Taking Actions Separately: A Bidirectionally-Adaptive Transfer Learning Method for Low-Resource Neural Machine Translation

Xiaolin Xing, Yu Hong[*], Minhan Xu, Jianmin Yao, Guodong Zhou
School of Computer Science and Technology, Soochow University, China
{actuallyxxl, tianxianer, cosmosbreak5712}@gmail.com
{jyao, gdzhou}@suda.edu.cn

Abstract

Training Neural Machine Translation (NMT) models suffers from sparse parallel data, in the infrequent translation scenarios towards low-resource source languages. The existing solutions primarily concentrate on the utilization of Parent-Child (**PC**) transfer learning. It transfers well-trained NMT models on high-resource languages (namely Parent NMT) to low-resource languages, so as to produce Child NMT models by fine-tuning. It has been carefully demonstrated that a variety of PC variants yield significant improvements for low-resource NMT. In this paper, we intend to enhance PC-based NMT by a bidirectionally-adaptive learning strategy. Specifically, we divide inner constituents (6 transformers) of Parent encoder into two “teams”, i.e., T1 and T2. During representation learning, T1 learns to encode low-resource languages conditioned on bilingual shareable latent space. Generative adversarial network and masked language modeling are used for space-shareable encoding. On the other hand, T2 is straightforwardly transferred to low-resource languages, and fine-tuned together with T1 for low-resource translation. Briefly, T1 and T2 take actions separately for different goals. The former aims to adapt to characteristics of low-resource languages during encoding, while the latter adapts to translation experiences learned from high-resource languages. We experiment on benchmark corpora SETIMES, conducting low-resource NMT for Albanian (Sq), Macedonian (Mk), Croatian (Hr) and Romanian (Ro). Experimental results show that our method yields substantial improvements, which allows the NMT performance to reach BLEU4-scores of 62.24%, 56.93%, 50.53% and 54.65% for Sq, Mk, Hr and Ro, respectively.

1 Introduction

NMT has achieved significant improvements (Bahdanau et al., 2015; Vaswani et al., 2017) in recent years. Nevertheless, It heavily relies on large-scale observable parallel corpora. As a result, NMT generally fails to perform perfectly in an infrequent translation scenario, where the available parallel data for training is sparse. For example, the size of training data for NMT between English (En) and Macedonian (Mk) is about 200K, which is significantly smaller than that (582M) between English (En) and German (De). The issue has been widely known as low-resource NMT.

The existing studies attempt to overcome the issue primarily by 1) producing cross-language embeddings, 2) constructing bilingual shareable latent space for encoding, and 3) transferring well-trained models to low-resource languages. We overview the studies in Section 2. Within the aforementioned arts, nowadays, Parent-Child (PC) transfer learning (Zoph et al., 2016) represents a considerable advance in our knowledge. It allows a Parent NMT model to be fully trained and developed over high-resource languages (e.g., that for De→En), and transfers it to low-resource languages (e.g., that for Mk→En) for fine-tuning. This contributes to the construction of a Child NMT model that inherits the translation experiences of Parent model.

The recent experimental results suggest that PC transfer learning suffers from the weak perception of semantics in low-resource languages, at the very beginning of encoding. In other words, fine-tuning Parent NMT model over low-resource languages is unavoidably started from scratch. This results in less effective and inefficient representation learning. Though, it is proven that duplicating embeddings (Aji et al., 2020; Xu and Hong, 2022) of cross-language shareable tokens, synonyms and mutually-aligned tokens helps to alleviate the cold-start fine-tuning problem. This also implies that conventional methods of constructing shareable latent space (Artetxe et al., 2018; Lample et al., 2018) may produce similar but general effects.

In this paper, we intend to strengthen PC trans-

[*] Corresponding author.

Proceedings of the 29th International Conference on Computational Linguistics, pages 4481–4491
October 12–17, 2022.

fer learning by coupling it with space-shareable encoding. Different from the previous work, we neither use an unabridged Parent encoder for space-shareable encoding, nor directly transfer it towards low-resource translation. Instead, we divide the unabridged Parent encoder into two parts. One part engages in space-shareable encoding for alleviating cold-start fine-tuning problems. The other is straight transferred without being "brainwashed" for the pre-existing translation experiences (i.e., the ones learned during high-resource NMT). The goal is to fulfill bidirectional adaptation, i.e., 1) establishing the encoding mode that adapts to linguistic characteristics of low-resource languages, conditioned on the shareable latent space; and 2) preserving the encoding mode that adapts to original translation experiences of the Parent NMT model.

In our experiments, we follow Vaswani et al. (2017) to build a transformer-based NMT model within the encoder-decoder architecture, where both encoder and decoder comprise 6 transformer layers. We intensively train it on large-scale high-resource language pairs to produce a knowledgeable Parent NMT model. On the basis, we take Parent's encoder, and divide it into two teams: T1 (1^{st} transformer layer) and T2 (2^{nd}-6^{th} transformer layers). We train T1 to perform space-shareable encoding for low-resource languages. And we carry out monolingual unsupervised learning when training T1, where the generative adversarial network and masked language model are used. When moulding the Child NMT model, we transfer Parent to low-resource languages as usual. The difference lies in that T2 in Parent performs hot-start encoding by absorbing "home-made" hidden states from T1, i.e., the ones fabricated by T1 in terms of both monolingual features of low-resource languages and distributions in shareable latent space.

We conduct experiments on corpora SETIMES, where the low-resource MT scenarios of Sq-En, Mk-En, Hr-En and Ro-En are considered. Experimental results show that our method yields substantial improvements, and achieves competitive performance compared to the state of the art.

2 Related Work

There are a variety of advanced methodologies proposed for tackling low-resource NMT. Due to page limitation, we merely overview the closely-related arts reported in recent five years.

• Shareable Latent Space

Recently, the impressive hypothesis is that embeddings of both high-resource and low-resource source languages can be produced conditioned on the distributions in the same latent space. This undoubtedly contributes to the construction of versatile NMT models towards different language pairs, frankly, including those in the low-resource NMT scenarios. The key issue, in this case, is to establish a shareable latent space.

Artetxe et al. (2018) and Lample et al. (2018) design unsupervised learning approaches to construct shareable latent space. The approaches actually enable an encoder to properly project low-resource languages into the latent space of high-resource languages. Iterative back translation (Sennrich et al., 2016) and denoising auto-encoder (Vincent et al., 2008) are used to fulfill space-shareable encoding. The studies demonstrate the versatility of space-shareable encoding for multilingual translation in simulated experiments, where high-resource language pairs (e.g., En-De) are used though the size of training data is reduced. Soon after, Guzmán et al. (2019) prove that space-shareable encoding fails to obtain promising performance for authentic low-resource scenarios (e.g., En-Nepali and En-Sinhala). Marchisio et al. (2020) suggest that weak isomorphism between non-family languages results in the performance degradation. To strengthen space-shareable encoding, recently, multilingual BART (Liu et al., 2020a) is used in the unsupervised NMT framework, together with denoising autoencoder (Üstün et al., 2021) and multitask learning (Ko et al., 2021).

• Constructing Shareable Vocabulary

The first study for alleviating weak isomorphism most probably derives from Lakew et al. (2019)'s effort, where perplexity-based similarity computation is utilized to automatically select most relevant high-resource languages for space-shareable encoding. The obtained improvements in this study imply that common linguistic units (isomorphic constituents) between high-resource and low-resource languages serve as informative seeds for harvesting embeddings of heterogeneous constituents. It raises the interest in building shareable vocabulary.

Kim et al. (2018) construct a synthetic dictionary by iteratively updating linear mapping relationships between bilingual embeddings. Aji et al. (2020) build a joint vocabulary where the matched tokens are assigned with the same embeddings, while the mismatched the randomly-initialized em-

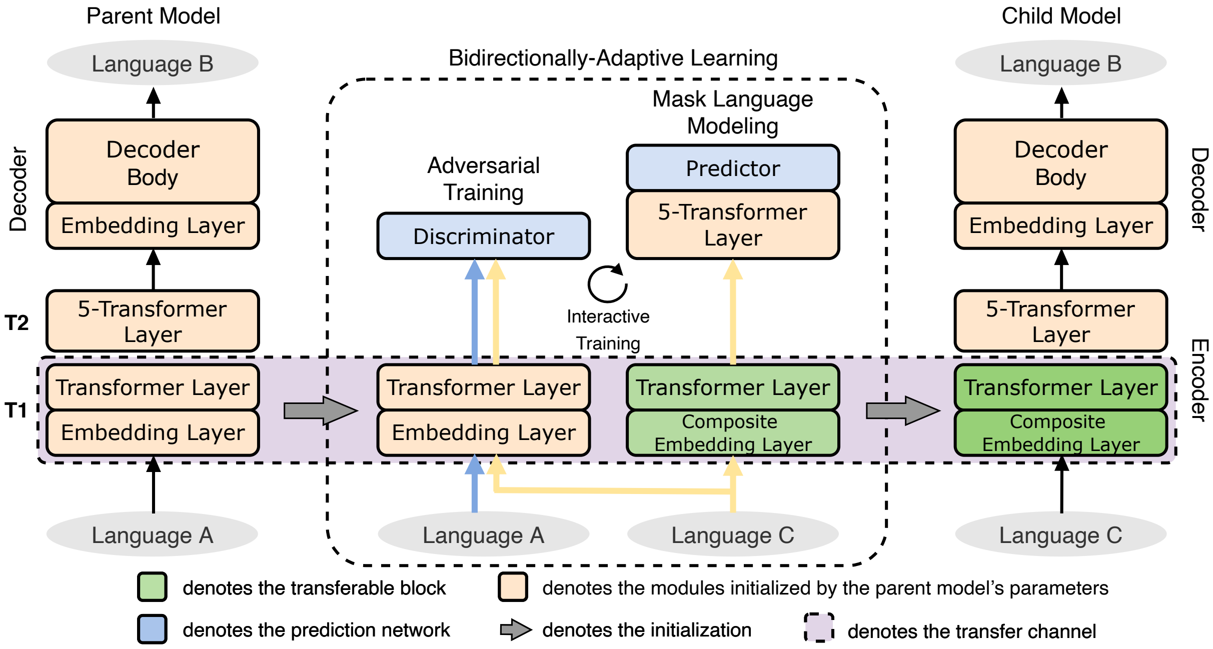

Figure 1: Architecture of bidirectionally-adaptive transfer learning for PC-based NMT, which conducts bilingual translation from high-resource language pairs (A→B) to low-resource language pairs (C→B).

beddings. Chronopoulou et al. (2021) retrain Byte-Pair-Encoding (BPE) over bilingual hybrid corpus, and use the segmented tokens by BPE to build the shareable vocabulary. Xu and Hong (2022) carry out word alignment between low-resource and high-resource languages, and share embeddings among aligned sub-tokens. This effectively expands the existing shareable vocabularies.

• **Transfer Learning for NMT**

Transfer learning approaches for NMT are primarily developed within Parent-Child (PC) framework (Zoph et al., 2016; Zhang et al., 2021). PC allows an NMT model to be trained on large-scale high-resource parallel data, and fine-tunes it on a small quantity of low-resource parallel data. It is proven that PC produces significant and increasing improvement with less warming-up time. Nowadays, PC has been successfully coupled with the aforementioned shareable vocabulary construction (Aji et al., 2020; Chronopoulou et al., 2021; Xu and Hong, 2022).

• **Pretrained Langauge Models for NMT**

Pretrained langauge models, such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019), have been demonstrated to be effective for natural language processing. They enable the deep perception and encoding of semantics by learning that from large-scale monolingual data. Recently, Liu et al. (2020b) develop a multilingual BART, whose encoder significantly contributes to the enhancement of low-resource NMT.

3 Approach

We show the architecture of our low-resource NMT model in Figure 1. The 12-layer transformer-based encoder-decoder network (at the left side in Figure 1) serves as Parent NMT model, which contains six layers of transformer encoder and six layers of transformer decoder. From here on, we refer them as encoder and decoer layers respectively. The Parent NMT model has been intensively trained to perform translation for high-resource language pairs A and B (i.e., NMT for A→B).

We divide Parent encoder layers into two teams T1 and T2, where T1 is constituted merely with the 1^{st} encoder layer, while T2 the rest five encoder layers. On the basis, T1 is pushed into the bidirectionally-adaptive learning channel (which is marked by the dotted rectangular box with a purple background in Figure 1). During the adaptation process, T1 engages in the bilingual encoding program launched by the generative adversarial network, so as to learn space-shareable encoding mode for both high-resource language A and low-resource language C. Besides, T1 engages in the masked language modeling program for low-resource language C, with the aim to learn language-specific encoding mode in terms of distinct characteristics of C. T1 will be trained iteratively and alternately in the two programs. This results in a refined T1.

We construct Child NMT model by transferring Parent decoder and T2 to low-resource language pairs C and B (NMT for C→B), and connect T2 with the refined T1. On the basis, we fine-tune Child NMT model using small-scale low-resource parallel data. The Child model is deployed at the right side in Figure 1. Instead, the T1 and T2 we used come from the decoder of the Parent model (A→B) if the child translation direction is A→C. In the rest of this section, we detail all the components of our model.

3.1 Baseline Low-resource NMT

We follow Xu and Hong (2022)'s work to construct the baseline transferable NMT model for low-resource language pairs, where Parent-Child (PC) transfer mechanism (Zoph et al., 2016) is used, and Xu and Hong (2022)'s expanded version of Aji et al. (2020)'s joint vocabulary is adopted.

- **NMT Framework** We performance sentence-level NMT. Given a source language sentence s, we convert each token in it into the real-valued embedding $v_s^i \in \mathbb{R}^{512}$. This results in the distributed representations V_s of s (V_s=$\{v_s^1 ... v_s^l\}$), where "l" denotes the maximum length of input sequence. A trainable embedding layer is used for obtaining token-level embeddings, which possesses a dynamic **source-language** vocabulary mapping from tokens to embeddings.

We feed V_s into the 1^{st} encoder layer, the one deployed ahead of other five sequentially-connected encoder layers. We use the encoder layers to obtain deep representations H_s of V_s as follows:

$$\begin{cases} H_s^{(i)} = f_e^{(i)}\left(H_s^{(i-1)}\right), & 1 < i \leq 6 \\ H_s^{(i)} = f_e^{(i)}\left(V_s\right), & i = 1 \end{cases} \tag{1}$$

where, $f_e^{(i)}$ is the i-th encoder layer of vanilla transformer (Vaswani et al., 2017; Al-Rfou et al., 2019).

Conditioned on the representations $H_s^{(6)}$ output by the encoder stack, we conduct decoding using six successively-connected decoder layers:

$$\begin{cases} h_t^{(i)} = f_d^{(i)}\left(H_s^{(6)}, h_t^{(i-1)}\right), & 1 < i \leq 6 \\ h_t^{(i)} = f_d^{(i)}\left(H_s^{(6)}, v_{t-1}\right), & i = 1 \end{cases} \tag{2}$$

where, $f_d^{(i)}$ denotes the i-th decoder layer of vanilla transformer, $h_t^{(i)}$ is the hidden state output by $f_d^{(i)}$ at a certain decoding time step t, and v_{t-1} is the embedding of the (t-1)-th token predicted at the earlier time step. Note that we obtain v_{t-1} using the trainable **target-language** embedding layer. Each target-language token is predicted by a linear layer with Softmax normalization, conditioned on $h_t^{(6)}$.

- **PC Transfer Learning** We train the aforementioned encoder-decoder network for NMT on high-resource language pairs A and B, i.e., learning to encode $H_A^{(6)}$ and decode $H_B^{(6)}$. This allows Parent NMT model to be formed. We transfer this well-trained Parent NMT model to low-resource language pairs C and B, and fine-tune it over the parallel data between C and B. By parametric inheritance and adaptive training (i.e., fine-tuning), transfer learning enables the generation of Child NMT model. Ideally, it learns to encode $H_C^{(6)}$ and decode $H_B^{(6)}$ to some extent.

- **Joint Vocabulary** During the transfer learning towards C→B NMT, the embedding layer of source language C is enhanced using the expanded joint vocabulary. In terms of the vocabulary, both morphologically-identical sub-tokens (Aji et al., 2020) and aligned sub-tokens (Xu and Hong, 2022) (between source language C and target language B) share the same embeddings, i.e., the ones learned from the process of training Parent NMT. When conducting bilingual embedding sharing between aligned sub-tokens, Xu and Hong (2022)'s element-wise mean aggregation (namely Mean-PC) is used for N-to-1 alignment cases.

3.2 Bidirectionally-adaptive Transfer Learning

We strengthen the baseline NMT model using a Bidirectionally-Adaptive Transfer Learning strategy (BATL for short). BATL adopts T1 (i.e., 1^{st} encoder layer) of Parent NMT model, and exclusively trains it for activating its bidirectional adaptability, including the adaptation to Parent's encoding mode towards high-resource NMT (A→B), as well as that to monolingual linguistic characteristics of the low-resource language C. Generative Adversarial Network (GAN) and Masked Language Modeling (MLM) are utilized for BATL.

- **GAN-based Backward Adaptation** We construct a discriminator, and couple it with T1 for adversarial training. T1 encodes a source-language sentence, which may derive from high-resource language pairs or low-resource. Conditioned on the representation (of a sentence) output by T1, the discriminator determines whether the sentence is of high-resource language or low-resource, within a binary classification task.

More importantly, T1 plays the role of a "counterfeiter". It produces the representation according as closely as possible to distributions in the semantic space of high-resource language, i.e., the one learned during the training for Parent NMT. Briefly, T1 counterfeits the high-resource sentence representation even if the sentence is actually of low-resource language. By contrast, the discriminator is trained to perform for anti-counterfeiting, determining the provenance of a sentence as precisely as possible. Repeatedly training T1 and the discriminator within the adversarial framework (counterfeiting versus anti-counterfeiting) will enhance both themselves. In particular, T1 learns to encode the low-resource language in the way of encoding the high-resource language, conditioned on a shareable semantic space. Coupling such a T1 into Child NMT model, frankly, contributes to the enhancement of its adaptation to Parent's translation experience, during the process of tackling the low-resource language.

• **MLM-based Forward Adaptation** T1 appears as a junior encoder when dealing with sentences of low-resource language at the very beginning, due to a lack of pragmatic and semantic knowledge in it. Consequently, the aforementioned adversarial training that directly utilizes such a junior T1, most probably, fails to form a reliable bilingual semantic space. In other words, although T1 learns to encode low-resource language in the mode of high-resource language (by GAN), it is grounded on a shallow or even inexact understanding of the former's pragmatics and semantics.

To address the issue, we construct a Masked Language Modeling channel (MLM) to enhance the capacity of T1 in encoding low-resource languages. It enables the forward adaptation of T1 to inherent linguistic characteristics of low-resource language.

MLM is conducted with the task of predicting masked tokens. Given a sentence of low-resource language, we mask about 10% tokens in it. The Masking strategy is implemented by substituting the randomly selected tokens with the special token "UNK", and initializing them with the unified embeddings. We feed the partially-masked sentence into T1 to encode each token in it, where the parameters of T1 are learnable during training. T2 is used for further encoding over the output of T1, where the parameters of T2 are frozen. Freezing T2 prompts T1 to learn low-resource languages as actively as possible.

• **Collaborative Training** We train T1 by GAN and MLM, alternatively and iteratively. First, T1 is trained by GAN, where 781 batches of hybrid monolingual data are used (i.e., 781 batches of monolingual sentences selected from the low-resource dataset, as well as 781 batches of high-resource cases). Secondly, T1 is further trained by MLM, where 1,562 batches of monolingual low-resource instances are used. This alternative training is carried out iteratively for 100 times within 25 epochs.

3.3 Shaping Low-resource NMT by BATL

We utilize the aforementioned BATL as a midway stage of transfer learning. In order to shape a concrete Child NMT, we still need to transfer T1 and the accompanying networks to low-resource language pairs, and fine-tune them on the parallel data.

• **Components of Child** comprises T1, T2 and the decoder of Parent NMT model. During assembling the components, T2 is connected behind T1, and both act as an encoder. The decoder of Parent with an embedding layer is coupled with the encoder. Briefly, Child inherits Parent's architecture.

• **Transfer Learning** includes the stages of BATL and transfer to low-resource language pairs. Within Child's components, only T1 is considered during BATL. All the components are fine-tuned on the parallel data of low-resource language pairs. It is noteworthy that although T2 is used for MLM in BATL, its parameters are frozen at the stage. During fine-tuning towards low-resource languages, all the components are trainable.

• **Loss of NMT** is calculated as follows, where the cross-entropy estimation is used:

$$\mathcal{L}_{MT} = \mathbb{E}_{x_i \in S}[-log\, p(y_i|x_i)] \qquad (3)$$

where, $p(y_i|x_i)$ denotes the conditional probability that the ground-truth target-language sentence y_i is predicted given the source-language sentence x_i.

4 Experimentation

4.1 Datasets and Evaluation Metric

We experiment on SETIMES (Tiedemann, 2012)[1]. To facilitate the comparison with the previous work, we concentrate on the low resource translation tasks of Sq↔En, Mk↔En, Hr↔En and Ro↔En, where

[1] http://opus.nlpl.eu/SETIMES.php

#HighR	Fr-En	Es-En	De-En	Ru-En
Train.	747M	952M	582M	217M

Table 1: Statistics in high-resource (#HighR) parallel datasets. Note that we won't report the development and test results of Parent NMT models, and therefore the statistics for high-resource validation and test sets are omitted in this study. The high-resource NMT performance has been discussed in Tiedemann's work[3].

#Languages	Datasets	Train.
De, Sq, Mk, Hr, Ro	SETIMES	200k
Fr, Es, Ru	Europarl	200k

Table 2: Statistics in monolingual datasets.

Sq, Mk, Hr, Ro and En refer to Albanian, Macedonian, Croatian, Romanian and English, respectively. We utilize 200K sentences for training, 1K sentences for validation, and 3K sentences for testing for each language. We take into consideration various Parent NMT models for shaping Child in a series of separate experiments, where four classes of high-resource language pairs are used, including Es→En, Fr→En, De→En and Ru→En (Spanish: Es, French: Fr, German: De, Russian: Ru). All the high-resource parallel data is derived from Tatoeba[2]. Table 1 shows the scales of high-resource parallel training data.

In addition, we introduce different monolingual datasets into our experiments, which are used for GAN and MLM during the stage of BATL. The monolingual data of De, Sq, Mk, Hr and Ro is taken from the parallel data in SETIMES, while that of Fr, Es and Ru is selected from Europarl (Koehn, 2005)[4]. Table 2 shows the statistics in monolingual datasets.

We follow the previous work to evaluate all NMT models with SacreBLEU (Post, 2018).

4.2 Hyperparameter Settings

We directly use the off-the-shelf transformer-base NMT models (Tiedemann, 2020) as Parents, and the newly-developed Child NMT models inherit all the configurations and hyperparameters of Parent.

First of all, all the sentences are tokenized using SentencePiece (Kudo and Richardson, 2018) with a 100k vocabulary size.

Secondly, we use monolingual datasets to train T1. We train T1 by GAN and MLM in 25 epochs, using NVIDIA RTX 2080Ti 11GB GPU. The optimizer is set to Adam (Kingma and Ba, 2015), and the learning rate is set to 10^{-4}.

Finally, we shape a Child NMT model using the well-trained T1 as well as Parent's T2 and decoder. Fine-tuning Child NMT is conducted on the low-resource parallel data. During fine-tuning, HuggingFace Transformers library (Wolf et al., 2020) and AdamW (Loshchilov and Hutter, 2019) optimizer are used. The latter runs with a weight decay rate of 0.1. We carry out grid search in the learning rates of $\{10^{-4}, 5 \times 10^{-5}\}$ for each translation task, and adopt the best model occurred during the development process. All fine-tuning is conducted on NVIDIA RTX 3090 24GB GPU.

4.3 Models for Comparison

We compare with two baseline models, which are denoted as Baseline1 and Baseline2. Baseline1 acts as a 12-layer transformer-based encoder-decoder NMT. It is randomly initialized and trained on low-resource parallel data. Baseline2 is a variant of Baseline1 since it is enhanced by transfer learning within the Parent-Child (PC) framework.

Besides, we compare our model to different state-of-the-art NMT models, including:

• **XLM** (Conneau and Lample, 2019) is a transferable language model. It is obtained by cross-language pretraining, where parallel sentences are concatenated for joint encoding, within a masked language modeling process.

• **RE-LM** (Chronopoulou et al., 2020) learns to reuse language models across different monolingual datasets. An extended bilingual vocabulary is constructed to enhance cross-language pretraining. The obtained language models are transferred to low-resource NMT.

In addition, we involve Xu and Hong (2022)'s **Mean-PC** into the discussion, which recently improves low-resource NMT using shareable embeddings of aligned sub-tokens. Nevertheless, we fail to directly compare with it because its performance, as reported, is obtained on different corpora and source languages. We discuss Mean-PC in a separate ablation experiment, where it is reproduced and equipped with our BATL.

4.4 Main Result

We show the primary test results in Table 3. It can be observed that BATL produces substantial improvements compared to both baselines. It is note-

[2] https://opus.nlpl.eu/Tatoeba.php

[3] https://huggingface.co/Helsinki-NLP

[4] https://statmt.org/europarl/

Model	Sq-En		Mk-En		Hr-En		Ro-En	
	→	←	→	←	→	←	→	←
Baseline1	32.50	52.03	30.13	50.62	25.90	38.90	28.92	46.39
Baseline2 (Zoph et al., 2016)	38.15	54.50	33.32	54.57	29.93	41.34	32.92	48.71
XLM (Conneau and Lample, 2019)	60.90	55.10	55.00	55.50	-	-	-	-
RE-LM (Chronopoulou et al., 2020)	61.10	54.80	55.20	55.30	-	-	-	-
BATL (Ours)	**62.24**	**56.82**	**56.93**	**56.15**	**50.53**	**45.21**	**54.65**	**52.19**

Table 3: Low-resource NMT performance. BLEU scores (%) are reported in both translation directions (← and →) for each low-resource language pairs. The performance of previous work is quoted from the published literature (instead of reproduction) due to the use of the same test sets.

Model	Sq→En	Mk→En
Unabridged.	**62.24**	**56.93**
−GAN	61.95	56.45
−MLM	60.61	52.81
−Mean-PC	62.06	56.34
−BATL	61.45	56.42
−All	38.15	33.09

Table 4: Verifying the effectiveness of different components of our NMT model in ablation experiments.

worthy that we conduct transfer learning within the same framework with Baseline2, i.e., PC transfer. The additional components we use include Mean-PC (for embedding sharing among matched or aligned sub-tokens), as well as GAN and MLM (for bidirectional adaptation to low-resource and high-resource languages). This demonstrates that the considerable performance gains benefit from the collaboration between bilingual commonality perception and bidirectionally-adaptive encoding.

Compared to the state-of-the-art low-resource NMT models, our BATL-based models achieve better performance. The possible reasons behind the advantages may include the following aspects:

• The commonly-used cross-language pretraining in the previous work (XLM, RE-LM and the variant) is proceeded with a task-irrelevant scenario, where a knowledgeable multilingual pretrained model may be used for initialization, though it fails to learn the experience in MT. By contrast, we take out part of encoder of the well-trained Parent NMT, and train it to adapt different source languages during encoding. This allows the resultant representations of new languages to be compatible with the pre-existing translation mode, i.e., ensuring the task-specific cross-language training.

• MLM is used alone for cross-language pretraining in the previous work, where bilingual source sentences are concatenated, masked and encoded thereafter. This contributes to the encoding within a shareable latent space, though the exclusive linguistic characteristics of a specific source language are neglected to some extent. By contrast, we simultaneously pursue commonality and exclusive characteristics using MLM and GAN, where MLM intently encodes low-resource source languages in terms of their natures, while GAN is utilized to explore shareable encoding mode.

4.5 Ablation study

In a series of ablation experiments, we verify the effectiveness of different individual components of BATL, including GAN and MLM. Considering that we expand PC-based NMT using both Mean-PC (Xu and Hong, 2022) and BATL, we also ablate them alternatively to examine their influences. Table 4 shows the experimental results. Note that, hereafter, we merely report the forward NMT performance (i.e., that of "→" NMT), where low-resource languages (Sq and Mk) are considered as source languages.

It can be found that ablating MLM results in more significant performance degradation, compared to GAN. This implies that MLM plays a dominant role in BATL, or in other words, the adaptation to low-resource languages during the "preheated" transfer process is crucial. Note that we merely push part of encoder (i.e., T1) into MLM-based cross-language transfer learning. It is different from the previous work which uses the whole encoder. In fact, by comparing the performance of XLM and RE-LM to our model that ablates GAN (i.e., mere use of MLM) across Table 3 and 4, we can find that our local transfer strategy produces the positive effects (better performance is obtained even if GAN is disabled).

Compare to Mean-PC, ablating BATL (i.e., ablating both GAN and MLM) causes more substantial

Parent Model	Sq→En	Mk→En
De→En	61.34	55.28
Fr→En	62.18	56.24
Es→En	**62.24**	**56.93**
Ru→En	61.03	55.86

Table 5: BLEU scores (%) of Child NMT models transferred from different Parent models.

performance reduction for the translation scenario of Sq→En, while relatively comparable reduction for Mk→En. This illustrates that learning-centered strategy of BATL has an advantage over the knowledge sharing mechanism of Mean-PC. Frankly, both are non-negligible. It is proven by the severe performance degradation caused by disabling both BATL and Mean-PC (see the performance obtained when "ALL" is ablated in Table 4).

4.6 Discussion and Analysis

• Effects of Different Parent Models

We construct four Parent NMT models using different source languages, including De, Fr, Es and Ru. On the basis, we verify the effects of such Parents on Child NMT models. The low-resource NMT performance resulted from different Parents is shown in Table 5. It can be observed that the Parents of Fr→En and Es→En are more beneficial to transfer, helping to produce higher BLEU scores.

Ideally, the PC-based transfer learning ought to benefit from high-resource languages that derive from the consistent or similar language family, such as the relatively closer relationship between Ru and Mk. However, the experimental results fail to support this hypothesis. Our findings show that the size of high-resource training data plays a more crucial role in improving the performance of PC transfer. As shown in Table 1, the most knowledgeable Parent, i.e., that of Es→En, is obtained on 952M training data. The scale of training data is much larger than that used for Parent of Ru→En. The latter fails to obtain an equivalently strong Child model. These findings are consistent with the conclusion of Kocmi and Bojar (2018).

• Is it Necessary to Construct Multi-layer T1

Our BATL is performed merely using T1, i.e., a single encoder layer, while the considered encoder is actually constituted with 6 transformer layers. It may be questioned whether BATL induces varying effects when T1 is expanded with more layers. The following experiment demonstrates that a larger T1 negatively influences NMT performance.

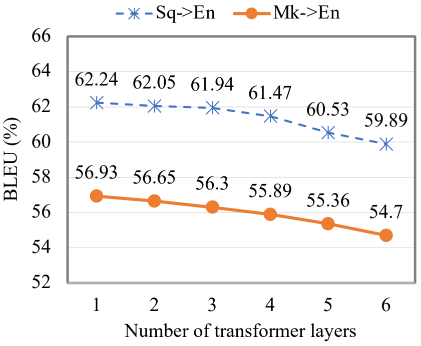

Figure 2: BLEU scores of our approach with respect to the number of self-attention layers in T1.

In a separate experiment, we split different numbers of transformer layers from Parent's encoder, and use them to form different depths of T1s. There are six T1s constructed in total using 1 to 6 encoder layers respectively. We conduct BATL for each of them separately, and reform Child NMT models accordingly. Figure 3 shows performance curves of such models in the MT scenarios of Sq↔En and Mk↔En, where the horizontal axis indicates the number of encoder layers in T1. It can be observed that performance degrades gradually when T1 is enlarged using more encoder layers.

The experimental results imply that overly shuffling and remodeling Parent's encoder for adaptation enhancement is risky. Most of well-trained parameters (translation experience) of Parent need to be directly inherited by Child.

• Compatibility with DAE and BT

Both denoising autoencoder (DAE) and back-translation (BT) have been proven effective in low-resource NMT. They were known as data augmentation methods that build synthetic corpora using monolingual data (Artetxe et al., 2018; Lample et al., 2018). We attempt to combine DAE and BT with our BATL, and rerun all the experiments to verify whether compatibility can be achieved.

For verifying the compatibility with DAE, we combine it with BATL from behind. During training, BATL is first used to optimize T1 and then DAE runs. Within the process, there are 3M[5] monolingual sentences (of low-resource language) used for DAE, where the model is additionally trained to assist revivification of all the falsified sentences, with the role of autoencoder. The effect of combining BATL and DAE is negative, as shown in Table 6, where the low-resource NMT performance is reduced severely. Our analysis suggests that, worse

Model	Sq→En	Mk→En
Ours.	62.24	56.93
+DAE	60.97	55.76
+BT	**62.75**	**57.32**

Table 6: BLEU scores obtained in the study of compatibility, where our BATL is respectively combined with DAE and BT.

than incompatibility and redundancy, the additional utilization of DAE leads to catastrophic forgetting of translation experience. In fact, DAE nearly plays the same role as the MLM module of our BATL, and thus, the posteriori-prompted DAE breaks the compromise effects reached by MLM and GAN.

For verifying the compatibility with BT, we expand the training data of low-resource language pairs using parallel instances. As usual, such pseudo instances are obtained by translating En to a certain low-resource language (e.g., Sq or Mk) forwardly and backwardly, where an off-the-shelf NMT model[6] is used. In our experiments, there are 2M pseudo instances created by BT for expansion. We use the expanded low-source data to fine-tune the Child NMT model. The rest configuration remains unchanged. Combining BATL and BT yields additional performance gains, as shown in Table 6, where BLEU scores are up to 62.75% and 57.32% for Sq→En and Mk→En. It demonstrates that BATL can be jointly used with BT safely.

5 Conclusion

We propose a Bidirectionally-Adaptive Transfer Learning (BATL) approach to enhance low-resource NMT models. Experimental results show that our approach yields substantial improvements, compared to the state of the art. In addition, it is demonstrated that BATL is compatible with BT-based data augmentation. Combining BATL and BT obtains additional performance gains. In a series of auxiliary experiments, we analyze the effects of various Parent NMT models and multi-layer BATL on low-resource NMT, some of which are negative and therefore noteworthy for risks in real applications.

The Commonality of linguistics stands for the fundamental principle in prompting Parent-Child transfer for low-resource NMT. Different family languages hold inconsistent commonalities with a specific low-resource language. Considering this phenomenon, in the future, we will study on a multilingual Parent-Child transfer learning. A selective transfer will be developed, in terms of case-specific adhesion to different high-resource family languages. The adhesion will be perceived by modeling the relevance of topics, provenances and domains, as well as document-level structure information (e.g., monolingual discourse relationships and rhetorics).

Acknowledgements

The research is supported by National Key R&D Program of China (2020YFB1313601) and National Science Foundation of China (62076174, 62076175).

References

Alham Fikri Aji, Nikolay Bogoychev, Kenneth Heafield, and Rico Sennrich. 2020. In neural machine translation, what does transfer learning transfer? In *Proceedings of ACL*, pages 7701–7710, Online. Association for Computational Linguistics.

Rami Al-Rfou, Dokook Choe, Noah Constant, Mandy Guo, and Llion Jones. 2019. Character-level language modeling with deeper self-attention. In *Proceedings of AAAI, Honolulu, Hawaii, USA*, pages 3159–3166. AAAI Press.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *Proceedings of ICLR, Vancouver, BC, Canada*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR, San Diego, CA, USA*.

Alexandra Chronopoulou, Dario Stojanovski, and Alexander Fraser. 2021. Improving the lexical ability of pretrained language models for unsupervised neural machine translation. In *Proceedings of NAACL*, pages 173–180. Association for Computational Linguistics.

Alexandra Chronopoulou, Dario Stojanovski, and Alexander M. Fraser. 2020. Reusing a pretrained language model on languages with limited corpora for unsupervised NMT. In *Proceedings of EMNLP*, pages 2703–2711. Association for Computational Linguistics.

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

[5] We follow the previous work to adopt a similar number of instance for running DAE (Chronopoulou et al., 2020).

[6] https://huggingface.co/Helsinki-NLP

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

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of EMNLP*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Yunsu Kim, Jiahui Geng, and Hermann Ney. 2018. Improving unsupervised word-by-word translation with language model and denoising autoencoder. In *Proceedings of EMNLP, Brussels, Belgium*, pages 862–868. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of ICLR, San Diego, CA, USA*.

Wei-Jen Ko, Ahmed El-Kishky, Adithya Renduchintala, Vishrav Chaudhary, Naman Goyal, Francisco Guzmán, Pascale Fung, Philipp Koehn, and Mona T. Diab. 2021. Adapting high-resource NMT models to translate low-resource related languages without parallel data. In *Proceedings of ACL*, pages 802–812. Association for Computational Linguistics.

Tom Kocmi and Ondřej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. In *Proceedings of WMT*, pages 244–252.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MTSummit, Phuket, Thailand*, pages 79–86.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of EMNLP*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Surafel M. Lakew, Alina Karakanta, Marcello Federico, Matteo Negri, and Marco Turchi. 2019. Adapting multilingual neural machine translation to unseen languages. In *Proceedings of IWSLT*, Hong Kong. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *Proceedings of ICLR 2018, Vancouver, BC, Canada*.

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

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

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of ICLR, New Orleans, LA, USA*.

Kelly Marchisio, Kevin Duh, and Philipp Koehn. 2020. When does unsupervised machine translation work? In *Proceedings of WMT*, pages 571–583, Online. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of WMT*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of ACL 2016*. The Association for Computer Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of LREC*, pages 2214–2218. European Language Resources Association.

Jörg Tiedemann. 2020. The Tatoeba Translation Challenge – Realistic data sets for low resource and multilingual MT. In *Proceedings of WMT*, pages 1174–1182, Online. Association for Computational Linguistics.

Ahmet Üstün, Alexandre Berard, Laurent Besacier, and Matthias Gallé. 2021. Multilingual unsupervised neural machine translation with denoising adapters. In *Proceedings of EMNLP*, pages 6650–6662, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of ICML 2008*, volume 307 of *ACM International Conference Proceeding Series*, pages 1096–1103. ACM.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of EMNLP*, pages 38–45, Online. Association for Computational Linguistics.

Minhan Xu and Yu Hong. 2022. Sub-word alignment is still useful: A vest-pocket method for enhancing low-resource machine translation. In *Proceedings of ACL*, pages 613–619, Dublin, Ireland. Association for Computational Linguistics.

Meng Zhang, Liangyou Li, and Qun Liu. 2021. Two parents, one child: Dual transfer for low-resource neural machine translation. In *Findings of ACL, Online*, pages 2726–2738. Association for Computational Linguistics.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of EMNLP*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.