Leveraging Cross-Lingual Knowledge from Pre-Trained Models for Low-Resource Neural Machine Translation

Anonymous ACL submission

Abstract

Neural machine translation (NMT) quality significantly depends on large parallel corpora, making low-resource language translation a challenge. This paper introduces a novel approach that leverages cross-lingual alignment knowledge from multilingual pre-trained language models (PLMs) to enhance low-resource NMT. Our method segments the translation model into source encoding, target encoding, and alignment modules, each initialized with different pre-trained BERT models. Experiments on four translation directions with two low-resource language pairs demonstrate significant BLEU score improvements, validating the efficacy of our approach.

1 Introduction

The quality of neural machine translation (NMT) heavily depends on rich parallel corpora, making NMT perform poorly with low-resource languages (Arivazhagan et al., 2019; Haddow et al., 2022). The key challenge in handling low-resource languages lies in acquiring monolingual semantics and bilingual alignment knowledge. Traditional NMT systems, reliant on large parallel datasets, often fail to capture these knowledge under data scarcity. Pre-trained language models (PLMs), by acquiring knowledge from extensive monolingual corpora, offer a promising solution to this problem (Liu et al., 2020; Baziotis et al., 2020). By leveraging PLMs pre-trained on large monolingual corpora, we can inject valuable linguistic knowledge into NMT systems, indirectly alleviating the lack of resources.

Previous research has explored combining PLMs with translation models to better utilize the prior knowledge in PLMs. Guo et al. (2020) proposed to use BERT models for source and target languages as the encoder and decoder respectively, and employ adapters to learn bilingual alignment for high-quality non-autoregressive translation. Weng et al. (2022) initialized the encoder of the translation model with mBERT, and used a Layer-wise Coordination Structure and multi-task learning to enhance autoregressive translation. Duan and Zhao (2023) split the decoder into separate history encoding and generation prediction modules to effectively utilize target language BERT for improved autoregressive translation. Pang et al. (2024) modularized the translation model into encoder, decoder, and transfer modules, and explored to efficiently use monolingual and bilingual knowledge while mitigating catastrophic forgetting.

However, these methods primarily focus on monolingual knowledge from PLMs, failing to effectively utilize cross-lingual alignment knowledge from multilingual PLMs (Muller et al., 2021). To address this issue, we propose a low-resource NMT model that leverages cross-lingual alignment knowledge learned from multilingual PLMs to improve translation quality. This knowledge is crucial in resource-scarce settings as models struggle to learn high-quality alignments from limited parallel corpora. Specifically, we partition the translation model into source encoding, target encoding, and alignment modules, initializing them with different pre-trained models according to their functions. Source and target encoding modules are initialized with respective language BERT models to obtain monolingual encoding capabilities, while the alignment module is initialized with multilingual BERT to utilize cross-lingual alignment knowledge. Experiments on four translation directions of two low-resource parallel corpora show significant BLEU score improvements, validating the effectiveness of our approach.

2 Related Work

2.1 Two-part Decoder

Previous efforts combining PLMs with NMT models have primarily focused on utilizing monolingual knowledge from the source language, with limited success in using target language PLMs to improve translation quality (Weng et al., 2022). To address this issue, the Two-part Decoder method (Duan and Zhao, 2023) reconstructs the translation model's decoder into two independent components: a history encoding module and a generation module. The history encoding module encodes previously generated information, while the generation module generates translations token by token. This approach aligns the history encoding module more closely with the target language BERT, enabling the model to better utilize monolingual knowledge from target language, thereby improving translation quality. Additionally, auxiliary tasks like MLM (Devlin et al., 2018) and knowledge distillation (Yang et al., 2020) provide extra training signals to reinforce learned representations, further enhancing model performance.

2.2 MoNMT

Fine-tuning PLMs can lead to catastrophic forgetting, where models lose previously learned domain-specific and monolingual knowledge (French, 1999). To mitigate this, the MoNMT approach (Pang et al., 2024) modularizes the translation model into encoder, decoder, and transfer modules. The encoder and decoder are trained on monolingual data to learn monolingual encoding and generation knowledge, while the transfer module is trained on parallel corpora to learn bilingual alignment knowledge. This modular approach helps retain pre-trained knowledge and allows independent updates and improvements for each module, which is particularly beneficial for low-resource languages by enabling models to adapt and integrate new data without extensive retraining, maintaining efficiency and effectiveness.

3 Methodology

3.1 Model Architecture

To better leverage cross-lingual knowledge from PLMs, we propose a low-resource NMT model that utilizes bilingual knowledge from pre-trained models. Inspired by the Two-part Decoder method (Duan and Zhao, 2023), our architecture partitions the translation model into source encoding, target encoding, and alignment modules. As shown in Figure 1, both source and target encoding modules consist of multiple layers, each containing a self-attention sublayer and a feed-forward sublayer. The alignment module also consists of multiple layers, each containing a cross-attention sublayer and a feed-forward sublayer. The source and target encoding modules focus on obtaining the monolingual knowledge of their respective languages, while the alignment module ensures effective alignment of representations learned from both languages. This architecture ensures that each part of the model is dedicated to its specific task, thereby improving overall performance.

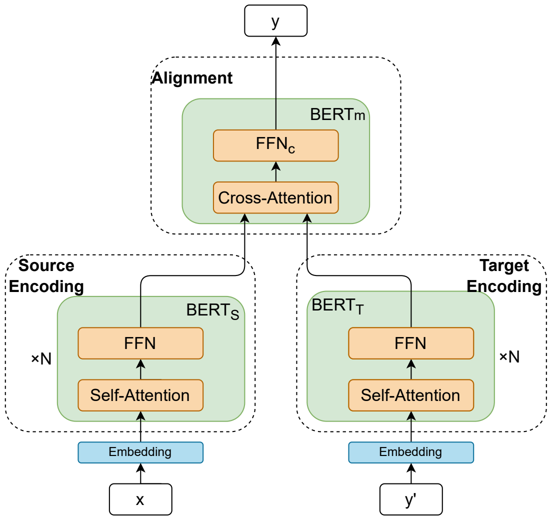

Figure 1: Architecture of the proposed low-resource NMT model, partitioned into three modules: source encoding, target encoding, and alignment. These modules are initialized with $\mathrm{BERT_S}$, $\mathrm{BERT_T}$, and $\mathrm{BERT_m}$ respectively.

3.2 Initialization with Pre-Trained BERT Models

We use different BERT models to provide the necessary prior knowledge for each module. Specifically, source language BERT($\mathrm{BERT_S}$) and target language BERT($\mathrm{BERT_T}$) initialize the source and target encoding modules, respectively, capturing richer contextual information and semantic relationships for better monolingual representations. The alignment module is initialized with multilingual BERT($\mathrm{BERT_m}$), whose cross-lingual alignment knowledge serves as prior knowledge for translation alignment, improving low-resource translation quality. This initialization strategy ensures that each module is equipped with the most relevant linguistic knowledge from the start, enabling the model to effectively utilize this knowledge during training and translation. Using multilingual BERT for the alignment module is particularly important as it brings valuable cross-lingual alignment knowl-

Dataset	Train	Valid	Test
En-Nb	142,906	2,000	2,000
De-Nb	110,248	2,000	2,000

Table 1: The size of datasets

edge critical for low-resource translation tasks.

3.3 Training Objective

We fine-tune our model on bilingual parallel corpora, focusing on comparing the impact of cross-lingual alignment knowledge from $\mathrm{BERT_m}$ on translation performance. Therefore, we did not incorporate complex multi-task training like (Duan and Zhao, 2023). The training objective is defined as:

$$L = -\log P(y|x, \theta_{\mathrm{BERT_S}}, \theta_{\mathrm{BERT_T}}, \theta_{\mathrm{BERT_m}})$$

where (x, y) denotes a pair of parallel sentences.

4 Experiments

4.1 Datasets

We evaluate our model on two low-resource language pair datasets. For English-Norwegian(en-nb), we use OPUS-100 data (Zhang et al., 2020), following the default data split. For German-Norwegian(de-nb), we use the KDE4 dataset (Tiedemann, 2012). Since KDE4 does not divide the default test set, we randomly selected 2000 items as the validation set and 2000 items as the test set. Table 1 provides detailed data statistics.

4.2 Model Configurations

For the monolingual BERT models, we use *bert-base-cased* [1] for English, *bert-base-german-cased* [2] for German, and *nb-bert-base* [3] for Norwegian. For the multilingual BERT model, we use *bert-base-multilingual-cased* [1].

Our model parameters are consistent with those of the pre-trained models, using their tokenizers and vocabularies without modification. Note that when we initialize the alignment module with mBERT, we will replace the vocabulary used for the final prediction with the vocabulary of $\mathrm{BERT_m}$ to ensure that cross-lingual knowledge is fully utilized.

[1]https://github.com/google-research/bert
[2]https://www.deepset.ai/german-bert
[3]https://github.com/NBAiLab/notram

The consistency in model parameters and tokenization ensures that our initialization process is seamless and that the pre-trained knowledge is effectively transferred to the translation model. This setup also facilitates reproducibility and comparability of results across different experiments.

4.3 Results

We compared the BLEU values of the randomly initialized alignment module (Random Init) and the alignment module initialized with $\mathrm{BERT_m}$ ($\mathrm{BERT_m}$ Init). For the baseline model, we built a Transformer (Transformer) (Vaswani et al., 2017) based on the hyper-parameters of BERT-base and modified the number of Decoder layers from 12 to 24 to keep the parameter scale close.

Our experimental results are shown in Table 2. The results show that module partitioning and initialization of source and target encoding modules can effectively improve the quality of low-resource translation, even if the alignment module is randomly initialized, because it can learn alignment knowledge from bilingual data. This indicates that the monolingual knowledge from source BERT and target BERT effectively improves the encoding representation quality of both languages, showcasing the effectiveness of module partitioning. On this basis, using $\mathrm{BERT_m}$ to initialize the alignment module further improves the translation quality. This shows that our model can effectively utilize the cross-language alignment knowledge from $\mathrm{BERT_m}$, indicating the importance of utilizing prior alignment knowledge for low-resource translation.

To further verify the effectiveness of cross-lingual knowledge from $\mathrm{BERT_m}$, we initialized the alignment module with English BERT ($\mathrm{BERT_S}$ Init) and Norwegian BERT ($\mathrm{BERT_T}$ Init) separately for the en-nb task.

The results shown in Table 3. Using the $\mathrm{BERT_S}$ to initialize the alignment module is even harmful to the model, because the knowledge of the source language is not helpful for the generation of the target language. Using the $\mathrm{BERT_T}$ to initialize the alignment module can also help the model because it can provide knowledge of generating the target language, indicating the rationality of decomposing the Decoder into two parts: encoding and generation, which verifies the view of (Duan and Zhao, 2023). However, it is still lower than the result of initialization with $\mathrm{BERT_m}$, indicating that alignment knowledge is more important for low-resource translation tasks because it is difficult

Model	En ⇔ Nb		De ⇔ Nb	
	En ⇒ Nb	Nb ⇒ En	De ⇒ Nb	Nb ⇒ De
Transformer	12.69	23.20	23.59	21.98
Random Init	18.04	30.67	25.13	24.23
BERT_m Init	27.79	35.58	31.41	29.59

Table 2: BLEU scores of the baseline and our model on the OPUS-100 En-Nb and the KDE4 De-Nb task. *Random Init* and *$BERT_m$ Init* represent initializing the alignment module randomly or using BERT_m, respectively.

Model	En ⇒ Nb
Random Init	18.04
BERT_S Init	17.15
BERT_T Init	26.99
BERT_m Init	27.79

Table 3: BLEU scores of our model in the En-Nb direction. *Random Init*, *$BERT_S$ Init*, *$BERT_T$ Init*, and *$BERT_m$ Init* represent different ways to initialize the alignment module. When using *$BERT_S$ Init*, the final predicted vocabulary is the vocabulary of BERT_m vocabulary.

for the model to learn them from resource-scarce bilingual data.

5 Conclusion

Existing methods of enhancing NMT with PLMs fail to effectively utilize cross-lingual alignment knowledge from multilingual PLMs. To address this, we propose a low-resource NMT model that leverages bilingual knowledge from pre-trained models. By initializing different parts of the model according to the functions of BERT, our approach effectively utilizes monolingual semantic knowledge and cross-lingual alignment knowledge from PLMs, significantly improving translation quality for low-resource languages. Our method not only demonstrates the potential of cross-lingual alignment knowledge but also lays the foundation for future research in effectively combining different types of PLMs for various NLP tasks.

6 Limitations

Although our work has achieved some success, there are still existing the following limitations:

- **Model Variety** Our current approach is limited to BERT-type pre-trained models, which may not be easily adaptable to seq2seq pre-trained models like BART. Future work will explore ways to utilize knowledge from various PLMs, maximizing both monolingual and bilingual knowledge.

- **Dataset Variety** Due to constraints on dataset availability and PLM accessibility, our experiments are currently limited to low-resource languages within specific language families. Further validation is needed to determine the effectiveness of our approach across different language families and cross-language translation tasks.

- **Large Models** Large models contain richer knowledge and possess capabilities not found in smaller models. However, due to computational resource limitations, we have yet to explore enhancing low-resource translation with large models. Future research will investigate leveraging large models to further improve low-resource translation if conditions permit.

References

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, et al. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

Christos Baziotis, Barry Haddow, and Alexandra Birch. 2020. Language model prior for low-resource neural machine translation. *arXiv preprint arXiv:2004.14928*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Sufeng Duan and Hai Zhao. 2023. Encoder and decoder, not one less for pre-trained language model sponsored nmt. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3602–3613.

Robert M French. 1999. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135.

Junliang Guo, Zhirui Zhang, Linli Xu, Hao-Ran Wei, Boxing Chen, and Enhong Chen. 2020. Incorporating bert into parallel sequence decoding with adapters. *Advances in Neural Information Processing Systems*, 33:10843–10854.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pretraining for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamé Seddah. 2021. First align, then predict: Understanding the cross-lingual ability of multilingual bert. *arXiv preprint arXiv:2101.11109*.

Jianhui Pang, Baosong Yang, Derek F Wong, Dayiheng Liu, Xiangpeng Wei, Jun Xie, and Lidia S Chao. 2024. Monmt: Modularly leveraging monolingual and bilingual knowledge for neural machine translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11560–11573.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218. Citeseer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Rongxiang Weng, Heng Yu, Weihua Luo, and Min Zhang. 2022. Deep fusing pre-trained models into neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11468–11476.

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 *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9378–9385.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. *arXiv preprint arXiv:2004.11867*.