenters has been actively involved in. We will introduce several important work related to the monolingual, the multi-lingual and the multi-modal pre-training for NMT.

5 Diversity

In the tutorial, some multilingual pre-training methods will scale to over 50 to 100 different languages. Researchers working on the diverse language pairs might find this tutorial relevant and useful.

6 Prerequisites

The tutorial is self-contained. We will address the background, the technical details and the examples. Basic knowledge about neural networks are required, including word embeddings, attention, and encoder-decoder models. Prior NLP courses and familarity with the machine translation task are preferred.

It is recommended (and optional) that audience to read the following papers before the tutorial:

1. Basic MT model: Attention is all you need (Vaswani et al., 2017).
2. Google's multilingual neural machine translation system (Johnson et al., 2017).
3. Text pre-training with BERT (Devlin et al., 2019) and GPT (Radford et al., 2019).
4. Audio pre-training with Wav2vec and Wav2vec2.0 (Schneider et al., 2019; Baevski et al., 2020b).
5. Pre-training multilingual NMT (Lin et al., 2020; Liu et al., 2020).

7 Target Audience

This tutorial will be suitable for researchers and practitioners interested in pre-training applications and multilingual NLP, especially for machine translation.

To the best of our knowledge, this is the first tutorial that focuses on the pre-training methods and practice for NMT.

8 Technical Requirements

The tutorial will be online. Internet connection with proper live video device is needed.

9 Open access

Our slides and video is open to public, available at `https://lileicc.github.io/TALKS/2021-ACL/`.

10 Tutorial Presenters

Mingxuan Wang (ByteDance AI Lab)
Google Scholar

Dr. Mingxuan Wang is a senior researcher at ByteDance AI Lab. He received his PhD degree from the Chinese Academy of Sciences Institute of Computing Technology in 2017. His research focuses on natural language processing and machine translation. He has published over 20 papers in leading NLP/AI journals and conferences such as ACL, AAAI and EMNLP. He has served in the Program Committee for ACL/EMNLP 2016-2020, AAAI/IJCAI 2018/2019, NeurIPS 2020. He achieved outstanding results in various machine translation evaluation competitions, including the first place of Chinese-to-English translation at at the WMT 2018, the third place of Chinese-to-English translation at NIST 2015, etc. Together with Dr. Lei Li, he is leading a team developing the VolcTrans machine translation system.

He has given a tutorial about Machine Translation at CCMT 2017 and was an guest lecturer for 2016 Machine Translation for University of Chinese Academy of Sciences (UCAS).

Lei Li (ByteDance AI Lab)
https://lileicc.github.io/

Dr. Lei Li is Director of ByteDance AI Lab, leading the research and product development for NLP, robotics, and drug discovery. His research interests are machine translation, speech translation, text generation, and AI powered drug discovery. He received his B.S. from Shanghai Jiao Tong University and Ph.D. from Carnegie Mellon University, respectively. His dissertation work on fast algorithms for mining co-evolving time series was awarded ACM KDD best dissertation (runner up). His recent work on AI writer Xiaomingbot received 2nd-class award of Wu Wen-tsün AI prize in 2017. He is a recipient of CCF distinguished speaker in 2017, and CCF Young Elite award in 2019. His team won first places for five language translation directions in WMT 2020 and the best in corpus filtering challenge. Before ByteDance, he worked at EECS department of UC Berkeley and Baidu's Institute of Deep Learning in Silicon Valley. He has served organizers and area chair/senior PC for multiple conferences including KDD, EMNLP, NeurIPS, AAAI, IJCAI, and CIKM. He has published over 100 technical papers in ML, NLP and data mining and holds more than 10 patents. He has started and is developing ByteDance's machine translation system, VolcTrans and many of his algorithms have been deployed.

He has delivered four tutorials at EMNLP 2019, NLPCC 2019, NLPCC 2016, and KDD 2010. He was an lecturer for 2014 Probabilistic Programming for Advancing Machine Learning summer school at Portland, USA.

11 Other Information

Prior Related Tutorials Neural Machine Translation, presented by Thang Luong, Kyunghyun Cho, and Christopher Manning at ACL 2016. This tutorial is related but different from ACL 2016 NMT tutorial. It focuses on pre-training methods for both bilingual, multi-lingual, and multi-modal neural machine translation.

Unsupervised Cross-Lingual Representation Learning, presented by Sebastian Ruder, Anders Søgaard, and Ivan Vulić at ACL 2019. This tutorial is related in concerning multi-lingual NLP. However, their tutorial was on representation learning, while our tutorial is on neural machine translation.

References

Alexei Baevski, Steffen Schneider, and Michael Auli. 2020a. vq-wav2vec: Self-supervised learning of discrete speech representations. In *Proc. of ICLR*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020b. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proc. of NeurIPS*.

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. 2019. Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. In *Proc. of NAACL-HLT*, pages 58–68.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. In *Proc. of EMNLP*, pages 1538–1548.

Yen-Chun Chen, Zhe Gan, Yu Cheng, Jingzhou Liu, and Jingjing Liu. 2020. Distilling knowledge learned in BERT for text generation. In *Proc. of ACL*, pages 7893–7905.

Yung-Sung Chuang, Chi-Liang Liu, and Hung-Yi Lee. 2020. SpeechBERT: An audio-and-text jointly learned language model for end-to-end spoken question answering. In *Proc. of INTERSPEECH*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Proc. of NeurIPS*, pages 7057–7067.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL-HLT*, pages 4171–4186.

Qianqian Dong, Mingxuan Wang, Hao Zhou, Shuang Xu, Bo Xu, and Lei Li. 2021a. Consecutive decoding for speech-to-text translation. In *Proc. of AAAI*.

Qianqian Dong, Rong Ye, Mingxuan Wang, Hao Zhou, Shuang Xu, Bo Xu, and Lei Li. 2021b. Listen, understand and translate: Triple supervision decouples end-to-end speech-to-text translation. In *Proc. of AAAI*, volume 35, pages 12749–12759.

Chi Han, Mingxuan Wang, Heng Ji, and Lei Li. 2021. Learning shared semantic space for speech-to-text translation. In *Proc. of ACL - Findings*.

Kaiming He, Ross B. Girshick, and Piotr Dollár. 2019. Rethinking imagenet pre-training. In *Proc. of ICCV*, pages 4917–4926.

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. In *Proc. of EMNLP*, pages 2485–2494.

Haoyang Huang, Lin Su, Di Qi, Nan Duan, Edward Cui, Taroon Bharti, Lei Zhang, Lijuan Wang, Jianfeng Gao, Bei Liu, et al. 2021. M3p: Learning universal representations via multitask multilingual multimodal pre-training. In *Proc. of CVPR*.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *TACL*, 5:339–351.

Jianze Liang, Chengqi Zhao, Mingxuan Wang, Xipeng Qiu, and Lei Li. 2021. Finding sparse structure for domain specific neural machine translation. In *Proc. of AAAI*.

Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. 2020. Pre-training multilingual neural machine translation by leveraging alignment information. In *Proc. of EMNLP*, pages 2649–2663.

Zehui Lin, Liwei Wu, Mingxuan Wang, and Lei Li. 2021. Learning language specific sub-network for multilingual machine translation. In *Proc. of ACL*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *TACL*, 8:726–742.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke S. Zettlemoyer, and Veselin Stoyanov. 2019a. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Yuchen Liu, Hao Xiong, Jiajun Zhang, Zhongjun He, Hua Wu, Haifeng Wang, and Chengqing Zong. 2019b. End-to-end speech translation with knowledge distillation. In *Proc. of INTERSPEECH*, pages 1128–1132.

Quanyu Long, Mingxuan Wang, and Lei Li. 2021. Generative imagination elevates machine translation. In *Proc. of NAACL-HLT*, pages 5738–5748.

Xiao Pan, Liwei Wu, Mingxuan Wang, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. In *Proc. of ACL*.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. In *Proc. of INTERSPEECH*.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL-HLT*, pages 2227–2237.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proc. of ACL*, pages 4996–5001.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proc. of NAACL-HLT*, pages 529–535.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. In *Proc. of INTERSPEECH*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: masked sequence to sequence pre-training for language generation. In *Proc. of ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 5926–5936.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proc. of NeurIPS*, pages 5998–6008.

Chengyi Wang, Yu Wu, Shujie Liu, Ming Zhou, and Zhenglu Yang. 2020. Curriculum pre-training for end-to-end speech translation. In *Proc. of ACL*, pages 3728–3738.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proc. of ICCV*, pages 4580–4590.

Qizhe Xie, Minh-Thang Luong, Eduard H. Hovy, and Quoc V. Le. 2020. Self-training with noisy student improves imagenet classification. In *Proc. of CVPR*, pages 10684–10695.

Jiacheng Yang, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Weinan Zhang, Yong Yu, and Lei Li. 2020. Towards making the most of BERT in neural machine translation. In *Proc. of AAAI*.

Rong Ye, Mingxuan Wang, and Lei Li. 2021. End-to-end speech translation via cross-modal progressive training. In *Proc. of INTERSPEECH*.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. 2020. Incorporating BERT into neural machine translation. In *Proc. of ICLR*.