Breaking Language Barriers: Enhancing Multilingual Representation for Sentence Alignment and Translation

Zhuoyuan Mao

Jun 20, 2024

Zhuoyuan Mao, 毛 卓遠

Education

- 2013/8 – 2017/7: BSc, Mathematics, East China University of Science and Technology
- 2019/4 – 2021/3: MInf, Intelligence Science and Technology, Kyoto University
- 2021/4 – 2024/3: PhD, Intelligence Science and Technology, Kyoto University

Research Career

- 2022/2 – 2022/3: Research Internship, SenseTime Japan
- 2022/4 – 2024/3: JSPS Research Fellowship for Young Scientists, DC2
- 2022/8 – 2022/10: Student Researcher, Google Japan
- 2023/6 – 2023/9: Research Internship, Apple Japan
- 2024/4 – Present: Full-time Researcher, Sony Group Corporation

Outline

- Introduction to Multilingual Representation Learning
- Key Challenges and Our Proposals
  - High Computation Demands: *EMS* and *LEALLA*.
  - Data Scarcity: *JASS*+*ENSS*, *WCL*, and *AlignInstruct*.
  - Limitations in Transformer Architecture: *INTL* and *LayerNorm*.
- Contributions, Discussions, Conclusion and Future Prospects

Introduction to Multilingual Representation Learning

The Diversity of Global Languages

- Our world is home to over 7,100 languages and 147 language families, each representing a unique culture and community.
- However, this diversity often leads to communication barriers among people who speak different languages.

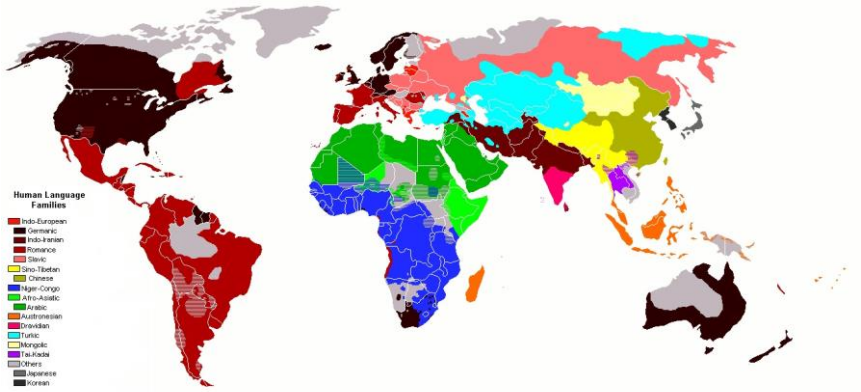
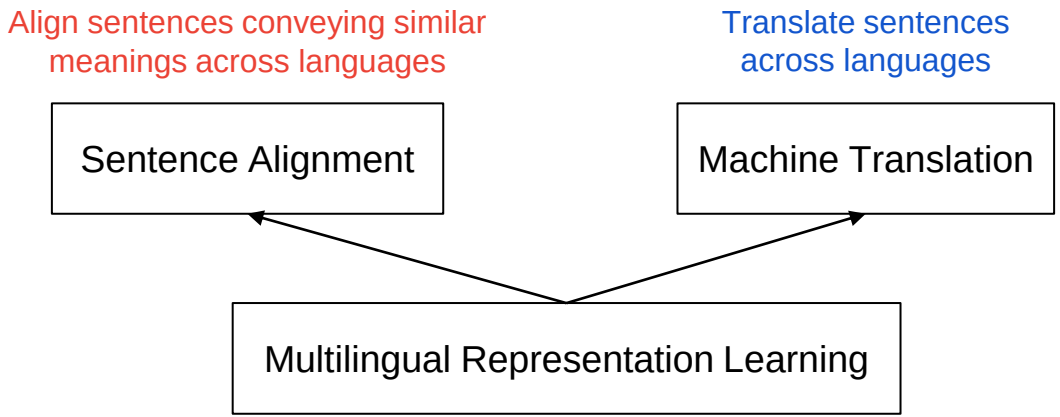


Figure from Wikipedia.

Breaking Language Barriers

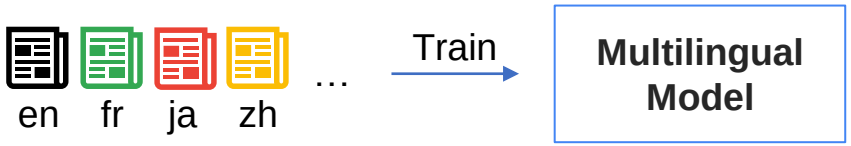
- To bridge these linguistic divides, **sentence alignment** and **machine translation** techniques emerge as essential tools, enabling clearer and more effective cross-lingual communication.



- **Multilingual representation learning** is the core NLP technique that supports the sentence alignment and translation tasks.

What is Multilingual Representation Learning?

- Train one universal neural model for multiple languages
  (Johnson et al., 2017, Artetxe and Schwenk, 2019)
  - No need to train separate models for each language
  - Benefit from knowledge transfer across languages



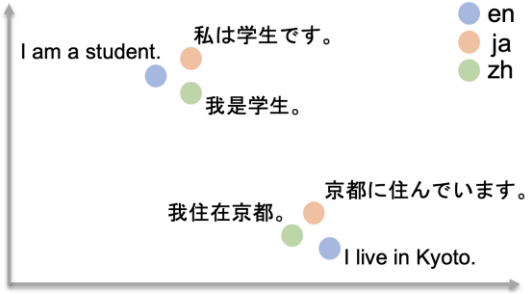
Multilingual Sentence Embedding Learning

Sentence embedding:

I am a student . -> [0.013, -0.065, …, 0.043]

Multilingual sentence embedding:

I am a student. -> [0.013, -0.065, …, 0.043]
私は学生です. -> [0.014, -0.063, …, 0.041]



Aligning sentences in a shared embedding space

Initialize

Zero-shot cross-lingual transfer on downstream tasks (document classification, etc.):

Training: US Professional Sports Teams Promote Vaccinations

Inference: 新奇超薄膜「二次元材料」の成長と評価

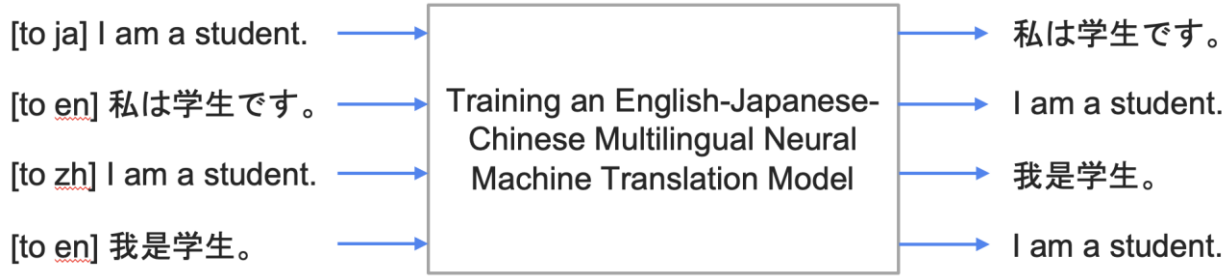
Model for downstream task e.g. classification

News

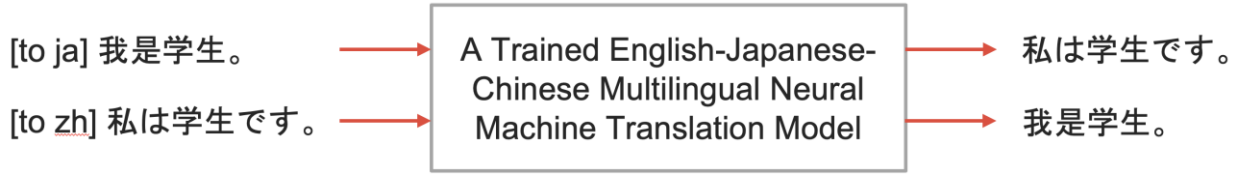
Science

Multilingual Neural Machine Translation

- Translation for multiple language pairs with a single neural model. (Johnson et al., 2016)
- The target language can be specified with a language tag (e.g., [to ja], [to en]).
- Multilingual NMT can benefit low-resource languages through cross-lingual transfer.

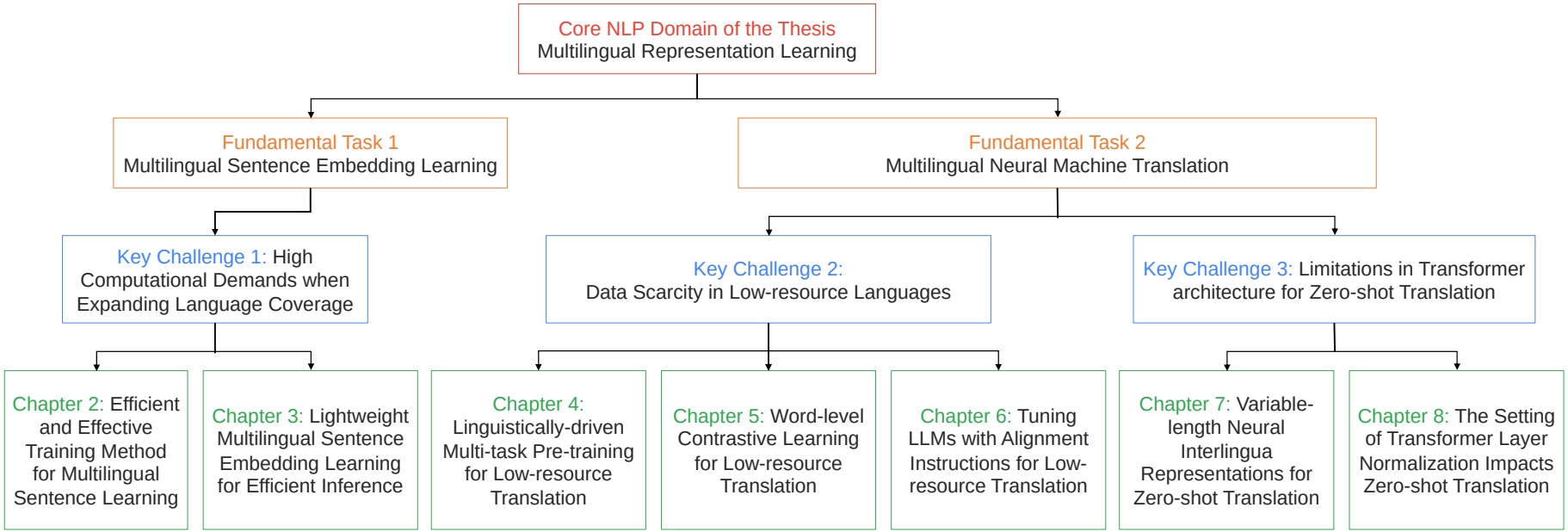


- **Zero-shot translation**
  - Translation for unseen language pairs without training data.
  - Promising for its low latency compared with pivot-based translation.



Key Challenges and Our Proposals in Multilingual Representation Learning

An Overview


Core NLP Domain of the Thesis
Multilingual Representation Learning
Fundamental Task 1
Multilingual Sentence Embedding Learning
Fundamental Task 2
Multilingual Neural Machine Translation
Key Challenge 1: High Computational Demands when Expanding Language Coverage
Key Challenge 2: Data Scarcity in Low-resource Languages
Key Challenge 3: Limitations in Transformer architecture for Zero-shot Translation
Chapter 2: Efficient and Effective Training Method for Multilingual Sentence Learning
Chapter 3: Lightweight Multilingual Sentence Embedding Learning for Efficient Inference
Chapter 4: Linguistically-driven Multi-task Pre-training for Low-resource Translation
Chapter 5: Word-level Contrastive Learning for Low-resource Translation
Chapter 6: Tuning LLMs with Alignment Instructions for Low-resource Translation
Chapter 7: Variable-length Neural Interlingua Representations for Zero-shot Translation
Chapter 8: The Setting of Transformer Layer Normalization Impacts Zero-shot Translation
ACL 2021, IEEE/ACM TASLP
EACL 2023
LREC 2020, ACM TALLIP
NAACL 2022 Findings
Submitted to LoResMT 2024 Workshop
Multi3Generation 2023 Workshop
ACL 2023

Key Challenges in Multilingual Representation Learning

① High computational demands

- Computational demands escalate as we extend multilingual models to **accommodate additional languages**.
- How to address this challenge **has not been explored so far** in the context of multilingual sentence embedding learning.

EMS: Efficient and Effective Massively Multilingual Sentence Representation Learning (ACL 2021 & IEEE/ACM TASLP)

Goal

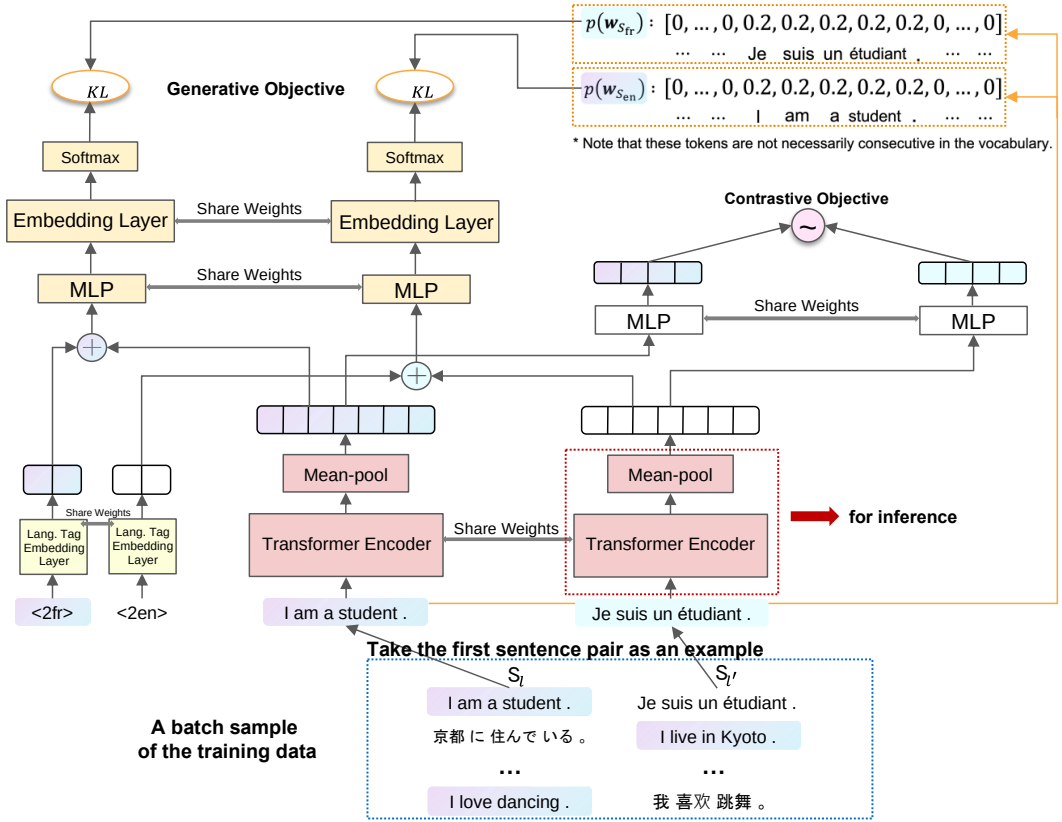
- Develop an **efficient and effective** method for **training** multilingual sentence embedding.

Proposals

- Proposed **effective generative and contrastive objectives** with **decreased amount** of **data** and **computation overhead**.

Results

- Achieved significant **training acceleration** compared with previous work.
- Obtained **SOTA performance** on **six sentence alignment tasks**.



LEALLA: Learning Lightweight Language-agnostic Sentence Embeddings with Knowledge Distillation (EACL 2023)

Goal

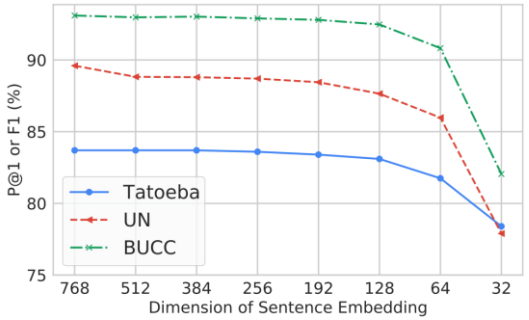
- Develop **lightweight** language-agnostic sentence embedding models based on LaBSE (Feng et al., 2022), for **faster inference**.

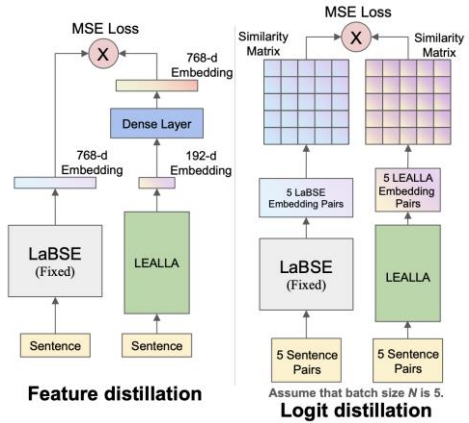
Explorations and Proposals

- Explored the **optimal dimension** of language-agnostic sentence embeddings.
- Explored the **optimal model architecture**.
- Proposed **two distillation methods** to enhance the lightweight model.

Results

- Obtained **SOTA** performance for cross-lingual sentence retrieval for **109 languages**.
- Released LEALLA models on TFHub: https://tfhub.dev/google/collections/LEALLA/1

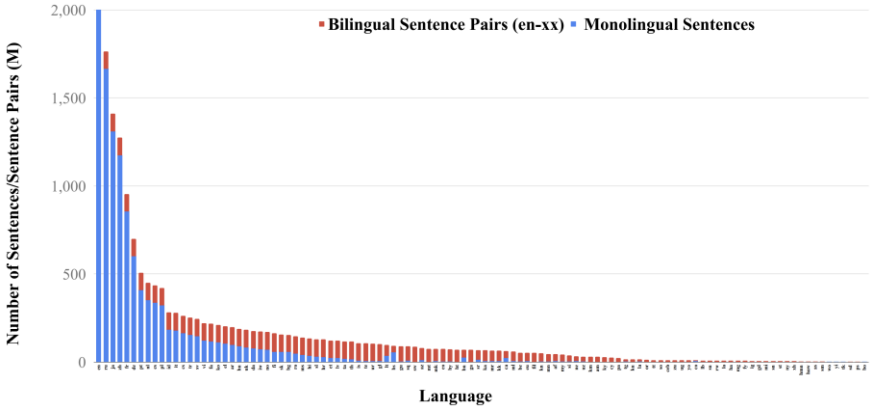




Model	Languages	Dimension	Parameters	Tatoeba	UN	BUCC
LASER (2019b)	93	1024	154M	65.5	-	**93.0**
m-USE (2020)	16	512	**85M**	-	84.3	87.7
SBERT (2020)	50	768	270M	67.1	-	88.6
EMS (2022)	62	1024	148M	69.2	-	91.7
LaBSE (2022)	**109**	768	471M	**83.7**	**89.6**	**93.1**
LEALLA-small	**109**	**128**	**69M**	80.7	87.3	91.5
LEALLA-base	**109**	**192**	**107M**	**82.4**	**88.7**	92.4
LEALLA-large	**109**	**256**	147M	**83.5**	**89.3**	**92.8**

Key Challenges in Multilingual Representation Learning

② Data scarcity



- The scarcity of training data in **low-resource languages** often leads to inferior performance of multilingual models for these languages.
- Prior research has focused on pre-training, cross-lingual transfer, and data augmentation for low-resource languages. Our approach advances this by utilizing more **fine-grained linguistic features or alignments** to enhance performance for low-resource languages.

ENSS & JASS: Linguistically-driven Multi-task Pre-training for Low-resource Neural Machine Translation (LREC 2020 & ACM TALLIP)

Goal

- Improve sequence-to-sequence **pre-training for NMT in low-resource languages**.

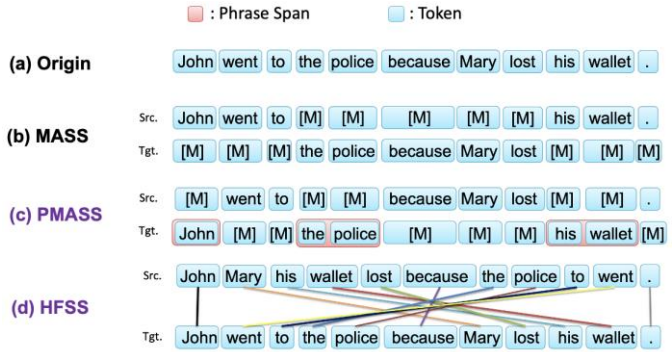
Proposals

- **PMASS** and **HFSS**: **Linguistically-driven** pre-training methods for NMT involving English.
- **BMASS** and **BRSS**: **Linguistically-driven** pre-training methods for NMT involving Japanese.

Results

- Achieved **up to 7.0 BLEU improvements** on 4 datasets in low-resource scenarios.

ENSS: PMASS & HFSS



JASS: BMASS & BRSS

WCL: When do Contrastive Word Alignments Improve Many-to-many Neural Machine Translation? (NAACL 2022 Findings)

Goal

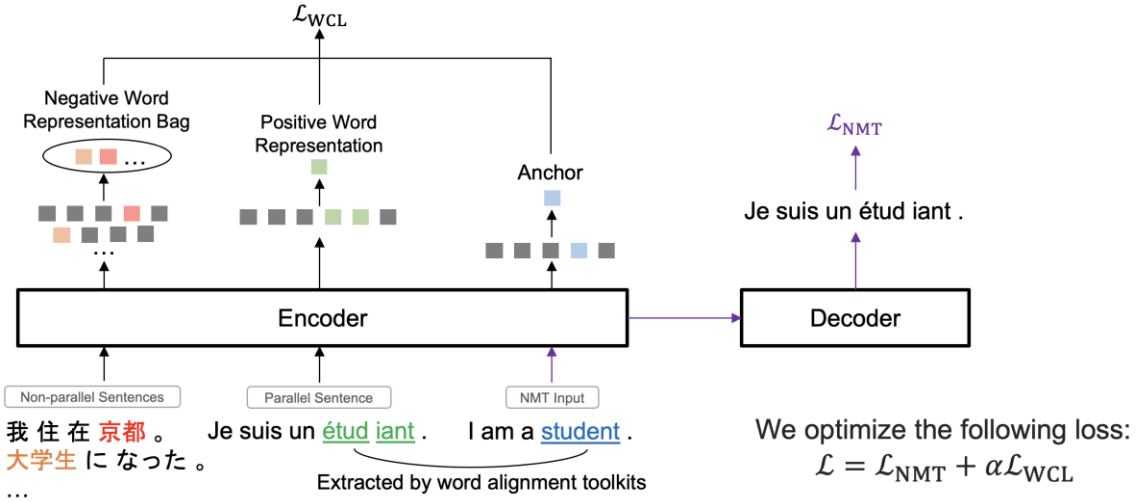
- Improve **encoder-side representations for low-resource languages** in multilingual NMT.

Proposals

- Proposed **word-level contrastive learning** to augment and improve the cross-lingual signals for multilingual NMT.

Results

- Improved the translation quality on **several low-resource language pairs**, which comes from **better aligned encoder representations**.



AlignInstruct: Tuning LLMs with Contrastive Alignment Instructions for Machine Translation in Unseen, Low-resource Languages (Submitted to LoResMT 2024)

Goal

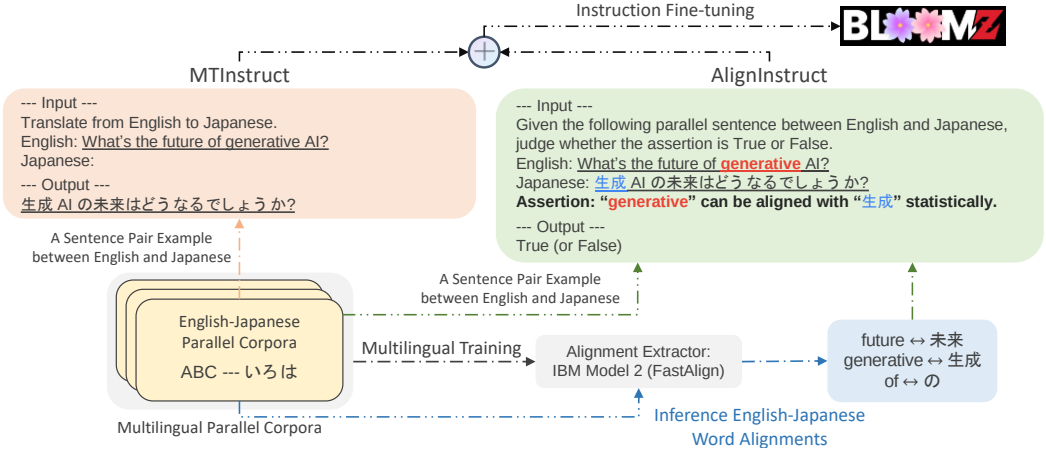
- Adapt multilingual **LLMs** to **unseen, low-resource languages** in **MT** tasks with **enhanced cross-lingual signals**.

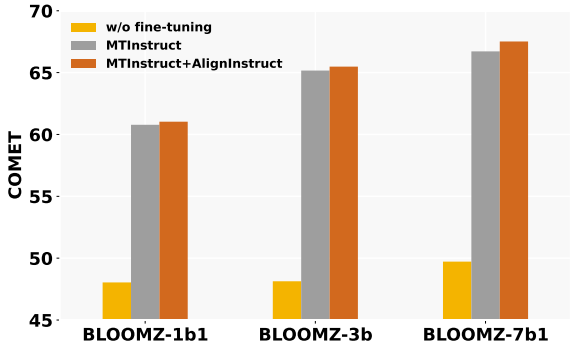
Explorations and Proposals

- Proposed **AlignInstruct for enhancing cross-lingual signals** for better aligning low-resource languages in **MT instruction fine-tuning**.

Results

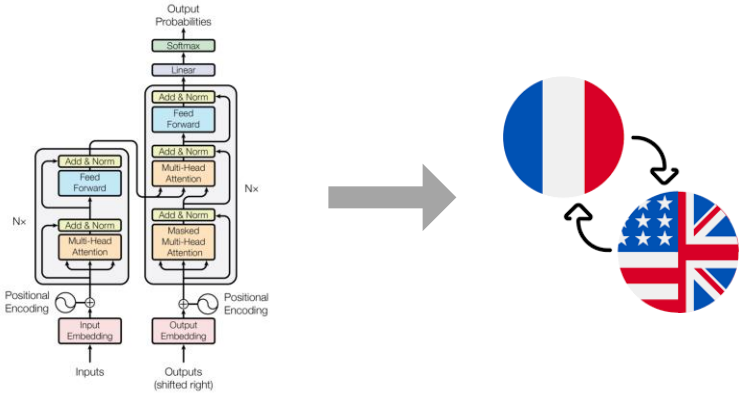
- **AlignInstruct** led to consistent improvements in translation quality across **48 translation directions involving English and 30 zero-shot directions**.





Key Challenges in Multilingual Representation Learning

③ Limitations in Transformer architecture



Transformer is initially designed for the English—French bilingual translation task.

- The Transformer architecture, initially designed for bilingual translation, **may not be ideally suited for multilingual contexts**.
- Addressing this architecture's limitations in zero-shot translation **was underexplored**, with only two studies by Zhu et al. (2020) and Liu et al. (2021) delving into it.

Variable-length Neural Interlingua Representations for Zero-shot Neural Machine Translation (Multi3Generation 2023 Workshop)

Goal

- Figure out **a better Transformer architecture** for zero-shot NMT.

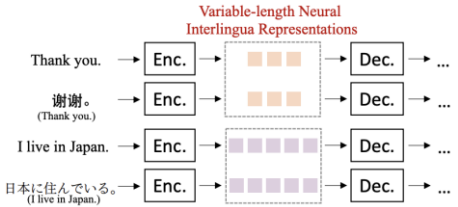
Proposals

- Proposed **variable-length interlingua representations** to construct language-agnostic representation sequence with variable length.

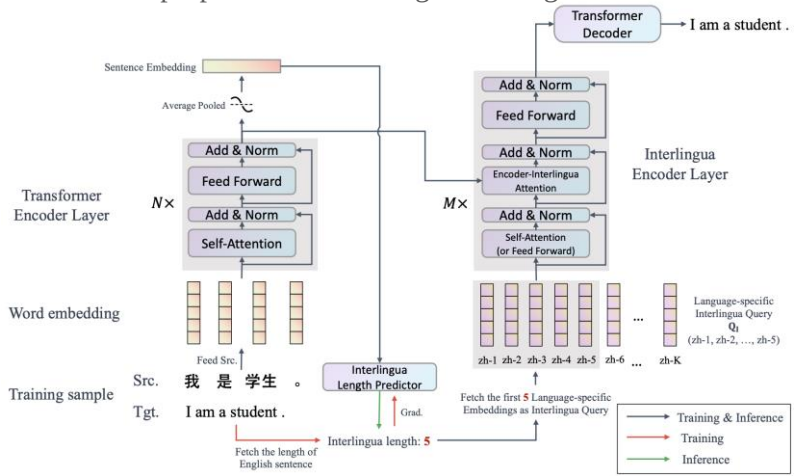
Results

- **Improved** and **stabilized** translation quality in **48 zero-shot translation directions** on three datasets.

An overview of the proposed variable-length interlingua representations.



The proposed variable-length interlingua module.



PostNorm: Exploring the Impact of Layer Normalization for Zero-shot Neural Machine Translation (ACL 2023)

Goal

- Figure out the **Transformer component that impacts zero-shot NMT**.

Explorations and Proposals

- Explored PostNorm in **different language tag and residual connection settings**.
- Proposed **LLR score** to analyze encoder and decoder representations for zero-shot NMT.

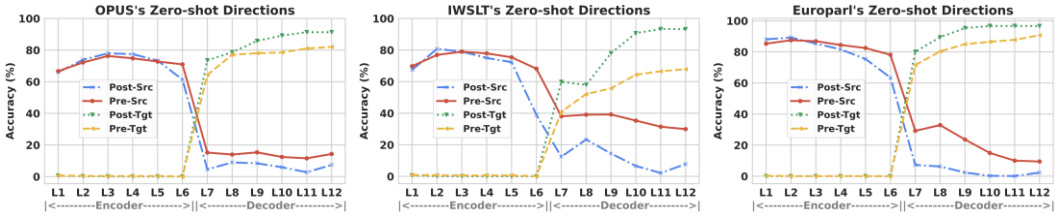
Results

- **PostNorm was consistently better** than PreNorm across various settings.
- LLR revealed that PostNorm leads to **better aligned encoder representations** and **more language-specific decoder representations**, which benefited zero-shot translation.

BLEU scores (and off-target rates) of PreNorm vs. PostNorm across different language tag and residual connection settings.

#	Layer Norm	Language Tag	Res.	Zero-shot OPUS	Zero-shot IWSLT	Zero-shot Europarl	Supervised OPUS	Supervised IWSLT	Supervised Europarl
0		*Pivot*		21.8	20.0	29.5	-	-	-
1	**PreNorm**	S-ENC-T-DEC	w/	10.1 (42.19%)	4.9 (64.84%)	24.9 (7.73%)	33.7	31.5	34.3
2	**PostNorm**	S-ENC-T-DEC	w/	**16.8** (**8.59%**)	**12.4** (**10.61%**)	**29.2** (**0.34%**)	33.9	31.5	34.5
3	**PreNorm**	T-ENC	w/	13.3 (22.99%)	13.7 (**3.98%**)	29.5 (0.23%)	33.7	31.6	34.4
4	**PostNorm**	T-ENC	w/	**14.0** (**22.86%**)	**15.5** (4.59%)	**30.8** (**0.11%**)	34.1	31.5	34.5
5	**PreNorm**	S-ENC-T-DEC	w/o	14.3 (20.67%)	8.0 (50.16%)	16.7 (41.87%)	33.6	30.9	34.3
6	**PostNorm**	S-ENC-T-DEC	w/o	**16.0** (**15.27%**)	**17.4** (**1.83%**)	**29.0** (**0.41%**)	33.8	30.7	34.4
7	**PreNorm**	T-ENC	w/o	13.4 (27.15%)	16.2 (1.54%)	29.9 (2.15%)	33.5	30.9	34.3
8	**PostNorm**	T-ENC	w/o	**13.9** (**26.68%**)	**17.8** (**1.50%**)	**30.8** (**0.13%**)	33.9	30.6	34.4

LLR scores for recognizing source or target languages using PreNorm or PostNorm.



Conclusion and Future Prospects

Conclusion

- We explored methods to tackle the three challenges in multilingual representation learning, focusing on the sentence alignment and translation tasks.
- The methods proposed offer strategies for training more efficient and effective multilingual models and can potentially benefit the research pertaining to multilingual LLMs, thus broadening the reach of NLP techniques to a wider audience.

Future Prospects

- Further expanding language coverage
  - Low-resource languages, regional dialects, etc.
- Confirming the effectiveness of the proposed methods for LLMs
- Integration with multimodal data
  - Images and videos as universal pivots for further improving multilingual models.
- Cross-culture understanding
  - Detecting and addressing cultural nuances.

Thanks for your listening!

Publication List

[1] Zhuoyuan Mao, Fabien Cromieres, Raj Dabre, Haiyue Song and Sadao Kurohashi. JASS: Japanese-specific Sequence to Sequence Pre-training for Neural Machine Translation. In Proceedings of the 12th Language Resources and Evaluation Conference, pages 3683–3691, 2020.

[2] Zhuoyuan Mao, Prakhar Gupta, Chenhui Chu, Martin Jaggi and Sadao Kurohashi. Lightweight Cross-Lingual Sentence Representation Learning. In Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 2902–2913, 2021.

[3] Zhuoyuan Mao, Chenhui Chu, Raj Dabre, Haiyue Song, Zhen Wan and Sadao Kurohashi. When do Contrastive Word Alignments Improve Many-to-many Neural Machine Translation? In Findings of the Association for Computational Linguistics: NAACL 2022, pages 1766–1775, 2022.

[4] Zhuoyuan Mao, Chenhui Chu and Sadao Kurohashi. Linguistically-driven Multi-task Pre-training for Low-resource Neural Machine Translation. In ACM Trans. Asian Low-Resour. Lang. Inf. Process., 21, 4, Article 68 (Jul. 2022), 29 pages.

[5] Zhuoyuan Mao and Tetsuji Nakagawa. LEALLA: Learning Lightweight Language-agnostic Sentence Embedding with Knowledge Distillation. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 1886–1894, 2023.

Publication List

[6] Zhuoyuan Mao, Haiyue Song, Raj Dabre, Chenhui Chu and Sadao Kurohashi. Variable-length Neural Interlingua Representations for Zero-shot Neural Machine Translation. In the 1st Workshop on Multilingual, Multimodal and Multitask Language Generation, pages 16–25, 2023.

[7] Zhuoyuan Mao, Raj Dabre, Qianying Liu, Haiyue Song, Chenhui Chu and Sadao Kurohashi. Exploring the Impact of Layer Normalization for Zero-shot Neural Machine Translation. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, pages 1300–1316, 2023.

[8] Zhuoyuan Mao and Yen Yu. Tuning LLMs with Contrastive Alignment Instructions for Machine Translation in Unseen, Low-resource Languages. Submitted to LoResMT 2024 Workshop, 2023.

[9] Zhuoyuan Mao, Chenhui Chu and Sadao Kurohashi. EMS: efficient and effective massively multilingual sentence representation learning. In IEEE ACM Trans. Audio Speech Lang. Process. 32: 2841-2856, 2024.