Variable-length Neural Interlingua Representations for Zero-shot Neural Machine Translation

Zhuoyuan Mao [1]   **Haiyue Song** [1]   **Raj Dabre** [2]   **Chenhui Chu** [1]   **Sadao Kurohashi** [1,3]

[1] Kyoto University, Japan   [2] NICT, Japan   [3] NII, Japan

{zhuoyuanmao, song, chu, kuro}@nlp.ist.i.kyoto-u.ac.jp
raj.dabre@nict.go.jp

Abstract

The language-independency of encoded representations within multilingual neural machine translation (MNMT) models is crucial for their generalization ability on zero-shot translation. Neural interlingua representations have been shown as an effective method for achieving this. However, fixed-length neural interlingua representations introduced in previous work can limit its flexibility and representation ability. In this study, we introduce a novel method to enhance neural interlingua representations by making their length variable, thereby overcoming the constraint of fixed-length neural interlingua representations. Our empirical results on zero-shot translation on OPUS, IWSLT, and Europarl datasets demonstrate stable model convergence and superior zero-shot translation results compared to fixed-length neural interlingua representations. However, our analysis reveals the suboptimal efficacy of our approach in translating from certain source languages, wherein we pinpoint the defective model component in our proposed method.

1 Introduction

Multilingual neural machine translation (MNMT) (Dong et al., 2015; Firat et al., 2016; Ha et al., 2016; Johnson et al., 2017; Dabre et al., 2021) systems enable translation between multiple language pairs within a single model by
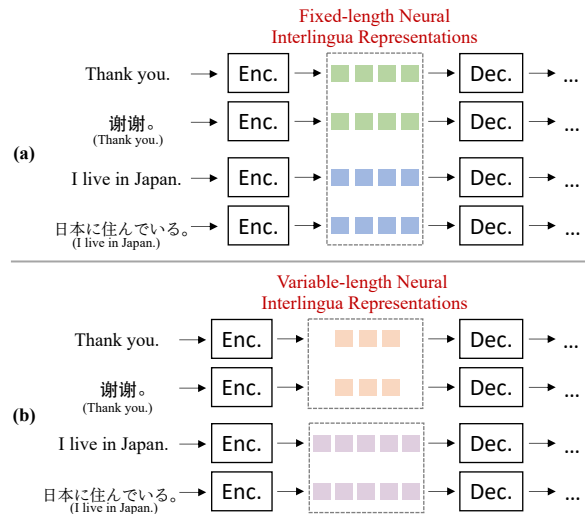
Figure 1: (a) Previous fixed-length neural interlingua representations; (b) Our proposed variable-length neural interlingua representations. Each colored box denotes the representation ($\mathbb{R}^{d \times 1}$) on the corresponding position. "Enc.", "Dec.", and "d" are encoder, decoder, and dimension of model hidden states.

learning shared representations across different languages. One of the key challenges in building effective MNMT systems is zero-shot translation performance involving unseen language pairs.

Previous work reveals that improving the language-independency of encoded representations is critical for zero-shot translation performance, with neural interlingua representations (Lu et al., 2018; Vázquez et al., 2019; Zhu et al., 2020) being proposed as an effective method for achieving this. Neural interlingua representations are shared, language-independent representations that behave as a neural pivot between different natural languages. As shown in Figure 1 (a), it enables sentences in different languages with the same meaning to have the same interlingua representations. Previous work has shown the effective-

ness of fixed-length neural interlingua representations for zero-shot translation. However, a fixed length can limit neural interlingua representations' flexibility and representation ability. It is highly model size and training data size-sensitive according to our experimental results for different settings of model and training data size.

This paper proposes a novel method for improving neural interlingua representations by making their length variable. As shown in Figure 1 (b), our method enables the length of the interlingua representations to vary according to different lengths of source sentences, which may provide more flexible neural interlingua representations. Specifically, we utilize the sentence length in the centric language[1] (e.g., English) as the length of neural interlingua representations. We propose a variable-length interlingua module to project sentences in different source languages with the same meaning into an identical neural interlingua representation sequence. To enable translating from noncentric language source sentences during inference, we also introduce a length predictor within the variable-length interlingua module. Moreover, as for the initialization of the interlingua module, we propose a novel method that facilitates knowledge sharing between different interlingua lengths, which can avoid introducing redundant model parameters. We expect that variable-length interlingua representations provide enhanced representations according to different source sentence lengths, which mitigates the model size and training data size-sensitive problem of previous work in low-resource scenarios and improves performance for zero-shot translation.

We conduct experiments on three MNMT datasets, OPUS (Zhang et al., 2020), IWSLT (Cettolo et al., 2017), and Europarl (Koehn, 2005) with different settings of training data size and model size. Results demonstrate that our proposed method yields superior results for zero-shot translation compared to previous work. Our method exhibits stable convergence in different settings while previous work (Zhu et al., 2020) is highly sensitive to different model and training data sizes. However, we also observe the inferior performance

of our method for translation from non-centric language source languages. We attribute it to the accuracy of the interlingua length predictor and point out the possible directions of this research line.

2 Related Work

This paper focuses on variable-length interlingua representations for zero-shot NMT.

2.1 Zero-shot Translation

In recent years, MNMT (Dong et al., 2015; Firat et al., 2016; Ha et al., 2016; Johnson et al., 2017; Aharoni et al., 2019; Tan et al., 2019; Dabre et al., 2021; Zhang et al., 2020) has been a popular research topic, where the generalization ability of MNMT models to zero-shot translation is a critical problem as obtaining sufficient training data for all translation directions is often impractical. An MNMT model's zero-shot translation performance usually benefits from the encoder-side representations being language-independent and decoder-side representations being language-specific. To achieve this, some studies have proposed removing encoder-side residual connections (Liu et al., 2021) or introducing language-independent constraints (Al-Shedivat and Parikh, 2019; Pham et al., 2019; Arivazhagan et al., 2019; Yang et al., 2021; Mao et al., 2023). Other methods involve decoder pre-training and back-translation (Gu et al., 2019; Zhang et al., 2020), denoising autoencoder objectives (Wang et al., 2021), and encoder-side neural interlingua representations (Lu et al., 2018; Vázquez et al., 2019; Zhu et al., 2020).

2.2 Neural Interlingua Representations for Zero-shot Translation

As mentioned above, constructing neural interlingua representations is a powerful method to improve shared encoder representations across various source languages and enhance zero-shot translation. Lu et al. (2018) first proposed the concept of neural interlingua representations for MNMT, intending to bridge multiple language-specific encoders and decoders using an intermediate interlingua attention module, which has a fixed sequence length. Vázquez et al. (2019) extended this approach with a universal encoder and decoder architecture for MNMT and introduced a regularization objective for the interlingua attention similarity matrix. More recently, Zhu et al. (2020) applied the neural interlingua approach in the Trans-

---

[1] In this work, we consider using an x-centric parallel corpus, wherein all sentence pairs within the corpus consist of sentences in language x paired with another language. It is noteworthy that the English-centric corpus is the most prevalent setting. We denote a language distinct from x as a "noncentric language" in the subsequent text.
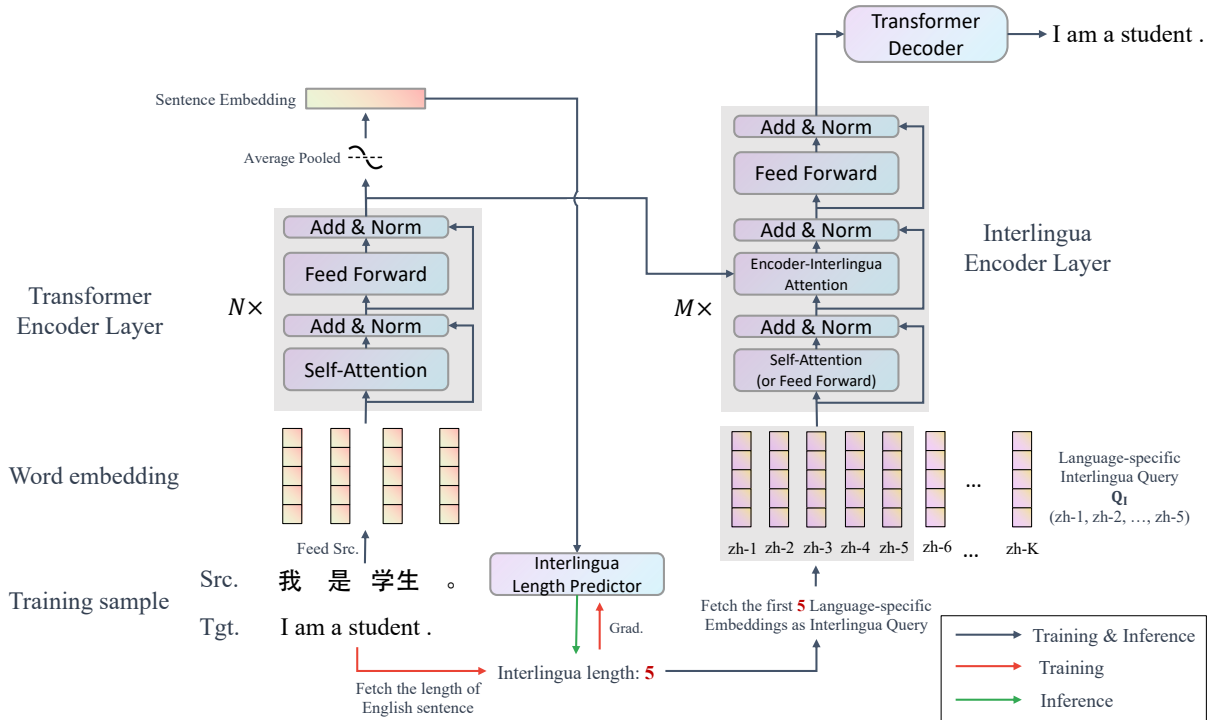
Figure 2: Variable-length interlingua module. "zh-x" denotes the x-th embedding of a Chinese-specific interlingua query.

former (Vaswani et al., 2017) model architecture and proposed a position-wise alignment objective to ensure consistent neural interlingua representations across different languages. However, these methods utilized fixed-length neural interlingua representations, which may reduce the model's representation ability for source sentences with different lengths. This paper focuses on revisiting and improving neural interlingua approaches.

3 Variable-length Neural Interlingua Representations

We present an MNMT model that comprises three distinct components: a source language encoder, a neural interlingua module, and a decoder. The source language encoder converts source sentences to language-specific representations, the neural interlingua module generates language-agnostic representations, and the decoder converts these representations into the target language translation. In this section, we introduce a novel neural interlingua module.

Specifically, we propose variable-length neural interlingua representations surpassing prior work's fixed-length constraint. To achieve this breakthrough, we have developed a module that includes interlingua encoder layers, an interlingua length predictor, and a language-specific interlin-

gua query. Our module uses an embedding sharing mechanism, as shown in Figure 2. Moreover, we introduce the objectives that guide the training of variable-length neural interlingua representations.

3.1 Variable-length Interlingua Module

Interlingua Encoder Layers In accordance with Zhu et al. (2020), we construct a variable-length interlingua module within a Transformer model architecture. Our model utilizes N Transformer encoder layers and 6 Transformer decoder layers, with M interlingua encoder layers introduced between them. To maintain consistency with a standard 6-layer Transformer encoder, we set $M + N = 6$, ensuring that the number of model parameters remains almost the same. Each interlingua encoder layer consists of a sequential series of operations, including self-attention mechanisms (or feed-forward networks),[2] encoder-interlingua attention, and feed-forward networks, as illustrated in Figure 2.

The input representations for interlingua encoder layers are denoted as $\mathbf{Q}_{\mathrm{I}} \in \mathbb{R}^{d \times \mathrm{len}_{\mathrm{I}}(X)}$, where d and $\mathrm{len}_{\mathrm{I}}(X)$ respectively indicates the di-

---

[2]We utilize feed-forward networks for the first interlingua encoder layer and employ a self-attention mechanism for subsequent layers. This is because the interlingua query is initially weak and unable to capture similarities through a self-attention mechanism. This design choice is similar to that of Zhu et al. (2020).

mension of hidden representations and the length of the neural interlingua representations given a source sentence $X = x_1, x_2, ..., x_k$. Specifically, we define $\text{len}_\text{I}(X)$ as follows:

$$\text{len}_\text{I}(X) = \begin{cases} \text{len}(X), & X \text{ is in centric} \\ \text{len}(\text{CT}(X)), & X \text{ is in non-centric} \end{cases},$$

(1)

where $\text{CT}(X)$ denotes the translation of X in the centric language. We use teacher forcing to generate interlingua length during training. For instance, if we use English-centric parallel sentences as training data, $\text{len}_\text{I}(X)$ for each sentence pair will be the length of English sentences. Thus, sentences that convey the same semantic meaning can have the same interlingua length, and interlingua length is variable according to different sentences. For the initialization of $\mathbf{Q}_\text{I}$, we will provide a detailed explanation of how to generate it later in this section.

Subsequently, $\mathbf{Q}_\text{I}$ undergoes self-attention (or feed-forward networks), and we obtain the output $\mathbf{Q}_\text{I}'$. Assume that the contextualized representations on top of N Transformer encoder layers are $\mathbf{H}_\text{S} \in \mathbb{R}^{d \times k}$. Then we establish an encoder-interlingua attention mechanism:

$$\mathbf{H}_\text{EI} = \text{Attn}(\mathbf{Q}_\text{I}', \mathbf{H}_\text{S}, \mathbf{H}_\text{S}), \quad (2)$$

where $\text{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V})$ indicates the multi-head attention mechanism (Vaswani et al., 2017). This encoder-interlingua attention inherits the design in previous studies of neural interlingua representations (Lu et al., 2018; Vázquez et al., 2019; Zhu et al., 2020).

Finally, we pass $\mathbf{H}_\text{EI}$ through position-wise feed-forward networks to obtain $\mathbf{H}_\text{I}$, the output of the interlingua encoder layers. $\mathbf{H}_\text{I}$ serves as a language-agnostic neural interlingua and can vary in length depending on the source sentence. Once we have $\mathbf{H}_\text{I}$, we feed it into a standard Transformer decoder to generate the translation.

Interlingua Length Predictor Length of interlingua representations is not readily available during inference when translating from non-centric source sentences (e.g., non-English source sentences) using Eq. (1). To address this, we propose using an interlingua length predictor to obtain $\text{len}_\text{I}(X)$ for inference. Specifically, we treat the length prediction of translation in the centric language as a classification task, addressed utilizing mean pooled contextualized representations atop

the Transformer encoder.[3] More precisely, we predict X's interlingua length as:

$$\text{len}_\text{I}(X) = \arg\max_i \text{ softmax}(\frac{\mathbf{1} \, \mathbf{H}_\text{S}^\text{T}}{k}\mathbf{W} + \mathbf{b})_i,$$

(3)

where k is the length of X, $\mathbf{1} \in \mathbb{R}^{1 \times k}$ denotes a vector with all the elements of 1, $\mathbf{W} \in \mathbb{R}^{d \times K}$ and $\mathbf{b} \in \mathbb{R}^{1 \times K}$ indicates the weight and bias of a linear layer, and K is the maximum sequence length allowed in the model.

Language-specific Interlingua Query Here, we present the method for obtaining input representations $\mathbf{Q}_\text{I}$ for the interlingua encoder layers. Initially, we randomly initialize an embedding matrix $\mathbf{E}_l \in \mathbb{R}^{d \times K}$ containing K embeddings for the source language l. Next, we extract the first $\text{len}_\text{I}(X)$ embeddings from $\mathbf{E}_l$ to obtain $\mathbf{Q}_\text{I}$.

$$\mathbf{Q}_\text{I} = \mathbf{E}_l \mathbf{I}_\text{S}, \quad (4)$$

where $\mathbf{I}_\text{S} \in \mathbb{R}^{K \times \text{len}_\text{I}(X)}$ has 1s as main diagonal elements and 0s for other elements. Note that the language-specific nature of $\mathbf{E}_l$ allows the model to learn a unique mapping from each language to the neural interlingua representations. Zhu et al. (2020) used the technique of language-aware positional embedding (Wang et al., 2019) for both the neural interlingua representations and the source and target sentences, resulting in ambiguity regarding whether the improvements were from the neural interlingua representations or not. In contrast, our proposed language-specific interlingua query clarifies whether a language-specific mapping to neural interlingua representations benefits zero-shot translation.

3.2 Training Objectives

Given a training sample sentence pair (X, Y), we introduce the following training objective, combining an NMT loss, an interlingua alignment loss, and a length prediction loss. The interlingua alignment loss is utilized to guarantee the consistency of the neural interlingua representations for each training sentence pair sample. In contrast, the length prediction loss ensures the generation of variable interlingua length during inference. Specifically, the training objective is defined as follows:

$$\mathcal{L}(X, Y) = \alpha\mathcal{L}_\text{NMT} + \beta\mathcal{L}_\text{IA} + \gamma\mathcal{L}_\text{LP}, \quad (5)$$

---

[3]We attempted to treat it as a regression task, but the performance of the regression model was notably inferior to that of the classifier-based predictor.

Datasets	Languages	# Sup.	# Zero.	# Train	# Valid	# Test
OPUS	ar, de, en, fr, nl, ru, zh	12	30	12,000,000	2,000	2,000
IWSLT	en, it, nl, ro	6	6	1,378,794	2,562	1,147
Europarl	de, en, es, fr, nl	8	12	15,782,882	2,000	2,000

Table 1: Statistics of the training data. "# Sup." and "# Zero." indicate the respective number of language pairs for supervised and zero-shot translation. "# Train" denotes the total number of the training parallel sentences while "# Valid" and "# Test" showcase the number per language pair.

where α, β, and γ are weight hyperparameters for each loss, $\mathcal{L}_{\text{LP}}$ is a cross-entropy loss computed from the softmax outputs from Eq. (3), and $\mathcal{L}_{\text{IA}}$ is a position-wise alignment loss using cosine similarity following Zhu et al. (2020):

$$\mathcal{L}_{\text{IA}} = 1 - \tfrac{1}{\text{len}_{\text{I}}(X)} \sum_i \cos < \mathbf{H}_{\text{I}}(X)_i, \mathbf{H}_{\text{I}}(Y)_i > . \quad (6)$$

Here $\mathbf{H}_{\text{I}}(\cdot)_i$ denotes the i-th column of $\mathbf{H}_{\text{I}}(\cdot)$.[4] Please note that during training, we always have $\text{len}_{\text{I}}(X) = \text{len}_{\text{I}}(Y)$ because we apply teacher forcing to generate the interlingua length for the sentence pair (X, Y). With $\mathcal{L}_{\text{IA}}$, different sentence pairs with varying lengths of translation in centric language can be represented using variable-length neural interlingua representations. This can enhance the bridging ability for zero-shot translation.

4 Experimental Settings

4.1 Datasets

Our study involves conducting experiments on zero-shot translation using three distinct datasets, OPUS (Zhang et al., 2020), IWSLT (Cettolo et al., 2017), and Europarl (Koehn, 2005), each comprising 7, 4, and 5 languages, respectively. For each dataset, we adopt the train, valid, and test splits following Zhang et al. (2020), Wu et al. (2021), and Liu et al. (2021). Table 1 presents each dataset's overall statistics. The training and validation data exclusively contains English-centric sentence pairs, indicating the centric language is English in all the experiments, leading to 12, 6, and 8 supervised directions, and 30, 6, and 12 zero-shot directions for each dataset. Refer to Appendix A for preprocessing details.

4.2 Overall Training and Evaluation Details

For the OPUS and IWSLT datasets, we utilize a `Transformer-base` model, while for Eu-

[4]To derive $\mathbf{H}_{\text{I}}(Y)$, it is necessary to feed the target sentence to both the encoder and interlingua encoder layers, which can potentially result in increased computational requirements.

roparl, we employ a `Transformer-big` model, to evaluate the performance of Transformer with both sufficient and insufficient training data. Regarding language tag strategies to indicate the source and target languages to the model, we adopt the method of appending the source language tag to the encoder input and the target language tag to the decoder input (Liu et al., 2020). This approach allows for the creation of fully language-agnostic neural interlingua representations in between.[5] The maximum sentence length is set as 256, which indicates that $K = 256$ (Section 3.1). Refer to Appendix B for other training details.

For evaluation, we choose the evaluation checkpoint based on the validation $\mathcal{L}_{\text{NMT}}$ with the lowest value. We use a beam size of 5 during inference on the trained models to conduct inference. We report SacreBLEU (Post, 2018).[6]

4.3 Baselines and Respective Training Details

To compare our variable-length neural interlingua method with previous fixed-length neural interlingua methods, we trained the following settings:

MNMT (Johnson et al., 2017) is a system trained with standard `Transformer-base` or `Transformer-big` for multiple language pairs. We applied the language tag strategy of source language tag for encoder input and target language tag for decoder input.

Pivot translation (Zoph and Knight, 2016) involves translating a source language into a pivot language, usually English, and then translating the pivot language into the target language. This system constitutes a robust baseline for zero-shot translation, which we include for reference. We implement this setting by feeding the pivot language output of the MNMT model to itself to generate the target language.

Len-fix. Uni. Intl. We follow the setting described by Zhu et al. (2020), but we remove its language-aware positional embedding to test whether a single interlingua module can improve zero-shot translation. Compared to our variable-length interlingua representations presented in Section 3.1, these fixed interlingua representations have a universal len_{I} (Eq. (1)) for different source

[5]We do not consider employing target language tag appending on the encoder-side (Johnson et al., 2017) in this work because it would require removing both the source and target language information after feeding the source sentence to obtain the neural interlingua representations.
[6]We utilize the "zh" tokenization mode for Chinese, and the "13a" tokenization mode for other languages.

Methods	Zero-shot			Supervised: From en			Supervised: To en		
	OPUS	IWSLT	Europarl	OPUS	IWSLT	Europarl	OPUS	IWSLT	Europarl
Pivot	*22.0*	*19.9*	*29.5*	-	-	-	-	-	-
MNMT	16.5	13.1	29.0	**31.2**	**29.6**	**32.9**	**36.8**	**33.5**	**36.1**
Len-fix. Uni. Intl.	18.2	12.7	17.4	29.6	19.6	20.1	35.3	22.2	21.8
Len-fix. LS. Intl.	18.4	4.7	5.8	30.1	7.3	6.7	35.7	12.9	7.1
Len-vari. Intl. (ours)	**18.9**†	**14.8**	**29.6**	30.2†	26.2	32.6	34.0	27.1	33.8

Table 2: Overall BLEU results on OPUS, IWSLT, and Europarl. The best result among all the settings except *Pivot* is in **bold**. We mark the results significantly (Koehn, 2004) better than "Len-fix. Uni. Intl." with † for OPUS dataset.

Methods	de–fr		ru–fr		nl–de		zh–ru		zh–ar		nl–ar		Zero-shot
	→	←	→	←	→	←	→	←	→	←	→	←	Avg.
Pivot	*23.4*	*21.2*	*31.0*	*26.0*	*21.8*	*23.6*	*24.8*	*37.9*	*24.0*	*38.9*	*7.4*	*17.4*	*22.0*
MNMT	17.6	15.0	21.5	17.7	17.9	21.4	15.3	27.6	18.0	28.6	5.3	13.3	16.5
Len-fix. Uni. Intl.	20.1	17.0	25.0	22.4	19.5	21.3	20.3	30.9	19.6	30.4	6.1	14.4	18.2
Len-fix. LS. Intl.	**20.7**	17.7	25.7	21.7	19.8	21.6	19.9	31.5	**20.1**	31.6	**6.5**	**14.5**	18.4
Len-vari. Intl. (ours)	20.6†	**18.3**†	**26.0**†	**23.4**†	**20.2**†	**22.1**†	20.8	**31.8**†	20.0	**31.9**†	6.3	**14.5**	**18.9**†

Table 3: BLEU results of zero-shot translation on OPUS. We randomly select six zero-shot language pairs and report the results. The best result among all the settings except "*Pivot*" is in **bold**. We mark the results significantly (Koehn, 2004) better than "Len-fix. Uni. Intl." with †.

sentences and a universal $\mathbf{E} \in \mathbb{R}^{d \times \text{len}_I}$ for different languages and without a $\mathbf{Q}_I$ (Eq. (4)). The fixed interlingua length is set to 17, 21, and 30, which are the average lengths of each dataset following Zhu et al. (2020) and Vázquez et al. (2019).

Len-fix. LS. Intl. The only difference between this system and the "Len-fix. Uni. Intl." system mentioned above is the initialization of the interlingua query. We use a language-specific $\mathbf{E}_l \in \mathbb{R}^{d \times \text{len}_I}$ for each source language l without a $\mathbf{Q}_I$ (Eq. (4)).

Len-vari. Intl. (ours) This refers to variable-length neural interlingua representations proposed in Section 3.

For the last three neural interlingua settings, we set M and N to 3 for both the Transformer encoder and interlingua encoder layers. The values of α, β, and γ (Eq. (5)) are set as 1.0, 1.0, and 0.1, respectively. We remove the first residual connection within the first interlingua encoder layer to improve the language-independency of the interlingua representations, inspired by Liu et al. (2021).

5 Results and Analysis

We now present in tables 2, 3, and 4 the results of our variable-length interlingua approach and compare them against several baselines.

5.1 Main results

Firstly, Tables 2 and 3 indicate that our proposed variable-length interlingua representations outper-

form previous work in zero-shot directions. The severe overfitting issue of "Len-fix. Uni. Intl." and "Len-fix. LS. Intl." on IWSLT and Europarl suggests that they are limited to model size and training data size settings, while our proposed method can converge stably on all three settings. These results demonstrate that our flexible interlingua length can benefit zero-shot translation more effectively. Secondly, our proposed method performs better than previous work in "from en" supervised directions as shown in Tables 2 and 4, but still falls short of the MNMT baseline. This may be attributed to the interlingua module's weak source-target awareness. Thirdly, our variable-length neural interlingua representations perform significantly worse on "to en" directions than "Len-fix." methods on OPUS and MNMT on all datasets. We provide analysis of this phenomenon next.

5.2 Validation NMT Loss

We investigate why variable-length neural interlingua representations perform poorly in "to en" supervised directions by analyzing the validation NMT loss, an approximate measure of NMT performance on the validation set. Figure 3 displays the validation NMT loss for all settings on OPUS. We observe that variable-length interlingua representations can converge well, even smaller than the validation loss of "Len-fix. Uni. Intl." and "Len-fix. LS. Intl." However, the interlingua length predictor was teacher-forced during training, indicat-

Methods	en–ar		en–de		en–fr		en–nl		en–ru		en–zh		Supervised Avg.	
	→	←	→	←	→	←	→	←	→	←	→	←	From en	To en
MNMT	**23.9**	**37.8**	**30.8**	**34.6**	**33.9**	**35.5**	**27.8**	**31.5**	29.4	**35.1**	**41.2**	**46.4**	**31.2**	**36.8**
Len-fix. Uni. Intl.	22.6	36.6	28.9	33.0	31.7	33.5	27.4	30.1	28.4	34.0	38.8	44.6	29.6	35.3
Len-fix. LS. Intl.	22.9	36.8	29.0	33.8	32.3	33.9	27.7	30.6	28.9	34.3	39.5	44.8	30.1	35.7
Len-vari. Intl. (ours)	23.3†	33.8	30.1†	32.3	32.9†	32.6	27.3	27.9	**29.5**†	32.2	38.0	45.3†	30.2†	34.0

Table 4: BLEU results of supervised translation on OPUS. The best result among all the settings is in **bold**. We mark the results significantly (Koehn, 2004) better than "Len-fix. Uni. Intl." with †.

	ar	de	fr	nl	ru	zh	Avg.
Acc. of Len. Pre.	20.6	26.5	17.6	19.3	21.1	13.8	19.8
Avg. of | Len. Pre. $-$ *gold* |	2.4	3.4	3.8	3.1	3.3	3.9	3.3
BLEU w/ Len. Pre.	33.8	32.3	32.6	27.9	32.2	45.3	34.0
BLEU w/ *gold*	35.5†	33.4†	33.3†	29.4†	33.4†	46.0†	35.2†

Table 5: Accuracy of the interlingua length predictor, averaged absolute difference between predicted length and *gold* length, and "to en" BLEU scores of each non-English source language on OPUS. "w/ Len. Pre." and "w/ *gold*" indicate using the predicted interlingua length and the correct interlingua length (length of the English translation), respectively. Accuracy of the length predictor and average abosulute difference are evaluated using OPUS's test set. We mark the results significantly (Koehn, 2004) better than "BLEU w/ Len. Pre." with †.
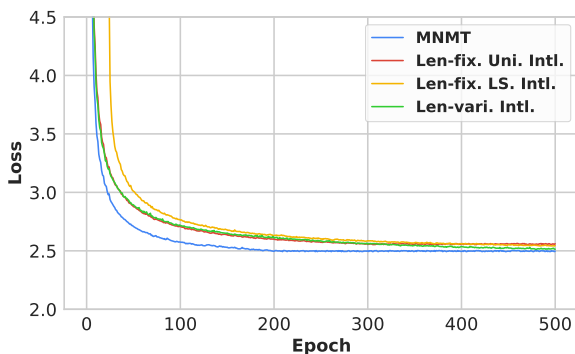

Figure 3: Validation NMT loss curve on OPUS.

ing the validation NMT loss was calculated with a 100% accurate interlingua length predictor. As a result, the inaccurate interlingua length predictor is likely the primary cause of our method's inferior performance in "to en" directions, despite its well-converged validation NMT loss.

5.3 Impact of the Interlingua Length Predictor

We analyze the interlingua length predictor and identify the reason for the subpar performance in "to en" translations. We input the source sentences of the test set in non-English languages into the model and check whether the predicted length in interlingua is identical to the length of its English reference. We present the accuracy on the OPUS dataset in Table 5. The results show that the accuracy for each language is approximately 20.0%, which can result in error propagation when trans-

lating from those languages. To further understand the impact of the length predictor quality on translation performance, we attempt to provide the model with the correct interlingua length instead of relying on the length predictor. As shown in Table 5, the results reveal significant BLEU improvements when the correct interlingua length is applied. This suggests that the performance issue encountered when translating from a non-centric source language can be addressed by upgrading the interlingua length predictor's accuracy. Furthermore, we can also enhance zero-shot translation performance if we have a better length predictor. Nevertheless, we observe that even with a low length prediction accuracy of approximately 20.0%, we can still achieve solid BLEU performance, averaging 34.0 BLEU points. This indicates that an incorrectly predicted length with just a trivial difference, as shown in Table 5, will not result in the enormous information loss required for translation.

6 Conclusion

This study introduced a novel variable-length neural interlingua approach that improved zero-shot translation results while providing a more stable model than previous fixed-length interlingua methods. Although our analysis revealed a performance downgrade in "to en" directions, we have identified the problematic model component and plan to address it in future studies.

Acknowledgements

References

Aharoni, Roee, 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, June. Association for Computational Linguistics.

Al-Shedivat, Maruan and Ankur Parikh. 2019. Consistency by agreement in zero-shot 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 1184–1197, Minneapolis, Minnesota, June. Association for Computational Linguistics.

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

Cettolo, Mauro, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuhito Sudoh, Koichiro Yoshino, and Christian Federmann. 2017. Overview of the IWSLT 2017 evaluation campaign. In *Proceedings of the 14th International Conference on Spoken Language Translation*, pages 2–14, Tokyo, Japan, December 14-15. International Workshop on Spoken Language Translation.

Dabre, Raj, Chenhui Chu, and Anoop Kunchukuttan. 2021. A survey of multilingual neural machine translation. *ACM Comput. Surv.*, 53(5):99:1–99:38.

Dong, Daxiang, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732, Beijing, China, July. Association for Computational Linguistics.

Firat, Orhan, 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, June. Association for Computational Linguistics.

Gu, Jiatao, 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, July. Association for Computational Linguistics.

Ha, Thanh-Le, Jan Niehues, and Alex Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. In *Proceedings of the 13th International Conference on Spoken Language Translation*, Seattle, Washington D.C, December 8-9. International Workshop on Spoken Language Translation.

Hu, Junjie, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multitask benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Johnson, Melvin, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

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

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

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

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers, MTSummit 2005, Phuket, Thailand, September 13-15, 2005*, pages 79–86.

Liu, Yinhan, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and

Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Liu, Danni, Jan Niehues, James Cross, Francisco Guzmán, and Xian Li. 2021. Improving zero-shot translation by disentangling positional information. 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 1259–1273, Online, August. Association for Computational Linguistics.

Lu, Yichao, Phillip Keung, Faisal Ladhak, Vikas Bhardwaj, Shaonan Zhang, and Jason Sun. 2018. A neural interlingua for multilingual machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 84–92, Brussels, Belgium, October. Association for Computational Linguistics.

Mao, Zhuoyuan, Raj Dabre, Qianying Liu, Haiyue Song, Chenhui Chu, and Sadao Kurohashi. 2023. Exploring the impact of layer normalization for zero-shot neural machine translation. *CoRR*, abs/2305.09312.

Micikevicius, Paulius, Sharan Narang, Jonah Alben, Gregory F. Diamos, Erich Elsen, David García, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. 2018. Mixed precision training. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Pham, Ngoc-Quan, 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, August. Association for Computational Linguistics.

Post, Matt. 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, October. Association for Computational Linguistics.

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

Tan, Xu, Jiale Chen, Di He, Yingce Xia, Tao Qin, and Tie-Yan Liu. 2019. Multilingual neural machine translation with language clustering. 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 963–973, Hong Kong, China, November. Association for Computational Linguistics.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Guyon, Isabelle, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Vázquez, Raúl, Alessandro Raganato, Jörg Tiedemann, and Mathias Creutz. 2019. Multilingual NMT with a language-independent attention bridge. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 33–39, Florence, Italy, August. Association for Computational Linguistics.

Wang, Xinyi, Hieu Pham, Philip Arthur, and Graham Neubig. 2019. Multilingual neural machine translation with soft decoupled encoding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Wang, Weizhi, 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, November. Association for Computational Linguistics.

Wu, Liwei, 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, August. Association for Computational Linguistics.

Yang, Yilin, Akiko Eriguchi, Alexandre Muzio, Prasad Tadepalli, Stefan Lee, and Hany Hassan. 2021. 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, November. Association for Computational Linguistics.

Zhang, Biao, 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, July. Association for Computational Linguistics.

Zhu, Changfeng, Heng Yu, Shanbo Cheng, and Weihua Luo. 2020. Language-aware interlingua for multilingual neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1650–1655, Online, July. Association for Computational Linguistics.

Zoph, Barret and Kevin Knight. 2016. Multi-source neural translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 30–34, San Diego, California, June. Association for Computational Linguistics.

A   Preprocessing Details

Jieba[7] is used to segment Chinese while Moses[8] (Koehn et al., 2007) is utilized to tokenize other languages. We employ BPE (Sennrich et al., 2016) with $50,000$, $40,000$, and $50,000$ merge operations to create a joint vocabulary for each dataset, resulting in the vocabulary sizes of $66,158$, $40,100$, and $50,363$, respectively.

B   Training Details

Our models are trained using Fairseq.[9] As the data size for each language pair is relatively similar, oversampling is not implemented for MNMT. The dropout rate was set to 0.1, 0.4, and 0.3 for each dataset, and we use the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 5e-4, 1e-3, and 5e-4, respectively, employing $4,000$ warm-up steps. The `Transformer-base` model was trained using four 32 GB V100 GPUs, and the `Transformer-big` model was trained using eight 32 GB V100 GPUs, with a batch size of $4,096$ tokens. To speed up training, mixed precision training (Micikevicius et al., 2018) is also employed. Each dataset is trained for 500, 200, and 500 epochs.

C   Limitations

While this study proposed a novel method for improving neural interlingua representations for zero-shot translation, the following limitations should be addressed in future work:

- The inaccurate interlingua length predictor currently leads to inferior performance for translation from non-centric languages. Therefore, a better predictor should be explored to improve the performance.

- We used the length of centric language sentences as the interlingua length, which may limit the application for using parallel sentences not involving the centric language. Therefore, a better way to generate variable lengths for neural interlingua representations should be developed in future work.

- We have yet to test whether the neural interlingua representations obtained in this study can act as a semantic pivot among all the languages. Thus, it would be interesting to evaluate the effectiveness of our variable-length interlingua representations on cross-lingual language understanding tasks (Hu et al., 2020).

---

[7]`https://github.com/fxsjy/jieba`
[8]`https://github.com/moses-smt/mosesdecoder`
[9]`https://github.com/facebookresearch/fairseq`