TẠP CHÍ KHOA HỌC
TRƯỜNG ĐẠI HỌC SƯ PHẠM TP HỒ CHÍ MINH
Tập 19, Số 10 (2022): 1749-1755
Website: https://journal.hcmue.edu.vn
ISSN: 2734-9918

HO CHI MINH CITY UNIVERSITY OF EDUCATION
JOURNAL OF SCIENCE
Vol. 19, No. 10 (2022): 1749-1755
https://doi.org/10.54607/hcmue.js.19.10.3633(2022)

Research Article

TOWARDS CROSS-ATTENTION PRE-TRAINING IN NEURAL MACHINE TRANSLATION

Pham Vinh Khang**, Nguyen Hong Buu Long
University of Science, Vietnam National University Ho Chi Minh City, Vietnam
**Corresponding author: Pham Vinh Khang– Email: vkhangpham@gmail.com*
Received: October 11, 2022; Revised: October 28, 2022; Accepted: October 29, 2022

ABSTRACT

The advent of pre-train techniques and large language models has significantly leveraged the performance of many natural language processing (NLP) tasks. However, pre-trained language models for neural machine translation remain a challenge as little information about the interaction of the language pair is learned. In this paper, we explore several studies trying to define a training scheme to pre-train the cross-attention module between the encoder and the decoder by using the large-scale monolingual corpora independently. The experiments show promising results, proving the effectiveness of using pre-trained language models in neural machine translation.

Keywords: cross-attention; cross-lingual; natural language processing; neural machine translation; pre-training, language model

1. Introduction

Recent studies have constantly demonstrated the effectiveness of using pre-trained language models in a wide range of NLP problems. Inspired by the seminal work of GPT (Radford, Narasimhan, Salimans, & Sutskever, 2018) and BERT (Devlin, Chang, Lee, & Toutanova, 2018), many approaches attempted to use the Transformer-based model (Vaswani, et al., 2017) to learn the universal representations of words in a self-supervised manner, then transfer this knowledge to downstream tasks. This training method allows the model to take advantage of large monolingual data, thereby learning better, especially for tasks suffering from low-resource data like neural machine translation (NMT). However, such a technique cannot capture the semantic interaction between the source and target languages, which is critical to the decoding phase of the translation model. Consequently, the cross-attention module between the Transformer encoder and decoder has to be randomly initialized and trained from scratch, leading to a bottleneck in the fine-tuning process.

Cite this article as: Pham Vinh Khang, & Nguyen Hong Buu Long (2022). Towards cross-attention pre-training in neural machine translation. *Ho Chi Minh City University of Education Journal of Science, 19*(10), 1749-1755.

Pre-training a whole sequence-to-sequence model (Lewis et al., 2019) was proposed to tackle the above problem. However, it is used as a pre-trained target-side language model, i.e., it still lacks a mapping between the two languages. One notable approach is to train a cross-lingual word embedding on the concatenated monolingual corpora of the language pair (Lample & Conneau, 2019), which achieved state-of-the-art performance in unsupervised neural machine translation at the time. A recent work (Ren et al., 2021) further extends this idea. They define a language-independent interface between the encoder and the decoder, which acts as a shared cross-lingual space between the two languages. The encoder and the decoder are pre-trained independently, but different from the methods above, they have an interface component in common. This interface will be removed in the fine-tuning phase. Therefore, we can naturally connect the encoder and the decoder with the pre-trained cross-attention weights.

Given the promising results from the above approach, especially on low-resource language pairs, we aim to investigate the effectiveness of such a semantic interface and conduct experiments following the original work. The results show that cross-lingual embedding space can be crucial for pre-training NMT.

2. Research Objectives and Methodology

2.1. Research Objectives

2.1.1 Pre-training neural machine translation

As stated above, how to leverage the knowledge learned on large-scale monolingual data in NMT remains a challenge due to the difference in the training objective between NMT and language modeling. In language modeling, state-of-the-art models like GPT or BERT aim to predict the word in the same language, even for multi-lingual settings. In contrast, in NMT, the target is to be able to translate words/sentences/documents between the language pair. The problem becomes even more complex for languages having significant differences in morphology and syntax, where previous contextual knowledge has little usage in finding a mapping between two languages (Yang et al., 2020). This problem is the main reason why directly transferring pre-trained embeddings to the translation model brings limited results, thus motivating researchers to search for a more efficient training paradigm.

One of the first lines of work in pre-training NMT was based on the idea of unsupervised machine translation (Lample, Conneau, Denoyer, & Ranzato, 2017), i.e., learning without any labeled data. Using a concatenated monolingual corpus, the model first extracts the features of the input sentences to a shared latent space, then learns to reconstruct the language from that space. The authors then continued to propose a cross-lingual language model (XLM) (Lample & Conneau, 2019), which achieves SOTA in both unsupervised and supervised machine translation. The main contribution of this study in pre-training NMT is the introduction of the translation language modeling (TLM) objective. The training sample

contains two parallel sentences from the source and target languages to leverage the available parallel data. Then, similar to BERT MLM, they randomly mask several tokens, but now in both sentences. By doing so, the model is encouraged to identify the current masked token based on the context of both the original sentence and the translated one, thus learning about the alignment of the two languages. While the results are promising, such a training objective does not guarantee that the model can learn meaningful interaction of the language pair.

A second approach attempts to pre-train the whole sequence-to-sequence (seq2seq) model, including the cross-attention module (Lewis et al., 2019; Liu et al., 2020; Song, Tan, Qin, Lu, & Liu, 2019). BART (Lewis et al., 2019) is a denoising autoencoder trained to reconstruct a corrupted document, and it is essentially a traditional seq2seq Transformer. Although the transformation of the language pair can be learned in the pre-training phase, in fine-tuning, the entire BART model is used only as a single pre-trained decoder; thus, it requires random initialization on the encoder side.

Besides two main approaches, another work (Weng, Yu, Huang, Cheng, & Luo, 2020) focuses on leveraging BERT-based pre-trained models for machine translation by defining a fusion mechanism to extract learned contextual knowledge, then distill them continuously during the fine-tuning process. A more recent study (Ren et al., 2021) proposes yet another idea for pre-training NMT with the main target of capturing the cross-lingual transformation of word representation during the pre-training phase. The details of this approach will be described in the next section.

For Vietnamese, PhoBERT (Nguyen & Nguyen, 2020) and BARTpho (Tran, Le, & Nguyen, 2022) are the two most prominent pre-trained language models. While PhoBERT takes a similar approach to RoBERTa (Liu et al., 2019), which optimizes the training scheme of BERT for more robust performance, BARTpho is a trained seq2seq model that has a significant performance gain compared to the baseline mBART (Liu et al., 2020) for Vietnamese. Both models consider many concerns regarding the morphology and syllable structure of Vietnamese, thereby improving the effectiveness of pre-training for downstream tasks.

2.1.2. Pre-training cross-attention module

Although there are plenty of studies on pre-trained NMT, few can learn the semantic interface between the encoder and decoder. To address this, Ren et al. (Ren et al., 2021) proposed a new training method with a semantic interface (SemFace) between the two components of a seq2seq model. This interface acts like a language-independent embedding space, similar to Lample et al. (Lample & Conneau, 2019), to which the encoder extracts features and the decoder gets the contextual information to decode the sample. More interestingly, during the pre-training phase of the decoder, the cross-attention weights can also be learned by interacting with the SemFace. After pre-training, we can remove the SemFace, and directly connect the encoder with the decoder without concerning whether the

cross-lingual mapping is retained. The complete training procedure includes two phases: first, we pre-train the encoder and the decoder separately with the SemFace on monolingual corpora; then, we fine-tune the seq2seq model, SemFace excluded, on parallel data. The model is essentially an encoder-decoder Transformer architecture.

When pre-train the encoder, we first concatenate two monolingual corpora to learn joint BPE embeddings. Unlike the original paper, we randomly initialize the final linear projection weights instead of using a pre-trained cross-lingual embeddings. The linear projection layer is also the semantic interface for this setting. The training objective includes the standard masked language modeling (MLM) (Vaswani et al., 2017), and the mean squared error (MSE) loss between the SemFace weights and the final hidden units of the encoder. Ren et al. (Ren et al., 2021) suggest that using MSE loss will encourage the encoder output to be similar to the cross-lingual embeddings, which is critical to help replicate the encoder output when pre-training the decoder.

The pre-training process for the decoder is similar to training a language model with causal language modeling (CLM) loss or MLM loss, except for the cross-attention module. Following the work by Lample et al. (Lample, Conneau, Denoyer, & Ranzato, 2017), to simulate the encoder output, we create a noisy sample by randomly dropping or shuffling the original input sentence. The hidden states of the decoder need to attend to the content simulated by the encoder output created by passing the such noisy sample through the semantic interface.

After the pre-training phase, we now have an encoder and a decoder with a fully trained cross-attention module. The encoder is trained to extract a context-free language-independent space from the input, and the decoder tries to leverage information from that space via the attention mechanism. By doing this, we can pre-train the entire seq2seq model utilizing monolingual data. Thus, we can remove the SemFace, connect the two components, and then fine-tune on parallel data using the standard cross-entropy loss.

2.2. Methodology

We run our experiments on the translation task for German-English. For the pre-training phase, we used the sub-sample of the Wikipedia dataset for said languages. For fine-tuning phase, we used the IWSLT'14[1] German to English dataset.

We implemented our method using the fairseq (Ott et al., 2019) library. We first learn 30,000 BPE codes on the concatenated monolingual dataset. Unlike Ren et al. (Ren et al., 2021), we did not learn cross-lingual embeddings to initialize for the semantic interface. We instead pre-trained the encoder with the randomly initialized weights, then used the trained weights of the linear embedding to calculate cross-attention when pre-train the decoder. We used a Transformer encoder-decoder with six layers, each with an embedding dimension of 1024, and a feed-forward size of 4096. We used Adam optimizer (Kingma & Ba, 2014) with

[1] https://workshop2014.iwslt.org

a learning rate of 0.0001 and 4,000 warm-up steps, decaying based on the inverse squared root of the update number. The batch size is 32 and the max number of tokens in one sample is 128.

3. Results and Discussion

3.1. Result

Table 1. *Experiment results*

Model	Cross-entropy loss	BLEU score
Transformer	3.625	**34.22**
SemFace small	**3.4**	27.9
SemFace	3.6	29.03

Table 1 displays the results. We compared our model with the Transformer baseline. The difference between the two SemFace models is the amount of monolingual data for pre-training. Due to hardware limitations, we only use 5 and 15 million sentences for each language in SemFace small and SemFace, respectively.

3.2. Discussion

The results from the investigated method are below that of the baseline, which is different from the results reported in the study by Ren et al. (Ren et al., 2021). Although both SemFace models have good performance in terms of cross-entropy loss, the BLEU scores could not match that of the baseline. One reason for this could come from the fact that we re-use the shared vocabulary for the fine-tuning task rather than using language-specific vocabulary, leading to an unstable translation when using fairseq. Another point that could contribute to such difference is that we did not use pre-trained cross-lingual embeddings as in the original model. Thus, our model does not benefit from context-free cross-lingual knowledge. These reasons will be further investigated in our future work, along with potential improvement for the training scheme.

The results also demonstrate the effectiveness of using monolingual data for pre-training. Specifically, the SemFace model, which was pre-trained with a larger monolingual dataset, also gains an increase of more than 1 BLEU compared to the SemFace small.

4. Conclusion and future work

While there are various approaches to pre-train an NMT model, few consider the semantic interface between the source and target languages, i.e., the cross-attention module between the encoder and the decoder. In this paper, we investigated and conducted experiments on one of such approaches. Due to several limitations, we cannot reproduce the results in the original paper. Nevertheless, we realize this can be an effective pre-training method for neural machine translation, especially for low-resource languages. We will continue to explore this idea and address current limitations in future work.

❖ ***Conflict of Interest:*** *Authors have no conflict of interest to declare.*

REFERENCES

Artetxe, M., Labaka, G., & Agirre, E. (2018). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 789-798.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* arXiv preprint arXiv:1810.04805.

Kingma, D. P., & Ba, J. (2014). *Adam: A Method for Stochastic Optimization.* arXiv preprint arXiv:1412.6980.

Lample, G., & Conneau, A. (2019). *Cross-lingual Language Model Pretraining.* arXiv preprint arXiv:1901.07291.

Lample, G., Conneau, A., Denoyer, L., & Ranzato, M. (2017). Unsupervised Machine Translation Using Monolingual Corpora Only.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., . . . Zettlemoyer, L. (2019). *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension.* arXiv preprint arXiv:1910.13461.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., . . . Zettlemoyer, L. (2020). Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics, 8*, 726-742.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., . . . Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach.* arXiv preprint arXiv:1907.11692.

Nguyen, Q. D., & Nguyen, T. A. (2020). *PhoBERT: Pre-trained language models for Vietnamese.* arXiv preprint arXiv:2003.00744.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., . . . Auli, M. (2019). fairseq: A Fast, Extensible Toolkit for Sequence Modeling. *Proceedings of NAACL-HLT 2019: Demonstrations*.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving Language Understanding by Generative Pre-Training.* OpenAI.

Ren, S., Zhou, L., Liu, S., Wei, F., Zhou, M., & Ma, S. (2021). SemFace: Pre-training Encoder and Decoder with a Semantic Interface for Neural Machine Translation. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 4518-4527.

Song, K., Tan, X., Qin, T., Lu, J., & Liu, T.-Y. (2019). *MASS: Masked Sequence to Sequence Pre-training for Language Generation.* arXiv preprint arXiv:1905.02450.

Tran, N. L., Le, D. M., & Nguyen, D. Q. (2022). *BARTpho: Pre-trained Sequence-to-Sequence Models for Vietnamese.* arXiv preprint arXiv:2109.09701.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention is All you Need. *NIPS*.

Weng, R., Yu, H., Huang, S., Cheng, S., & Luo, W. (2020). Acquiring Knowledge from Pre-trained Model to Neural Machine Translation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 9266-9273.

Yang, J., Wang, M., Zhou, H., Zhao, C., Zhang, W., Yu, Y., & Li, L. (2020). Towards making the most of bert in neural machine translation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 9378-9385.

HƯỚNG ĐẾN TIỀN HUẤN LUYỆN CROSS-ATTENTION TRONG DỊCH MÁY BẰNG NƠ-RON

Phạm Vĩnh Khang*, Nguyễn Hồng Bửu Long

Trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Thành phố Hồ Chí Minh, Việt Nam

**Tác giả liên hệ: Phạm Vĩnh Khang– Email: vkhangpham@gmail.com*

Ngày nhận bài: 11-10-2022; ngày nhận bài sửa: 28-10-2022; ngày duyệt đăng: 29-10-2022

TÓM TẮT

Sự xuất hiện của các kĩ thuật tiền huấn luyện (pre-training) và những mô hình ngôn ngữ đã cải thiện đáng kể nhiều giải pháp của các bài toán trong lĩnh vực xử lí ngôn ngữ tự nhiên (XLNNTN). Tuy nhiên, việc ứng dụng những mô hình ngôn ngữ đã được tiền huấn luyện (pre-trained language models) vào bài toán dịch máy vẫn còn là một vấn đề khó, vì mô hình ngôn ngữ không học được thông tin về sự tương tác giữa cặp ngôn ngữ trong quá trình tiền huấn luyện. Trong bài báo này, chúng tôi sẽ tìm hiểu một số công trình nghiên cứu về việc tiền huấn luyện mô-đun cross-attention giữa encoder và decoder bằng cách sử dụng ngữ liệu đơn ngữ lớn. Kết quả thí nghiệm đã chứng minh được sự hiệu quả của việc sử dụng mô hình ngôn ngữ được tiền huấn luyện cho bài toán dịch tự động)

Từ khóa: cross-attention; xuyên ngữ; xử lí ngôn ngữ tự nhiên; dịch máy bằng nơ-ron; tiền huấn luyện; mô hình ngôn ngữ