Bei Li

NO. 3-11, Wenhua Road, Heping District, Shenyang, P. R. China
+86 18704230711 • libei_neu@outlook.com •

EDUCATION

Northeastern University, Shenyang, P.R. China — Jul 2020 – present

- Ph.D in Computer Science and Technology
  - Supervisor: Prof. Tong Xiao
  - Research on machine translation, machine learning, architecture design and efficient network.

Northeastern University, Shenyang, P.R. China — Jul 2017 – Apr 2020

- M.S. in Computer Software and Theory
  - Supervisor: Prof. Tong Xiao
  - Research on machine translation, natural language processing, and deep neural network.
  - Ranked 1 / 230

Northeastern University, Shenyang, P.R. China — Sep 2013 – Jun 2017

- B.S. in Computer Science and Technology
  - Graduated with University Honors.

RESEARCH EXPERIENCE

Microsoft Research Asia (ML Group) — Dec 2022-present

- Connecting large language models with evolutionary algorithms yields powerful prompt optimizers
  This paper introduces EvoPrompt, a novel framework that automates the optimization of prompts for Large Language Models (LLMs) using evolutionary algorithms (EAs). The framework allows for the effective combination of EAs and LLMs to generate optimized, coherent, and human-readable prompts without relying on gradients or parameters. The study tests EvoPrompt on multiple datasets and types of LLMs, including both closed- and open-source models like GPT-3.5 and Alpaca. Results show that EvoPrompt substantially outperforms human-crafted prompts and existing automated methods, thereby demonstrating the synergistic benefits of integrating LLMs with conventional algorithms.
- Deliberate then Generate: Enhanced Prompting Framework for Text Generation
  We encourage the model to deliberate by proposing a novel Deliberate then Generate (DTG) prompting framework, which consists of error detection instructions and candidates that may contain errors. DTG is a simple yet effective technique that can be applied to various text generation tasks with minimal modifications. We conduct extensive experiments on 20+ datasets across 7 text generation tasks, including summarization, translation, dialogue, and more. We show that DTG consistently outperforms existing prompting methods and achieves state-of-the-art performance on multiple text generation tasks.

Microsoft Research Asia (NLC Group) — May 2022-Dec 2022

- ManagerTower: Aggregating the Insights of Uni-Modal Experts for Vision-Language Representation Learning
  We propose ManagerTower, a novel Vision-Language model architecture that gathers and combines the insights of pre-trained uni-modal experts at different levels. The managers introduced in each cross-modal layer can adaptively aggregate uni-modal semantic knowledge to facilitate more comprehensive cross-modal alignment and fusion. ManagerTower outperforms prior work.
- TranSFormer: Slow-Fast Transformer for Machine Translation
  Building upon our previous ICML work, we refine the extraction of fine-grained character-level features by developing a multiscale Transformer model with a two-branch architecture. The Slow-Fast framework effectively mitigates the computational overhead associated with capturing long-term dependencies among character-level sequences, while employing a cross-granularity attention mechanism to learn interactions between the fast and slow branches. Comprehensive experiments conducted on multiple machine translation benchmarks attest to the efficacy of our proposed TranSFormer model.

NEUNLP — Jul 2017-present

- Rethinking and Improving Multi-task Learning for End-to-end Speech Translation
  In this work, we find that the textual encoder primarily facilitates cross-modal conversion, but the presence of noise in speech impedes the consistency between text and speech representations. Furthermore, we propose an improved multi-task learning (IMTL) approach for the ST task, which bridges the modal gap by mitigating the difference in length and representation.
- Incorporating Probing Signals into Multimodal Machine Translation via Visual Question-Answering Pairs
  This work is a further exploration of our ACL 2022. We aim to address the insufficient cross-modal interaction by proposing to generation VQA-style pairs from the text and modeling the probing signals during training. Extensive experiments show that this kind of multi-task learning framework can indeed allievate the aforementioned issue.
- Augmenting Large Language Model Translators via Translation Memories

In-context learning (ICL) augments the capabilities of large language models (LLMs) in various downstream tasks by leveraging input and output exemplars. This paper explores the use of translation memory (TM) as a form of prompting to aid LLMs in machine translation tasks. Notably, the LLM's inherent ability to comprehend these prompts significantly bolsters the use of TM. Experimental results indicate that incorporating TM considerably enhances the translation proficiency of the LLM, elevating its BLEU score to levels commensurate with state-of-the-art neural machine translation systems.

- Learning Multiscale Transformer Models for Sequence Generation
  First, we re-define the concept of scale for NLP, including scales of sub-word, word and phrase. Our intention is to leverage the word boundaries and phrase-level prior knowledge to compensate for the sub-word features.. Then we establish the relationships among different scales. Ultimately, we built a multiscale Transformer model via making full use of the relationships.
- On Vision Features in Multimodal Machine Translation
  This work investigates the effect of vision features in multimodal machine translation (MMT) scenarios. We proposed three probing tasks to evaluate MMT systems which can help the following researchers. The main contribution is to reveal the importance of strong vision features.
- ODE Transformer: An Ordinary Differential Equation-Inspired Model for Neural Machine Translation
  This work establishes the relationship between ODE and the design of Transformer architecture. We also redesign the Transformer architecture inspired by the lower truncation error achieved by high-order solvers in ODE. ODE Transformer can deliver much better translation performance within the same model capacity. Experimental results on three sequence generation tasks demonstrate the effectiveness.
- Weight Distillation: Transferring the Knowledge in Neural Network Parameters
  This work attempts to further enhance the standard sequence-level KD method by taking full advantage of the teacher parameters and generate the parameters for student.
- Learning Light-Weight Translation Models from Deep Transformer
  This work attempts to learn a light-weight translation model from a deep Transformer teacher network. It introduces a group-permutation based knowledge distillation method to compressing a strong deep Transformer teacher into a much shallower counterpart with a minor BLEU degradation. Furthermore, to enhance the performance of the teacher network, we also propose a skipping sub-layer regularization training method to randomly omit some sub-layers vertically. Both methods can be well applicable into the teacher training process.
- Shallow-to-Deep Training for Neural Machine Translation
  Deep Transformer systems have been widely investigated in the MT community recently. However, with the model going deeper, a crucial challenge is the huge memory cost and extremely long training time. We investigate the behavior of trained systems and find that adjacent layers behave similarly. Thus, we proposed a shallow-to-deep training method instead of learning from scratch which speeds up the training process up to 1.5 times with no loss in BLEU.
- Does Multi-Encoder Help? A Case Study on Context-Aware Neural Machine Translation
  We investigate a general-used multi-encoder framework on document-level machine translation task. It utilizes an additional context-encoder to capture the relationship between the current sentence and its contextual information. However, through specially designed context inputs, we find that the context-encoder acts more like a noise generator instead of encoding the contextual information, which is similar with dropout.Especially when we turn off the context-encoder during inference, there is even slight improvements in terms of BLEU score.
- Learning Deep Transformer Models for Machine Translation
  It studies deep encoders in Transformer and mathematically explains the importance of the location of layer normalization for deep models. It also proposes a novel connection schema to successfully train a 30-layer Transformer system, which is the deepest encoder at that time. While, it is one of the most high cited NMT papers.
- The NiuTrans System for WNGT 2020 Efficiency Task
  It describes the submission of the NiuTrans systems for WNGT2020 efficiency task. We utilized deep encoder and shallow decoder for strong translation performance and fast inference speed on both GPU and CPU tasks.
- The NiuTrans Machine Translation Systems for WMT21
  It describes the submission of the NiuTrans systems for WMT2021. We employ the ODE Transformer as our backbone and enhance the single model by enlarging the model depth. Similar techniques are used with the previous submission, and we rank first on the Chinese-English task in terms of human evaluation.
- The NiuTrans Machine Translation Systems for WMT20
  It describes the submission of the NiuTrans systems for WMT2020, including English$\leftrightarrow$Japanese, English$\rightarrow$Chinese and $\{$Inuktitut, Tamil$\}$ language pairs. The main techniques are the same as described in WMT19 paper, and we attempted several fine-tune strategies for better domain adaptation. Also, pre-training and multilingual training are adopted in this work.
- The NiuTrans Machine Translation Systems for WMT19
  It describes the submission of the NiuTrans systems for WMT2019 on both supervised and unsupervised tasks, including 13 language directions. This paper shows the details about model architectures, data augmentation methods, ensemble knowledge distillation and system combination strategies.
- The NiuTrans Machine Translation System for WMT18
  It describes the submission of the NiuTrans neural machine translation system for the WMT2018 Chinese $\leftrightarrow$ English new translation tasks.

- On Ensemble Learning of Neural Machine Translation
  It demonstrates how to integrate several individual models to jointly predicted the probabilities of target words in each decoding step. It also gives some advice for better ensemble neural models.

PUBLICATIONS & SUBMISSIONS

- Qingyan Guo, Rui Wang, Junliang Guo, **Bei Li**, Kaitao Song, Xu Tan, Guoqing Liu, Jiang Bian, Yujiu Yang Connecting large language models with evolutionary algorithms yields powerful prompt optimizers (ICLR2024)
- Chenglong Wang, Hang Zhou, Yimin Hu, Yifu Huo, **Bei Li**, Tongran Liu, Tong Xiao, Jingbo Zhu. ESRL: Efficient Sampling-based Reinforcement Learning for Sequence Generation. (AAAI2024)
- Yuhao Zhang, Kaiqi Kou, **Bei Li**, Chen Xu, Chunliang Zhang, Tong Xiao, Jingbo Zhu. Soft Alignment of Modality Space for End-to-end Speech Translation. (ICASSP2024)
- **Bei Li***, Yuxin Zuo*, Chuanhao Lv, Tong Zheng, Tong Xiao and Jingbo Zhu. Incorporating Probing Signals into Multimodal Machine Translation via Visual Question-Answering Pairs. (Findings of EMNLP2023)
- Yuhao Zhang, Chen Xu, **Bei Li**, Tong Xiao and Jingbo Zhu. Rethinking and Improving Multi-task Learning for End-to-end Speech Translation. (EMNLP2023 Long Paper)
- **Bei Li**, Rui Wang, Junliang Guo, Kaitao Song, Xu Tan, Hany Hassan, Arul Menezes, Tong Xiao, Jiang Bian and JingBo Zhu. Deliberate then Generate: Enhanced Prompting Framework for Text Generation. (In progress)
- Yongyu Mu, Abudurexiti Reheman, Zhiquan Cao, Yuchun Fan, **Bei Li**, Yinqiao Li, Tong Xiao, Chunliang Zhang and Jingbo Zhu Augmenting Large Language Model Translators via Translation Memories (Findings of ACL2023)
- Xiao Xu, **Bei Li**, Chenfei Wu, Shao-Yen Tseng, Anahita Bhiwandiwalla, Shachar Rosenman, Vasudev Lal, Wanxiang Che and Nan Duan. ManagerTower: Aggregating the Insights of Uni-Modal Experts for Vision-Language Representation Learning. (ACL2023 Long Paper)
- **Bei Li**, Yi Jing, Xu Tan, Zhen Xing, Tong Xiao and Jingbo Zhu TranSFormer: Slow-Fast Transformer for Machine Translation. (Findings of ACL2023)
- **Bei Li**, Tong Zheng, Yi Jing, Chengbo Jiao, Tong Xiaoand Jingbo Zhu. Learning Multiscale Transformer Models for Sequence Generation. (ICML2022)
- **Bei Li**, Chuanhao Lv, Zefan Zhou, Tao Zhou, Tong Xiao, Anxiang Ma and Jingbo Zhu. On Vision Features in Multimodal Machine Translation. (ACL2022 Long Paper)
- **Bei Li**, Quan Du, Tao Zhou, Yi Jing, Shuhan Zhou, Xin Zeng, Tong Xiao,and Jingbo Zhu. ODE Transformer: An Ordinary Differential Equation-Inspired Model for Neural Machine Translation. (ACL2022 Long Paper)
- Ye Lin, Yanyangli, Ziyang Wang, **Bei Li**, Quan Du, Tong Xiao,and Jingbo Zhu. Weight Distillation: Transferring the Knowledge in Neural Network Parameters. (ACL2021 Long Paper)
- **Bei Li**, Ziyang Wang, Hui Liu, Quan Du, Tong Xiao, Chunliang Zhang and Jingbo Zhu. Learning Light-Weight Translation Models from Deep Transformer. (AAAI2021 Main Track)
- **Bei Li**, Ziyang Wang, Hui Liu, Yufan Jiang, Quan Du, Tong Xiao, Huizhen Wang and Jingbo Zhu. Shallow-to-Deep Training for Neural Machine Translation. (EMNLP2020 Long Paper)
- **Bei Li**, Hui Liu, Ziyang Wang, Yufan Jiang, Quan Du, Tong Xiao, Jingbo Zhu, Tongran Liu and Changliang Li. Does Multi-Encoder Help? A Case Study on Context-Aware Neural Machine Translation. (ACL2020 Short Paper)
- Chi Hu, **Bei Li**, Yinqiao Li, Ye Lin, Yanyang Li, Chenglong Wang, Tong Xiao, Jingbo Zhu. The NiuTrans System for WNGT 2020 Efficiency Task. (WNGT2020)
- Qiang Wang, **Bei Li**, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F. Wong, Lidia S. Chao. Learning Deep Transformer Models for Machine Translation. (ACL2019 Long Paper)
- **Bei Li**, Yinqiao Li, Chen Xu, Ye Lin, Jiqiang Liu, Hui Liu, Ziyang Wang, Yuhao Zhang, Nuo Xu, Zeyang Wang, Kai Feng, Hexuan Chen, Tengbo Liu, Yanyang Li, Qiang Wang, Tong Xiao and Jingbo Zhu. The NiuTrans Machine Translation Systems for WMT19. (WMT2019)
- **Bei Li**, Qiang Wang, Tong Xiao, Yufan Jiang, Zheyang Zhang, Jiqiang Liu and Qing Yu Meng. On Ensemble Learning of Neural Machine Translation(in Chinese). (Journal of Chinese Information Processing)
- Qiang Wang, **Bei Li**, Jiqiang Liu et al. The NiuTrans Machine Translation System for WMT18. (WMT2018)

- Qiang Wang, **Bei Li**, Jiqiang Liu et at. Toward Building Strong Neural Machine Translation Systems. (CWMT2018)
- Ziyang Wang, Hui Liu, Hexuan Chen, Kai Feng, Zeyang Wang, **Bei Li**, Chen Xu, Tong Xiao and Jingbo Zhu. NiuTrans Submission for CCMT19 Quality Estimation Task. (CCMT2019)
- Nuo Xu, Yinqiao Li, Chen Xu, Yanyang Li, **Bei Li**, Tong Xiao, Jingbo Zhu. Analysis of Back-Translation Methods for Low-Resource Neural Machine Translation. (NLPCC2019)

AWARDS & HONORS

- Baidu Scholarship Finalist — 2022
  Award nomination
- Top ten Graduate students of Northeastern University — 2022
  The May 4th medal
- National Scholarship — 2020 – 2022
  For the first 0.2% Ph.D students in China
- The Excellent Ph.D Student of Northeastern — 2020 – 2021
- The Excellent Master thesis of Liaoning Province — 2020
- The Excellent Master Graduate of Northeastern — 2020
- The Excellent Master Graduate of Liao Ning Province — 2020
- National Scholarship — 2017 – 2019
  For the first 0.2% graduate students in China
- The Excellent Graduate of Shenyang — 2017 – 2018
  For the first graduate in the CSE college.
- The Excellent Master Student of Northeastern — 2017 – 2019
- The First Prize of Scholarship of Northeastern University — 2017 – 2019
- The Third Prize of Scholarship of Northeastern University — 2013 – 2017

ACTIVITIES

- PC members of AAAI, IJCAI — 2020-2023
- PC members of ACL/EMNLP/COLING, ACL rolling review — 2020-2023
- PC members of ICML/Neurips/ICLR — 2022-2023
- Oral presentation on the 59th Annual Meeting of the Association for Computational Linguistics — Jul 2021
- Poster presentation on the 35th AAAI Conference on Artificial Intelligence — Feb 2021
- Poster presentation on the The 2020 Conference on EMNLP — Nov 2020
- Oral presentation on the 58th Annual Meeting of the Association for Computational Linguistics — Jul 2020
- Oral presentation on the 57th Annual Meeting of the Association for Computational Linguistics — Jul 2019
- Poster presentation on the 4th ACL 2019 Fourth Conference on Machine Translation (WMT19) — Jul 2019
- Oral presentation on 17th China National Conference on Computational Linguistics, CCL 2018 — Oct 2018

ACHIEVEMENT

- Third Conference on Machine Translation (WMT18) — 2018
  - 1st Rank in Chinese-English news translation (out of 16 systems, human-evaluation)
  - 3rd Rank in English-Chinese news translation (out of 16 systems, auto-evaluation)
- The 14th China Workshop on Machine Translation (CWMT 2018) — 2018
  - 1st Rank in Chinese-English news translation
  - 2nd Rank in English-Chinese news translation
- Fourth Conference on Machine Translation (WMT19) — 2019
  - 1st Rank in Kazakh-English, English-Kazakh, Gujarati-English news translation tasks (auto-evaluation)
  - 2nd Rank in Russian-English news translation tasks and German-Czech, Czech-German unsupervised tasks (auto-evaluation)
  - 3rd Rank in Chinese-English, German-English, English-German, English-Russian, Lithuanian-English news translation tasks (auto-evaluation)
- The 15th China Conference on Machine Translation (CCMT 2019) — 2019
  - 1st Rank in word-level quality estimation task
  - 1st Rank in sentence-level quality estimation task
- Fifth Conference on Machine Translation (WMT20) — 2020
  - 1st Rank in Japanese-English and English-Japanese (auto-evaluation)

- 2nd Rank in Tamil-English and Inuktitut-English
- 3rd Rank in English-Chinese

▪ Sixth Conference on Machine Translation (WMT21) 2021
- 1st Rank in Chinese-English (human evaluation)

LANGUAGES

▪ Mandarin: Native language.
▪ English: Fluent (speaking, reading, writing)

GOOGLE SCHOLAR

▪ https://scholar.google.com/citations?user=wzbJ5EIAAAAJ&hl=en
▪ 971 citations up to now 2018 – present

SKILLS

Programming Language
Python, C, C++, Shell,

Deep Learning Platform
Pytorch, Tensorflow, Keras, NiuTrans.Tensor, Theano

Others
LaTeX

RESEARCH INTERESTS

Natural Language Processing
Machine Translation, Language Modeling, Machine Learning, Presentation Learning and Pre-training.

Machine Learning
Structured Prediction, Neural Architecture Search, Reinforcement Learning,

REFERENCES

▪ **Prof. Tong Xiao** (Ph.D. Supervisor)
Professor & Chairman of the Natural Language Processing Lab in Northeastern University
Northeastern University
NO. 3-11, Wenhua Road, Heping District, Shenyang, P. R. China
xiaotong@mail.neu.edu.cn

[*Last updated on 17 January 2024*]