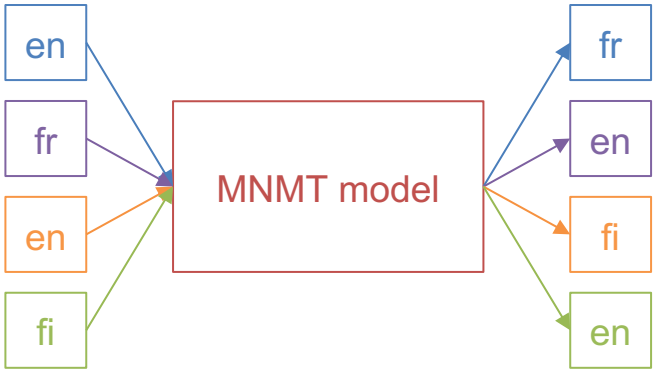
Variable-length Neural Interlingua Representations for Zero-shot Neural Machine Translation

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

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

Multilingual Neural Machine Translation

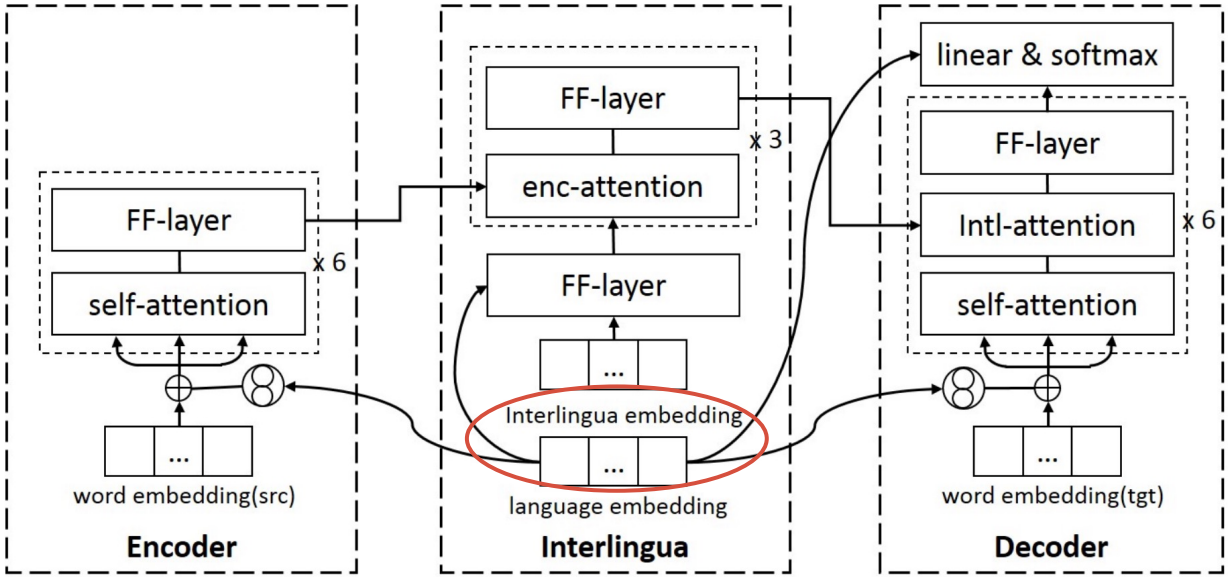
- Multilingual neural machine translation (MNMT) (Johnson et al., 2016)
  - Translation for multiple language pairs with a single model.
- Zero-shot translation (ZST)
  - Translation for unseen language pairs without training data.
  - Promising for its low latency compared with pivot-based translation.



An English-Finnish-French multilingual translation model, where en-to-fr, en-to-fi, fr-to-en, fi-to-en are supervised directions, and fi-fr, fr-fi are zero-shot directions.

Neural Interlingua Representations for ZST

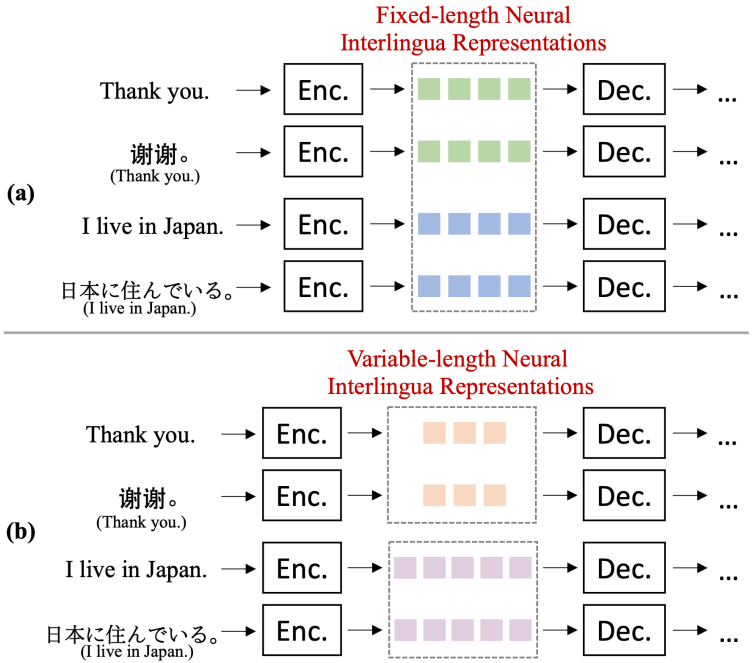
- Language-aware Interlingua for Multilingual Neural Machine Translation (Zhu et al., 2020)
  - Constructed **fixed-length interlingua** for MNMT with a interlingua module.



Motivation

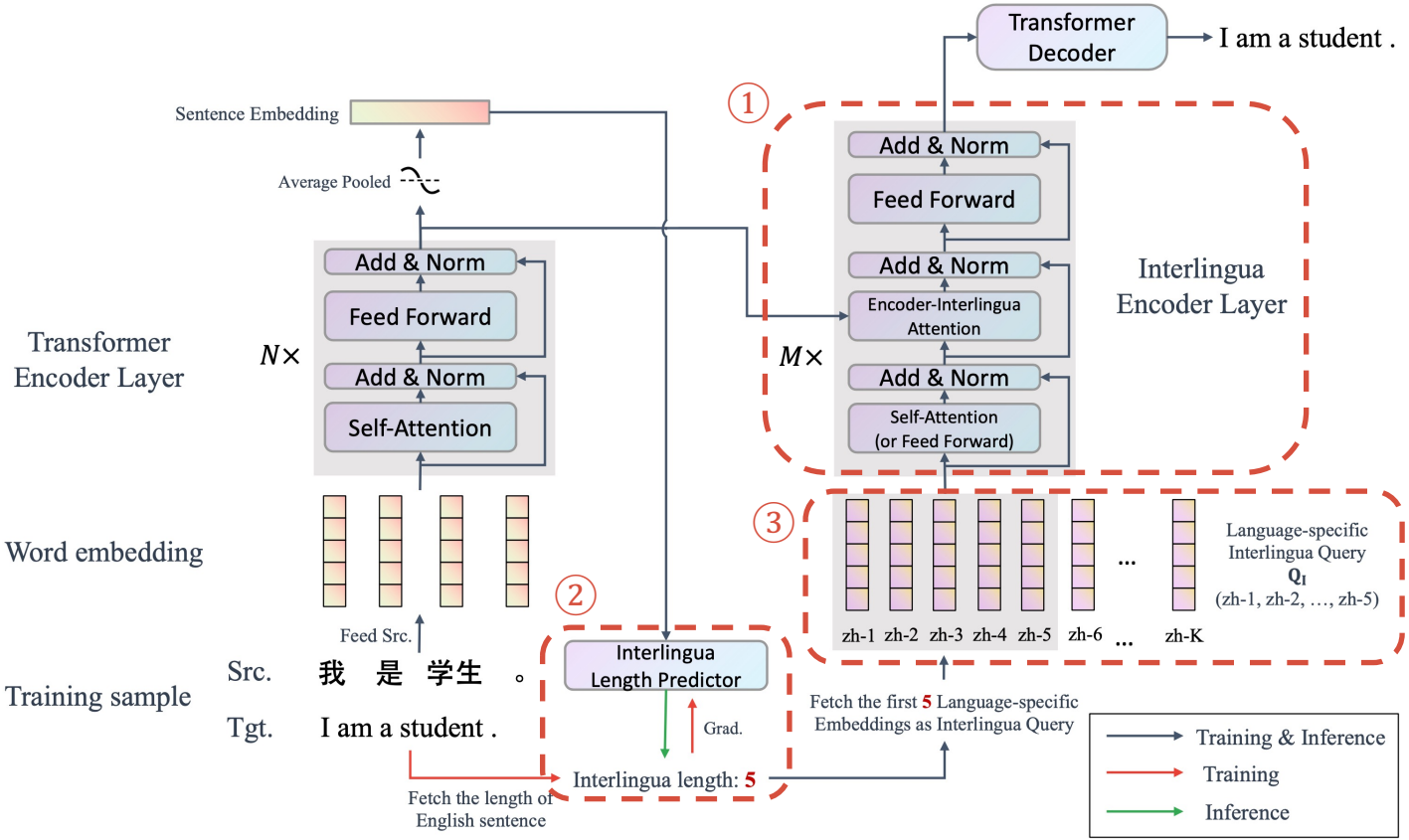
- Fixed-length neural interlingua representations in previous work (a) can limit its flexibility and representation ability.

- This can be improved by introducing variable-length interlingua representations (b).



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.

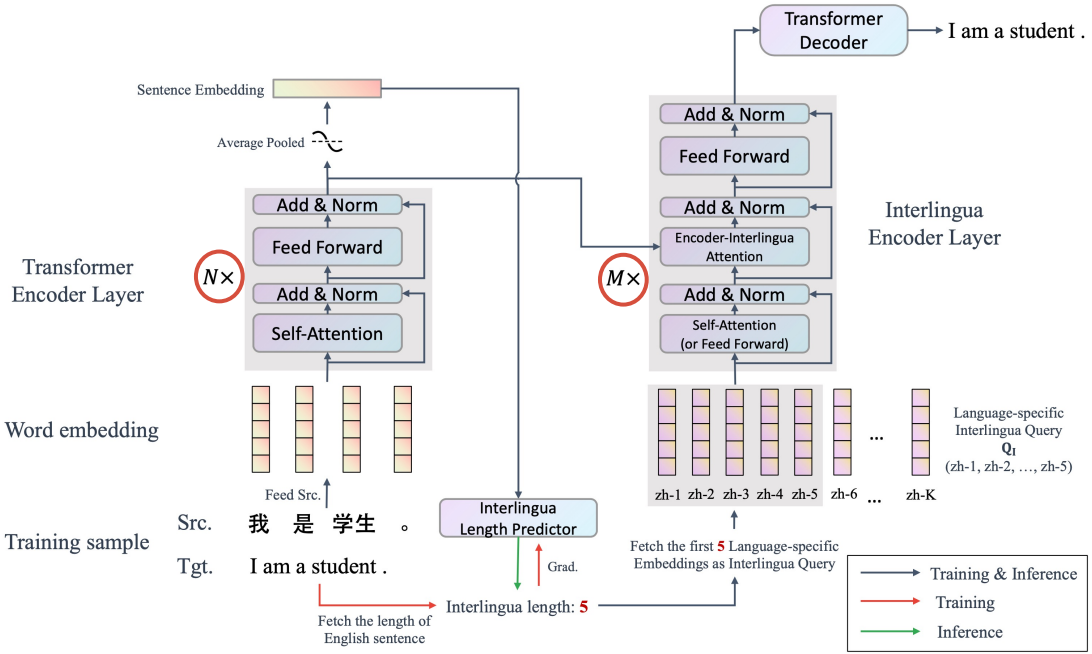
An Overview of Variable-length Interlingua Module (VIM)



"zh-x" denotes the x-th embedding of a Chinese-specific interlingua query.

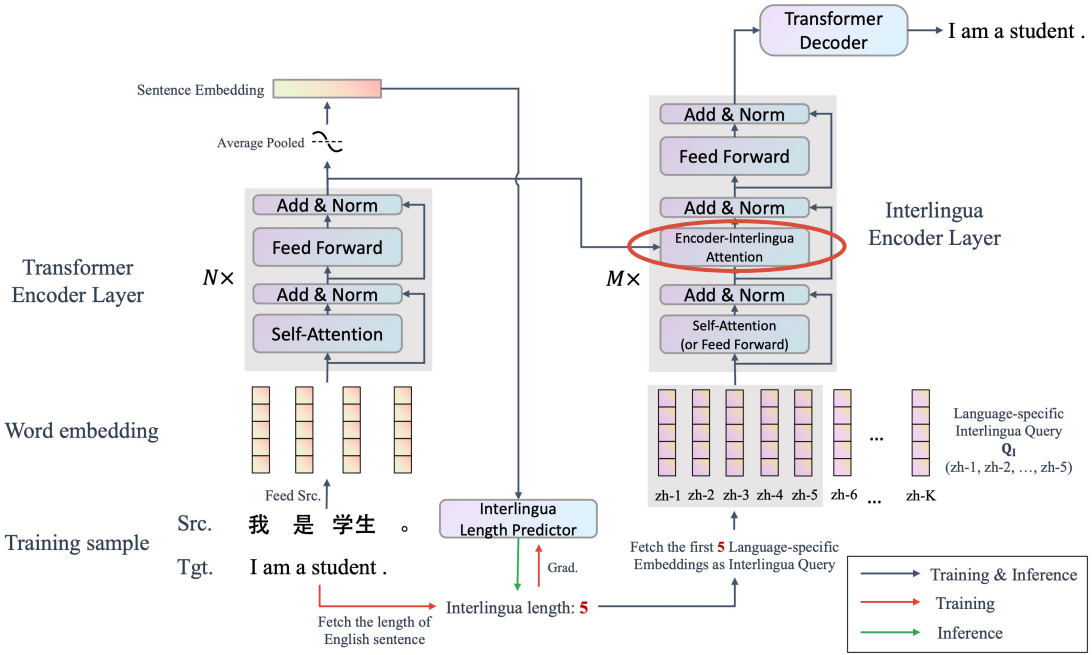
VIM: ① Interlingua Encoder Layers

- M interlingua encoder layers, N vanilla transformer encoder layers.
  - $M + N = 6$ in order to maintain the consistency with a standard Transformer model.



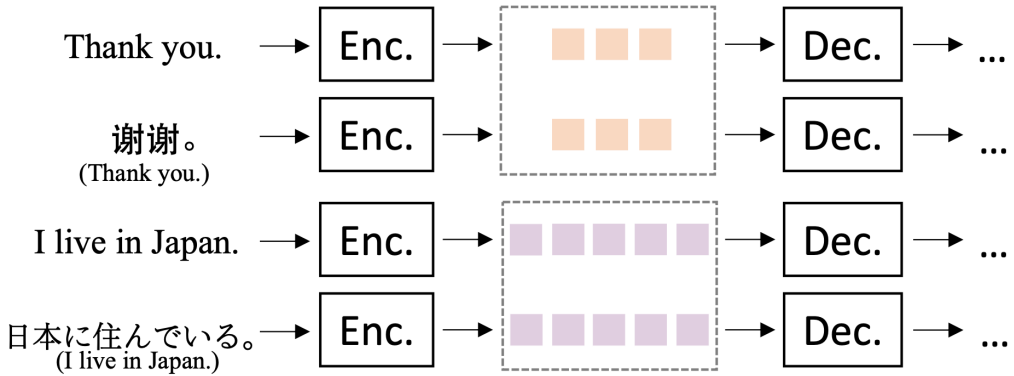
VIM: ① Interlingua Encoder Layers

- **Encoder-interlingua attention (cross-attention) modules** to bridge encoder layers and interlingua encoder layers, following Zhu et al. (2020).



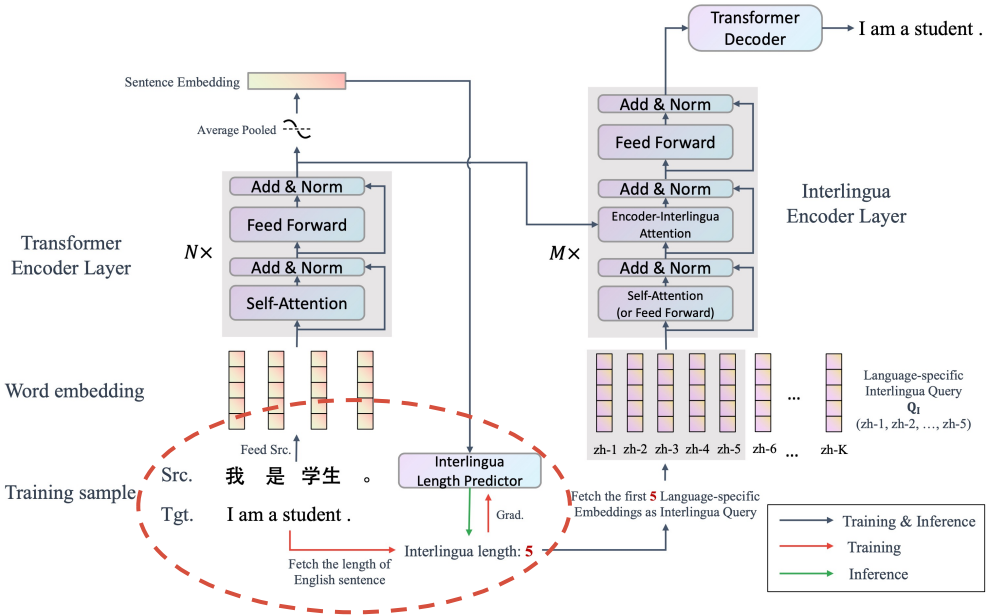
VIM: ① Interlingua Encoder Layers

- **An identical interlingua representation sequence** for sentences in different languages **with the same meaning**, where **length can vary** for different meanings.
- We use **length of sentence in centric language** as length of interlingua representations for simplicity.



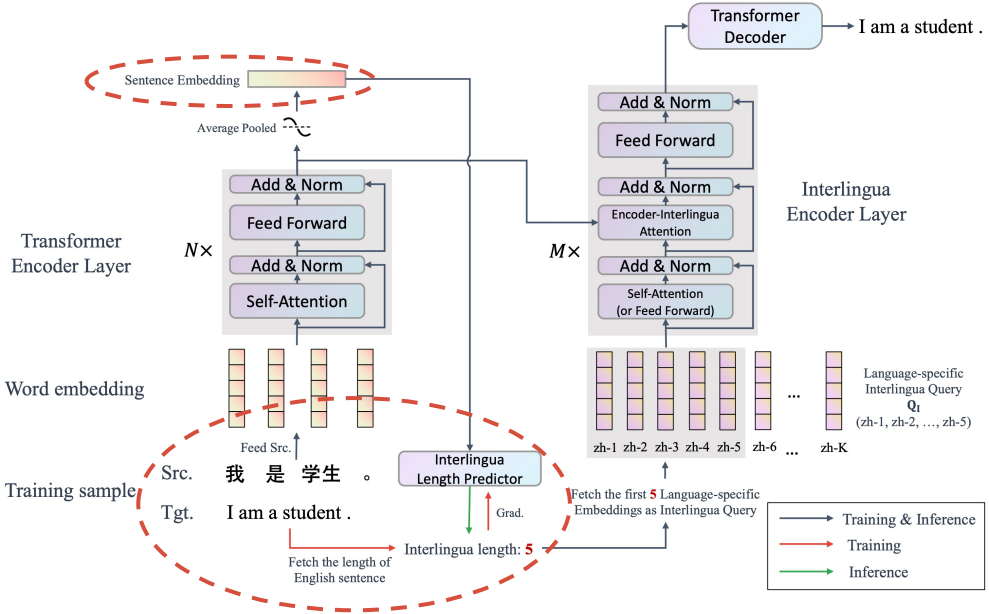
VIM: ② Interlingua Length Prediction

- We use English-centric datasets for training:
  - **Length of interlingua is defined as length of English.**
- **A interlingua length predictor** for predicting length of interlingua during inference.



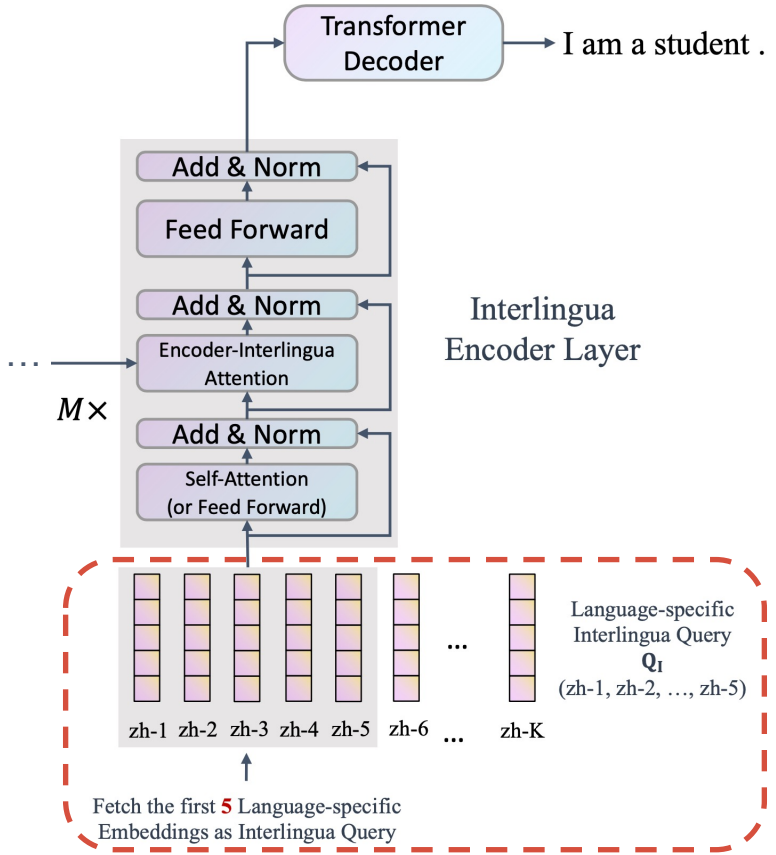
VIM: ② Interlingua Length Prediction

- Length predictor predicts interlingua length using **sentence embedding**.
- **Training**: Use length of English to calculate **a classification loss**.
- **Inference**: Use length predictor to **generate interlingua length**.



VIM: ③ Language-specific Interlingua Query

- Initialize **language-specific embeddings as interlingua queries** as inputs for interlingua encoder layers.
- Use the first x embeddings according to the intended interlingua length x.



Training Objectives

- Given a sentence pair (X, Y), we compute the loss by combining **NMT loss, interlingua alignment loss, and length prediction loss**.

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

- Interlingua alignment loss follows Zhu et al. (2020), which is a position-wise cosine similarity loss for **aligning interlingua embeddings for each sentence pair**:

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

where $H_I(\cdot)_i$ denotes the i-th column of interlingua encoder outputs.

Datasets

- 3 Datasets and 48 zero-shot translation directions in total.

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

"# 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.

Baselines

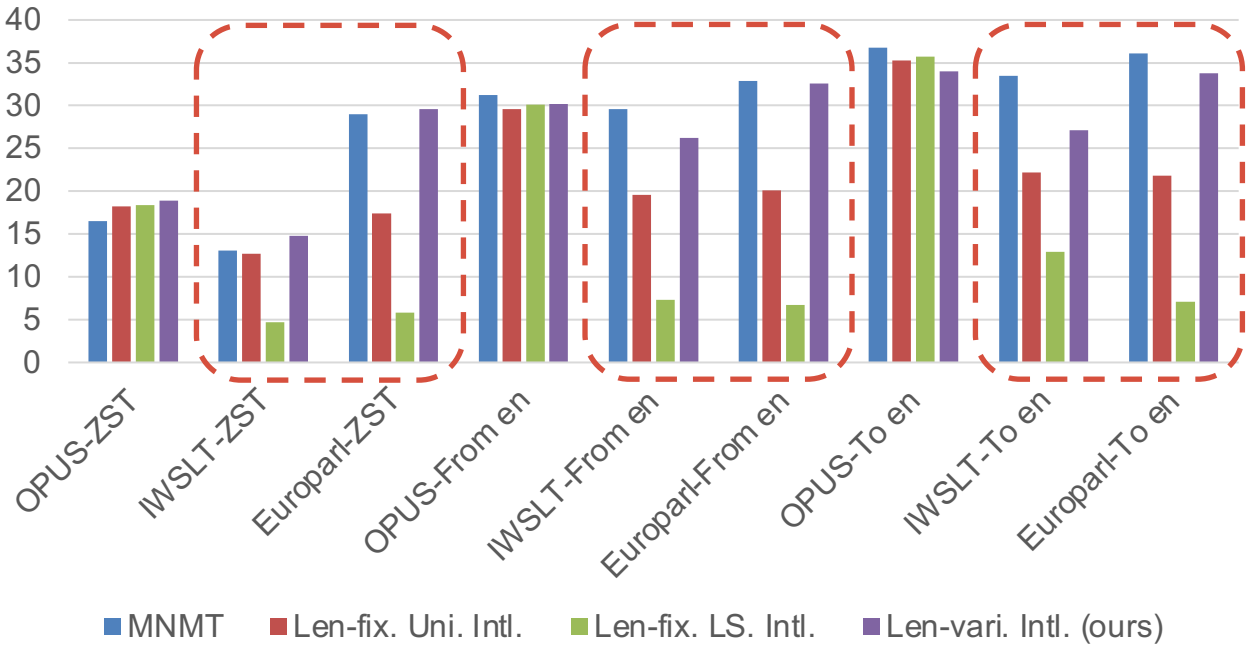
1. **MNMT**
   - Vanilla multilingual neural machine translation.
2. **Len-fix. Uni. Intl.** (Zhu et al., 2020)
   - Fixed-length neural interlingua method following Zhu et al. (2020).
3. **Len-fix. LS. Intl.**
   - We use language-specific interlingua queries for Len-fix. Uni. Intl.
4. **Len-vari. Intl.** (ours)
   - Our proposed variable-length interlingua method.

Results and Analysis

- BLEU results
- Validation NMT loss
- Impact of the interlingua length predictor

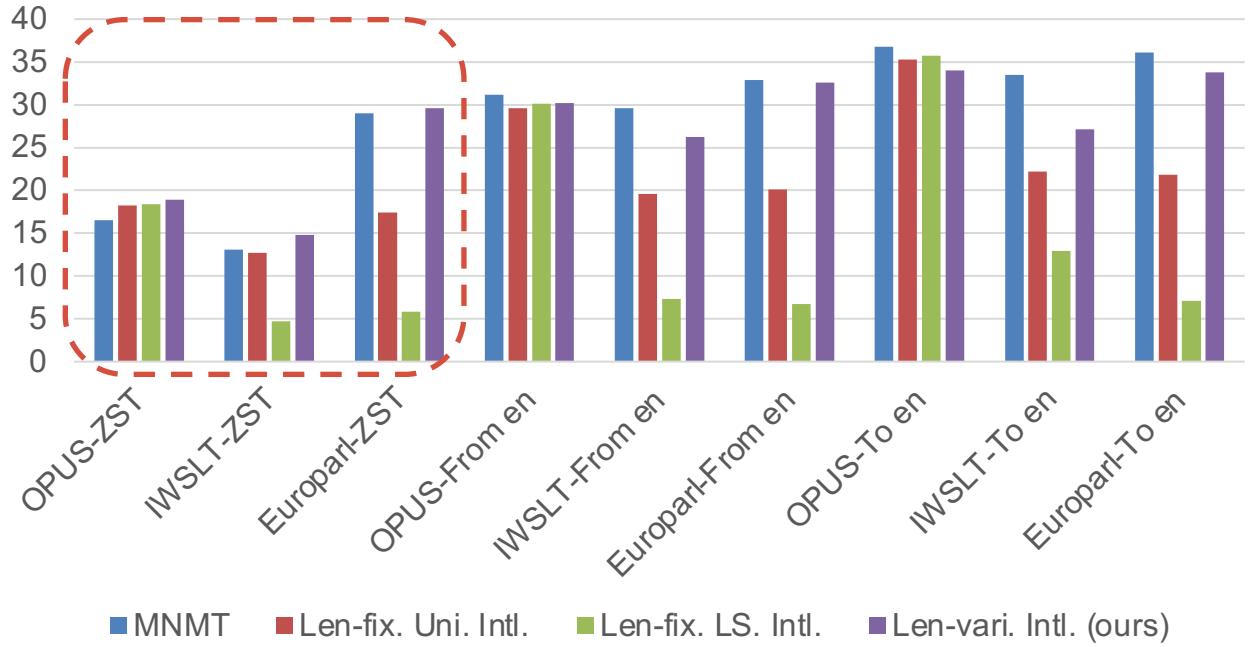
BLEU Results: An Overview

- **Previous interlingua-based methods are not stable** in low-resource datasets and Transformer-big model.



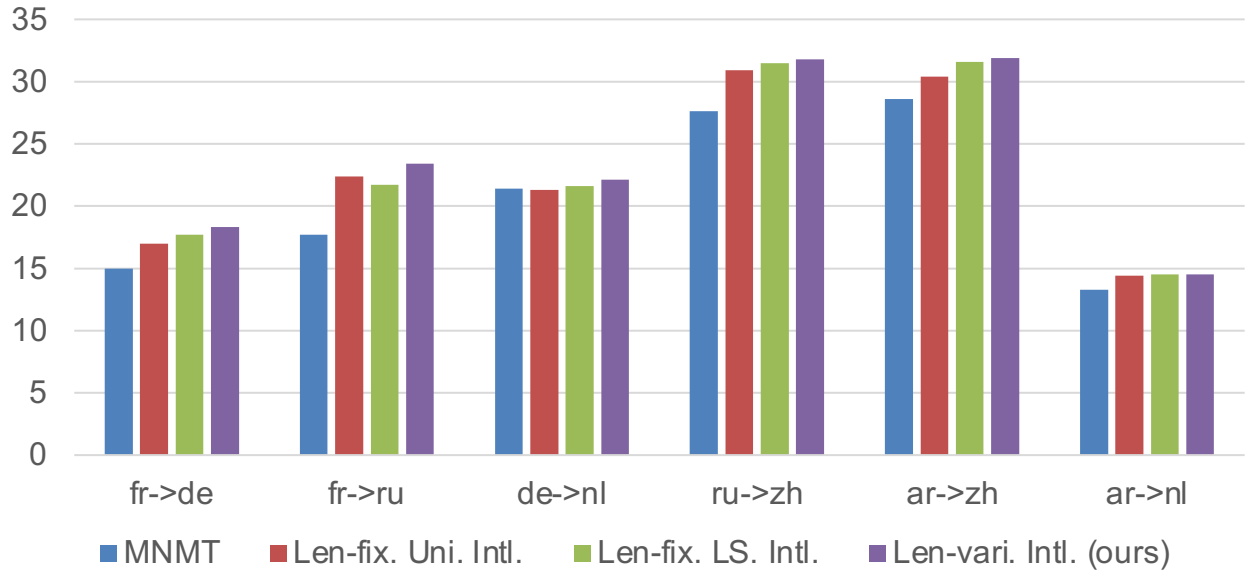
BLEU Results: An Overview

- Our method outperforms previous methods **for zero-shot translation**.



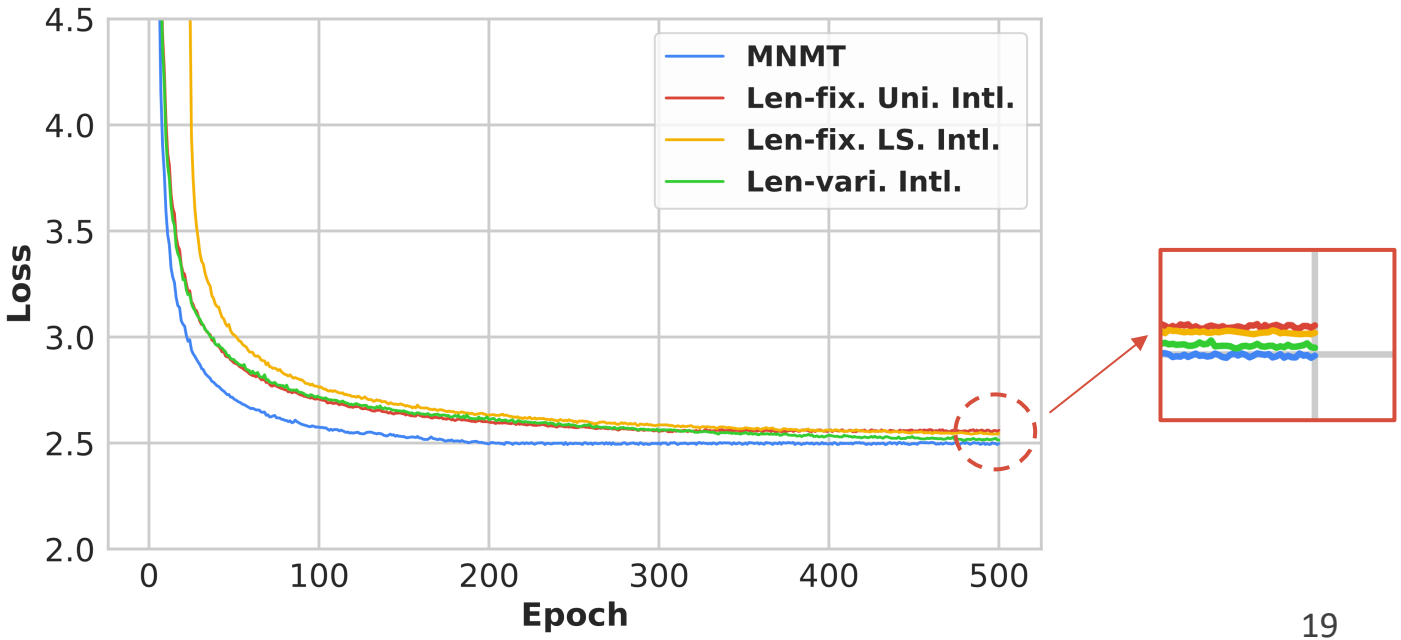
Some Specific BLEU Results on OPUS

- On OPUS, our proposed method performs significantly better than previous methods for zero-shot translation.



Validation NMT Loss

- Our method can achieve similar validation loss as a vanilla MNMT model. (validation loss is calculated for all the supervised directions)



Impact of the Interlingua Length Predictor

- Although our method can obtain improvements for zero-shot translation, we got **worse performance for “to en” translation**.

Impact of the Interlingua Length Predictor

- Although our method can obtain improvements for zero-shot translation, we got **worse performance for "to en" translation**.
- This is due to the incorrect length prediction of length predictor.
  - **The discrepancy of around 3 tokens** between the predicted lengths and the gold labels in average.

	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†

Impact of the Interlingua Length Predictor

- Although our method can obtain improvements for zero-shot translation, we got **worse performance for "to en" translation**.
- This can be attributed to the incorrect length prediction of length predictor.
  - **The discrepancy of around 3 tokens** between the predicted lengths and the gold labels in average.
  - Translation is improved by **giving model the correct interlingua length**.

	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†

Conclusion

- A novel **variable-length neural interlingua** approach.
- **Improved and stabilized zero-shot translation** results.
- Future work can focus on improving the accuracy of **length predictor**.