UM4: Unified Multilingual Multiple Teacher-Student Model for Zero-Resource Neural Machine Translation

Jian Yang[1 *], **Yuwei Yin**[2 *], **Shuming Ma**[2], **Dongdong Zhang**[2], **Shuangzhi Wu**[3], **Hongcheng Guo**[2], **Zhoujun Li**[1 †], **Furu Wei**[2]

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

[2]Microsoft Research

[3]Tencent Cloud Xiaowei

{jiaya, lizj}@buaa.edu.cn, frostwu@tencent.com,

{v-yuweiyin, shumma, dozhang, v-hongguo, fuwei}@microsoft.com

Abstract

Most translation tasks among languages belong to the zero-resource translation problem where parallel corpora are unavailable. Multilingual neural machine translation (MNMT) enables one-pass translation using shared semantic space for all languages compared to the two-pass pivot translation but often underperforms the pivot-based method. In this paper, we propose a novel method, named as **U**nified **M**ultilingual **M**ultiple teacher-student **M**odel for N**M**T (**UM4**). Our method unifies source-teacher, target-teacher, and pivot-teacher models to guide the student model for the zero-resource translation. The source teacher and target teacher force the student to learn the direct source→target translation by the distilled knowledge on both source and target sides. The monolingual corpus is further leveraged by the pivot-teacher model to enhance the student model. Experimental results demonstrate that our model of 72 directions significantly outperforms previous methods on the WMT benchmark.

1 Introduction

The encoder-decoder framework [Vaswani *et al.*, 2017; Gheini *et al.*, 2021] has gained outstanding performance on rich-resource machine translation tasks, such as English-German, English-French, and Chinese-English [Koehn *et al.*, 2019; Zhou *et al.*, 2021; Johnson *et al.*, 2017], where large-scale parallel corpora are available. However, it is incapable of directly modeling the zero-resource translation task when the parallel training data does not exist.

A straightforward solution for the zero-resource machine translation problem is the pivot translation approach [Bertoldi *et al.*, 2008; Wu and Wang, 2009; Zahabi *et al.*, 2013; Cheng *et al.*, 2017]. Bilingual pivot-based models perform the two-pass translation, which increases the computation cost and potentially suffers from the error propagation problem [Zhu *et al.*, 2013]. There are also some works [Chen *et al.*, 2017; Currey and Heafield, 2019; Kim *et al.*, 2019a] directly building the source→target model but limited by the bilingual translation task. Beyond pivot-based methods, the multilingual model trained over multiple pivot corpora with shared parameters utilizes the language symbol to infer the desired translation direction [Firat *et al.*, 2016; Johnson *et al.*, 2017; Lakew *et al.*, 2019; Currey and Heafield, 2019]. The multilingual model benefits from different language pairs and only requires one-step translation, which avoids error propagation and saves inference time. But the performance of this approach [Kim *et al.*, 2019b] is worse than pivot-based models.

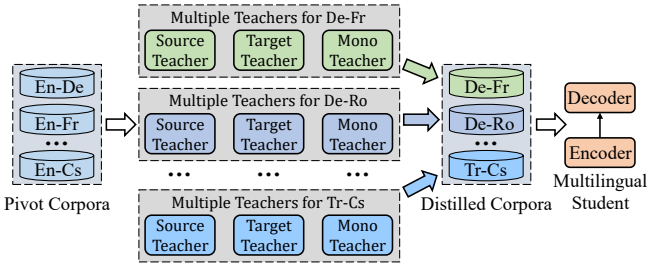

Figure 1: Framework of UM4. Unified multiple teachers with shared parameters trained on the original pivot corpora are used to guide the multilingual student model. English (En) is the pivot language.

Along the line of leveraging the multilingual model to address the zero-resource translation problem, we propose a novel method called **U**nified **M**ultilingual **M**ultiple teacher-student **M**odel for N**M**T (**UM4**). Given the available corpora of the pivot and other languages, we directly build the source→target student translation model guided by the multilingual multiple teachers as shown in Figure 1. The multiple teacher models can be decomposed into a source-teacher model, a target-teacher model, and a pivot-teacher model. The source-teacher model transfer the knowledge from the pivot to the source sentence. The target teacher distills pivot knowledge to the target side and boosts the capability of the target generation. The pivot-teacher model further enhances the student model by mining the potential of monolingual pivot corpora. The overall distilled corpora from unified teachers with normalized scores are used for the student.

Specifically, we first train a single multilingual model on

[*] Equal contribution. Work done during internship at Microsoft.

[†] Corresponding author.

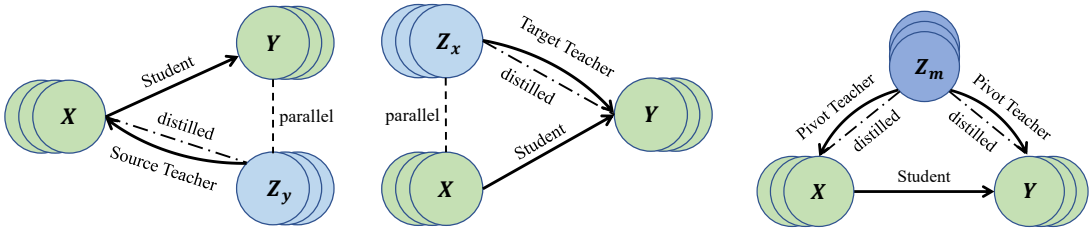

Figure 2: Overview of our unified multilingual multiple teacher-student model: (a) the source-teacher model, (b) target-teacher model, and (c) the pivot-teacher model. X, Y, Z separately denote the source, target, and pivot language. The dashed line "- -" denotes a real parallel corpus is available between the connected language pair and the dotted line ". ." denotes a distilled parallel corpus generated by the teacher model is available. Solid arrow lines represent translation directions. Our multiple teacher models consist of a source-teacher model, a target-teacher, and a pivot-teacher model, where Z_x and Z_y represent pivot language Z in the parallel corpus associated with source language X and target language Y respectively. The source teacher and target teacher transfers the knowledge from Z_y and Z_x separately. Given the monolingual corpus Z_m, the pivot-teacher model further enhances the multilingual student model by distilling knowledge to both source and target sides.

all pivot corpora as unified multiple teachers sharing all model parameters. Then, we construct the distilled multilingual corpora of all zero-resource directions using multiple teacher models. The distilled corpora with the normalized scores generated by the unified multiple teachers are used to guide a source→target student. We conduct experiments on the multilingual corpora from the WMT benchmark of 9 languages with 72 translation directions. The experimental results show that our method can significantly outperform multilingual baselines and pivot-based methods. Furthermore, we verify the effectiveness of our method by perturbation experiments and visualization of the multilingual sentence representations. Analytic results demonstrate that our UM4 student model with better crosslingual ability enhances zero-resource translations and avoids error propagation.

2 Our Approach

In this section, we introduce the unified multiple teacher-student model for zero-resource machine translation. As illustrated in Figure 2, our method simultaneously uses multiple teacher models to train a multilingual end-to-end translation model when the direct parallel data is unavailable.

2.1 Overview of UM4

Given the bilingual corpora $D_B = \{D_{B_n}\}_{n=1}^{N}$ of N languages, where one side is the pivot language L_z and the other side is the language $L_n \in \{L_n\}_{n=1}^{N}$, the multilingual model is trained on the available pivot corpora D_B to solve the zero-resource translation without direct parallel data between the zero-resource pair L_i and L_j ($1 \le i, j \le N$ and $i \ne j$):

$$\begin{aligned}\mathcal{L}_D = &\sum_{n=1}^{N} \mathbb{E}_{x,z_x \in D_{B_n}} [-\log P_\theta(z_x|x)] \\ &+ \sum_{n=1}^{N} \mathbb{E}_{y,z_y \in D_{B_n}} [-\log P_\theta(y|z_y)]\end{aligned} \tag{1}$$

where x, z_x denote the source and pivot language sentence in the bilingual corpus D_B. y, z_y denote the pivot and target sentence. $\mathcal{L}_D$ is the combined objective of the multilingual model. The multilingual model over the source-pivot and pivot-target corpora with shared parameters prepends the language symbol to indicate the zero-resource translation direction from language L_i to language L_j.

Without parallel training data for zero-resource language pairs, the multilingual model can easily translate into a wrong language and result in poor translation quality. Therefore, we introduce the synthetic multilingual multiple corpora of zero-resource language pairs $D_S = \{D_{S_m}\}_{m=1}^{M}$.

$$\mathcal{L}_S = \sum_{m=1}^{M} \mathbb{E}_{x,y \in D_{S_m}} [-w_{x,y} \log P_\theta(y|x)] \tag{2}$$

where x and y denote the source and target language sentence in the distilled multilingual corpora D_S. $w_{x,y}$ is the weight of the training sample from multilingual multiple teachers.

Our multilingual student model is trained on both original corpora D_B and distilled corpora D_S, which improves the translation quality under the supervised manner among zero-resource directions:

$$\mathcal{L}_T = \mathcal{L}_D + \mathcal{L}_S \tag{3}$$

where $\mathcal{L}_T$ is the total objective of our multilingual student model. $\mathcal{L}_D$ and $\mathcal{L}_S$ denote training objective over the original pivot corpora D_B and distilled corpora D_S respectively.

2.2 Multiple Teacher Models

Formally, given the source-pivot and target-pivot parallel corpus $D_{B_i} = \{x^{(k)}, z_x^{(k)}\}_{k=1}^{|D_{B_i}|}$ and $D_{B_j} = \{y^{(k)}, z_y^{(k)}\}_{k=1}^{|D_{B_j}|}$, we aim to build a source→target translation model $\theta_{x\to y}$ for the zero-resource translation task. x and y denote the source and target sentence respectively, z_x and z_y denote the pivot sentence from the source-pivot corpus D_{B_i} and pivot-target corpus D_{B_j} separately. $|D_{B_i}|$ and $|D_{B_j}|$ are the size of corpora D_{B_i} and D_{B_j}. θ denote model parameters.

Source-teacher Model If the target sentence y and pivot sentence z_y are parallel from the dataset D_{B_j}. The source teacher $\theta_{z_y\to x}$ is trained on the source-pivot corpus D_{B_i}. The source-teacher student training objective can be written as:

$$\mathcal{L}_S^{src} = -\mathbb{E}_{y,z_y \in D_{B_m}} \left[P(x|z_y;\theta_{z_y\to x}) \log P_\theta(y|x)\right] \tag{4}$$

where $P(x|z_y;\theta_{z_y\to x})$ is the weight generated by the source-teacher model $\theta_{z_y\to x}$.

Target-teacher Model If the source sentence x and pivot sentence z_x are parallel from the dataset D_{B_i}. The target teacher $\theta_{z_x \to y}$ on the pivot-target corpus D_{B_j}. The target-teacher student training objective can described as:

$$\mathcal{L}_S^{tgt} = -\mathbb{E}_{x,z_x \in D_{B_i}} \left[P(y|z_x; \theta_{z_x \to x}) \log P_\theta(y|x) \right] \quad (5)$$

where $P(y|z_x; \theta_{z_x \to y})$ is the weight generated by the target-teacher model $\theta_{z_x \to y}$.

Pivot-teacher Model Given the monolingual pivot corpus D_M, the pivot-teacher is used to guide the student model. The pivot-teacher model $\theta_{z_m \to x} \cup \theta_{z_m \to y}$ is trained on the pivot corpora D_{B_i} and D_{B_j}. The pivot-teacher student training objective can described as:

$$\mathcal{L}_S^{pivot} = -\mathbb{E}_{z_m \in D_M} \left[w_{x,y} \log P_\theta(y|x) \right] \quad (6)$$

where $w_{x,y} = P(y|z_m; \theta_{z_m \to y}) P(x|z_m; \theta_{z_m \to x})$ is the weight generated by the pivot-teacher model.

All teachers are based on the multilingual training on the available corpora D_B and shares the same semantic space for all languages. Therefore, the unified teacher is comprised of different teachers with respective functions simultaneously.

Combining the source-teacher model, the target-teacher model, and the pivot-teacher model, the training objective of our teacher-student training can be described as:

$$\mathcal{L}_S = \mathcal{L}_S^{src} + \mathcal{L}_S^{tgt} + \mathcal{L}_S^{pivot} \quad (7)$$

where the parameters of multiple teachers remain unchanged during the training process.

We adopt sequence-level knowledge distillation [Chen *et al.*, 2017] to distill the knowledge from teacher models to the student model in practice. Specifically, we use multiple teacher models to construct the distilled corpora of zero-resource language pairs $D_S = \{D_{S_1}, \ldots, D_{S_M}\}$ with corresponding normalized scores, combined with the original pivot corpora $D_B = \{D_{B_1}, \ldots, D_{B_N}\}$ to train the student model.

As shown in Figure 2, our method can utilize the source-teacher model, target-teacher model, and pivot-teacher model simultaneously to guide the source→target student model, resulting in a more powerful student model.

2.3 Teacher-Student Transfer

This section will introduce the knowledge distillation details of multilingual multiple teacher-student. Limited by the exponential search space of the source sentences x and y, we employ the beam search strategy to generate N-best translation candidates[1] and renormalize the probabilities to teach the student model to approximate the distribution of the teacher model as below:

$$w_{x,y} = \frac{exp(w_{x,y}/\tau)}{\sum_{s=1}^{S} exp(w_{x,y}^s/\tau)} \quad (8)$$

where S is the beam size of the fixed teacher model. $w_{x,y}^s$ is the probability of the s-th sentence generated by the teacher model. τ is the temperature. The temperature $\tau \to 0$ increases the weight for the top-selected distilled sentences. We set $\tau < 1.0$ to force the model to focus more on the best-distilled sentence pairs when training.

We first train a single multilingual model with all available pivot corpora D_B as the multiple teacher models for all languages instead of training different bilingual teacher models.

Source-teacher Transfer For the source-teacher model, we use the pivot→source model to translate the monolingual pivot sentences of the pivot-target corpus D_B into the distilled source sentences. In this way, we get a distilled corpus D_S^{src}. According to the Equation 4, the distilled corpora D_S^{src} with the score $w_{x,y} = P(x|z_y; \theta_{z_y \to x})$ is used to teach the student model.

Target-teacher Transfer We adopt the beam search strategies and translate the monolingual pivot language part in the source-pivot corpora D_B into target language sentences. Another distilled corpora D_S^{tgt} is obtained with the score $w_{x,y} = P(y|z_x; \theta_{z_x \to y})$ for knowledge transfer.

Pivot-teacher Transfer Given the additional monolingual pivot corpora, the pivot sentences are separately translated to the distilled source and target sentences by the pivot-teacher model. We obtain the distilled corpora D_S^{pivot} with the score $w_{x,y} = P(x|z_m; \theta_{z_m \to x}) P(y|z_m; \theta_{z_m \to y})$ from the monolingual corpus D_M.

Eventually, with the parameters of the multilingual teacher model fixed, we generate the distilled knowledge and combine them into a whole training dataset $D_S = D_S^{src} \bigcup D_S^{tgt} \bigcup D_S^{pivot}$ to train the multilingual source→target student model.

3 Experiments

We evaluate our method on the multilingual dataset including 9 languages and 56 zero-resource translation directions. English is the most popular language and there are extensive English-centric data in the real world compared to other languages. Therefore, English (En) is treated as the pivot language in all experiments.

3.1 Dataset

All experiments are conducted on the multilingual dataset of 9 languages extracted from the previous work [Wang *et al.*, 2020], including English (En), French (Fr), Czech (Cs), German (De), Finnish (Fi), Estonian (Et), Romanian (Ro), Hindi (Hi), and Turkish (Tr).

Bitext Data We collect the training data from the latest available year of each language between English and other languages on the WMT benchmark and exclude WikiTiles. The duplicated samples are removed and the number of parallel data of each language pair is limited to 10 million by randomly sampling from the whole corpus. For 72 translation directions of 9 languages, we use the same valid and test sets from TED Talks as the previous work[2] for evaluation.

[1] Sampling from synthetic data generated by the teacher model according to the probabilities is an easy way to force the student to approximate the teacher [Kim and Rush, 2016].

[2] http://phontron.com/data/ted_talks.tar.gz

Monolingual Data The English monolingual data is collected from NewsCrawl[3] and randomly sample 1 million English sentences. We use a multilingual NMT model to translate these English monolingual data to sentences of other languages as the augmented parallel data, which is utilized as the back-translation data for all baselines. Our method uses the pivot-teacher model to guide the training of the source→target student model with the monolingual data.

3.2 Evaluation

During the inference, the beam search strategy is performed with a beam size of 5 for the target sentence generation. We set the length penalty as 1.0. The last 5 checkpoints are averaged for evaluation. We report the case-sensitive detokenized BLEU using sacreBLEU[4].

3.3 Baselines

Our method is compared with pivot-based and multilingual baselines. **Bilingual Pivot-based** [Cheng *et al.*, 2017] translate source to target via pivot language using two single-pair NMT models trained on each pair. **Multilingual Pivot-based** [Lakew *et al.*, 2019] leverages a single multilingual NMT model trained in all available directions for pivot translation. The details of the multilingual baselines are described as follows. **Multilingual** [Johnson *et al.*, 2017] shares the same vocabulary of all languages and prepends the language symbol to the source sentence to indicate the translation directions. **Monolingual Adapter** [Philip *et al.*, 2020] tunes adapter of each language for zero-shot translation based on a pretrained multilingual model. **Teacher-Student** [Chen *et al.*, 2017] uses the pivot-target translation model to teach the source-target translation model. **MTL** [Wang *et al.*, 2020] proposes a multi-task learning (MTL) framework including the translation task and two denoising tasks.

3.4 Implementation Details

All experiments are conducted based on the `Transformer_big` architecture [Vaswani *et al.*, 2017]. Both the encoder and decoder contain 6 layers with 16 heads per layer. The word embedding size d_{model} is set to 1024 and the FFN (feed-forward network) size is 4096. The learning rate is set to 3e-4 with 4000 warmup steps on the multilingual dataset. Adam [Kingma and Ba, 2014] is used for updating the parameters. The model with a mini-batch size of 4096 tokens is trained on 64 Tesla V100 GPUs.

3.5 Experiment Results

The evaluation results over the test sets against the baselines are listed in Table 1 and Table 2. The source-pivot corpus is high-resource compared to the low-resource pivot-target corpus in Table 1, and the source-pivot corpus is low-resource compared to the high-resource pivot-target corpus in Table 2.

As observed in Table 1, pivot-based methods including **Multilingual Pivot** and **Bilingual Pivot** significantly outperforms the multilingual methods including **Multilingual**, **MTL**, and **Monolingual Adapter**. But pivot-based methods still suffer from error propagation and are computationally expensive by the two-pass translation. On the contrary, our **UM4** method is a unified multilingual source→target model and alleviates this problem. Compared with strong baselines **Teacher-Student**, our model achieves consistent improvements with $\geq$ 0.6 BLEU points gains ($\geq$ 0.6 with "MonoData") on $\text{Avg}^{>}_{28}$. It shows that our student model learns a high-quality representation space across multiple languages with guides of multiple teachers to enforce the capability of zero-resource translation directions. Given the monolingual pivot corpus, our method also beats the pivot-based and multilingual methods by the distilled knowledge from the pivot-teacher model.

In Table 2, the size of the source-pivot corpus is smaller than the pivot-target corpus, so the source teacher transfers more knowledge from the larger pivot-target corpus to guide the student model with more distilled sentences. Our **UM4** also beats all previous methods and gains $\geq$ 0.9 BLEU points gains ($\geq$ 0.5 with "MonoData") on $\text{Avg}^{<}_{28}$. It demonstrates the effectiveness and significance of the introduction of source-teacher model. Our UM4 method without monolingual data ($\text{Avg}^{>}_{28}$=11.8 and $\text{Avg}^{<}_{28}$=14.7) performs even better than all baselines with back-translation data.

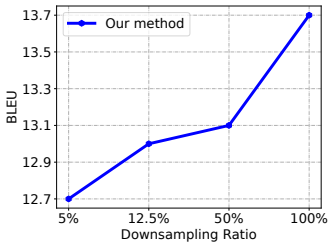

Figure 3: The overall average BLEU points of our method with different downsampling ratios.

3.6 Analysis

Effect of Different Teachers To investigate the effect of the different teachers, we train 7 student guided by all possible combinations of the source-teacher, target-teacher, and pivot-teacher model. Our method combines multiple teachers to direct the source-target student model simultaneously so that our method can improve performance. Table 3 shows ablation results guided by different teachers. Consistently, more teachers can lead to better results, which demonstrates that our proposed model can comprehensively manipulate the advantages of different teachers.

Size of Distilled Training Data Given the multilingual pivot corpora of the pivot language and other N languages, $N(N-1)$ distilled training sets of zero-resource directions are used to guide the multilingual student model. The overall scale of corpora contains $TN(N-1)$ sentence pairs, where T is the average size of the distilled corpora. To reduce complexity $\mathcal{O}(TN^2)$ to $\mathcal{O}(TN)$, we adopt a downsampling strategy with $\frac{1}{N}$ downsampling ratio as formulated below:

$$T' = \max\{T_m,\, T_m + (T - T_m)/N\} \tag{9}$$

[3]http://data.statmt.org/news-crawl

[4]BLEU+case.mixed+lang.{src}-{tgt}+numrefs.1+smooth.exp+tok.13a+version.1.3.1

X (High) → Y (Low)	Fr→Fi	Cs→Fi	Cs→Ro	Cs→Hi	De→Et	Fi→Et	Fi→Ro	Fi→Tr	Avg_8	$Avg_{28}^{>}$
Train on Parallel Data (Bitext).										
Bilingual Pivot [Cheng *et al.*, 2017]	13.5	13.4	15.2	2.6	13.4	12.7	13.1	3.2	10.9	9.5
Multilingual Pivot [Lakew *et al.*, 2019]	12.5	11.9	16.1	6.9	14.8	13.3	14.0	5.3	11.9	11.2
Multilingual [Johnson *et al.*, 2017]	3.8	10.2	12.6	5.1	12.5	12.0	10.7	4.0	8.9	8.1
Teacher-Student [Chen *et al.*, 2017]	13.0	13.6	16.4	7.1	15.6	14.6	14.6	5.0	12.5	10.9
Monolingual Adapter [Philip *et al.*, 2020]	8.2	10.7	14.3	5.9	12.1	12.6	12.4	4.8	10.1	9.2
MTL [Wang *et al.*, 2020]	6.0	9.0	13.0	6.0	14.3	12.0	11.7	4.6	9.6	8.9
UM4 w/o pivot-teacher model (our method)	**13.8**	**13.9**	**16.8**	**7.3**	**16.3**	**14.9**	**15.1**	**5.4**	**12.9**	**11.8**
Train on Parallel and Monolingual Data (Bitext + MonoData).										
Bilingual Pivot + BT [Cheng *et al.*, 2017]	13.9	13.4	16.3	6.9	15.3	13.7	13.6	4.8	12.2	11.0
Multilingual Pivot + BT [Lakew *et al.*, 2019]	13.5	12.6	16.0	6.7	14.8	13.3	14.0	5.6	12.1	11.2
Multilingual + BT [Johnson *et al.*, 2017]	7.5	10.2	14.4	5.7	12.5	12.9	10.7	5.3	9.9	9.4
Teacher-Student + BT [Chen *et al.*, 2017]	13.6	13.0	16.6	6.8	15.2	14.8	15.2	5.5	12.6	11.6
Monolingual Adapter + BT [Philip *et al.*, 2020]	10.8	7.6	15.1	5.0	15.4	14.1	14.1	5.4	10.9	10.0
MTL + BT [Wang *et al.*, 2020]	10.6	9.0	13.5	5.4	12.7	12.8	12.8	5.2	10.3	8.0
UM4 (our method)	**14.1**	**14.1**	**17.1**	**7.4**	**16.2**	**15.0**	**15.8**	**5.9**	**13.2**	**12.4**

Table 1: X→Y test results for bilingual and multilingual models of 9 language pairs on the WMT benchmark, where the source-pivot corpus is high-resource compared to the low-resource pivot-target corpus. Avg_8 is the average results of the listed directions and $Avg_{28}^{>}$ is the average BLEU points of all 28 directions under this setting.

X (Low) → Y (High)	Fi→De	Et→De	Et→Fi	Ro→Cs	Ro→De	Ro→Et	Tr→Fr	Tr→Et	Avg_8	$Avg_{28}^{<}$
Train on Parallel Data (Bitext).										
Bilingual Pivot [Cheng *et al.*, 2017]	15.5	15.3	11.0	14.6	16.8	11.8	10.0	5.8	12.6	11.1
Multilingual Pivot [Lakew *et al.*, 2019]	14.6	16.3	12.9	15.1	18.2	14.0	15.7	9.9	14.6	13.6
Multilingual [Johnson *et al.*, 2017]	11.4	12.5	10.1	12.1	15.6	10.7	7.2	5.2	10.6	9.2
Teacher-Student [Chen *et al.*, 2017]	16.0	17.9	14.1	16.0	19.1	15.1	16.4	11.0	15.7	13.6
Monolingual Adapter [Philip *et al.*, 2020]	11.8	14.7	11.5	13.1	16.4	12.2	11.7	7.8	12.4	10.4
MTL [Wang *et al.*, 2020]	11.7	15.1	10.1	13.0	16.1	12.5	10.4	7.0	12.0	10.4
UM4 w/o pivot-teacher model (our method)	**16.6**	**18.5**	**14.2**	**16.3**	**19.9**	**15.4**	**17.1**	**11.3**	**16.2**	**14.7**
Train on Parallel and Monolingual Data (Bitext + MonoData).										
Bilingual Pivot + BT [Cheng *et al.*, 2017]	15.0	17.0	12.3	16.0	18.6	13.9	14.6	9.0	14.6	13.8
Multilingual Pivot + BT [Lakew *et al.*, 2019]	16.2	17.4	12.8	15.8	19.4	14.2	16.7	10.4	15.4	14.1
Multilingual + BT [Johnson *et al.*, 2017]	13.6	16.3	12.3	14.9	16.1	12.7	12.1	8.6	13.3	11.3
Teacher-Student + BT [Chen *et al.*, 2017]	16.6	19.0	13.8	16.5	20.0	15.0	16.8	10.9	16.1	14.3
Monolingual Adapter + BT [Philip *et al.*, 2020]	13.8	13.8	11.6	15.6	11.7	13.7	13.4	9.6	12.9	10.8
MTL + BT [Wang *et al.*, 2020]	12.8	16.6	11.5	13.9	17.0	13.0	14.2	8.7	13.5	11.7
UM4 (our method)	**17.6**	**19.6**	**14.3**	**17.2**	**20.7**	**15.6**	**17.5**	**11.5**	**16.8**	**15.1**

Table 2: X→Y test results for bilingual and multilingual models of 9 language pairs on the WMT benchmark, where the source-pivot corpus is low-resource compared to the high-resource pivot-target corpus. Avg_8 is the average results of the listed directions and $Avg_{28}^{<}$ is the average BLEU points of all 28 directions under this setting.

Source	Target	Mono	Fr→De	De→Ro	Et→Ro	Avg_{56}
✓			21.3	17.0	14.5	12.3
	✓		21.4	16.2	15.2	13.0
		✓	22.5	17.2	15.4	12.7
	✓	✓	22.4	17.5	15.8	13.4
✓		✓	22.3	16.5	14.6	12.6
✓	✓		21.7	17.5	15.6	13.3
✓	✓	✓	**22.8**	**17.7**	**16.4**	**13.7**

Table 3: Ablation study on different teachers. Avg_{56} denotes the average BLEU points of 56 zero-resource translation directions.

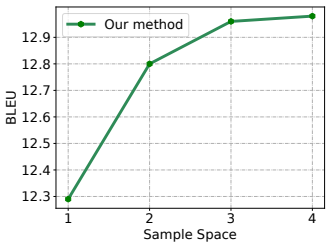

Figure 4: Effect of different beam sizes. We plot the curve of the average BLEU points of all directions with different beams sizes.

where T' is the size of the downsampled corpus. $T_m = 1M$ is the threshold to avoid undersampling the low-resource pairs.

In our work, 16 parallel corpora and 56 (8×7) distilled corpora are used for training. The results with different sampling ratios are listed in Figure 3, which indicates the proper downsampling ratio ($\frac{1}{N} = \frac{1}{8} = 12.5\%$) can simultaneously reduce computation cost and get comparable performance. The training time cost of our model is acceptable in a practical scenario. Therefore, our proposed method has been evaluated on the 72 translation directions and can be easily extended to more languages.

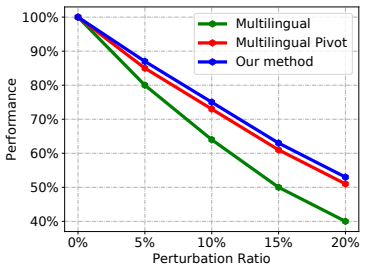

Figure 5: Comparison of different methods on the perturbation experiments with different corrupting probabilities. We display the average performance of all 56 zero-resource translation directions.

#Pairs	Fr→De	Ro→De	Tr→Cs	Avg_{16}	Avg_{56}
Supervised	11.7	16.1	9.6	22.8	8.7
Zero-resource	22.0	19.8	11.5	-	13.0
Both	**22.8**	**20.7**	**12.3**	**23.1**	**13.7**

Table 4: Comparison among multilingual student models trained with the original corpora ("Surpervised"), the distilled training corpora ("Zero-resource"), and their combination ("Both").

Sample Space of Possible Sequences We employ the sequence-level knowledge distillation as formulated in Equation 8 and examine our method with different sample space (beam size) settings, where S denotes beam size. Limited by the exponential search space, we use the beam search strategy with a beam size of S ($S \in [1, 4]$) to guide the student model. Figure 4 shows that the multilingual student model gains the best performance when $S = 3$ or $S = 4$ on the zero-resource translation tasks. Considering the computation cost and model performance, we set $S = 4$ in our work.

Robustness against Input Errors To further test the robustness of different methods, we add perturbations with different ratios to the source sentence of the test set in Figure 5. The input sentences are randomly corrupted with four types of perturbations including (1) deletion (drop words), (2) masking (replace words with "`[unk]`"), (3) swap (swap words), and (4) substitution (replace words with random words in the vocabulary). For the test set, we randomly perturb source sentences by a fixed corrupting probability.

Given the perturbed input sentence with different corrupting probabilities in Figure 5, the multilingual model [Johnson *et al.*, 2017] (green line) is easily influenced by the noisy input and degrades to the worst performance. It indicates the multilingual model delivers unstable translation performance for the zero-resource directions that are unseen at training time. The performance of the multilingual pivot-based method [Lakew *et al.*, 2019] (red line) also consistently drops more than our method due to error propagation introduced by the two-pass translation procedure. The results demonstrate that the multilingual student guided by multiple teachers performs better and avoids error propagation.

Number of Training Language Pairs Our student model is trained on the original parallel corpora D_B and the distilled training corpora D_S generated by multiple teachers described in Equation 3. The zero-resource translation ability of our student model benefits from the shared semantic space. In Table 4, "Supervised" denotes the multilingual model trained only with the original corpora of 16 directions, "Zero-resource" denotes the student multilingual model trained only with distilled corpora of 56 directions, and "Both" denotes our method trained on both the original corpora and distilled corpora. Our UM4 model trained jointly with 72 directions gets the best performance by transferring the knowledge among different languages.

4 Related Work

Zero-Resource NMT Zero-resource neural machine translation (NMT) is a challenging task since the source-target parallel corpus is not available. A feasible solution is the pivot-based NMT [Zhu *et al.*, 2013; Firat *et al.*, 2016; Cheng *et al.*, 2017; Zheng *et al.*, 2017; Currey and Heafield, 2019], where the source language is translated to the pivot language followed by translating the pivot language into the target language. This two-pass translation procedure both increases the complexity and potentially suffers from the error propagation problem because the errors made by the source→pivot model will be introduced to the pivot→target model [Lakew *et al.*, 2019]. Recent works [Chen *et al.*, 2017; Zheng *et al.*, 2017; Currey and Heafield, 2019] apply explorations into using the available parallel corpus and the additional monolingual corpus to improve zero-resource performance but limited by the bilingual setting.

Multilingual NMT Multilingual neural machine translation (MNMT) [Firat *et al.*, 2016; Johnson *et al.*, 2017; Lakew *et al.*, 2019; Tan *et al.*, 2019; Garcia *et al.*, 2020; Yang *et al.*, 2021] provides an alternative manner for zero-resource translation without any source-target parallel data but the performance is worse than pivot-based models. The multilingual models with language-aware module [Bapna and Firat, 2019; Zhang *et al.*, 2020; Philip *et al.*, 2020] are used to translate in zero-resource directions which are unseen at training time. However, the multilingual models often underperform the pivot-based models and deliver poor zero-resource translations. Multilingual pretraining method [Kim *et al.*, 2019a] are used to obtain the crosslingual encoder and then finetunes on the pseudo data. Inspired by previous works [Chen *et al.*, 2017; Zheng *et al.*, 2017], we employ multilingual multiple teachers to guide the multilingual source→target student to enhance the zero-resource translation.

5 Conclusion

In this paper, we propose a novel method called **U**nified **M**ultilingual **M**ultiple teacher-student **M**odel for N**MT** (**UM4**) to ameliorate the translation of zero-resource directions. Our method unifies the source-teacher model, target-teacher model, and pivot-teacher model to guide the multilingual source→target student model, alleviating the error propagation problem caused by two-pass translation. Experimental results on the multilingual dataset of the WMT benchmark corroborate the effectiveness of our method in leveraging the distilled knowledge from the unified teachers.

References

[Bapna and Firat, 2019] Ankur Bapna and Orhan Firat. Simple, scalable adaptation for neural machine translation. In *EMNLP 2019*, pages 1538–1548, 2019.

[Bertoldi *et al.*, 2008] Nicola Bertoldi, Madalina Barbaiani, Marcello Federico, and Roldano Cattoni. Phrase-based statistical machine translation with pivot languages. In *IWSLT 2008*, pages 143–149, 2008.

[Chen *et al.*, 2017] Yun Chen, Yang Liu, Yong Cheng, and Victor O. K. Li. A teacher-student framework for zero-resource neural machine translation. In *ACL 2017*, pages 1925–1935, 2017.

[Cheng *et al.*, 2017] Yong Cheng, Qian Yang, Yang Liu, Maosong Sun, and Wei Xu. Joint training for pivot-based neural machine translation. In *IJCAI 2017*, pages 3974–3980, 2017.

[Currey and Heafield, 2019] Anna Currey and Kenneth Heafield. Zero-resource neural machine translation with monolingual pivot data. In *EMNLP 2019*, pages 99–107, 2019.

[Firat *et al.*, 2016] Orhan Firat, Baskaran Sankaran, Yaser Al-Onaizan, Fatos T. Yarman-Vural, and Kyunghyun Cho. Zero-resource translation with multi-lingual neural machine translation. In *EMNLP 2016*, pages 268–277, 2016.

[Garcia *et al.*, 2020] Xavier Garcia, Pierre Foret, Thibault Sellam, and Ankur P. Parikh. A multilingual view of unsupervised machine translation. In *EMNLP 2020*, pages 3160–3170, 2020.

[Gheini *et al.*, 2021] Mozhdeh Gheini, Xiang Ren, and Jonathan May. Cross-attention is all you need: Adapting pretrained transformers for machine translation. In *EMNLP 2021*, pages 1754–1765, 2021.

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

[Kim and Rush, 2016] Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *EMNLP 2016*, pages 1317–1327, 2016.

[Kim *et al.*, 2019a] Yunsu Kim, Petre Petrov, Pavel Petrushkov, Shahram Khadivi, and Hermann Ney. Pivot-based transfer learning for neural machine translation between non-english languages. In *EMNLP 2019*, pages 866–876, 2019.

[Kim *et al.*, 2019b] Yunsu Kim, Petre Petrov, Pavel Petrushkov, Shahram Khadivi, and Hermann Ney. Pivot-based transfer learning for neural machine translation between non-english languages. In *EMNLP 2019*, pages 866–876, 2019.

[Kingma and Ba, 2014] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[Koehn *et al.*, 2019] Philipp Koehn, Francisco Guzmán, Vishrav Chaudhary, and Juan Pino. Findings of the WMT 2019 shared task on parallel corpus filtering for low-resource conditions. In *WMT 2019*, pages 54–72, 2019.

[Lakew *et al.*, 2019] Surafel Melaku Lakew, Marcello Federico, Matteo Negri, and Marco Turchi. Multilingual neural machine translation for zero-resource languages. *CoRR*, abs/1909.07342, 2019.

[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.

[Tan *et al.*, 2019] Xu Tan, Yi Ren, Di He, Tao Qin, Zhou Zhao, and Tie-Yan Liu. Multilingual neural machine translation with knowledge distillation. In *ICLR 2019*, 2019.

[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 6000–6010, 2017.

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

[Wu and Wang, 2009] Hua Wu and Haifeng Wang. Revisiting pivot language approach for machine translation. In *ACL 2009*, pages 154–162, 2009.

[Yang *et al.*, 2021] Yilin Yang, Akiko Eriguchi, Alexandre Muzio, Prasad Tadepalli, Stefan Lee, and Hany Hassan. Improving multilingual translation by representation and gradient regularization. In *EMNLP 2021*, pages 7266–7279, 2021.

[Zahabi *et al.*, 2013] Samira Tofighi Zahabi, Somayeh Bakhshaei, and Shahram Khadivi. Using context vectors in improving a machine translation system with bridge language. In *ACL 2013*, pages 318–322, 2013.

[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.

[Zheng *et al.*, 2017] Hao Zheng, Yong Cheng, and Yang Liu. Maximum expected likelihood estimation for zero-resource neural machine translation. In *IJCAI 2017*, pages 4251–4257, 2017.

[Zhou *et al.*, 2021] Chunting Zhou, Daniel Levy, Xian Li, Marjan Ghazvininejad, and Graham Neubig. Distributionally robust multilingual machine translation. In *EMNLP 2021*, pages 5664–5674, 2021.

[Zhu *et al.*, 2013] Xiaoning Zhu, Zhongjun He, Hua Wu, Haifeng Wang, Conghui Zhu, and Tiejun Zhao. Improving pivot-based statistical machine translation using random walk. In *EMNLP 2013*, pages 524–534, 2013.