Improving Multilingual Neural Machine Translation by Utilizing Semantic and Linguistic Features

Mengyu Bu[1,3], Shuhao Gu[1,3†], Yang Feng[1,2,3*]
[1]Key Laboratory of Intelligent Information Processing,
Institute of Computing Technology, Chinese Academy of Sciences (ICT/CAS)
[2] Key Laboratory of AI Safety, Chinese Academy of Sciences
[3] University of Chinese Academy of Sciences, Beijing, China
{bumengyu23z, fengyang}@ict.ac.cn, shuhaog515@gmail.com

Abstract

The many-to-many multilingual neural machine translation can be regarded as the process of integrating semantic features from the source sentences and linguistic features from the target sentences. To enhance zero-shot translation, models need to share knowledge across languages, which can be achieved through auxiliary tasks for learning a universal representation or cross-lingual mapping. To this end, we propose to exploit both semantic and linguistic features between multiple languages to enhance multilingual translation. On the encoder side, we introduce a disentangling learning task that aligns encoder representations by disentangling semantic and linguistic features, thus facilitating knowledge transfer while preserving complete information. On the decoder side, we leverage a linguistic encoder to integrate low-level linguistic features to assist in the target language generation. Experimental results on multilingual datasets demonstrate significant improvement in zero-shot translation compared to the baseline system, while maintaining performance in supervised translation. Further analysis validates the effectiveness of our method in leveraging both semantic and linguistic features.[1]

1 Introduction

The many-to-many multilingual neural machine translation (NMT) enables translation across multiple languages in a single model (Firat et al., 2016; Fan et al., 2021; Siddhant et al., 2020). This process involves integrating the semantic information from the source sentence and the linguistic features from the target sentence. Specifically, the encoder captures the semantic information of the source sentence and maps it to the representation space, when the decoder integrates this representation with the inherent features of the target language to generate the target language sentences. By sharing parameters, multilingual NMT models enable knowledge transfer across languages, which significantly benefits low-resource machine translation (Aharoni et al., 2019), especially zero-shot translation. However, previous multilingual NMT models do not explicitly differentiate between semantic and linguistic features, resulting in the entanglement of knowledge and linguistics within the model. This entanglement negatively affects the performance of zero-shot translation in two main aspects. Firstly, the linguistic features of the source language interfere with the encoder's ability to learn shared semantic information, causing the encoder to encode different languages into different representation subspaces (Kudugunta et al., 2019), which hinders knowledge transfer across multiple languages. Secondly, during the generation of the target language, the decoder is influenced by spurious correlations due to insufficient target language features, leading to off-target issues (Zhang et al., 2020).

Numerous methods have been proposed to address these challenges. Some methods focus on learning a universal cross-lingual representation space at the encoder (Arivazhagan et al., 2019; Pan et al., 2021; Pham et al., 2019; Gu and Feng, 2022), while others aim to facilitate target language generation at the decoder (Yang et al., 2021a; Gao et al., 2023). For example, Pan et al. (2021) employ contrastive learning to minimize the distance between parallel sentence pairs and maximize the distance between irrelevant sentence pairs. Gu and Feng (2022) leverage optimal transport theory to learn a universal representation and introduce an agreement-based training approach to make consistent predictions. Gao et al. (2023) promote consistent cross-lingual representation by cross-lingual consistency regularization. However, these methods primarily concentrate on learning

---

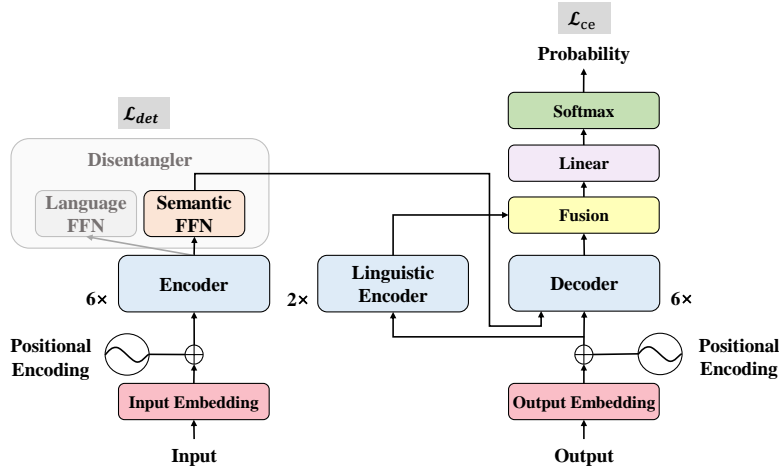[1]The code is available at https://github.com/ictnlp/SemLing-MNMT.

Figure 1: The framework of our method. We propose a disentangler to learn a universal semantic representation via disentangling and utilize a linguistic encoder to fuse low-level linguistic features.

language-agnostic semantic features while neglecting the language-specific linguistic features. Such incompleteness of information can lead to less accurate semantic mapping and alignment degradation, which can harm the supervised and zero-shot translation performance.

In this paper, we propose to utilize both semantic and linguistic features for multilingual NMT. Our insight is to learn a semantic representation space non-destructively and utilize linguistic features to guide target language generation, thereby improving the performance of zero-shot translation and maintaining supervised translation performance lossless. Specifically, at the encoder, we introduce a disentangler and disentangling learning task to facilitate harmless cross-lingual semantic alignment. At the decoder, we utilize a linguistic encoder to integrate low-level linguistic features from lower layers and high-level semantic features from higher layers. This fusion provides implicit guidance for the target language generation.

Experiments on multilingual datasets show that our method brings an average of 0.18+ BLEU in the supervised translation direction and an average of 3.74+ BLEU in the zero-shot translation direction compared with the baseline system. Furthermore, the analysis demonstrates that our method can alleviate the problems caused by the entanglement of semantic and linguistic features, obtain a better semantic representation, and reduce the off-target rate in zero-shot translation.

2 Background

In this section, we will give a brief introduction to the Transformer (Vaswani et al., 2017) model for NMT and many-to-many multilingual NMT.

2.1 Transformer for NMT

Define the set $\mathcal{D}$ of parallel corpus with input sentence $\mathbf{x}_i^u$ and reference $\mathbf{y}_i^v$, where u and v denote language type and i denotes the sentence id, which is the same for a sentence pair. The model first maps the source sentence $\mathbf{x}_i^u$ from the token representation to the token vector representation $embed(\mathbf{x}_i^u)$. After that, the encoder encodes the token vector sequence into the hidden state representation $\mathbf{h}(\mathbf{x}_i^u)$. Similarly, the model maps the predicted $k-1$ target words into a sequence of token vectors, and the decoder decodes the kth target word based on the sequence of hidden states $\mathbf{h}(\mathbf{x}_i^u)$ from the encoder and the predicted sequence of $k-1$ target words. The model repeats this process until the $\langle eos \rangle$ token is predicted. The model is optimized by minimizing the cross-entropy (CE) loss between the predicted sentence and the reference as follows:

$$\mathcal{L}_{ce} = - \sum_{\mathbf{x}_i^u, \mathbf{y}_i^v \in \mathcal{D}} \log(P_\theta(\mathbf{y}_i^v | \mathbf{x}_i^u)) \qquad (1)$$

where θ denotes the model parameters.

2.2 Many-to-Many Multilingual NMT

The many-to-many multilingual NMT model supports translation between multiple languages. For-
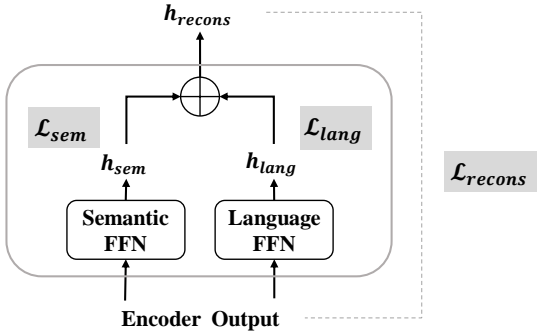
Figure 2: The architecture of disentangler. $\oplus$ denotes the summation of the feature dimensions.

mally, we define the set of languages $L = \{L_1, ..., L_M\}$, where M is the number of languages. We prepend a language token at the beginning of the source sentence and the target sentence to indicate the language type respectively. For example, the following sentence pair translated from English to French: "Hello world! $\rightarrow$ Bonjour le monde!" is transformed into "$\langle en \rangle$ Hello world! $\rightarrow$ $\langle fr \rangle$ Bonjour le monde!".

3 Method

The core idea of our approach is to improve multilingual NMT by exploiting semantic and linguistic features at the same time. To achieve this, we disentangle semantic and linguistic features for the encoder and utilize low-level linguistic features for the decoder. Specifically, we propose a disentangler at the encoder to facilitate lossless semantic alignment. Meanwhile, we introduce a linguistic encoder at the decoder to fuse low-level features to guide target language generation. Figure 1 illustrates the framework of our approach.

3.1 Disentangling Semantic and Linguistic Features

Empirically, sentences possess both semantic and linguistic properties, which can be expressed as "sentence = semantics + linguistics". Semantic features encompass the meanings of language units that are shared across languages, while linguistic features comprise the rules for constructing sentences, such as lexical labels and morphological information. By disentangling linguistic features from sentences, we can obtain a universal semantic representation shared by different languages. Referring to the design of Tiyajamorn et al. (2021) on multilingual language understanding tasks, we

extract semantic and linguistic features separately and implement interactive decomposition using reconstruction constraints.

The structure of our disentangler is shown in Figure 2. The disentangler consists of two feed-forward networks (FFN), a semantic FFN and a language FFN, respectively. For the encoder representations, the former extracts language-agnostic semantic information, and the latter extracts language-specific linguistic features. Then, the two outputs are summed to reconstruct the original encoder representations. The extracted semantic features are fed into the decoder.

We define disentangling loss $\mathcal{L}_{det}$ to optimize disentangler by multi-task learning:

$$\mathcal{L}_{det} = \mathcal{L}_{sem} + \mathcal{L}_{lang} + \lambda_1 \mathcal{L}_{recons} \qquad (2)$$

where $\mathcal{L}_{sem}$ assists the semantic FFN to extract semantic information, $\mathcal{L}_{lang}$ assists the language FFN to extract linguistic features, $\mathcal{L}_{recons}$ is the reconstruction constraint, and λ_1 is the hyperparameter to balance the losses.

Semantic Learning To learn a universal representation space across multiple languages, we explicitly minimize the distance of parallel sentence pairs and maximize the distance of irrelevant sentence pairs. For the multilingual parallel dataset $\mathcal{D}$, we make the parallel sentence pair $(\mathbf{x}_i^u, \mathbf{y}_i^v) \in \mathcal{D}$ a positive example and randomly choose a sentence $\mathbf{y}_j^w$ from language L_w as a negative example $(\mathbf{x}_i^u, \mathbf{y}_j^w)$, where L_u can be the same as L_w. We optimize by minimizing the following loss function:

$$\begin{aligned}\mathcal{L}_{sem} =& \mathbb{E}_{\mathbf{x}_i^u, \mathbf{y}_i^v \in \mathcal{D}}(1 - sim^+(\mathcal{R}(\mathbf{x}_i^u), \mathcal{R}(\mathbf{y}_i^v)) + \\ & \lambda_2 \mathbb{E}_{\mathbf{x}_i^u, \mathbf{y}_j^w \in \mathcal{D}}(1 + sim^-(\mathcal{R}(\mathbf{x}_i^u), \mathcal{R}(\mathbf{y}_j^w)))\end{aligned}$$
$$(3)$$

where $sim(\cdot)$ calculates the similarity of different sentences, and we use cosine similarity to evaluate the similarity. $+$ and $-$ denote positive and negative samples, respectively. $\mathcal{R}(\cdot)$ calculates the average-pooled hidden state representation. λ_2 is the hyperparameter that balances positive and negative samples. To simplify the implementation in the training, we sample the positive and negative samples within every training batch.

Language Learning Empirically, different languages have different linguistic features, such as lexical and syntactic patterns. Compared with sentences in different languages, the linguistic features are more similar between sentences in the same

10412
3

language. To extract linguistic features of multiple languages, we design the training objective similar to semantic learning. We minimize the distance of sentences in the same language and maximize the distance of sentences in different languages. We make the same language sentence pair $(\mathbf{x}_i^u, \mathbf{y}_j^u) \in \mathcal{D}$ a positive example and choose a random sentence $\mathbf{y}_k^v$ from any other language L_k as a negative example $(\mathbf{x}_i^u, \mathbf{y}_k^v)$. We optimize by minimizing the following loss function:

$$
\begin{aligned}
\mathcal{L}_{lang} =& \mathbb{E}_{\mathbf{x}_i^u, \mathbf{y}_j^u \in \mathcal{D}}(1 - sim^+(\mathcal{R}(\mathbf{x}_i^u), \mathcal{R}(\mathbf{y}_j^u)) + \\
& \lambda_2 \mathbb{E}_{\mathbf{x}_i^u, \mathbf{y}_k^v \in \mathcal{D}}(1 + sim^-(\mathcal{R}(\mathbf{x}_i^u), \mathcal{R}(\mathbf{y}_k^v)))
\end{aligned}
\tag{4}
$$

Where $sim(\cdot)$, $+$, $-$ and $\mathcal{R}(\cdot)$ are defined the same as semantic learning. λ_2 is a shared hyperparameter. We sample the positive and negative samples within every training batch.

Reconstruction Constraint To make the semantic and linguistic features fully interact and learn a more reasonable semantic representation, we design the reconstruction constraint to utilize the linguistic features. We input the encoded hidden state $\mathbf{h}(x_i^u)$ from the encoder into the disentangler, extract the semantic information $\mathbf{h}_{sem}(x_i^u)$ and the language features $\mathbf{h}_{lang}(x_i^u)$ respectively, and reconstruct the original encoder representation by summation. We perform optimization by minimizing the following loss function:

$$
\begin{aligned}
\mathcal{L}_{recons} =& \mathbb{E}_{\mathbf{x}_i^u \in \mathcal{D}}(\frac{1}{d}||\mathbf{h}(\mathbf{x}_i^u)) - \\
& \mathbf{h}_{sem}(\mathbf{x}_i^u) - \mathbf{h}_{lang}(\mathbf{x}_i^u)||_2)
\end{aligned}
\tag{5}
$$

Where d denotes the average sequence length.

3.2 Integrating Linguistic Features

In multilingual NMT models, decoders need to map the semantic representation to the target language representation. For zero-shot translation, since the model has not seen the translation direction during training, it is difficult to model the mapping relationships well, leading to the off-target phenomenon (Gu et al., 2019). Previous works show that the off-target rate is strongly correlated with the zero-shot translation performance (Zhang et al., 2020; Wu et al., 2021; Wang et al., 2022). Therefore, we hope to reduce the off-target rate by fusing additional information to guide the decoder to generate the target language.

Zhang et al. (2021) show that the lower encoder layers will focus more on linguistic information, which is crucial to distinguishing different languages. Therefore, we introduce a linguistic encoder with two encoder layers, which has the same input as the decoder. We utilize a fusion layer to integrate the low-level representations containing more linguistic features and the high-level representations containing more semantic features. We denote the linguistic encoder output as $\mathbf{h}_{lingEnc}$ and the decoder output as $\mathbf{h}_{dec}$. We concatenate these features in the feature dimension and utilize a two-layer FFN for fusion:

$$
\mathbf{h} = \mathbf{W}_1(\text{ReLU}(\mathbf{W}_2([\mathbf{h}_{dec}; \mathbf{h}_{lingEnc}])))
\tag{6}
$$

Where $\mathbf{W}_1 \in \mathbb{R}^{d \times d_h}$ and $\mathbf{W}_2 \in \mathbb{R}^{2d_h \times d}$ are two feed-forward layers.

3.3 Joint Training Strategy

We optimize the final loss function using a joint training strategy, including cross-entropy loss and disentangling loss:

$$
\mathcal{L} = \mathcal{L}_{ce} + \lambda d \mathcal{L}_{det}
\tag{7}
$$

where λ is the hyperparameter that controls the contribution of $\mathcal{L}_{det}$. Since $\mathcal{L}_{ce}$ is calculated at the token level and $\mathcal{L}_{det}$ is calculated at the sentence level, we multiply $\mathcal{L}_{det}$ by the average sequence length d.

4 Experiment

We evaluate the performance of our model on the multilingual benchmarks, including supervised and zero-shot translation performance.

4.1 Dataset Description

We conduct experiments on the IWSLT2017, OPUS-7, and PC-6 datasets. The brief information about these datasets is shown in Appendix A.

IWSLT2017 The IWSLT2017 benchmark (Cettolo et al., 2017) contains five languages: German (De), English (En), Dutch (Nl), Romanian (Ro), and Italian (It). We experiment on the English-centric dataset consisting of eight supervised translation directions and twelve zero-shot translation directions. The training set contains 0.22M∼0.26M parallel sentence pairs per language direction. We use the official validation set and test set.

OPUS-7 The OPUS-7 dataset is a subset of the OPUS-100 dataset (Pan et al., 2021). The OPUS-7 dataset is English-centric and contains seven languages: Arabic (Ar), German (De), English (En),

French (Fr), Dutch (Nl), Russian (Ru), and Chinese (Zh). This dataset contains twelve supervised translation directions and thirty zero-shot translation directions. The training set contains 1M parallel sentence pairs per language direction. We use the official validation set and test set.

PC-6 We construct the PC-6 dataset following (Gu and Feng, 2022). PC-6 is a dataset of six languages extracted from PC32 (Pan et al., 2021), including Czech (Cs), Kazakh (Kk), Romanian (Ro), Russian (Ru), and Turkish (Tr). The PC-6 dataset contains ten supervised translation directions and twenty zero-shot translation directions. The amount of data in each language direction varies in size from 0.12M to 1.84M sentence pairs. For the supervised translation, we use the official WMT2016~2019 validation set and test set. For the zero-shot translation, we extract from WikiMatrix (Schwenk et al., 2021) to obtain the test set.

We directly apply the Unigram Model algorithm to preprocess the original multilingual corpus using *Sentencepiece* toolkit (Kudo and Richardson, 2018). We build a shared dictionary across multiple languages, which contains 32K tokens.

4.2 Model Configuration

We use the base Transformer configuration (Vaswani et al., 2017) for both the encoder and decoder. For the disentangler and the fusion layer, we set $ffn\_dim = 2048$. For the linguistic encoder, we apply two encoder layers of the base Transformer configuration. We use the Adam optimizer and set $\beta = (0.9, 0.98)$. We set $warmup = 4000$ and use the inverse square root learning scheduler with the learning rate set to $7e^{-4}$. We apply label-smoothed cross-entropy loss with a smoothing rate of 0.1. We set $dropout = 0.3$ for the IWSLT2017 dataset and $dropout = 0.1$ for the OPUS-7 and PC-6. We set $\lambda = 0.05$, $\lambda_1 = 0.2$ and $\lambda_2 = 0.2$ for our system. We train the models on 4 RTX3090 GPUs.

In the inference stage, we use beam search with beam size 5 and set the length penalty to 1. We report the BLEU (Papineni et al., 2002), ChrF (Popović, 2015) and COMET (Rei et al., 2022)[2] scores for comprehensive evaluation. The BLEU and ChrF scores are computed using the *SacreBLEU* toolkit (Post, 2018). We select the best-performing three checkpoints on the valida-

tion sets from the last six checkpoints and report the average scores on the test sets.

4.3 Contrast Systems

We compare our model with the following multilingual NMT systems. We prepend a language token at the beginning of the sentences for NMT systems.

m-Transformer (Johnson et al., 2017) This method utilizes cross-entropy loss to train a Transformer model on multilingual datasets.

Pivot Translation (Cheng et al., 2017) This method transforms zero-shot translation into English-centric two-stage translation. The translation model is m-Transformer.

Contrastive Learning (mRASP2 w/o AA) (Pan et al., 2021) This method introduces contrastive learning to learn a universal representation space for the encoder. For a fair comparison, we do not use the aligned augmentation (AA).

Target Language Prediction (TLP) (Yang et al., 2021a) This method leverages a target language prediction task to help the decoder retain information about the target language.

Denoising Training (DT) (Wang et al., 2021) This method introduces a denoising auto-encoder task during training. Specifically, they utilize all English sentences to construct the denoising corpus via text infilling operation.

CrossConST (Gao et al., 2023) This approach adds KL divergence loss to the Softmax layer to promote consistent cross-lingual representation:

$$\mathcal{L}_{kl} = \gamma \mathbf{KL}(P_\theta(\mathbf{x}, \mathbf{y}) || P_\theta(\mathbf{y}, \mathbf{y})) \quad (8)$$

Where γ is the hyperparameter controlling the proportion of $\mathcal{L}_{kl}$. We set $\gamma = 0.25$ on IWSLT2017, $\gamma = 0.06$ on OPUS-7 and $\gamma = 0.1$ on PC-6. We jointly train $\mathcal{L}_{ce}$ and $\mathcal{L}_{kl}$.

Alpaca (Taori et al., 2023) Alpaca is fine-tuned by Stanford from LLaMA-7B (Touvron et al., 2023) with 52k instruction data. We reproduce Alpaca using the LoRA (Hu et al., 2022) setting.

ChatGPT We use the GPT-3.5-Turbo API[3] for translation experiments. Our translation prompt for Alpaca and ChatGPT is reported in Appendix B.

---

[2] https://huggingface.co/Unbabel/wmt22-comet-da

[3] The API version is GPT-3.5-Turbo-0613.

IWSLT2017	De ↔ It	De ↔ Nl	De ↔ Ro	It ↔ Nl	It ↔ Ro	Nl ↔ Ro	Zero-shot Average	Supervised Average
m-Transformer	16.94	20.89	17.02	17.58	17.32	17.83	17.93	**34.32**
mRASP2 w/o AA	21.39	24.67	21.40	22.43	**22.70**	21.87	22.41	34.06
TLP	19.52	23.41	19.08	20.75	20.34	20.29	20.56	34.29
DT	18.76	22.05	18.65	20.04	20.31	19.83	19.94	34.17
CrossConST	20.93	24.10	20.65	21.98	21.68	21.40	21.73	34.10
Pivot	21.80	24.14	21.23	22.63	22.19	22.27	22.37	-
Alpaca	14.50	17.94	12.88	14.46	13.44	12.18	14.23	27.37
ChatGPT	23.85	27.01	23.21	25.17	24.72	24.83	24.79	35.26
Ours	**21.55**	**25.18**	**21.50**	**22.84**	22.66	**22.46**	**22.70**	34.25

OPUS-7	x → Ar	x → De	x → Fr	x → Nl	x → Ru	x → Zh	Zero-shot Average	Supervised Average
m-Transformer	12.73	14.58	23.11	15.82	17.24	16.44	16.65	31.91
mRASP2 w/o AA	14.20	16.54	25.67	**17.46**	19.29	17.72	18.48	31.48
TLP	13.10	14.32	22.71	15.82	17.29	16.29	16.59	31.90
DT	13.84	15.06	23.95	16.24	18.29	**17.85**	17.54	31.96
CrossConST	13.68	15.04	23.67	16.50	17.67	16.03	17.10	32.12
Pivot	15.79	17.48	27.81	18.50	20.98	15.71	19.38	-
Ours	**14.79**	**16.64**	**26.20**	17.42	**20.11**	17.40	**18.76**	**32.13**

PC-6	x → Cs	x → Kk	x → Ro	x → Ru	x → Tr	Zero-shot Average	Supervised Average
m-Transformer	8.05	4.27	11.05	10.91	5.45	7.94	20.38
mRASP2 w/o AA	**14.27**	2.85	15.67	16.12	8.22	11.42	19.93
TLP	9.67	4.27	11.35	10.79	6.17	8.45	20.68
DT	12.57	**4.85**	14.32	15.63	7.78	11.03	20.22
CrossConST	13.03	2.29	14.56	13.41	7.41	10.14	20.50
Pivot	13.73	2.13	15.91	15.20	8.30	11.05	-
Ours	14.13	4.20	**16.92**	**16.72**	**9.44**	**12.28**	**20.77**

Table 1: Overall performance on the multilingual test sets. We report the BLEU scores. "Zero-shot Average" and "Supervised Average" denote the average BLEU scores on the zero-shot and supervised directions. The "x" in the last two tables denotes all languages except for the target language. We bold the highest BLEU scores except for Alpaca, ChatGPT and Pivot.

4.4 Main Results

The results of the main experiment are shown in Table 1. For OPUS-7 and PC-6, we report the average BLEU scores for the same target language for convenience, and detailed results for the zero-shot translation directions are shown in Appendix C. The ChrF and COMET results are presented in Appendix D.

The results show that our system significantly improves zero-shot translation performance and maintains supervised translation performance. Specifically, in the supervised translation direction, the performance of our system is comparable to that of m-Transformer, which suggests that our disentangling approach mitigates alignment degradation and maintains the supervised translation perfor-

mance. In the zero-shot translation direction, our system significantly outperforms all the contrast models, indicating that our additional modeling of linguistic features can assist the model in learning better semantic representation and language generation. Moreover, our method even outperforms the Pivot system on IWSLT2017 and PC-6, which we believe is probably because Pivot accumulates too many errors on these two datasets.

The results of large language models (LLMs) indicate that our task-specific model still has advantages in the era of LLMs. Our model significantly outperforms Alpaca-7B in supervised and zero-shot translation. Compared to the ChatGPT, our model achieves more than 90% performance with just about 0.1% of the parameters. Furthermore, our model employs significantly less data

Model	Disentangler	Disentangling Objective	Linguistic Encoder	Zero-shot Average	Supervised Average
m-Transformer	✗	✗	✗	17.93	34.32
m-Transformer[†]	✗	✗	✗	17.03	34.33
①	✔	✗	✗	19.88	34.39
②	✔	✔	✗	21.86	**34.45**
③	✗	✗	✔	21.27	34.34
④	✔	✗	✔	21.92	34.23
⑤	✔	✔	✔	**22.70**	34.25

Table 2: The results of the ablation study. The markers ✔ and ✗ indicate the component is involved or not involved, respectively. The m-Transformer[†] has eight encoder layers and eight decoder layers, similar to our model size.

and inference is much faster.

5 Analysis

In this section, we will analyze what contributes to the performance. We first analyze the effects of each module via an ablation study. Then, we evaluate the off-target rate and visualize the encoder and decoder sentence representations, demonstrating that our approach learns universal representations and generates target language more accurately. Finally, we examine some translation cases to show the usefulness of our approach.

5.1 Ablation Study

To further analyze the effectiveness of our approach, we conducted experiments on different variants of our system on the IWSLT2017. We briefly report the experimental results in Table 2.

Compared with the m-Transformer, ③ can significantly improve the zero-shot translation. The gap in zero-shot translation performance remains even when scaling up the m-Transformer (m-Transformer[†]). This indicates that it is effective to leverage the linguistic encoder to learn low-level linguistic features and integrate different levels of representations. We present detailed results and analysis of the scaling experiments in Appendix E for parameter comparisons.

Compared with ①, ② provides a significant improvement in zero-shot translation while maintaining supervised translation performance. This indicates that the disentangling learning task is effective. Disentangling learning can significantly improve the performance of zero-shot translation while mitigating the performance loss in supervised translation caused by semantic alignment. The analysis between ④ and ⑤ similarly supports this view.

Compared with ② and ③, ⑤ shows that the dis-

	IWSLT2017	OPUS-7	PC-6
m-Transformer	6.76%	11.58%	21.02%
mRASP2 w/o AA	2.84%	8.91%	8.68%
CrossConST	3.07%	11.71%	7.94%
Ours	**2.46%**	**8.64%**	**7.25%**

Table 3: The average off-target rate for the test sets.

	IWSLT2017	OPUS-7	PC-6
m-Transformer	18.39	15.76	7.91
mRASP2 w/o AA	22.42	17.31	11.48
Ours	**22.69**	**17.57**	**11.82**

Table 4: The average in-target BLEU scores for the test sets.

entangler and the linguistic encoder can collaborate to improve the zero-shot translation performance.

5.2 Off-target Issue

The many-to-many multilingual NMT models face the off-target issue (Gu et al., 2019) in zero-shot translation, which seriously impacts the zero-shot translation performance. The off-target issue refers to the problem that the model simply copies the source sentences or generates incorrect target sentences. Related work shows that multilingual NMT models have difficulty establishing mapping relations for the zero-shot translation directions, which may be overfitted to other language directions. To test whether our approach can alleviate the off-target issue, we use *langid* toolkit (Lui and Baldwin, 2012) to identify the target language and define the off-target rate as the proportion of off-target sentences. We conduct experiments on the contrast systems and our method. Table 3 shows that our method can effectively reduce the off-target rate compared to other systems, proving that our method can provide an effective guide for target language generation.

(a) m-Transformer Encoder   (b) mRASP2 w/o AA Encoder   (c) Semantic FFN   (d) Language FFN

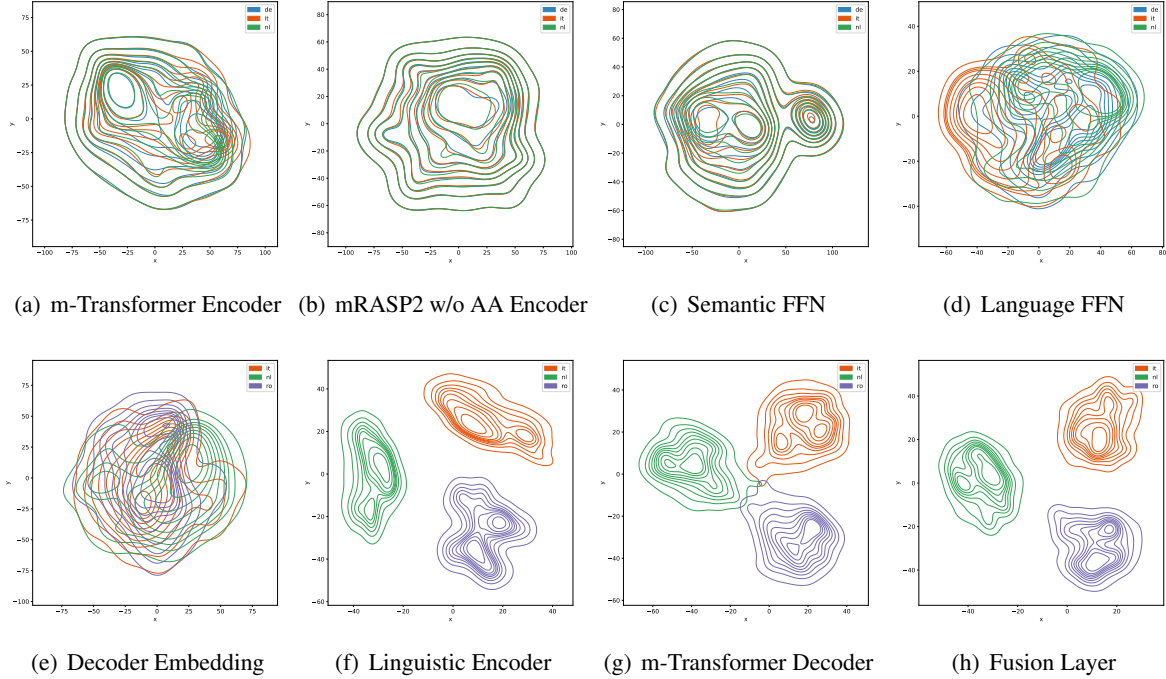(e) Decoder Embedding   (f) Linguistic Encoder   (g) m-Transformer Decoder   (h) Fusion Layer

Figure 3: Visualization of the m-Transformer, mRASP2 w/o AA and our system after dimension reduction. The subfigure captions describe the model modules, and we do dimension reduction on the outputs of these modules. The blue line denotes German (De), the orange line denotes Italian (It), the green line denotes Dutch (Nl), and the purple line denotes Romanian (Ro).

Furthermore, we test the performance of in-target sentences to demonstrate that our approach improves the in-target translation quality. Considering that there is no comparability between the BLEU scores of different test sets, we extract the part where all model outputs are in-target for testing. Table 4 presents the results. Compared with Table 1, the performance gap between different models in the in-target part is smaller, which indicates the importance of mitigating the off-target issue. Meanwhile, the zero-shot performance of our models in the in-target part still significantly outperforms m-Transformer, which implies that our approach achieves better knowledge transfer and improves the zero-shot translation quality.

5.3 Visualization

We visualize the sentence representations of the encoder and decoder to verify, respectively, that our disentangler aids semantic alignment and our linguistic encoder achieves more accurate target language generation. We conduct experiments on a subset of IWSLT2017 test sets. We use t-SNE (Van der Maaten and Hinton, 2008) to reduce the 512-dim representations to 2-dim.

Figures 3(a) - 3(d): The m-Transformer cannot

align different languages well, which is caused by the entanglement of semantic and linguistic features at the representation level. For semantic features, our semantic FFN pulls the semantic representation closer, achieving a similar effect as mRASP2 w/o AA. For linguistic features, our language FFN pushs the different language representations apart, which proves that the language FFN can extract linguistic features.

Figures 3(e) - 3(h): The representation after decoder embedding is language-mixed, while the linguistic encoder can encode different languages into different subspaces, even though no explicit task is designed to guide this. By introducing low-level linguistic features, our system can better distinguish the target language and thus generate the target sentences more accurately.

5.4 Case Study

To further demonstrate the usefulness of our system, we analyze some translation cases and compare the outputs of our model with that of the m-Transformer in Figure 4. While the m-Transformer translates whole or partial sentences to the wrong language, our model can translate accurately, proving the advantages of our method.

Direction: Ro-It
Input: Discută despre ea, își imaginează cum va arăta, se luptă pentru putere, apoi își petrec ceva timp planificând, organizând.
Reference: Ne parlano, si immaginano che aspetto potrebbe avere, competono per il potere, poi passano il tempo pianificando, organizzando.
Output (m-Transformer): Discussionate, immaginate what it's going to look like, combattimento per il potere, e poi spend some time pianificating, organizzando. (Mid-range off-target)
Output (Ours): Parlando di essa, immagina come sarà, lotta per il potere, e poi trascorrono un po' di tempo pianificando, organizzando.
Direction: Ro-Nl
Input: Vă mulțumesc foarte mult.
Reference: Hartelijk bedankt.
Output (m-Transformer): Thank you very much. (All off-target)
Output (Ours): Heel erg bedankt.
Direction: Nl-Ro
Input: Ongelooflijk. Je ziet het team en denkt: "die gaan winnen."
Reference: E incredibil. Vedeți, vă uitați în jur și spuneți: "Asta e echipa câștigatoare".
Output (m-Transformer): Uimitor. You see echipa și you think, "They're going to win." (Final off-target)
Output (Ours): Incredibil. Vezi echipa și te gândești, "Acestea vor câștiga."
Direction: It-Nl
Input: e costruiscono prototipi, uno dopo l'altro,
Reference: Ze bouwen prototypes, en nieuwe prototypes...
Output (m-Transformer): Ze bouwen prototypes, one by one. (Final off-target)
Output (Ours): Ze bouwen prototypes, één na de andere.
Direction: Fr-Zh
Input: Beth, s'il te plaît. Tu n'as pas à faire ça.
Reference: 贝丝 求你了 你没必要这么做
Output (m-Transformer): Beth, please, you don't have to. (All off-target)
Output (Ours): 贝丝,拜托 你不必这么做

Figure 4: Case study of zero-shot translation. We identify three off-target types which are categorized by off-target position and ratio.

6 Related Work

6.1 Zero-shot Machine Translation

To improve the zero-shot translation capability of multilingual NMT, researchers have proposed a series of auxiliary training objectives. These works can be divided into two categories. The first category is designed directly at the encoder so that the encoder outputs language-agnostic representations (Pham et al., 2019; Wei et al., 2021). These works are motivated by the hypothesis that a universal semantic representation can realize better knowledge transfer. Pan et al. (2021) propose a contrastive learning task to encourage the model to learn a universal representation space. Gu and Feng (2022) use optimal transport theory to close the representation of sentence pairs. However, these works focus only on semantic features and neglect linguistic features, which leads to performance degradation in the supervised translation direction. By integrating both semantic and linguistic features, our approach significantly improves the performance of zero-shot translation while effectively maintaining the supervised direction performance.

The second category is designed at the decoder to provide explicit guidance for target language generation. These works are motivated to improve multilingual representation. Yang et al. (2021a) propose the target language prediction task at the sequence level and the target gradient regularization at the gradient level. Gao et al. (2023) introduce the KL regularization term for the Softmax layer to promote cross-lingual consistent representation. In contrast to these approaches, we do not explicitly design auxiliary tasks, but utilize a linguistic encoder to fuse low-level representations containing more linguistic features and high-level representations containing more semantic features.

6.2 Disentangling for Cross-lingual Alignment

Cross-lingual alignment is a classic problem in multilingual tasks. Although the "pretraining-finetuning" paradigm can improve the cross-lingual ability of the model (Liu et al., 2020; Xue et al., 2021), the entangling of semantics and linguistics can impair the model performance. To address this problem, many researchers have improved it from the perspective of disentangling. Tiyajamorn et al. (2021) design auto-encoder to disentangle semantic and linguistic representation for natural language understanding tasks. Yang et al. (2021b) remove language identity information from pre-trained multilingual representations by singular value decomposition (SVD) and orthogonal projection. Zhao et al. (2021) explore three settings for removing language identity signals from multilingual representations, including vector space re-alignment, vector space normalization and input normalization. Wu et al. (2022) use SVD to extract linguistic features, use language-agnostic semantic information for task-specific training, and re-entangle semantic and linguistic features in the generation phase. We refer to this disentangling idea and utilize disentangling to achieve lossless semantic alignment for multilingual NMT.

7 Conclusion

In this paper, we propose to exploit both semantic and linguistic features to improve multilingual NMT. We design a disentangler and disentangling learning task at the encoder side to achieve lossless semantic alignment. Meanwhile, we introduce a linguistic encoder at the decoder side to fuse the low-level representation containing more linguistic features and the high-level representation containing more semantic features to improve the target language generation. Experiments on multilingual datasets indicate that our approach can significantly improve the performance of zero-shot translation while effectively maintaining the supervised translation performance. The analysis further demonstrates that utilizing both semantic and linguistic features can help the model learn semantic representation non-destructively and achieve more accurate target language generation.

Limitations

Although our approach can significantly improve zero-shot translation performance while maintaining supervised translation performance, the performance improvement in the supervised direction is limited. On the one hand, the datasets of our experiments contain only several languages, and the baseline model can already model the supervised translation well. On the other hand, our approach to modeling linguistic features is still preliminary and implicit. Future work could consider using larger multilingual datasets as well as designing better methods to model linguistic features.

Acknowledgments

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Roee Aharoni, Melvin Johnson, and Wolfgang Macherey. 2019. The missing ingredient in zero-shot neural machine translation. *arXiv preprint arXiv:1903.07091*.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Niehues Jan, Stüker Sebastian, Sudoh Katsuitho, Yoshino Koichiro, and Federmann Christian. 2017. Overview of the iwslt 2017 evaluation campaign. In *Proceedings of the 14th International Workshop on Spoken Language Translation*, pages 2–14.

Yong Cheng, Qian Yang, Yang Liu, Maosong Sun, and Wei Xu. 2017. Joint training for pivot-based neural machine translation. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 3974–3980. AAAI Press.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *The Journal of Machine Learning Research*, 22(1):4839–4886.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875, San Diego, California. Association for Computational Linguistics.

Pengzhi Gao, Liwen Zhang, Zhongjun He, Hua Wu, and Haifeng Wang. 2023. Improving zero-shot multilingual neural machine translation by leveraging cross-lingual consistency regularization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12103–12119, Toronto, Canada. Association for Computational Linguistics.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor O.K. Li. 2019. Improved zero-shot neural machine translation via ignoring spurious correlations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1258–1268, Florence, Italy. Association for Computational Linguistics.

Shuhao Gu and Yang Feng. 2022. Improving zero-shot multilingual translation with universal representations and cross-mapping. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6492–6504, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

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. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Trans. Assoc. Comput. Linguistics*, 5:339–351.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*, pages 66–71.

Sneha Kudugunta, Ankur Bapna, Isaac Caswell, and Orhan Firat. 2019. Investigating multilingual NMT representations at scale. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1565–1575, Hong Kong, China. Association for Computational Linguistics.

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

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 244–258, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318.

Ngoc-Quan Pham, Jan Niehues, Thanh-Le Ha, and Alexander Waibel. 2019. Improving zero-shot translation with language-independent constraints. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 13–23, Florence, Italy. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Aditya Siddhant, Ankur Bapna, Yuan Cao, Orhan Firat, Mia Chen, Sneha Kudugunta, Naveen Arivazhagan, and Yonghui Wu. 2020. Leveraging monolingual data with self-supervision for multilingual neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2827–2835, Online. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Nattapong Tiyajamorn, Tomoyuki Kajiwara, Yuki Arase, and Makoto Onizuka. 2021. Language-agnostic representation from multilingual sentence encoders for cross-lingual similarity estimation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7764–7774, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

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

Weizhi Wang, Zhirui Zhang, Yichao Du, Boxing Chen, Jun Xie, and Weihua Luo. 2021. Rethinking zero-shot neural machine translation: From a perspective of latent variables. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4321–4327, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wenxuan Wang, Wenxiang Jiao, Shuo Wang, Zhaopeng Tu, and Michael R Lyu. 2022. Understanding and mitigating the uncertainty in zero-shot translation. *arXiv preprint arXiv:2205.10068*.

Xiangpeng Wei, Rongxiang Weng, Yue Hu, Luxi Xing, Heng Yu, and Weihua Luo. 2021. On learning universal representations across languages. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Liwei Wu, Shanbo Cheng, Mingxuan Wang, and Lei Li. 2021. Language tags matter for zero-shot neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3001–3007, Online. Association for Computational Linguistics.

Xianze Wu, Zaixiang Zheng, Hao Zhou, and Yong Yu. 2022. Laft: Cross-lingual transfer for text generation by language-agnostic finetuning. In *Proceedings of*

the 15th International Conference on Natural Language Generation, pages 260–266.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Yilin Yang, Akiko Eriguchi, Alexandre Muzio, Prasad Tadepalli, Stefan Lee, and Hany Hassan. 2021a. Improving multilingual translation by representation and gradient regularization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7266–7279, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ziyi Yang, Yinfei Yang, Daniel Cer, and Eric Darve. 2021b. A simple and effective method to eliminate the self language bias in multilingual representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5825–5832, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Biao Zhang, Ankur Bapna, Rico Sennrich, and Orhan Firat. 2021. Share or not? learning to schedule language-specific capacity for multilingual translation. In *Ninth International Conference on Learning Representations 2021*.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Wei Zhao, Steffen Eger, Johannes Bjerva, and Isabelle Augenstein. 2021. Inducing language-agnostic multilingual representations. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 229–240, Online. Association for Computational Linguistics.

A Statistics of the Datasets

Table 5 summarizes information about the multilingual training datasets.

Dataset	Language Pairs	Size
IWSLT2017	En ↔ {De, It, Nl, Ro}	1.8M
OPUS-7	En ↔ {Ar, De, Fr, Nl, Ru, Zh}	12.0M
PC-6	En ↔ {Cs, Kk, Ro, Ru, Tr}	7.9M

Table 5: A brief description of the datasets.

B Translation Prompt for Alpaca and ChatGPT

For Alpaca, we follow the official template. We set *Instruction* to *"Translate input from <src> to <tgt>."* and set *Input* to the source sentence. For ChatGPT, we use the following format to structure the inputs: *"Translate the sentences from <src> to <tgt>: <input>."*. *<src>* denotes the source language, *<tgt>* denotes the target language and *<input>* denotes the source sentence.

C Detailed Results on the OPUS-7 and PC-6 Test Sets

The detailed results on the OPUS-7 and PC-6 test sets are shown in Table 6 and Table 7 respectively.

D ChrF and COMET Metrics

For the ChrF and COMET metrics, we evaluate the m-Transformer, mRASP2 w/o AA and our model. The results are presented in Table 8 and Table 9 respectively.

E Scaling Experiments on the IWSLT2017 Test Set

Considering that our approach introduces additional parameters, we scale the model sizes of the m-Transformer and mRASP2 w/o AA for a fair comparison. As shown in Table 10, our model outperforms these variant models. We note a significant performance degradation of the m-Transformer when the model size reaches 100.4M. We attribute this to the limited size of the IWSLT2017 training set, thus the model performance largely depends on the method, rather than the model size. This further demonstrates the effectiveness of our method.

m-Transformer	Ar	De	Fr	Nl	Ru	Zh	x (avg)
Ar	-	12.13	24.96	14.33	21.31	27.12	19.97
De	5.74	-	19.42	21.15	13.26	5.55	13.02
Fr	16.80	17.97	-	21.57	21.77	22.46	20.11
Nl	4.88	20.02	22.60	-	10.69	2.78	12.19
Ru	18.72	13.58	25.36	13.52	-	24.30	19.09
Zh	17.50	9.23	23.22	8.51	19.18	-	15.53
x (avg)	12.73	14.58	23.11	15.82	17.24	16.44	16.65

Pivot	Ar	De	Fr	Nl	Ru	Zh	x (avg)
Ar	-	15.67	29.81	18.16	26.00	22.67	22.46
De	7.30	-	23.07	23.16	16.02	5.13	14.94
Fr	20.32	20.90	-	23.80	25.01	25.20	23.05
Nl	6.55	21.40	25.47	-	13.40	3.34	14.03
Ru	22.56	16.72	30.56	16.27	-	22.21	21.66
Zh	22.23	12.70	30.13	11.10	24.47	-	20.13
x (avg)	15.79	17.48	27.81	18.50	20.98	15.71	19.38

mRASP2 w/o AA	Ar	De	Fr	Nl	Ru	Zh	x (avg)
Ar	-	14.74	27.06	15.71	22.86	26.05	21.28
De	6.48	-	22.09	23.06	14.91	6.82	14.67
Fr	19.00	20.01	-	23.85	23.89	26.30	22.61
Nl	6.08	21.75	24.86	-	12.83	3.90	13.89
Ru	20.02	15.51	28.10	14.84	-	25.52	20.80
Zh	19.43	10.67	26.22	9.82	21.97	-	17.62
x (avg)	14.20	16.54	25.67	17.46	19.29	17.72	18.48

TLP	Ar	De	Fr	Nl	Ru	Zh	x (avg)
Ar	-	11.59	23.91	13.91	20.91	25.03	19.07
De	6.24	-	19.72	21.40	13.29	4.60	13.05
Fr	17.52	17.47	-	21.64	22.22	24.81	20.73
Nl	4.93	19.87	22.47	-	10.53	2.96	12.15
Ru	19.18	13.60	24.90	13.52	-	24.05	19.05
Zh	17.65	9.08	22.56	8.65	19.52	-	15.49
x (avg)	13.10	14.32	22.71	15.82	17.29	16.29	16.59

DT	Ar	De	Fr	Nl	Ru	Zh	x (avg)
Ar	-	12.80	25.81	14.77	22.27	28.52	20.83
De	6.76	-	20.62	22.01	13.88	6.01	13.85
Fr	18.08	18.11	-	21.98	23.08	25.47	21.34
Nl	5.22	20.83	23.31	-	11.52	3.33	12.84
Ru	20.13	14.06	26.21	13.78	-	25.93	20.02
Zh	19.00	9.47	23.79	8.68	20.71	-	16.33
x (avg)	13.84	15.06	23.95	16.24	18.29	17.85	17.54

CrossConST	Ar	De	Fr	Nl	Ru	Zh	x (avg)
Ar	-	13.02	25.88	14.95	21.70	25.48	20.21
De	6.30	-	19.68	21.94	13.74	5.09	13.35
Fr	18.09	18.22	-	22.56	21.91	23.17	20.79
Nl	5.26	20.45	22.96	-	11.28	2.51	12.49
Ru	19.76	14.18	26.17	14.12	-	23.89	19.62
Zh	19.01	9.35	23.67	8.92	19.70	-	16.13
x (avg)	13.68	15.04	23.67	16.50	17.67	16.03	17.10

Ours	Ar	De	Fr	Nl	Ru	Zh	x (avg)
Ar	-	14.60	27.78	15.41	24.50	26.49	21.75
De	7.04	-	22.31	23.04	14.84	6.37	14.72
Fr	19.89	20.00	-	23.90	24.92	25.10	22.76
Nl	6.06	21.61	24.87	-	12.81	3.58	13.79
Ru	20.84	15.85	28.66	14.91	-	25.46	21.15
Zh	20.11	11.14	27.38	9.83	23.48	-	18.39
x (avg)	14.79	16.64	26.20	17.42	20.11	17.40	18.76

Table 6: The performance of the contrast models and our model on the OPUS-7 dataset.

m-Transformer	Cs	Kk	Ro	Ru	Tr	x (avg)
Cs	-	1.54	14.63	12.93	6.16	8.81
Kk	1.73	-	2.61	10.24	2.99	4.40
Ro	11.11	1.87	-	12.98	7.25	8.31
Ru	10.53	11.10	14.76	-	5.40	10.45
Tr	8.80	2.57	12.21	7.47	-	7.76
x (avg)	8.05	4.27	11.05	10.91	5.45	7.94

Pivot	Cs	Kk	Ro	Ru	Tr	x (avg)
Cs	-	1.01	21.4	20.08	10.19	13.17
Kk	1.49	-	2.09	5.07	2.29	2.74
Ro	21.1	1.32	-	24.1	11.43	14.49
Ru	19.08	4.99	23.42	-	9.27	14.19
Tr	13.23	1.19	16.74	11.54	-	10.68
x (avg)	13.73	2.13	15.91	15.20	8.30	11.05

mRASP2 w/o AA	Cs	Kk	Ro	Ru	Tr	x (avg)
Cs	-	1.11	20.70	22.09	9.94	13.46
Kk	2.01	-	2.35	7.65	2.33	3.59
Ro	20.32	1.60	-	24.42	11.94	14.57
Ru	21.88	6.53	23.51	-	8.65	15.15
Tr	12.86	2.16	16.09	10.31	-	10.36
x (avg)	14.27	2.85	15.67	16.12	8.22	11.42

TLP	Cs	Kk	Ro	Ru	Tr	x (avg)
Cs	-	1.37	14.50	12.75	7.14	8.94
Kk	1.96	-	2.53	10.92	3.06	4.62
Ro	13.36	2.07	-	12.25	7.88	8.89
Ru	13.71	10.83	15.96	-	6.58	11.77
Tr	9.66	2.82	12.41	7.25	-	8.03
x (avg)	9.67	4.27	11.35	10.79	6.17	8.45

DT	Cs	Kk	Ro	Ru	Tr	x (avg)
Cs	-	1.53	19.03	18.83	8.21	11.90
Kk	2.13	-	2.81	12.43	3.58	5.24
Ro	17.34	2.27	-	20.63	10.99	12.81
Ru	18.75	12.23	19.51	-	8.35	14.71
Tr	12.06	3.35	15.95	10.64	-	10.50
x (avg)	12.57	4.85	14.32	15.63	7.78	11.03

CrossConST	Cs	Kk	Ro	Ru	Tr	x (avg)
Cs	-	0.89	20.44	18.13	9.50	12.24
Kk	1.75	-	1.93	5.20	1.82	2.68
Ro	19.96	1.19	-	20.98	10.48	13.15
Ru	18.45	5.23	21.08	-	7.84	13.15
Tr	11.95	1.86	14.80	9.34	-	9.49
x (avg)	13.03	2.29	14.56	13.41	7.41	10.14

Ours	Cs	Kk	Ro	Ru	Tr	x (avg)
Cs	-	1.37	22.11	21.63	10.93	14.01
Kk	1.87	-	2.45	9.48	2.73	4.13
Ro	20.22	1.91	-	24.05	13.94	15.03
Ru	20.66	11.03	23.94	-	10.15	16.45
Tr	13.76	2.50	19.18	11.70	-	11.79
x (avg)	14.13	4.20	16.92	16.72	9.44	12.28

Table 7: The performance of the contrast models and our model on the PC-6 dataset.

IWSLT2017	De ↔ It	De ↔ Nl	De ↔ Ro	It ↔ Nl	It ↔ Ro	Nl ↔ Ro	Zero-shot Average	Supervised Average
m-Transformer	43.09	46.50	42.65	43.63	43.72	43.53	43.85	58.39
mRASP2 w/o AA	48.36	50.93	47.59	49.42	49.54	48.86	49.11	58.47
Ours	48.83	51.12	47.97	49.63	49.90	49.11	49.43	58.46

OPUS-7	x → Ar	x → De	x → Fr	x → Nl	x → Ru	x → Zh	Zero-shot Average	Supervised Average
m-Transformer	37.69	37.08	48.16	37.60	41.00	25.22	37.79	53.77
mRASP2 w/o AA	39.76	40.55	50.83	39.68	43.78	26.03	40.10	53.28
Ours	40.40	40.67	51.34	39.55	44.53	26.95	40.57	53.86

PC-6	x → Cs	x → Kk	x → Ro	x → Ru	x → Tr	Zero-shot Average	Supervised Average
m-Transformer	30.82	16.66	36.03	30.38	28.07	28.39	49.66
mRASP2 w/o AA	38.71	17.25	42.02	41.69	35.65	35.06	49.64
Ours	38.84	19.44	43.41	43.17	37.17	36.40	49.80

Table 8: The ChrF scores on the IWSLT2017, OPUS-7 and PC-6 test sets.

IWSLT2017	De ↔ It	De ↔ Nl	De ↔ Ro	It ↔ Nl	It ↔ Ro	Nl ↔ Ro	Zero-shot Average	Supervised Average
m-Transformer	70.14	74.01	71.54	71.73	74.07	73.59	72.51	83.82
mRASP2 w/o AA	77.01	79.27	78.14	78.49	81.24	80.16	79.05	83.80
Ours	77.21	79.45	78.19	78.60	81.33	80.11	79.15	83.83

OPUS-7	x → Ar	x → De	x → Fr	x → Nl	x → Ru	x → Zh	Zero-shot Average	Supervised Average
m-Transformer	72.33	65.77	70.67	68.97	73.48	71.52	70.46	79.45
mRASP2 w/o AA	74.36	68.41	72.77	70.89	76.15	74.02	72.77	79.25
Ours	74.64	68.74	73.33	71.02	76.65	74.18	73.09	79.66

PC-6	x → Cs	x → Kk	x → Ro	x → Ru	x → Tr	Zero-shot Average	Supervised Average
m-Transformer	58.10	46.51	62.63	62.25	54.84	56.86	76.36
mRASP2 w/o AA	68.98	50.74	69.19	70.95	64.85	64.94	76.32
Ours	70.13	52.99	72.32	72.98	66.86	67.05	77.31

Table 9: The COMET scores on the IWSLT2017, OPUS-7 and PC-6 test sets.

Models	l_{enc}	l_{dec}	Model Size	Zero-shot Average	Supervised Average
m-Transformer	6	6	79.4M	17.93	**34.32**
m-Transformer[†]	6	8	87.8M	18.10	34.23
m-Transformer[‡]	8	8	100.4M	16.84	34.21
mRASP2 w/o AA	6	6	79.4M	22.41	34.06
mRASP2 w/o AA[†]	8	8	100.4M	22.46	34.19
Ours	6	6	99.4M	**22.70**	34.25

Table 10: The scaling experiments on the IWSLT2017 test set. l_{enc} and l_{dec} denote layers of encoder and decoder respectively. The m-Transformer[†], m-Transformer[‡] and mRASP2 w/o AA[†] denote scaled variants of the original model.