MoNMT: Modularly Leveraging Monolingual and Bilingual Knowledge for Neural Machine Translation

Jianhui Pang[1*], Baosong Yang[2†], Derek Fai Wong[1†], Dayiheng Liu[2], Xiangpeng Wei[2], Jun Xie[2] and Lidia Sam Chao[1]
[1]University of Macau, [2]Alibaba Group
{nlp2ct.pangjh3, pemywei}@gmail.com
{yangbaosong.ybs, liudayiheng.ldyh, qingjing.xj}@alibaba-inc.com
{derekfw, lidiasc}@um.edu.mo

Abstract

The effective use of monolingual and bilingual knowledge represents a critical challenge within the neural machine translation (NMT) community. In this paper, we propose a modular strategy that facilitates the cooperation of these two types of knowledge in translation tasks, while avoiding the issue of catastrophic forgetting and exhibiting superior model generalization and robustness. Our model is comprised of three functionally independent modules: an encoding module, a decoding module, and a transferring module. The former two acquire large-scale monolingual knowledge via self-supervised learning, while the latter is trained on parallel data and responsible for transferring latent features between the encoding and decoding modules. Extensive experiments in multi-domain translation tasks indicate our model yields remarkable performance, with up to 7 BLEU improvements in out-of-domain tests over the conventional pretrain-and-finetune approach. Our codes are available at `https://github.com/NLP2CT/MoNMT`.

Keywords: Machine Translation, Monolingual and Bilingual Knowledge, Catastrophic Forgetting

1. Introduction

The Neural Machine Translation (NMT) models have exhibited impressive performance in translation tasks (Vaswani et al., 2017), yet their effectiveness is heavily dependent on the availability of bilingual data. To address this limitation, recent research has started to exploit monolingual knowledge derived from Pretrained Language Models (PLMs) (Rothe et al., 2020; Zhu et al., 2020; Liu et al., 2020, 2021b; Üstün et al., 2021; Zhu et al., 2023; Liu et al., 2023; Pang et al., 2024a). [1] Monolingual data can be effortlessly amassed for different domains and languages, whereas bilingual data consists of one-to-one translation examples that are indispensable in improving the translation models. Although recent research has demonstrated some translation capabilities in Large Language Models using large-scale monolingual data, they are prone to off-target translation, hallucination, and monotonic errors, and may exhibit performance gaps in comparison to strong supervised models (Zhu et al., 2023; Pang et al., 2024b). The conventional method for utilizing both monolingual data and bilingual data is the pretrain-and-finetune (PF) paradigm, which has proven effective in enhancing the performance of translation models (Liu et al., 2020; Lewis et al., 2020). However, the fine-tuning step involves adjusting the entire network by completely or partially updating model parameters. This can erase domain-specific and cross-lingual monolingual knowledge acquired by the model, resulting in a translation model susceptible to insufficient generalization and robustness capabilities, commonly referred to as catastrophic forgetting (French, 1999; Thompson et al., 2019; Yang et al., 2020). Hence, a natural question arises on *how to successfully synergize monolingual and bilingual knowledge to further enhance the translation capacity of NMT models.*

To approach this problem, we shift our attention to the translation process of the human being. The human translator does not "forget" language understanding and generation abilities while learning new translation tasks. This stems from that the human brain has a hierarchical modular organization and is able to functionally learn and memorize different tasks (Graziano and Aflalo, 2007; Zhang et al., 2023). On the contrary, the current NMT model is a functional coupling system. Although its encoder-decoder abstraction is conceptually functional independent (Sutskever et al., 2014; Vaswani et al., 2017), the translating function is coupled with the encoding and decoding functions. Therefore, learning each of the monolingual and bilingual knowledge affects and covers the other at the training time. Inspired by the human translator, a potential solution to this problem is to decompose the NMT model into function-specific modules.

∗This research was accomplished when Jianhui Pang was interning at Alibaba DAMO Academy.

†Baosong Yang and Derek Fai Wong are co-corresponding authors.

[1]This paper primarily focuses on the monolingual usage of PLMs. Back-translation is another option that requires a qualified reverse-translation model.

LREC-COLING 2024, pages 11560–11573

Accordingly, we propose a **Mo**dular **N**eural **M**achine **T**ranslation model (MoNMT), which makes the encoding, transferring, and decoding functions contribute to the translation task independently. Our model consists of three modules: 1) The encoding module (Enc) is to encode source text into source-oriented representations; 2) The transferring module (Trans) is responsible for transferring them into target-oriented representations; and 3) The decoding module (Dec) generates the target sentence. A major challenge is how to link up the three modules well, especially when they are not jointly optimized. To approach this issue, we first choose the monolingual sentence denoising as the training objective of Enc and Dec rather than the other self-supervised learning methods. The two modules are therefore more in line with the sequence-to-sequence task and Dec can generate sentences conditional to the source features. Then, we build Trans upon Enc, and train it to generate target-oriented representations by feeding the source ones. Trans is trained on the parallel corpus and optimized by assigning translation cross-entropy loss with freezing Enc and Dec, for which we also propose an optimization alternative with an auxiliary loss in the ablation study.

To evaluate the efficacy of MoNMT in leveraging both monolingual and bilingual knowledge, our study encompasses in-domain and out-of-domain translation tasks to assess the model's performance. During the training phase, we initially train the Enc and Dec modules using multi-domain monolingual data, followed by training the Trans module with domain-specific bilingual data. Our findings consistently demonstrate the exemplary performance of MoNMT in both in-domain and out-of-domain translation tasks. Notably, when exclusively trained on bilingual knowledge from the Subtitles domain, the model shows a substantial improvement of up to 7.0 BLEU in German-to-English multi-domain tasks, showcasing its enhanced generalization and robustness. Moreover, MoNMT exhibits effectiveness across diverse corpus sizes and translation directions, and shows an approximate 1.0 BLEU enhancement in low-resource translation tasks. Beyond improving translation abilities, MoNMT offers several desirable practical features:

- **Simple**: The proposed method is straightforward and readily implementable, utilizing the existing NMT architecture with minimal alterations required. Additionally, the training process for each module remains uncomplicated.

- **Parameter-Efficient**: The reusability of encoding and decoding modules for subsequent tasks significantly improves the efficiency of computational resources in the practical deployment of a translation system.

- **Scalable**: The scalability of each module can be dynamically adjusted to accommodate data volume requirements. Rather than fine-tuning the entire model, users can tailor the capacity of the transfer module based on the bilingual dataset size, thereby preventing overfitting or underfitting issues. This results in a more robust and customized approach.

2. Related Works

Monolingual data can be utilized for pretraining language models (PLMs), thus facilitating the development of enhanced translation models. PLMs are trained on large volumes of monolingual text using self-supervised training objectives (Devlin et al., 2019; Brown et al., 2020; Lewis et al., 2020; Liu et al., 2020), which equips these models with significant linguistic and domain knowledge. However, recent studies have indicated that large language models (LLMs) trained without parallel data may exhibit translation errors such as Off-target translation, Hallucination, and Monotonic translation, and potentially underperform compared to supervised methods (Zhu et al., 2023; Pang et al., 2024b). Additionally, Jiao et al. (2023) discovered that ChatGPT, a powerful LLM, lacks domain robustness when it comes to translation tasks.

The pretrain-and-finetune (PF) method, which combines both monolingual and bilingual knowledge, is a conventional approach that effectively enhances in-domain tasks (Liu et al., 2020, 2021a). However, directly fine-tuning the entire model using in-domain bilingual data may result in catastrophic forgetting, leading to the loss of monolingual knowledge and poor performance in out-of-domain scenarios (Thompson et al., 2019). In addition, existing research in the multilingual machine translation field employs strategies such as integrating adapters into encoders and decoders. These approaches, however, continuously merge translation functions into encoding and decoding processes by adding new parameters to original networks, aligning with traditional NMT models (Guo et al., 2020; Üstün et al., 2021). Consequently, fine-tuning adapters may alter encoder output distribution and potentially disrupt pretrained monolingual knowledge, similar to the PF method. For instance, Üstün et al. (2021) fine-tune adapters and cross-attention networks of decoders on parallel data to accommodate translation functions, whereas our method exclusively trains the transferring module. This distinction highlights our approach's ability to separate translation functions from encoding and decoding processes, facilitating a more efficient and flexible use of monolingual and bilingual knowledge. Moreover, our study's primary contribution lies in proposing a novel modular NMT framework featuring relatively independent functional components, rather than solely concentrating on multilingual translation models.

3. Modular Neural Machine Translation

Given a translation pair sentence $\{x, y\}$, a translation model is to model the joint probability $p(x, y)$, which maximizes the log-likelihood, $\bar{y} = \arg\max \log P(y|x)$, of a target sequence y conditioned on a source sequence x. The conventional NMT model is an encoder-to-decoder framework and couples the translating capability within both the encoder and decoder. In contrast to coupling functions, we propose a novel approach called Modular Neural Machine Translation (MoNMT) model, which comprises three function-independent modules. First of all, we introduce two latent semantic variables for the source sentence and the target sentence, z_x and z_y, and rewrite the joint probability of a translation pair as follows,

$$\begin{aligned} & p(x, y, z_x, z_y) \\ & = p(y|z_y, z_x, x)p(z_y|z_x, x)p(z_x|x)p(x) \\ & \propto \underbrace{p(z_x|x)}_{\text{encode}}\underbrace{p(z_y|z_x)}_{\text{transfer}}\underbrace{p(y|z_y)}_{\text{decode}}, \end{aligned} \tag{1}$$

where z_y is the sole guidance for generating target sentences. By then, the joint probability of the translation model is composed of three conditional probability distributions, which are for the Enc (encode), Trans (transfer), and Dec (decode), respectively. In that case, the Enc and Dec are responsible for encoding and decoding functions, integrating the translation capacity into the Trans.

Specifically, the Enc and Dec are conditional to the monolingual knowledge distribution. Rather than respectively denoting them by Masked Language Modeling (MLM) like BERT (Devlin et al., 2019) and Casual Language Modeling (CLM) like GPT (Brown et al., 2020), we denote them together by Denoising Auto-Encoding (DAE) (Lewis et al., 2020) for the reasons of 1) DAE is in line with sequence-to-sequence learning; 2) its decoder is conditional to the encoder outputs, which meet the need of the Dec; and 3) with a denoising decoder as the Dec, the Trans only needs to transfer the source-oriented representations into the target-oriented representation, then the Dec generates the translation hypothesis in a denoising manner. With z_x and z_y, we reformulate DAE as follows:

$$\begin{aligned} p(x, \hat{x}, z_x) &= p(x|z_x, \hat{x})p(z_x|\hat{x})p(\hat{x}) \\ &\propto \underbrace{p(z_x|\hat{x})}_{\text{encode}}\underbrace{p(x|z_x)}_{\text{decode}}, \end{aligned} \tag{2}$$

$$\begin{aligned} p(y, \hat{y}, z_y) &= p(y|z_y, \hat{y})p(z_y|\hat{y})p(\hat{y}) \\ &\propto \underbrace{p(z_y|\hat{y})}_{\text{encode}}\underbrace{p(y|z_y)}_{\text{decode}}, \end{aligned} \tag{3}$$

where $\hat{x}$ and $\hat{y}$ are the noising version of x and y, respectively. By then, the probability distributions of $p(z_x|x)$ and $p(y|z_y)$ are determinant, Equation 1 is further reformulated as:

$$p(x, y, z_x, z_y) \propto \underbrace{p(z_y|z_x)}_{\text{transfer}}, \tag{4}$$

where the translation process is modeled by transferring latent variables z_x to z_y. By integrating the translating function into the Trans module, our approach enables the retention of monolingual knowledge in the Enc and Dec modules while acquiring bilingual knowledge in Trans.

4. A Modularized Learning Strategy

This section illustrates the modularized training strategy for MoNMT as indicated in Figure 1. Given two languages x and y, we denote the monolingual sentences as $\mathbf{x}_{\text{mono}}$ and $\mathbf{y}_{\text{mono}}$, the translation pairs as $\mathbf{x}_{\text{para}}$ and $\mathbf{y}_{\text{para}}$ and the model parameters as Θ.

4.1. Encoding and Decoding Modules

Given language x, we firstly apply the noising function on x_{mono} and get the noise sentence $\hat{x}_{\text{mono}}$ following the default setting of Lample et al. (2017). Then, an encoder-to-decoder model is trained to recover the corrupted sentence $\hat{x}_{\text{mono}}$ with cross-entropy loss, of which the encoder and the decoder are adopted as the Enc and Dec of language x. So does the language y. The learning objectives are:

$$\begin{aligned} & \Theta^{\text{x}}_{\text{enc}}, \Theta^{\text{x}}_{\text{dec}} \\ & \quad = \mathop{\arg\max}_{\Theta^{\text{x}}_{\text{enc}}, \Theta^{\text{x}}_{\text{dec}}} \log P(x_{\text{mono}}|\hat{x}_{\text{mono}}, (\Theta^{\text{x}}_{\text{enc}}, \Theta^{\text{x}}_{\text{dec}})), \end{aligned} \tag{5}$$

$$\begin{aligned} & \Theta^{\text{y}}_{\text{enc}}, \Theta^{\text{y}}_{\text{dec}} \\ & \quad = \mathop{\arg\max}_{\Theta^{\text{y}}_{\text{enc}}, \Theta^{\text{y}}_{\text{dec}}} \log P(y_{\text{mono}}|\hat{y}_{\text{mono}}, (\Theta^{\text{y}}_{\text{enc}}, \Theta^{\text{y}}_{\text{dec}})), \end{aligned} \tag{6}$$

where Θ^{*}_{enc} and Θ^{*}_{dec} are the Enc and the Dec of an arbitrary language $*$. Note that both modules include the embedding layer.

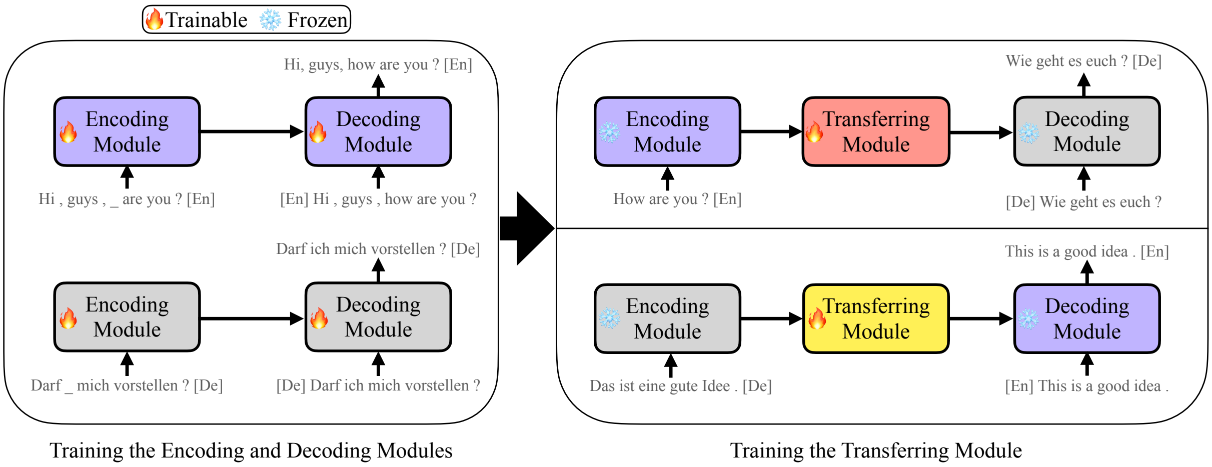


Figure 1: The training strategy for the Modular Neural Machine Translation model. The encoding module (Enc) and decoding module (Dec) are pretrained on large-scale monolingual data (left), while the transferring module (Trans) is trained on bilingual data (right). Modules with the same function are depicted using the same color.

In our implementation, we share the Enc and Dec of each translation language pair for both x-to-y and y-to-x translation directions. According to Equation 5 and 6, the DAE model (Θ_{enc}, Θ_{dec}) is optimized by the following reconstruction loss:

$$\begin{aligned}\text{Loss}_{\text{dae}} = &-\log p(x_{\text{mono}}|\hat{x}_{\text{mono}}) \\ &-\log p(y_{\text{mono}}|\hat{y}_{\text{mono}}).\end{aligned} \tag{7}$$

At this point, the Enc and Dec are ready for the encoding and decoding functions, respectively.

4.2. Transferring Module

Trans is an extra network connected in series upon the Enc, which transfers z_x to z_y in Equation 4. Given the frozen Enc $\Theta^{\text{x}}_{\text{enc}}$ and Dec $\Theta^{\text{y}}_{\text{dec}}$, we train Trans $\Theta^{\text{x2y}}_{\text{trans}}$ with bilingual data $\{\mathbf{x}_{\text{para}}, \mathbf{y}_{\text{para}}\}$ for the x-to-y translation direction:

$$\begin{aligned}&\Theta^{\text{x2y}}_{\text{trans}} \\ &= \arg\max_{\Theta^{\text{x2y}}_{\text{trans}}} \log P(y_{\text{para}}|x_{\text{para}}, (\Theta^{\text{x}}_{\text{enc}}, \Theta^{\text{x2y}}_{\text{trans}}, \Theta^{\text{y}}_{\text{dec}})),\end{aligned} \tag{8}$$

so does the y-to-x translation direction.

In our implementation, The Trans consists of K stacked layers, which are similar to the encoder layer of the Transformer model (Vaswani et al., 2017). By incorporating with the Frozen Θ_{enc} and Θ_{dec}, the Trans Θ_{trans} is optimized by the cross-entropy loss as follows:

$$\text{Loss}_{\text{mt}} = -\log p(y_{\text{para}}|x_{\text{para}}), \tag{9}$$

then we combine these three modules for the x-to-y MoNMT model.

Optimization: Gram matrix loss. To further reveal the potential of MoNMT, we study an optimizing alternative by employing the existing method, Gram matrix loss (Gatys et al., 2016), as an auxiliary term for training the Trans. In primary, the hidden size and sentence length are denoted as H and L, respectively.

The Gram matrix represents the covariance of a feature map and is used for transferring styles between two images in the computer vision community (Gatys et al., 2016; Li et al., 2017). In MoNMT, the Trans is expected to output representations, denoted as $A_{H\times L_{\text{s}}}$, that are close to the real target-oriented representations, denoted as $B_{H\times L_{\text{t}}}$, so that Dec may simply recover the target sentence as the denoising process of DAE. However, the length of a source sentence L_{s} is usually different from that of its target sentence L_{t}, so it is intractable to directly compute the difference between $A_{H\times L_{\text{s}}}$ and $A_{H\times L_{\text{t}}}$. As an alternative, we imitate the style transfer process and consider the sentence representations as "images", then reduce the difference of Gram metrics between $A_{H\times L_{\text{s}}}$ and $B_{H\times L_{\text{t}}}$, expecting to assist in training the Trans, as follows:

$$\text{Loss}_{\text{gram}} = \text{MSE}(AA^{\text{T}}, BB^{\text{T}}), \tag{10}$$

$$\text{Loss} = \text{Loss}_{\text{mt}} + \lambda\text{Loss}_{\text{gram}}, \tag{11}$$

where MSE is the mean square error following Gatys et al. (2016) and λ is a weight. Thus, we directly fit the transferred features to the real target-oriented features to assist in training the transferring module. An ablation study conducted in section 6.5 shows how MoNMT is improved by fitting additional Gram matrix loss using Equation 11.

	News			Medical			Law			Koran			IT			Subtitles		
	RD	**PF**	**Ours**	**RD**	**PF**	**Ours**	**RD**	**PF**	**Ours**	**RD**	**PF**	**Ours**	**RD**	**PF**	**Ours**	**RD**	**PF**	**Ours**
News	33.0	33.4	$33.9^{\Uparrow}$	7.2	8.1	$17.4^{\Uparrow}$	12.0	13.0	$20.3^{\Uparrow}$	1.4	3.6	$7.3^{\Uparrow}$	7.8	17.3	$18.2^{\Uparrow}$	14.5	19.1	$23.3^{\Uparrow}$
Medical	34.8	36.4	$37.7^{\Uparrow}$	51.1	52.6	52.5	18.6	24.6	$29.1^{\Uparrow}$	0.0	1.0	$6.9^{\Uparrow}$	10.8	24.4	$26.9^{\Uparrow}$	4.9	13.1	$24.4^{\Uparrow}$
Law	39.9	41.1	$41.4^{\Uparrow}$	18.6	24.6	$29.1^{\Uparrow}$	57.3	58.2	57.2	0.6	1.4	$6.7^{\Uparrow}$	7.2	17.4	$18.8^{\Uparrow}$	4.2	7.9	$18.4^{\Uparrow}$
Koran	12.5	12.7	$15.1^{\Uparrow}$	2.7	2.7	$6.5^{\Uparrow}$	3.2	3.5	$6.9^{\Uparrow}$	13.7	20.9	$21.3^{\Uparrow}$	3.4	9.0	$9.4^{\Uparrow}$	6.7	8.6	$11.1^{\Uparrow}$
IT	31.1	31.8	$32.1^{\Uparrow}$	10.0	11.2	$22.5^{\Uparrow}$	11.6	14.7	$23.0^{\Uparrow}$	0.6	1.5	$4.5^{\Uparrow}$	39.7	41.8	$42.7^{\Uparrow}$	6.1	8.7	$18.6^{\Uparrow}$
Subtitles	22.3	22.9	$23.1^{\Uparrow}$	3.2	3.6	$8.7^{\Uparrow}$	4.0	4.2	$7.4^{\Uparrow}$	1.5	3.0	$4.8^{\Uparrow}$	8.4	15.1	14.3	30.7	32.2	31.2
Average	28.9	29.7	**30.6**	14.8	16.0	**22.5**	17.8	19.7	**24.0**	3.0	5.2	**8.6**	12.9	20.9	**21.7**	11.2	13.9	**21.2**

(a) The BLEU scores for German-to-English on multi-domain translation tasks.

	News			Medical			Law			Ted		
	RD	**PF**	**Ours**	**RD**	**PF**	**Ours**	**RD**	**PF**	**Ours**	**RD**	**PF**	**Ours**
News	31.0	35.9	$36.3^{\Uparrow}$	5.5	7.2	$16.6^{\Uparrow}$	8.8	10.1	$20.6^{\Uparrow}$	14.5	20.9	$25.4^{\Uparrow}$
Medical	21.6	30.3	$33.6^{\Uparrow}$	82.5	82.3	$83.1^{\Uparrow}$	13.3	16.1	$24.8^{\Uparrow}$	6.1	14.3	$22.1^{\Uparrow}$
Law	33.2	39.3	$40.6^{\Uparrow}$	12.6	17.0	$26.3^{\Uparrow}$	61.2	63.4	62.5	7.3	7.4	$19.2^{\Uparrow}$
Ted	22.6	28.5	$29.0^{\Uparrow}$	4.1	4.5	$11.8^{\Uparrow}$	6.3	7.5	$13.5^{\Uparrow}$	19.1	41.9	$42.5^{\Uparrow}$
Average	31.0	33.4	**34.9**	26.2	27.6	**34.5**	22.7	24.3	**30.4**	16.8	21.1	**27.3**

(b) The BLEU scores for Romanian-to-English on multi-domain translation tasks.

Table 1: Main Results, where the methods are the Transformer model with a random initialization (**RD**), the pretrain-and-finetune paradigm (**PF**), and the MoNMT model (**Ours**). Noted that All the models are trained on training sets in the first row and tested on the test sets in the first column. **Bold** entries denote the best average performance. $^{\Uparrow}$ denotes statistically significant differences with $p \leq 0.05$ in the paired bootstrap resampling test compared to the baselines (Koehn, 2004).

5. Experiment

5.1. Settings

Data As for monolingual knowledge, we use the monolingual data from the public-available News-Crawl corpus, 36M (millions) for Romanian and Turkish, 100M for English and German.[2] Then, we add and upsample the English-side texts of multi-domain datasets into the monolingual data (Hu et al., 2019). As for bilingual knowledge, we include the multi-domain datasets (Medical, Law, IT, Koran, and Subtitles) for German-to-English translation (Koehn and Knowles, 2017; Aharoni and Goldberg, 2020), and the dataset from OPUS (Medical, Law, and Ted) for Romanian-to-English translation (Tiedemann, 2012). Besides, we further employ four widely-used benchmarks of translation tasks, which are WMT14 English-French (En-Fr), WMT14 English-German (En-De), WMT16 English-Romanian (En-Ro), and WMT18 English-Turkish (En-Tr), and consist of 36M, 4.5M, 600k, and 200k training pairs, respectively. Note that the datasets of En-De and En-Ro are adopted as the News domain datasets in the multi-domain tasks.

Model Following Vaswani et al. (2017), we adopt the Transformer architecture for all the models. We control the layer number of the Trans for training bilingual data of different sizes. Specifically, we set it to 1 for the MoNMT model of Koran, which only contains around 18k training pairs. Unless specified otherwise, each module of the MoNMT consists of 6 layers. For all language pairs, we apply subword-nmt to learn bpe subwords and form a joint dictionary (Sennrich et al., 2016b).[3] tDuring the training process, we use Adam (Kingma and Ba, 2015) to optimize the model parameters, with $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 10^{-9}$. For the Enc and Dec, we train them on monolingual corpus about 10 epochs. The transferring module, Trans, is trained on parallel data with 8192 max tokens (<200k) for small-size datasets and $32k$ max tokens for large-size datasets (>$500k$). Unless otherwise specified, Equation 7 and Equation 9 are adopted for optimizing Enc/Dec and Trans, respectively. The training procedure of all translation tasks is early-stop with 20 patience and $300k$ max steps. All the experiments are conducted on 4 Nividia Tesla V100 32GB GPUs.

Baselines 1) Transformer models (**RD**): We include the Transformer model with a random initialization without pretraining for each translation task. To obtain the best non-pretrained model performance, for a dataset size larger than 500k, we use the Transformer-Big architecture. Otherwise, we use the Transformer-Base architecture. 2) The pretrain-and-finetune paradigm (**PF**): is a widely employed and influential technique applied in var-

[2] https://data.statmt.org/news-crawl/

[3] https://github.com/rsennrich/subword-nmt

ious studies. In our experiments, we adopt the Transformer-Big architecture (Vaswani et al., 2017) for all translation tasks. The PF model is first pretrained on the same monolingual data as MoNMT, then finetuned on the parallel datasets.

5.2. Main Results

Results in Table 1 show our method (**Ours**) consistently outperforms other competing approaches across multi-domain translation tasks, as evidenced by the higher BLEU scores it achieves. These findings highlight that the MoNMT model successfully synergizes monolingual and bilingual knowledge, and improves generalization and robustness.

Comparison to the RD method. Our approach surpasses the RD method by achieving improvements ranging from 1 to 20 BLEU scores across multiple tasks. This provides strong evidence of the effective utilization of monolingual knowledge from extensive monolingual data in our method.

Comparison to the PF method. In comparison to the PF method, our approach demonstrates superior performance in out-of-domain tasks and similar performance in in-domain tasks, showcasing enhanced domain generalization and robustness. This is attributed to PF's susceptibility to the catastrophic forgetting problem and its tendency to overshadow monolingual knowledge when employed for translation tasks, ultimately resulting in suboptimal performance (McCloskey and Cohen, 1989; Kirkpatrick et al., 2017; Chen et al., 2020). Our method avoids this problem by modularly training monolingual and bilingual data.

Comparison on in-domain tests. In both the German-to-English and Romanian-to-English translation directions, results show that both the PF method and the MoNMT model outperform the RD method, while also possessing comparable performance to each other. This highlights the value of leveraging monolingual knowledge from large-scale datasets to improve translation proficiency. The distinction between the PF method and the MoNMT model barely exceeds a difference of 1 BLEU score, signifying that integrating bilingual knowledge into the transferring module is a viable alternative for machine translation, without compromising the importance of monolingual knowledge acquisition during parallel data training.

Comparison on out-of-domain tests. The MoNMT method exhibits superior performance compared to the RD and PF methods in out-of-domain tasks, with improvements ranging from 1 to 20 BLEU scores across various domains, underscoring its domain robustness and generalization capabilities. For example, in German-to-English translation, our method shows a noteworthy improvement of 12.5 and 11.3 BLEU scores in the medical domain compared to the RD and PF methods, respectively, as demonstrated in Table 1a. Similarly, in Romanian-to-English translation, our method achieves up to 7.7 and 7.3 BLEU scores above the RD and PF methods, respectively, in the medical domain, as shown in Table 1b. Note that the PF method generally outperforms the RD method. These results provide convincing evidence that 1) pretraining models on large-scale monolingual data can effectively enhance the domain robustness of translation models, and 2) our approach, MoNMT, effectively exploits both monolingual and bilingual knowledge by training dedicated function-independent modules for the encoding, transferring, and decoding functions.

6. Analysis

In this section, we begin our evaluation of the MoNMT model by conducting translation tasks across various language directions and dataset sizes. Following this, we provide a comprehensive analysis with a strong focus on evaluating the impact of monolingual and bilingual data volumes, as well as model dimensions. Ultimately, we present an interpretability analysis that aims to offer valuable insights into the inner workings of the model.

6.1. Influence of Bilingual Data Scales

To assess the effectiveness of our model across varying dataset sizes, we conduct evaluations on several translation tasks: En-Fr, En-De, En-Ro, and En-Tr. Our experimental results, presented in Table 2, show that the MoNMT-big model performs commendably in both translation directions across all four benchmarks. In scenarios where resources are abundant, we observe that MoNMT-Big competes favorably with the PF-Big method, lagging only 0.3 BLEU in English-to-French translation, a negligible variation. Conversely, the MoNMT-base model is not as successful, lagging about 2.0 BLEU compared to the PF-Base. This is owing to the insufficient capacity of a Trans model, with only 19M parameters, to train 36M bilingual pairs. Moreover, in cases where data is scarce, a significant challenge as far as machine translation is concerned, both MoNMT-base and MoNMT-big demonstrate significant improvement, with a rise of 1.6 BLEU score in English-to-Turkish translation. These findings suggest that our model has the capacity to tackle the data scarcity issue and improve its performance in low-resource settings. Collectively, our

Model	WMT14 En ⇔ Fr		WMT14 En ⇔ De		WMT16 En ⇔ Ro		WMT18 En ⇔ Tr		#Trained Parameters
	En ⇒ Fr	Fr ⇒ En	En ⇒ De	De ⇒ En	En ⇒ Ro	Ro ⇒ En	En ⇒ Tr	Tr ⇒ En	
RD-Base	40.9	36.9	27.3	31.9	33.9	29.8	9.4	15.3	61M
PF-Base	41.3	37.4	27.9	32.5	35.4	34.5	11.1	17.6	61M
MoNMT-Base	39.7	35.9	27.9	32.2	36.2	35.3	12.7	19.3	19M
RD-Big	42.2	38.4	27.9	33.0	34.2	31.0	1.3	3.8	211M
PF-Big	**42.6**	38.7	29.1	33.4	37.4	35.9	13.0	20.7	211M
MoNMT-Big	42.3	**38.8**	**29.4**	**33.9**	**37.6**	**36.3**	**13.8**	**20.9**	76M

Table 2: Results on common-used translation tasks. "Base" and "Big" indicate that the model layer settings are the same as those of Transformer-Base and Transformer-Big (Vaswani et al., 2017). The high- and low-resource tasks are arranged in a left-to-right manner for ease of comparison.

MoNMT model exhibits strong robustness and generalization while handling different dataset sizes and language directions for translation tasks.

6.2. Influence of Monolingual Data Scales

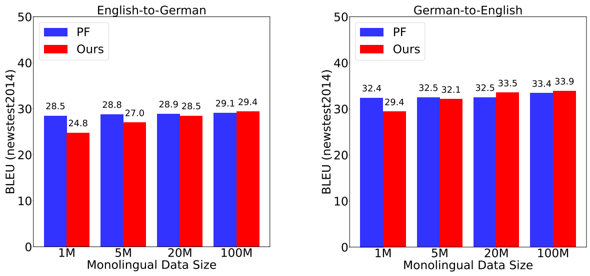


Figure 2: BLEU scores on WMT14 En-De with different sizes of monolingual data.

As shown in Figure 2, we conduct an experiment to evaluate the impact of different sizes of monolingual data on translation tasks. Specifically, we utilize the newstest2014 dataset. The Enc and Dec are trained on varying data volumes of 1M, 5M, 20M, and 100M, while the Trans is trained on WMT14 En-De. Our results indicate that our model performs worse than the PF method when trained with data volumes of 1M, 5M, and 20M. This can be attributed to the fact that the PF method fine-tunes the entire model on bilingual datasets, enhancing its encoding and decoding abilities, whereas our Enc and Dec are trained only on insufficient monolingual data. However, with a data volume of 100M, our MoNMT method surpasses the performance of the PF method, achieving BLEU scores of 29.4 for en2de and 33.9 for de2en. This suggests that the success of the MoNMT relies on the performance of both Enc and Dec, in addition to Trans.

6.3. Influence of Model Sizes

In Table 2, the final column displays the number of parameters trained for downstream tasks. Both the base and big architectures of the MoNMT model

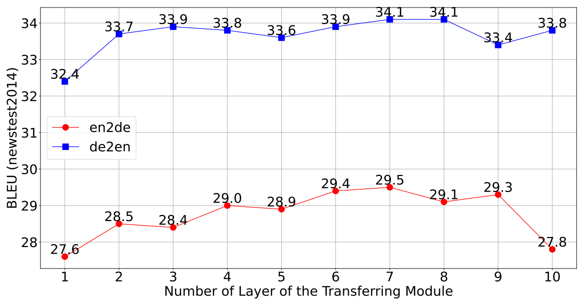
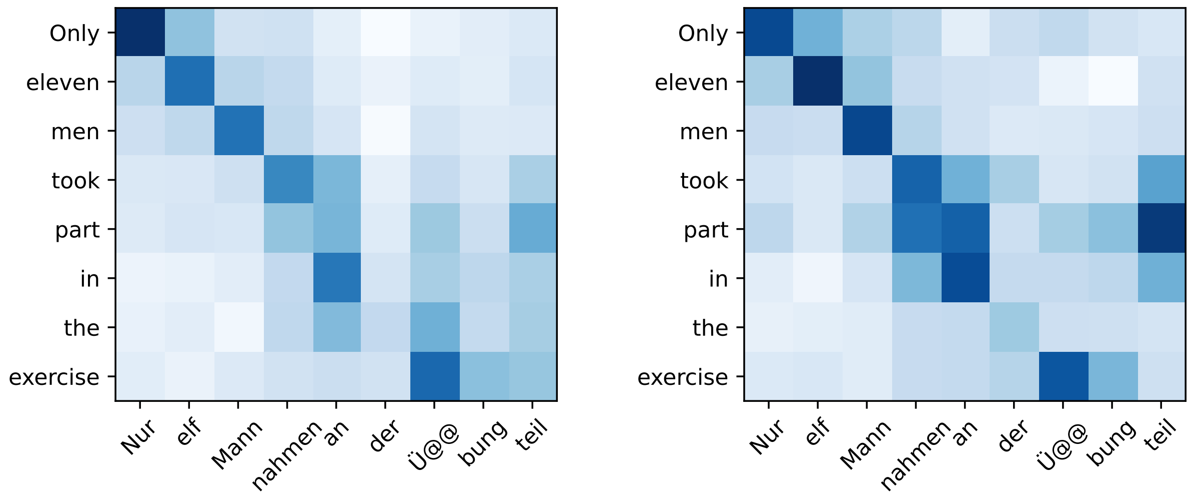


Figure 3: BLEU score on WMT14 En-De with various layer numbers of the transferring module

only require training of the Trans, which comprises one-third of the parameters in the baseline models. Despite this, the MoNMT model consistently delivers excellent performance. Notably, the MoNMT-big model has 76 million parameters for the Trans, which is similar to the Base model (61M) and considerably less than the Big model (211M). The results demonstrate that the MoNMT-Big model significantly outperforms the PF-Base models, presenting an improvement of over 2.0 BLEU points.

In Figure 3, we conduct an investigation into the effect of model capacity on the Trans architecture by manipulating the number of layers from 1 to 10, while keeping the Enc and Dec architectures constant. To ensure sufficient encoding and decoding abilities, we employ the large Transformer architecture and train it on a corpus of 100M monolingual data. Our research presents two translation curves, one for German-to-English (de2en) and another for English-to-German (en2de). Our findings show that a single-layer Trans performs comparably to the large RD-Big model (as presented in Table 2), achieving a BLEU score of 27.6 with just 13M parameters for fitting translation tasks. When increasing the number of layers to two, the performance is nearly equivalent to the PF-Big method, thus demonstrating the effectiveness of our MoNMT model for users with limited computing resources. Additionally, our results reveal that peak performance is achieved when using 7 layers for both translation directions.



(a) Correlation of the Enc output source and target representations.

(b) Correlation between the Enc output source representations and Trans output target representations.

Figure 4: Heap maps of word-level correlation coefficient metrics of a German-to-English translation case.

6.4. Interpretability

To delve deeper into the functionality of Trans on MoNMT, we undertake a thorough case study focusing on the word-level correlation between the representations of source and target sentences. Specifically, we calculate the correlation coefficient for each word pair across the two sentences, resulting in a correlation matrix represented as heat maps in Figure 4. Figure 4a depicts the correlation between the Enc output representations of source and target sentences, while Figure 4b demonstrates the correlation between the Trans output representations of the source sentence and the Enc output representations of the target sentence. Our findings demonstrate that the Enc output representations of the source and target sentences exhibit the correct correlation for words with similar semantic meanings, such as "Only" and "Nur". This indicates that the model trained with nonparallel data is capable of aligning the semantic information of two distinct languages. This finding aligns with prior studies (Conneau et al., 2020; Chi et al., 2021; Tan et al., 2022). Furthermore, we observe that the word-level correlations are enhanced for words with similar semantic meanings after the source sentence is processed by Trans, as evidenced by the deeper colors in Figure 4b. This finding suggests that the Trans effectively generates sentence representations that closely approximate the Enc output representations of the target sentence, making it possible that the Dec generates a translation in a denoising manner.

In order to further confirm our observation, we utilize an English-to-German bilingual word alignment test set (Vilar et al., 2006) for quantitative analysis. This test set is comprised of 508 sentence pairs along with their corresponding ground truth word alignments. The evaluation metric utilized in this analysis is the Alignment Error Rate (AER). For the sake of simplicity, we refer to the setting illustrated in Figure 4a as "enc2enc" and the setting depicted in Figure 4b as "enc2trans". In these two settings, we employ the prediction of the word alignment that possesses the highest correction score. Consequently, the AER scores for enc2enc and enc2trans are recorded as 28.0% and 24.6%, respectively. It is worth noting that lower scores indicate better performance in alignment (Vilar et al., 2006). Evidently, there is a noticeable difference of 3.4% between the results of enc2enc and enc2trans. This discrepancy suggests that the Trans enhances the alignment information.

6.5. Ablation Study

To enhance optimization, we incorporate the Gram matrix loss (in Equation 10) and the cross-entropy loss (in Equation 9) as the final loss (in Equation 11) to train the Trans. In our settings, the Gram matrix loss is weighted by $1e3$. These modifications enable us to achieve 29.7 BLEU scores for English-to-German translation and 34.2 BLEU scores for German-to-English translation in the newstest2014, which signifies a noteworthy improvement compared to the results obtained by the MoNMT-Big model (in Table 2). This observation suggests that MoNMT could be further improved by directly optimizing the output feature distribution of the transferring module, such as latent space regularization (Zhang et al., 2016), distribution transformation (Liu et al., 2022; Mahajan et al., 2020; Li et al., 2022) and so on, in future research.

7. Conclusion

This paper introduces a novel modular neural machine translation (MoNMT) model, modularly leveraging monolingual and bilingual knowledge. Distinct from traditional models, our method employs separate modules for utilizing monolingual and bilingual data, effectively addressing catastrophic forgetting of pretrained monolingual knowledge. Experimental results demonstrate that our approach achieves outstanding performance in both in-domain and out-of-domain tasks, showcasing superior model robustness and generalization. Furthermore, it proves highly effective in enhancing translation quality in low-resource scenarios. Notably, the MoNMT model is easy to implement, parameter-efficient, and scalable for practical applications. Future research should consider training unified encoding and decoding modules and extending our method to multilingual and multi-domain translation tasks. For industry applications, users can develop a translation system requiring fewer computational resources, as the encoding and decoding modules are reusable.

8. Ethics Statement

The main contributions of this research are methodological. We propose the Modular Neural Machine Translation model (MoNMT) along with its modular training strategy. Our experimental results offer compelling evidence for the effectiveness of our approach in enhancing model robustness and generalization. However, it is worth noting that the datasets employed in our experiments, although publicly accessible, may contain certain gender and social biases. We acknowledge these potential concerns that our work may encounter. Consequently, we recommend that users exercise caution and take appropriate measures to mitigate these risks according to their specific requirements.

9. Acknowledgements

This work was supported in part by the Science and Technology Development Fund, Macau SAR (Grant Nos. FDCT/0070/2022/AMJ, FDCT/060/2022/AFJ), Ministry of Science and Technology of China (Grant No. 2022YFE0204900), National Natural Science Foundation of China (Grant No. 62261160648), and the Multi-year Research Grant from the University of Macau (Grant No. MYRG-GRG2023-00006-FST-UMDF). This work was performed in part at SICC which is supported by SKL-IOTSC, and HPCC supported by ICTO of the University of Macau.

10. Bibliographical References

Roee Aharoni and Yoav Goldberg. 2020. Unsupervised domain clusters in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7747–7763, Online. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Sanyuan Chen, Yutai Hou, Yiming Cui, Wanxiang Che, Ting Liu, and Xiangzhan Yu. 2020. Recall and learn: Fine-tuning deep pretrained language models with less forgetting. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7870–7881, Online. Association for Computational Linguistics.

Zewen Chi, Li Dong, Bo Zheng, Shaohan Huang, Xian-Ling Mao, Heyan Huang, and Furu Wei. 2021. Improving pretrained cross-lingual language models via self-labeled word alignment. In *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)*, pages 3418–3430, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

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

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Michael S. A. Graziano and Tyson Aflalo. 2007. Mapping behavioral repertoire onto the cortex. *Neuron*, 56:239–251.

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.

Junjie Hu, Mengzhou Xia, Graham Neubig, and Jaime Carbonell. 2019. Domain adaptation of neural machine translation by lexicon induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2989–3001, Florence, Italy. Association for Computational Linguistics.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*, 1(10).

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *CoRR*, abs/1711.00043.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori Hashimoto. 2022. Diffusion-lm improves controllable text generation. *ArXiv*, abs/2205.14217.

Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. 2017. Universal style transfer via feature transforms. *Advances in neural information processing systems*, 30.

Xuebo Liu, Longyue Wang, Derek F. Wong, Liang Ding, Lidia S. Chao, Shuming Shi, and Zhaopeng Tu. 2021a. On the complementarity between pre-training and back-translation for neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2900–2907, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xuebo Liu, Longyue Wang, Derek F. Wong, Liang Ding, Lidia S. Chao, Shuming Shi, and Zhaopeng Tu. 2021b. On the copying behaviors of pre-training for neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4265–4275, Online. Association for Computational Linguistics.

Xuebo Liu, Yutong Wang, Derek F. Wong, Runzhe Zhan, Liangxuan Yu, and Min Zhang. 2023. Revisiting commonsense reasoning in machine translation: Training, evaluation and challenge. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15536–15550, Toronto, Canada. Association for Computational Linguistics.

Yihong Liu, Haris Jabbar, and Hinrich Schuetze. 2022. Flow-adapter architecture for unsupervised machine translation. In *Proceedings of the*

60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1253–1266, Dublin, Ireland. Association for Computational Linguistics.

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. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Shweta Mahajan, Iryna Gurevych, and Stefan Roth. 2020. Latent normalizing flows for many-to-many cross-domain mappings. In *International Conference on Learning Representations*.

Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.

Jianhui Pang, Baosong Yang, Derek Fai Wong, Yu Wan, Dayiheng Liu, Lidia Sam Chao, and Jun Xie. 2024a. Rethinking the exploitation of monolingual data for low-resource neural machine translation. *Computational Linguistics*, pages 1–23.

Jianhui Pang, Fanghua Ye, Longyue Wang, Dian Yu, Derek F Wong, Shuming Shi, and Zhaopeng Tu. 2024b. Salute the classic: Revisiting challenges of machine translation in the age of large language models. *arXiv preprint arXiv:2401.08350*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 371–376, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Zhixing Tan, Xiangwen Zhang, Shuo Wang, and Yang Liu. 2022. MSP: Multi-stage prompting for making pre-trained language models better translators. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6131–6142, Dublin, Ireland. Association for Computational Linguistics.

Brian Thompson, Jeremy Gwinnup, Huda Khayrallah, Kevin Duh, and Philipp Koehn. 2019. Overcoming catastrophic forgetting during domain adaptation of neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2062–2068, Minneapolis, Minnesota. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Ahmet Üstün, Alexandre Berard, Laurent Besacier, and Matthias Gallé. 2021. Multilingual unsupervised neural machine translation with denoising adapters. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6650–6662, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *ArXiv*, abs/1706.03762.

David Vilar, Maja Popovic, and Hermann Ney. 2006. AER: do we need to "improve" our alignments? In *Proceedings of the Third International Workshop on Spoken Language Translation: Papers*, Kyoto, Japan.

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, Deyi Xiong, Jinsong Su, Hong Duan, and Min Zhang. 2016. Variational neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 521–530, Austin, Texas. Association for Computational Linguistics.

Zhengyan Zhang, Zhiyuan Zeng, Yankai Lin, Chaojun Xiao, Xiaozhi Wang, Xu Han, Zhiyuan Liu, Ruobing Xie, Maosong Sun, and Jie Zhou. 2023. Emergent modularity in pre-trained transformers. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4066–4083, Toronto, Canada. Association for Computational Linguistics.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tieyan Liu. 2020. Incorporating bert into neural machine translation. In *International Conference on Learning Representations*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*.

A. Appendix

A.1. Evaluation Details

The metric of BLEU scores (Papineni et al., 2002) is adopted to evaluate the model performance for all our tasks. The details are as follows:

- X-to-English and English-to-Turkish: we adopt the Sacrebleu to calculate the BELU scores.[4]
- English-to-German: Following (Vaswani et al., 2017), we adopt the fairseq toolkit script the compute the BELU score for German texts.[5]
- English-to-Romanian: we follow (Liu et al., 2020) to post-process the Romanian text with Moses tokenization and normalization.[6]
- English-to-French: We use Moses language tokenizer to tokenize the texts, and calculate the tokenized BLEU scores.[7]

A.2. Back-Translation Versus Multi-Domain Translation Tasks

Back-Translation (BT) improves the translation model by enriching the bilingual data with synthetic pseudo bitexts, which requires a reverse translation model (Sennrich et al., 2016a). To evaluate BT on multi-domain translation tasks, we design two settings, one consists of 3M English-side multi-domain monolingual data from the multi-domain datasets and German News-Crawl monolingual data, and the other includes 8M monolingual data which includes 5M additional News-Craw data for both languages. Results are shown in Table 4. Specifically, the BT method trains the models on a small dataset of the Medical domain (about 250k) and the synthetic bitexts. In this instance, the reverse translation model utilized by the BT method is constricted by the scarcity of bilingual data, resulting in poor quality of synthetic pseudo bitexts (Edunov et al., 2018). Results show that both BT-3M and BT-8M consistently underperform the base model which is trained on the Medical training set. Besides, the performance degenerates as the monolingual data increases from 3M to 8M, about 2.5 BLEU lower in Average (AVG). On the other hand, our proposed MoNMT model demonstrates consistent improvements in the translation quality of out-of-domain tasks. Specifically, it achieved an increase of about 8.9 and 10.1 BLEU scores in the IT domain test for MoNMT-3M and MoNMT-8M, respectively. On the other hand, MoNMT consistently improves the translation quality of out-of-domain tasks, such as increasing by about 8.9 and 10.1 BLEU in the IT domain test of MoNMT-3M and MoNMT-8M, respectively. And its performance is improved as the data size increases for both in-domain and out-of-domain tests, resulting in favorable average performance. In a nutshell, the performance degeneration informs that the BT synthetic data is too detrimental for this brittle low-resource translation task, as the low-resource reverse translation model is not capable of producing qualified bitexts.

A.3. LLMs Versus Multi-Domain Translation Tasks

Table 5 presents three commonly used Language Model Machines (LLMs) and their respective performance in multi-domain translation tasks. The prompts used for LLMs are listed in Table 3. The results indicate that the Prompts LLMs still lag behind the supervised method, as claimed by Zhu et al. (2023). Among the three LLMs, ChatGPT outperforms Bloomz and Alpaca-LoRA significantly, indicating LLMs are heavily influenced by the model size and the training data However, although it is not a fair comparison, as a translation model, LLMs underperform our supervised method in the average performance, which only consists of 0.3B parameters compared to the 175B parameters of ChatGPT. Besides, Jiao et al. (2023) find that ChatGPT lacks domain robustness compared to existing translation systems.

Bloomz	Given the following source text in {src}: {src sentence}, a good {tgt} translation is:
Alpaca-LoRA	Translate the following {src} text into {tgt}: {src sentence}
ChatGPT	You are a faithful translator. Please translate the {src} sentence into {tgt}. [{src}]: {src sentence}\n[{tgt}]:

Table 3: Prompts used for LLMs, where src and tgt represent the source and target language.

A.4. Pearson Correlation Coefficient

Figure 4 presents the correlation of sentence representations as heat maps. The Correlation Coefficient is calculated between each word pair of the source sentence and the target sentence, in turn, using Equation 12:

$$P_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\Sigma(x_i - \bar{x})^2)(\Sigma(y_i - \bar{y})^2)}}, \quad (12)$$

where x and y present the word vectors of the source and target sentence representations.

[4]nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp

[5]https://github.com/facebookresearch/fairseq/blob/main/scripts/compound_split_bleu.sh

[6]https://github.com/rsennrich/wmt16-script

[7]https://github.com/moses-smt/mosesdecoder

	News	Medical	Law	Koran	IT	Subtitles	Average
base	7.2	51.1	18.6	2.7	10.0	3.2	15.5
+BT-3M	6.7	49.4	13.1	1.7	9.1	2.4	13.7
+BT-8M	2.3	45.7	9.9	0.6	6.9	1.8	11.2
MoNMT-3M	8.1	49.1	24.2	3.4	18.9	6.4	18.3
MoNMT-8M	13.7	50.3	25.7	4.5	20.1	7.3	**20.3**

Table 4: The BLEU scores of models trained with Medical training sets on multi-domain translation tasks. #M means the model is additionally trained with # millions synthetic bitexts or monolingual data. The performance degeneration indicates the negative effect of BT synthetic bitext for the translation models.

Model	**German-to-English**							**Romanian-to-English**					#Model Parameters
	News	Medical	Law	Koran	IT	Subtitles	Average	News	Medical	Law	Ted	Average	
Bloomz	20.8	28.0	20.9	8.6	15.6	17.3	18.5	12.0	16.6	16.3	5.4	12.6	7B
Alpaca-LoRA	29.4	31.5	26.0	13.0	26.0	20.8	24.5	31.4	30.3	28.3	22.1	28.0	7B
ChatGPT	35.2	38.9	35.7	16.3	31.5	28.1	31.0	39.6	36.5	37.4	32.0	36.4	175B
Ours	33.9	52.5	57.2	21.3	42.7	31.2	**39.8**	36.3	83.1	62.5	42.5	**56.1**	0.3B

Table 5: Results of LLMs on multi-domain translation tasks. Ours contains the results in Table 1a and 1b. These results indicate the LLMs still lag behind the strong supervised methods.