High-resource Language-specific Training for Multilingual Neural Machine Translation

Jian Yang[1 *], **Yuwei Yin**[2 *], **Shuming Ma**[2], **Dongdong Zhang**[2], **Zhoujun Li**[1 †], **Furu Wei**[2]

[1]State Key Lab of Software Development Environment, Beihang University

[2]Microsoft Research

{jiaya, lizj}@buaa.edu.cn, {v-yuweiyin, shumma, dozhang, fuwei}@microsoft.com

Abstract

Multilingual neural machine translation (MNMT) trained in multiple language pairs has attracted considerable attention due to fewer model parameters and lower training costs by sharing knowledge among multiple languages. Nonetheless, multilingual training is plagued by language interference degeneration in shared parameters because of the negative interference among different translation directions, especially on high-resource languages. In this paper, we propose the multilingual translation model with the high-resource language-specific training (HLT-MT) to alleviate the negative interference, which adopts the two-stage training with the language-specific selection mechanism. Specifically, we first train the multilingual model only with the high-resource pairs and select the language-specific modules at the top of the decoder to enhance the translation quality of high-resource directions. Next, the model is further trained on all available corpora to transfer knowledge from high-resource languages (HRLs) to low-resource languages (LRLs). Experimental results show that HLT-MT outperforms various strong baselines on WMT-10 and OPUS-100 benchmarks. Furthermore, the analytic experiments validate the effectiveness of our method in mitigating the negative interference in multilingual training.

1 Introduction

Recent advances in multilingual neural machine translation (MNMT) aim to build and deploy a single universal model in real industrial scenarios, which supports multiple translation directions by sharing model parameters [Firat *et al.*, 2016; Johnson *et al.*, 2017; Aharoni *et al.*, 2019; Fan *et al.*, 2020; Lin *et al.*, 2020]. Furthermore, parameter sharing across various languages encourages knowledge transfer, especially from the high-resource language (HRL) to low-resource language (LRL) and even enables zero-shot translation [Aharoni *et al.*, 2019; Zhang *et al.*, 2020].

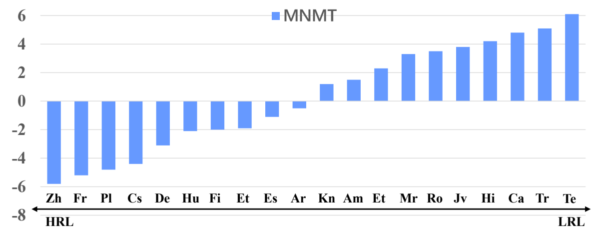


Figure 1: Results of the multilingual translation model are reported as ΔBLEU relative to the corresponding bilingual counterpart. The languages are arranged from high-resource languages (HRLs) to low-resource languages (LRLs).

While having attracted many interests, MNMT is still beleaguered by the *negative language interference* residing in the multilingual parameter sharing [Conneau *et al.*, 2020; Wang *et al.*, 2020b; Gong *et al.*, 2021], where the multiple translation directions degrade performance on high-resource languages. In Figure 1, the multilingual model outperforms the bilingual model on low-resource translations profited by knowledge transferability. But the multilingual model performs worse than the bilingual counterpart when it comes to high-resource translations, which is exactly the manifestation of the negative interference caused by the severe competition in multiple training directions among HRLs and LRLs.

To mitigate the negative interference, the previous works [Wang *et al.*, 2019; Philip *et al.*, 2020] propose the language adapter owing to its simplicity. The adapter modules are usually added to the encoder and decoder layers to indicate the source and target language. With the whole multilingual model frozen, only the parameters of the adapter continue to be tuned. But when the number of languages is large, the adapter methods suffer from the sharp increasing parameters and extra inference time. Another line of research is to extend the depth and width of the MNMT model and construct large-scale corpora to increase model capacity and include more languages [Zhang *et al.*, 2020; Kong *et al.*, 2021]. It still rises in substantial training and inference costs caused by numerous extra parameters.

In this work, we propose a novel multilingual translation model with the high-resource language-specific training (HLT-MT) for one-to-many and many-to-many translation directions in a two-stage training framework. In the first stage, our model is trained only on high-resource languages with

[*] Contribution during internship at Microsoft Research.
[†] Corresponding author.

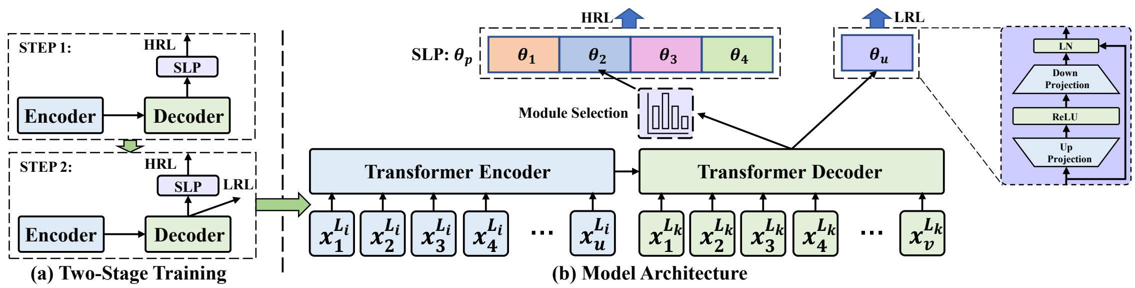


Figure 2: (a) is the two-stage training framework and (b) is the model architecture with the selective language-specific pool (SLP) for HRL and the universal layer for LRL. We first train the multilingual model on the high-resource pairs with the languages-specific pool and then continue tuning the model on all multilingual corpora to help transfer the knowledge from the high-resource to the low-resource languages. Given different translation directions of K languages $L_{all} = \{L_k\}_{k=1}^{K}$, we first employ language-specific training for the high-resource translation directions and then continue to train on all available multilingual corpora. If L_k is a HRL, we use the selection function $g(L_k)$ to decide which language-specific module will be used. Otherwise, we use the universal layer θ_u for the low-resource language L_k.

the language-specific modules to avoid the negative impact caused by LRLs. To address the negative interference problem among HRLs, we introduce a language-specific pool containing a sequence of independent modules for HRLs. Considering the increasing number of the languages, we apply the selection mechanism to the language-specific pool with the constrained size, denoted as selective language-specific pool (SLP), which enables different groups of certain languages to share the same module from SLP. After pretraining with the high-resource languages, we extract the shared representations of the decoder to transfer the fine-trained knowledge to the low-resource languages.

Our method aims to enhance translation quality of high-resource directions with the selective language-specific pool compared to the bilingual counterpart and then transfer the knowledge to the low-resource directions based on the bottom shared features. We conduct experiments on the WMT-10 benchmark of 11 languages and OPUS-100 benchmark of 95 languages. Experimental results demonstrate that our method significantly outperforms previous bilingual and multilingual baselines. Besides, extensive probing experiments are performed for the multilingual baseline and HLT-MT, helping further analyze how our method can benefit the multilingual machine translation. Empirical studies show that HLT-MT maintains a balance between language-agnostic and language-distinct features and thus helps to alleviate the negative language interference among various languages.

2 Negative Language Interference

Multilingual translation model aims at transferring knowledge across languages to boost performance on low-resource languages, where the multilingual model is trained in multiple translation directions simultaneously to enable cross-lingual transfer through parameter sharing. However, different groups of languages have heterogeneous characteristics, such as different dictionaries and grammars. The previous works have shown [Wang *et al.*, 2020b; Yu *et al.*, 2020] that knowledge transfer is not beneficial for all languages by sharing all parameters. To analyze the effect of mutual influence among different languages, we calculate the cosine similarities between gradients of two translation directions. In Figure 3, we observe that the certain high-resource language sustains negative interference from other HRLs and LRLs. For example, En→Cs acutely conflicts with En→De and En→Hi in the second row of Figure 3. This shows that HRLs are conducive to LRLs but may be hindered by other HRLs and LRLs in turn. To prevent the HRLs from negative interference introduced by LRLs, we focus on sufficiently training the high-resource directions and then continue tuning on all directions. To further address the conflicts among HRLs, we propose the selective language-specific pool for different high-resource languages. Our method effectively mitigates the negative interference in the analytic experiments compared to baselines.

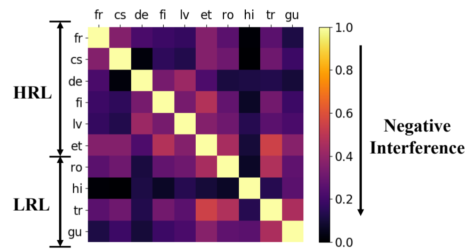


Figure 3: Negative interference among multiple languages.

3 Our Method

In this section, we introduce HLT-MT for multilingual translation. We propose the two-stage training framework, where the selective languages-specific pool (SLP) and the universal layer are applied for HRLs and LRLs respectively.

3.1 Overview of HLT-MT

Given different translation directions of K languages $L_{all} = \{L_k\}_{k=1}^{K}$, we employ SLP for high-resource directions and the universal layer for low-resource directions. Our method is illustrated in Figure 2, where the selective language-specific pool (SLP) for the high-resource languages is inserted into the top of the decoder denoted by $\theta_p = \{\theta_t\}_{t=1}^{T}$. T is the constrained size of the selective language-specific pool and

$T \leq K$ since the number of languages K can be numerous. Thus, SLP contains T individual modules with the same architecture, and each θ_t is a sub-network. If the translation direction $L_i \to L_k$ is the high-resource direction (from the source language L_i to target language L_k), we only activate the relevant language-specific module from SLP for L_k. Otherwise, the universal layer is triggered for L_k.

3.2 Multilingual Machine Translation

Given the high-resource bilingual corpora $D^h = \{D_m^h\}_{m=1}^M$ and low-resource corpora $D^l = \{D_n^l\}_{n=1}^N$, where M and N separately represent the number of the high-resource and low-resource training corpora of K languages $L_{all} = \{L_k\}_{k=1}^K$. The multilingual model is jointly trained on the union of the high- and low-resource training corpora $D^h \cup D^l$:

$$\begin{aligned}\mathcal{L}_{MT} = &-\sum_{m=1}^{M} \mathbb{E}_{x,y\sim D_m^h}\left[\log P(y|x;\Theta)\right] \\ &-\sum_{n=1}^{N} \mathbb{E}_{x,y\sim D_n^l}\left[\log P(y|x;\Theta)\right]\end{aligned} \tag{1}$$

where the first and second term denote objective of the high- and low-resource training corpora respectively. Θ are shared parameters for all languages. x and y are the sentence pair.

3.3 High-resource Language-specific Training

To prevent the high-resource languages from the negative interference caused by low-resource languages, we only train the model with SLP on high-resource directions, which effectively ameliorates translation quality of high-resource translation directions with slight extra parameters.

To take advantages of the cross-lingual pretrained encoder to boost model performance, our multilingual model is initialized by XLM-R [Conneau *et al.*, 2020]. Besides, we verify the effectiveness of our method on the Transformer model [Vaswani *et al.*, 2017] without any pretrained model. Given the source sentence $x^{L_i} = \{x_1^{L_i}, \dots, x_u^{L_i}\}$ with u words and target sentence $x^{L_k} = \{x_1^{L_k}, \dots, x_v^{L_k}\}$ with v words, the shared features $h_s^{L_k}$ of the target language L_k at the top of the decoder are obtained by the Transformer model:

$$h_s^{L_k} = \text{Transformer}(x^{L_i}, x^{L_k}; \Theta) \tag{2}$$

where $h_s^{L_k} = \{h_{s_1}^{L_k}, \dots, h_{s_j}^{L_k}, \dots, h_{s_v}^{L_k}\}$ and $h_{s_j}^{L_k}$ is the j-th representation of the target token $h_j^{L_k}$ generated by the single shared Transformer encoder and decoder.

After obtaining a sequence of decoder representations $h_s^{L_k}$ of the high-resource language L_k, we project the language-agnostic representations to the language-distinct ones via the language-specific pool with the selective mechanism. Given the translation direction $L_i \to L_k$ $(1 \leq i, k \leq K \wedge i \neq k)$ and the selective language-specific pool $\theta_p = \{\theta_t\}_{t=1}^T$, the corresponding module $\theta_{g(L_i,L_k)}$ is used to generate the language-distinct representations. $g(\cdot)$ is a map function that maps the language index to the corresponding module index: $L_k \in \{1, \dots, K\} \longmapsto t \in \{1, \dots, T\}$. $g(L_i, L_k)$ is the map function only depending on the target language and thus can be simplified into $g(L_k)$. Therefore, we project the language-agnostic representations $h_s^{L_k}$ to the language-specific features $h_b^{L_k}$ using function $\mathcal{F}_{\theta_{g(L_k)}}$ as below:

$$h_b^{L_k} = \mathcal{F}_{\theta_{g(L_k)}}(h_s^{L_k}) \tag{3}$$

where $h_s^{L_k}$ is the representations generated by the shared parameters. $\mathcal{F}_{\theta_{g(L_k)}}$ is a function defined as below:

$$\mathcal{F}_{\theta_{g(L_k)}}(h_s^{L_k}) = f(W_{g(L_k)}^d \sigma(W_{g(L_k)}^u h_s^{L_k}) + h_s^{L_k}) \tag{4}$$

where $f(\cdot)$ is the layer normalization and $\sigma(\cdot)$ is the ReLU activation function. $W_{g(L_k)}^u \in R^{d_e \times d_h}$ is the up-projection matrix and $W_{g(L_k)}^d \in R^{d_h \times d_e}$ is the down-projection matrix as shown in the right part of Figure 2, where d_e and d_h are the embedding size and hidden size of SLP ($d_e < d_h$).

Another issue is how to design a proper map function $g(\cdot)$ with an appropriate selection mechanism for the translation direction $L_i \to L_k$. In our work, each source sequence is prefixed with a special target language symbol to indicate the translation direction, which enables the decoder to correctly generate the target sentence with the shared decoder parameters. Therefore, the embedding of the target language symbol is used to select the language-specific module from SLP. The selection function $g(\cdot)$ is defined as:

$$g(L_k) = \arg\max_{1 \leq t \leq T} \frac{exp(e_t^{L_k})}{\sum_{i=1}^{T} exp(e_i^{L_k})} \tag{5}$$

where $e^{L_k} = W_g E[L_k]$ of T dimensions. $E[L_k]$ denotes the embedding of the target language L_k symbol. $W_g \in R^{d_e \times T}$ maps the target embedding to the vector e^{L_k}, where $e_i^{L_k}$ is the i-th element of e^{L_k} and SLP is comprised of T sub-networks. The sub-network with the highest probability will be selected to produce the language-specific features.

Equation 3 and 5 are only related to $\theta_{g(L_k)}$ and thus can not propagate gradients to all language-specific parameters. The selective language-specific pool $\theta_p = \{\theta_t\}_{t=1}^T$ contains a set of modules described in Equation 4. To tackle the undifferentiable problem of SLP, we use the weighted average to ensure gradients to be propagated to all language-specific modules:

$$h_b^{L_k} = \sum_{t=1}^{T} \alpha_t^{L_k} \mathcal{F}_{\theta_t}(h_s^{L_k}) \tag{6}$$

where $\alpha_t^{L_k}$ is calculated by the target embedding and softmax function. We project the target embedding $E[L_k]$ the probability vector $e^{L_k} = W_g E[L_k]$ with the learned matrix W_g.

$$\alpha_t^{L_k} = \frac{exp(e_t^{L_k})}{\sum_{i=1}^{T} exp(e_i^{L_k})} \tag{7}$$

where $e^{L_k} = W_g E[L_k]$ of T dimensions. $W_g \in R^{d_e \times T}$ and $E[L_k]$ denote the language embedding of L_k. d_e is the embedding size. $\alpha_t^{L_k}$ is the t-th element of the vector α^{L_k}.

In the practice training, we alternately leverage the Equation 3 and 6 with equal probabilities to learn the map function and generate the language-distinct representations $h_b^{L_k}$. Finally, the representations are used to generate the target sentence x_{L_k}:

$$x^{L_k} = softmax(W^o h_b^{L_k}) \tag{8}$$

where x^{L_k} is the target sentence and $W^o \in d_e \times V$ is the output matrix, where V is the vocabulary size.

3.4 Low-resource Transfer

After training the multilingual model on the high-resource language pairs, the bottom features generated by the shared parameters are utilized for the low-resource target sentence generation. Then, our multilingual model is jointly trained on the high-resource bilingual corpora D^h and low-resource bilingual corpora D^l. Given the low-resource translation direction $L_i \rightarrow L_j$, the shared representations $h_s^{L_j}$ generated by the shared parameters Θ are fed into a universal layer θ_u:

$$h_b^{L_j} = \mathcal{F}_{\theta_u}(h_s^{L_j}) \tag{9}$$

where $h_b^{L_j}$ are features generated by the shared parameters in Equation 2. θ_u is a sub-network same as the θ_t ($1 \leq t \leq T$) of SLP $\theta_p = \{\theta_t\}_{t=1}^{T}$. All low-resource languages share the same universal layer to project the shared features $h_s^{L_j}$ to the last representations $h_b^{L_j}$. Then the target sentence is produced by $h_b^{L_j}$ and output matrix W^o similar to Equation 8.

3.5 Training Strategy

Our model first accumulates the cross-entropy loss on the high-resource pairs and the disparity loss. Then, the multilingual model trained on multilingual corpora to maintain the performance of high-resource languages and meanwhile transfer the knowledge to the low-resource languages.

Disparity Loss To encourage different languages to select different language-specific modules of SLP, we minimize the disparity loss L_d, which measures the similarity of language-specific layer selection among languages.

$$\mathcal{L}_d = \sum_{i=1}^{N} \sum_{j=i+1}^{N} (\alpha^{L_i} \cdot \alpha^{L_k}) \tag{10}$$

where $\alpha^{L_i}, \alpha^{L_k} \in R^T$ denote the selection probabilities generated by Equation 7, where SLP contains T modules.

High-resource Language-specific Training The objective is to minimize the cross-entropy loss of high-resource training corpora and the auxiliary disparity loss jointly as below:

$$\mathcal{L}_{high} = -\sum_{m=1}^{M} \mathbb{E}_{x,y \sim D_m^h} [\log P(y|x; \Theta, \theta_p)] + \mathcal{L}_d \tag{11}$$

where Θ denotes all shared parameters and θ_p are parameters of SLP. x and y are sentence pair.

Multilingual Training After trained on the high-resource directions, the multilingual model is continued to be tuned on the union of the high- and low-resource corpora $D_h \cup D_l$ with the extra SLP for high-resource languages and the universal layer for low-resource languages:

$$\begin{aligned} \mathcal{L}_{all} = &-\sum_{m=1}^{M} \mathbb{E}_{x,y \sim D_m^h} [\log P(y|x; \Theta, \theta_p)] \\ &-\sum_{n=1}^{N} \mathbb{E}_{x,y \sim D_n^l} [\log P(y|x; \Theta, \theta_u)] \end{aligned} \tag{12}$$

where Θ are shared parameters for all languages. SLP contains a list of language-specific layers for HRLs and θ_u is a universal layer for LRLs.

4 Experiments

4.1 Datasets

To evaluate our method, we conduct experiments on the WMT-10 and the OPUS-100 dataset.

WMT-10 We use a collection of parallel data in different languages from the WMT datasets to evaluate the models [Wang *et al.*, 2020a], The parallel data is between English and other 10 languages, including French (Fr), Czech (Cs), German (De), Finnish (Fi), Latvian (Lv), Estonian (Et), Romanian (Ro), Hindi (Hi), Turkish (Tr) and Gujarati (Gu).

OPUS-100 We use the OPUS-100 corpus [Zhang *et al.*, 2020] for massively multilingual machine translation. OPUS-100 is an English-centric multilingual corpus covering 100 languages, which is randomly sampled from the OPUS collection. We obtain 94 English-centric language pairs after dropping out 5 languages, which lack corresponding test sets.

4.2 Baselines

Our method is compared to the bilingual and multilingual methods. For a fair comparison, **XLM-R** and **LS-MNMT** are initialized by XLM-R [Conneau *et al.*, 2020]. **BiNMT** [Vaswani *et al.*, 2017] is the bilingual Transformer model. **MNMT** [Johnson *et al.*, 2017] is jointly trained on all directions, where the target language symbol is prefixed to the input sentence. **mBART** [Liu *et al.*, 2020] is an encoder-decoder pretrained model and then is finetuned on all corpora. **XLM-R** [Conneau *et al.*, 2020] is initialized by the pretrained model XLM-R [Ma *et al.*, 2020]. **LS-MNMT** [Fan *et al.*, 2020] integrates the language-specific layers of all languages into the end of the decoder.

4.3 Training and Evaluation

We adopt Transformer as the backbone model for all experiments. We train multilingual models with Adam ($\beta_1 = 0.9$, $\beta_2 = 0.98$). The learning rate is set as 5e-4 with a warm-up step of 4,000. The models are trained with the label smoothing cross-entropy with a smoothing ratio of 0.1. The batch size is 4096 tokens on 64 Tesla V100 GPUs. For WMT-10, we first train the multilingual model with 6 languages and then finetunes on all languages. For OPUS-100, the model is trained in the languages where the number of pairs exceeds 10K. The evaluation metric is the case-sensitive detokenized sacreBLEU [Post, 2018].

4.4 Main Results

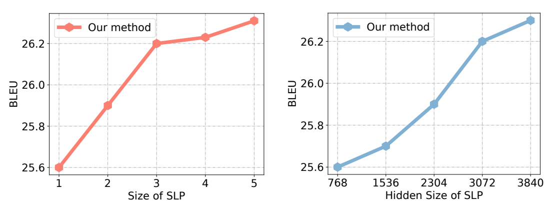
WMT-10 As shown in Table 1, our method clearly improves multilingual baselines by a large margin in 10 translation directions. Previously, multilingual machine translation underperforms the bilingual translation model in rich-resource scenarios. It is worth noting that our multilingual machine translation baseline XLM-R is already very competitive initialized by the cross-lingual pretrained model. Interestingly, our method outperforms the bilingual baseline in the high-resource translation direction, such as En→De translation direction (+1.5 BLEU points). Our method consistently outperforms the multilingual baseline on all language pairs, confirming that using HLT-MT to alleviate negative interference can help boost performance.

En→X test sets		#Params	Fr	Cs	De	Fi	Lv	Et	Ro	Hi	Tr	Gu	Avg_{all}
1→1	BiNMT [Vaswani *et al.*, 2017]	242M/10M	36.3	22.3	40.2	15.2	16.5	15.0	23.0	12.2	13.3	7.9	20.2
1→N	MNMT [Johnson *et al.*, 2017]	242M	34.2	20.9	40.0	15.0	18.1	20.9	26.0	14.5	17.3	13.2	22.0
	mBART [Liu *et al.*, 2020]	611M	33.7	20.8	38.9	14.5	18.2	20.5	26.0	15.3	16.8	12.9	21.8
	XLM-R [Conneau *et al.*, 2020]	362M	34.7	21.5	40.1	15.2	18.6	20.8	26.4	15.6	**17.4**	**14.9**	22.5
	LS-MNMT [Fan *et al.*, 2020]	409M	35.0	21.7	40.6	15.5	18.9	21.0	26.2	14.8	16.5	12.8	22.3
	HLT-MT (Our method)	381M	**36.2**	**22.2**	**41.8**	**16.6**	**19.5**	**21.1**	**26.6**	**15.8**	17.1	14.6	**23.2**
N→N	MNMT [Johnson *et al.*, 2017]	242M	34.2	21.0	39.4	15.2	18.6	20.4	26.1	15.1	17.2	13.1	22.0
	mBART [Liu *et al.*, 2020]	611M	32.4	19.0	37.0	13.2	17.0	19.5	25.1	**15.7**	16.7	14.2	21.0
	XLM-R [Conneau *et al.*, 2020]	362M	34.2	21.4	39.7	15.3	18.9	20.6	26.5	15.6	17.5	14.5	22.4
	LS-MNMT [Fan *et al.*, 2020]	409M	34.8	21.1	39.3	15.2	18.7	20.5	26.3	14.9	17.3	12.3	22.0
	HLT-MT (Our method)	381M	**35.8**	**22.4**	**41.5**	**16.3**	**19.6**	**21.0**	**26.6**	**15.7**	**17.6**	**14.7**	**23.1**

Table 1: En→X evaluation results for bilingual (1→1), one-to-many (1→N), and many-to-many (N→N) models on WMT-10. The languages are ordered from high-resource languages (left) to low-resource languages (right).

Models (N→N)	#Params	X→En					En→X				
		$High_{45}$	Med_{21}	Low_{28}	Avg_{94}	WR	$High_{45}$	Med_{21}	Low_{28}	Avg_{94}	WR
Previous Best System [Zhang *et al.*, 2020]	254M	30.3	32.6	31.9	31.4	-	23.7	25.6	22.2	24.0	-
MNMT [Johnson *et al.*, 2017]	242M	32.3	35.1	35.8	33.9	*ref*	26.3	31.4	31.2	28.9	*ref*
XLM-R [Conneau *et al.*, 2020]	362M	33.1	35.7	36.1	34.6	-	26.9	31.9	31.7	29.4	-
LS-MNMT [Fan *et al.*, 2020]	456M	33.4	35.8	35.9	34.7	-	27.5	31.6	31.5	29.6	-
HLT-MT (Our method)	391M	**34.1**	**36.6**	**36.1**	**35.3**	72.3	**27.6**	**33.3**	**31.8**	**30.1**	77.7

Table 2: X→En and En→X test BLEU for high/medium/low resource language pairs in many-to-many setting on OPUS-100 test sets. The BLEU scores are average across all language pairs in the respective groups. "WR": win ratio (%) compared to *ref* (MNMT).



(a) Number of Module (b) Hidden size

Figure 4: Average results of En→X high-resource directions (Fr, Cs, De, Fi, Lv, and Et) on the WMT-10 benchmark.

OPUS-100 In Table 2, we observe that HLT-MT achieves reasonable results on 94 translation directions. The improvement can be attributed to the high-resource training with the selective language-specific pool, which avoids the competition between high-resource and low-resource training directions. Another benefit of our approach is light and convenient to be applied to different backbone models since the parameters of the selective language-specific pool on the top of the decoder are tiny compared to all parameters.

5 Analysis

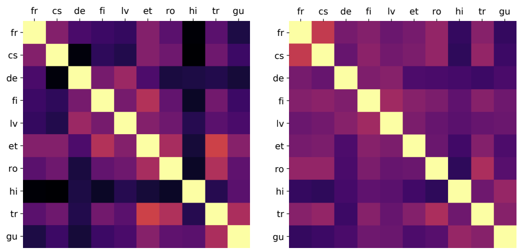
Size of Language-specific Parameters The size of the selective language-specific pool depends on the two key factors, namely the hidden size d_h and the number of modules T described in Equation 4 and 3. We tune the different values of d_h and T in Figure 4(a) and 4(b) on the WMT-10 dataset. Naturally, the selective language-specific pool with a larger capacity leads to better performance. Increasing the number of the selective pool brings more improvement than the improvement of hidden size. Our method can efficiently reduce the language-specific parameters ($T = 3$) using the selection mechanism and get comparable results compared to the baseline, where each high-resource language has the independent language-specific layer ($T = 6$).

XLM-R	Two-stage Training	SLP	Avg_{high}	Avg_{low}	Avg_{all}
			24.9	17.8	22.0
	✓		25.4	18.0	22.4
	✓	✓	26.0	18.1	22.8
✓			25.2	18.5	22.5
✓	✓		26.0	17.9	22.8
✓	✓	✓	**26.2**	**18.5**	**23.2**

Table 3: Ablation study of our proposed approach on the WMT-10 benchmark. Our method can be easily initialized by the cross-lingual pretrained model XLM-R to enhance the performance.

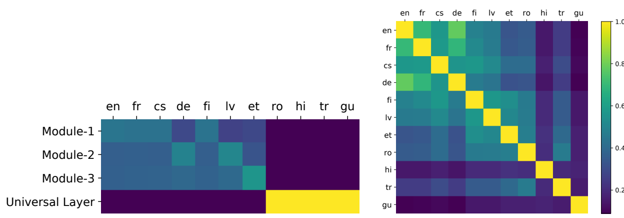
Ablation Study In Table 3, we empirically validate our approach on the different backbone models including Transformer [Vaswani *et al.*, 2017] without any pretrained model and XLM-T [Ma *et al.*, 2020] initialized by the cross-lingual pretrained model XLM-R [Conneau *et al.*, 2020]. High-resource training significantly helps improve the model performance but has trouble in effectively handling the low-resource translation directions merely by sharing all parameters, which is caused by the competition in the shared parameters between high-resource and low-resource languages. By introducing the selective language-specific pool (SLP) and extracting the bottom representations for low-resource languages, our approach ameliorates all translation directions.

Conflicting Gradient To delve into the function of the language-specific module for multilingual training [Yu *et al.*, 2020], we define $\Phi(L_a, L_b) = \frac{g_{L_a} \cdot g_{L_b}}{\|g_{L_a}\| \|g_{L_b}\|}$ as the cosine similarity between two task gradients g_{L_a} and g_{L_b}, where g_{L_a} and g_{L_b} separately denote the gradients of the En→ L_a and En→ L_b translation direction. $\Phi(L_a, L_b)$ determines whether g_{L_a} conflicts with g_{L_b} by computing the cosine similarity be-



(a) Baseline (b) Our method

Figure 5: Cosine similarities between two gradients of training directions in (a) the baseline and (b) our method. Lower similarity (darker color) means higher negative interference.



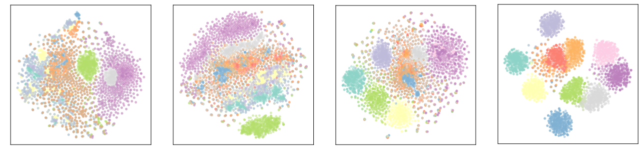
(a) Selection probabilities. (b) Dictionary overlaps.

Figure 6: (a) is the selection probabilities of different language-specific modules generated by the embedding of the target language symbol. The universal layer is only used for the low-resource languages. (b) is the overlapping ratio of the dictionaries between different languages. Lighter green means the higher selection probability of (a) and overlapping ratio of (b).

tween vectors g_{L_a} and g_{L_b}, where the small value indicate conflicting gradients. Figure 5(b) and 5(a) show the similarities of the baseline and our method between different training directions. Different training tasks of our method have similar optimization, where $\Phi(L_a, L_b)$ has larger value compared to the baseline scores. It corroborates that our language-specific training can effectively mitigate the conflicting gradients.

Language-specific Selection Mechanism The selective language-specific pool (SLP) of high-resource languages significantly contributes to translation quality. Equation 7 describes the selection of the module with the highest probabilities generated by the embedding of the target language symbol for the given translation direction. In Figure 6(a), θ_1 is used for En, Fr, Cs, and Fi. θ_2 is for De and Lv. θ_3 is for Et. The universal layer is used for all low-resource languages.

In Figure 6(b), we calculate the overlaps of dictionaries among multiple languages to measure the relationship of different languages. The language similarity between L_a and L_b is calculated by $\text{Sim}(L_a, L_b) = \frac{\|\mathcal{D}_{L_a} \cap \mathcal{D}_{L_b}\|}{\|\mathcal{D}_{L_a} \cup \mathcal{D}_{L_b}\|}$, where $\mathcal{D}_{L_a}$ and $\mathcal{D}_{L_b}$ are the dictionary of language L_a and L_b. Figure 6(b) shows that the language Et has the minimum overlap between other high-resource languages, where θ_3 is only used for Et. Therefore, we conclude that similar languages tend to select the same language-specific module from SLP.



(a) 2-th (b) 3-th (c) 6-th (d) 7-th

Figure 7: t-SNE visualization of the sentence representations from the bottom decoder layer (a), (b), (c), to the language-specific layer from SLP (d). Each color denotes one language.

Decoder Representation Visualization We randomly select 500 English sentences and visualize their representations [Maaten and Hinton, 2008] of the bottom decoder layers and the language-specific layer in Figure 7. The first hidden state of the decoder is regarded as the sentence representation. Compared to Figure 7(a), 7(b), and 7(c), different languages become more distinct and less likely to overlap with each other in Figure 7(d), proving that the selective language-specific pool (SLP) effectively projects the language-shared representations into language-distinct ones for better target generation of different target languages.

6 Related Work

Multilingual neural machine translation (MNMT) [Johnson *et al.*, 2017; Aharoni *et al.*, 2019; Fan *et al.*, 2020; Kong *et al.*, 2021; Tang *et al.*, 2021] enables numerous translation directions by shared encoder and decoder for all languages. The MNMT system can be categorized into one-to-many [Wang *et al.*, 2018], many-to-one [Tan *et al.*, 2019], and many-to-many [Pan *et al.*, 2021] translation. Previous studies utilize assisting high-resource languages to improve low-resource or even zero-shot translation.

While MNMT is promising, it often underperforms bilingual baselines due to the interference in shared parameters, especially on high-resource pairs [Wang *et al.*, 2020b]. To address this issue, language-specific modules are proposed to both enhance the low-resource translation and maintain the high-resource performance. Recent works mainly focus on designing language-specific components to boost the rich-resource translation quality [Vázquez *et al.*, 2019; Philip *et al.*, 2020; Gong *et al.*, 2021]. Further works discuss when and where language-specific capacity matters most in MNMT [Escolano *et al.*, 2021]. Our method finds a better balance between language-specific and language-agnostic models to mitigate negative interference.

7 Conclusion

In this work, we propose a novel multilingual translation model with the high-resource language-specific training called HLT-MT. The multilingual model is trained on multiple high-resource corpora with the selective language-specific pool, followed by continuing training on both high- and low-resource languages. Experimental results evaluated on WMT-10 and OPUS-100 benchmarks demonstrate that HLT-MT significantly outperforms all previous baselines.

References

[Aharoni *et al.*, 2019] Roee Aharoni, Melvin Johnson, and Orhan Firat. Massively multilingual neural machine translation. In *NAACL 2019*, pages 3874–3884, 2019.

[Conneau *et al.*, 2020] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *ACL 2020*, pages 8440–8451, 2020.

[Escolano *et al.*, 2021] Carlos Escolano, Marta R. Costa-jussà, José A. R. Fonollosa, and Mikel Artetxe. Multilingual machine translation: Closing the gap between shared and language-specific encoder-decoders. In *EACL 2021*, pages 944–948, 2021.

[Fan *et al.*, 2020] Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. Beyond english-centric multilingual machine translation. *CoRR*, abs/2010.11125, 2020.

[Firat *et al.*, 2016] Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *NAACL 2016*, pages 866–875, 2016.

[Gong *et al.*, 2021] Hongyu Gong, Xian Li, and Dmitriy Genzel. Adaptive sparse transformer for multilingual translation. *CoRR*, abs/2104.07358, 2021.

[Johnson *et al.*, 2017] Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's multilingual neural machine translation system: Enabling zero-shot translation. *TACL 2017*, 5:339–351, 2017.

[Kong *et al.*, 2021] Xiang Kong, Adithya Renduchintala, James Cross, Yuqing Tang, Jiatao Gu, and Xian Li. Multilingual neural machine translation with deep encoder and multiple shallow decoders. In *EACL 2021*, pages 1613–1624, 2021.

[Lin *et al.*, 2020] Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. Pre-training multilingual neural machine translation by leveraging alignment information. In *EMNLP 2020*, pages 2649–2663, 2020.

[Liu *et al.*, 2020] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *TACL*, 8:726–742, 2020.

[Ma *et al.*, 2020] Shuming Ma, Jian Yang, Haoyang Huang, Zewen Chi, Li Dong, Dongdong Zhang, Hany Hassan Awadalla, Alexandre Muzio, Akiko Eriguchi, Saksham Singhal, Xia Song, Arul Menezes, and Furu Wei. XLM-T: scaling up multilingual machine translation with pretrained cross-lingual transformer encoders. *CoRR*, abs/2012.15547, 2020.

[Maaten and Hinton, 2008] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(Nov):2579–2605, 2008.

[Pan *et al.*, 2021] Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. Contrastive learning for many-to-many multilingual neural machine translation. In *ACL 2021*, pages 244–258, 2021.

[Philip *et al.*, 2020] Jerin Philip, Alexandre Berard, Matthias Gallé, and Laurent Besacier. Monolingual adapters for zero-shot neural machine translation. In *EMNLP 2020*, pages 4465–4470, 2020.

[Post, 2018] Matt Post. A call for clarity in reporting BLEU scores. In *WMT 2018*, pages 186–191, 2018.

[Tan *et al.*, 2019] Xu Tan, Jiale Chen, Di He, Yingce Xia, Tao Qin, and Tie-Yan Liu. Multilingual neural machine translation with language clustering. In *EMNLP 2019*, pages 963–973, 2019.

[Tang *et al.*, 2021] Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. Multilingual translation from denoising pre-training. In *ACL 2021*, pages 3450–3466, 2021.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS 2017*, pages 5998–6008, 2017.

[Vázquez *et al.*, 2019] Raúl Vázquez, Alessandro Raganato, Jörg Tiedemann, and Mathias Creutz. Multilingual NMT with a language-independent attention bridge. In *ACL 2019*, pages 33–39, 2019.

[Wang *et al.*, 2018] Yining Wang, Jiajun Zhang, Feifei Zhai, Jingfang Xu, and Chengqing Zong. Three strategies to improve one-to-many multilingual translation. In *EMNLP 2018*, pages 2955–2960, 2018.

[Wang *et al.*, 2019] Yining Wang, Long Zhou, Jiajun Zhang, Feifei Zhai, Jingfang Xu, and Chengqing Zong. A compact and language-sensitive multilingual translation method. In *ACL 2019*, pages 1213–1223, 2019.

[Wang *et al.*, 2020a] Yiren Wang, ChengXiang Zhai, and Hany Hassan. Multi-task learning for multilingual neural machine translation. In *EMNLP 2020*, pages 1022–1034, 2020.

[Wang *et al.*, 2020b] Zirui Wang, Zachary C. Lipton, and Yulia Tsvetkov. On negative interference in multilingual models: Findings and A meta-learning treatment. In *EMNLP 2020*, pages 4438–4450, 2020.

[Yu *et al.*, 2020] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *NeurIPS 2020*, 2020.

[Zhang *et al.*, 2020] Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. Improving massively multilingual neural machine translation and zero-shot translation. In *ACL 2020*, pages 1628–1639, 2020.