Multi-Task Learning for Multilingual Neural Machine Translation

Presenter: Wenxiang Jiao

2021-06-08

The Ultimate Quest of Machine Translation

- # of human languages: > 6900
- How to build a universal MT system that is capable of translating any source language into a target one?

Multilingual NMT is Appealing

Advantages

- Multilingual NMT is model efficient.
- Parameter sharing across languages encourages knowledge transfer, benefiting low-resource and zero-shot translations.

Problems

- Existing multilingual NMT approaches often do not effectively utilize the abundance of monolingual data.
- No explicit signals for closing the representation gap between languages.

Recent Advances

Leveraging Monolingual Data with Self-Supervision for Multilingual Neural Machine Translation. ACL 2020 (Short). Google Research.

Contrastive Learning for Many-to-Many Multilingual Neural Machine Translation. ACL 2021. ByteDance AI Lab.

Leveraging Monolingual Data with Self-Supervision for Multilingual Neural Machine Translation

ACL 2020 (short)

Aditya Siddhant, Ankur Bapna, Yuan Cao, Orhan Firat, Mia Chen,
Sneha Kudugunta, Naveen Arivazhagan, Yonghui Wu
Google Research

Contributions and Findings

- Propose to train a multilingual NMT model with the translation objective on parallel data and the MASS objective on monolingual data.
- Using monolingual data significantly boosts the translation quality of low-resource languages in multilingual models.
- Self-supervision improves zero-shot translation quality in multilingual models.

Motivation

- Monolingual data is massively available, even for low-resource translation tasks.

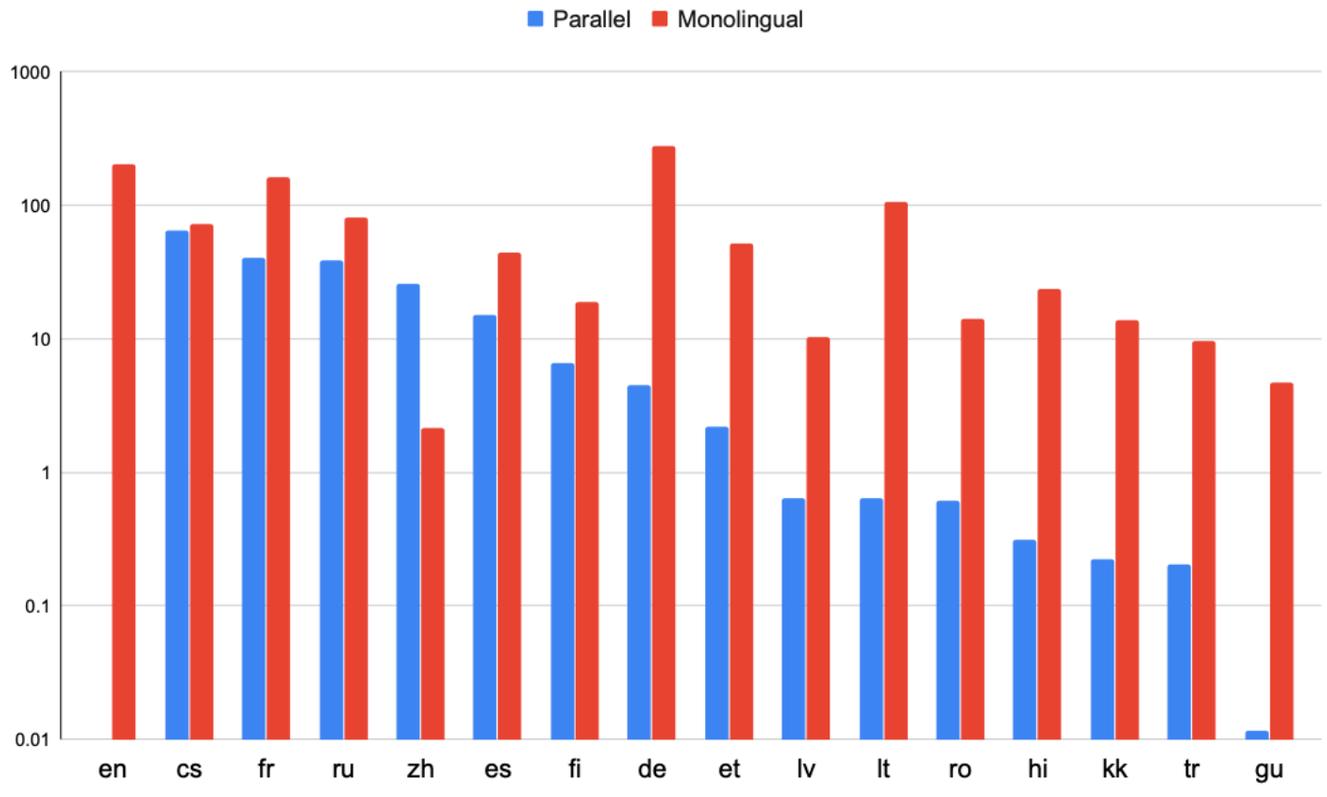


Figure 1: Number of parallel and monolingual training samples in millions for each language in WMT training corpora.

MASS Objective

- MASS: Masked sequence-to-sequence pre-training.

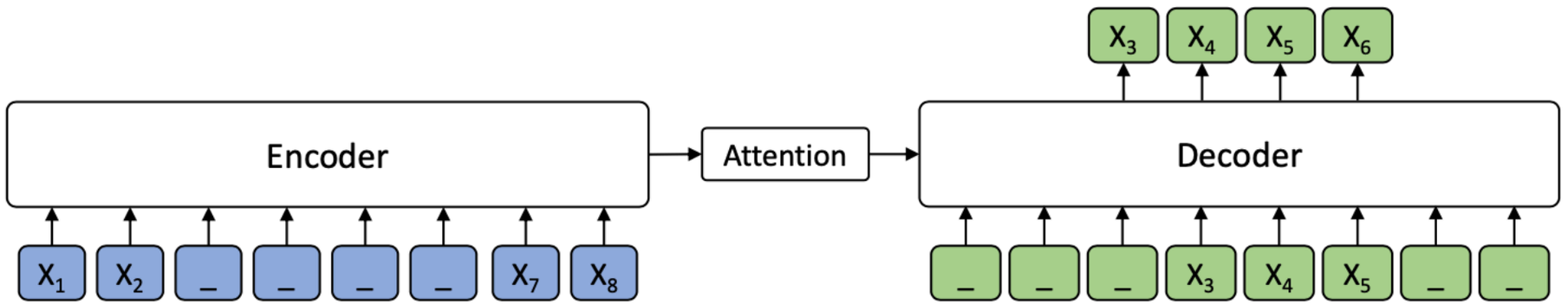


Figure 1. The encoder-decoder framework for our proposed MASS. The token "_" represents the mask symbol [M].

MASS objective

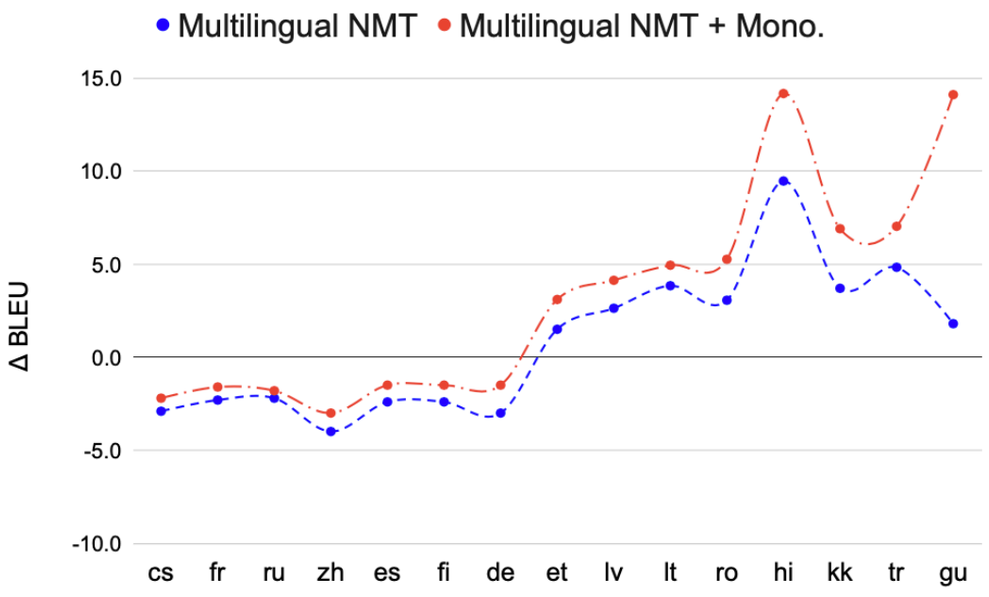
$X_1\ X_2\ \_\ \_\ \_\ \_\ X_7\ X_8 \longrightarrow \_\ \_\ X_3\ X_4\ X_5\ X_6\ \_\ \_$

Translation objective

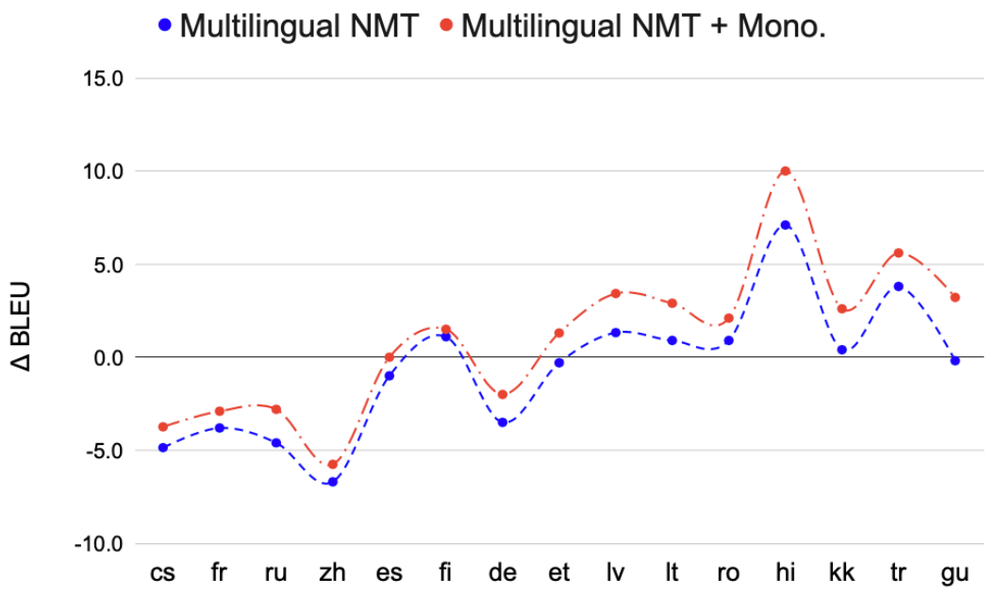
$X_1\ X_2\ X_3\ X_4\ X_5\ X_6\ X_7\ X_8 \longrightarrow Y_1\ Y_2\ Y_3\ Y_4\ Y_5\ Y_6\ Y_7\ Y_8$

Indigenous Language Pairs

- Adding additional monolingual data improves the multilingual model quality across the board, even for high-resource language pairs.



(a) Any-to-English (xx → en)



(b) English-to-Any (en → xx)

Figure 2: Translation quality of Multilingual NMT models relative to bilingual baselines with and without monolingual data. The left plot shows xx → en direction and right one shows en → xx direction. From left to right on x-axis, we go from high-resource to low-resource languages. The x-axis reflects the bilingual baselines.

Zero-Shot Language Pairs

- Adding additional monolingual data improves the multilingual model quality across the board, even for high-resource language pairs.

		fr_de	de_fr	cs_de	de_cs
4 lang.	w/ Parallel Data	27.7	35.3	—	—
	Translation via Pivot	21.9	29.2	20.4	19.0
	Arivazhagan et al. (2019a)	20.3	26.0	—	—
	Kim et al. (2019)	17.3	—	—	14.1
	Multilingual NMT	11.8	15.2	12.3	8.2
	Multilingual NMT + Mono.	18.5	27.2	16.9	12.6
30 lang.	Multilingual NMT	10.3	14.2	10.5	4.3
	Multilingual NMT + Mono.	16.6	22.3	14.8	7.9

Table 2: Zero-shot performance on non-English centric language pairs. We compare with pivot-based translation and two recent approaches from Arivazhagan et al. (2019a) and Kim et al. (2019). The translation quality between these language pairs when parallel data is available is also provided as a baseline. 4 lang. is a multilingual model trained on 4 language pairs (2 languages to and from English), while 30 lang. is our multilingual model trained on all English-centric language pairs.

Contrastive Learning for Many-to-Many Multilingual Neural Machine Translation

ACL 2021

Xiao Pan, Mingxuan Wang, Liwei Wu, Lei Li
ByteDance AI Lab

Contributions and Findings

- Propose a contrastive learning scheme to close the gap among representations of different languages.
- Propose aligned augmentation on both multiple parallel data and monolingual data.
- Achieve significant improvements on supervised, unsupervised, and zero-shot translations.

mCOLT Framework

- mCOLT: Multilingual contrastive learning framework for translation.

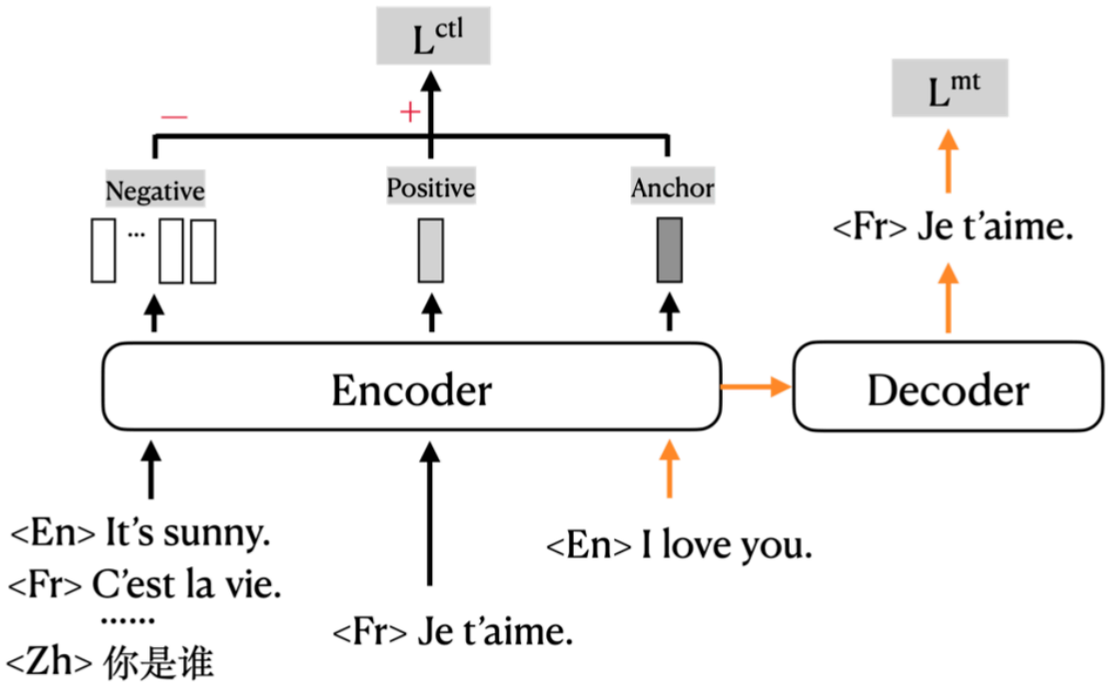


Figure 1: The proposed mCOLT. It takes a pair of sentences (or augmented pseudo-pair) and computes normal cross entropy loss with a multi-lingual encoder-decoder. In addition, it computes contrastive loss on the representations of the aligned pair (positive example) and randomly selected non-aligned pair (negative example).

Translation objective

$$\mathcal{L}^{\text{mt}} = \sum_{\mathbf{x}^i, \mathbf{x}^j \in \mathcal{D}} -\log P_\theta(\mathbf{x}^i | \mathbf{x}^j)$$

Contrastive learning objective

$$\mathcal{L}^{\text{ctl}} = -\sum_{\mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}} \log \frac{e^{\text{sim}^+(\mathcal{R}(\mathbf{x}^i), \mathcal{R}(\mathbf{x}^j))/\tau}}{\sum_{\mathbf{y}^j} e^{\text{sim}^-(\mathcal{R}(\mathbf{x}^i), \mathcal{R}(\mathbf{y}^j))/\tau}}$$

Aligned Augmentation

- Perturb the source sentence by replacing aligned words from a synonym dictionary.

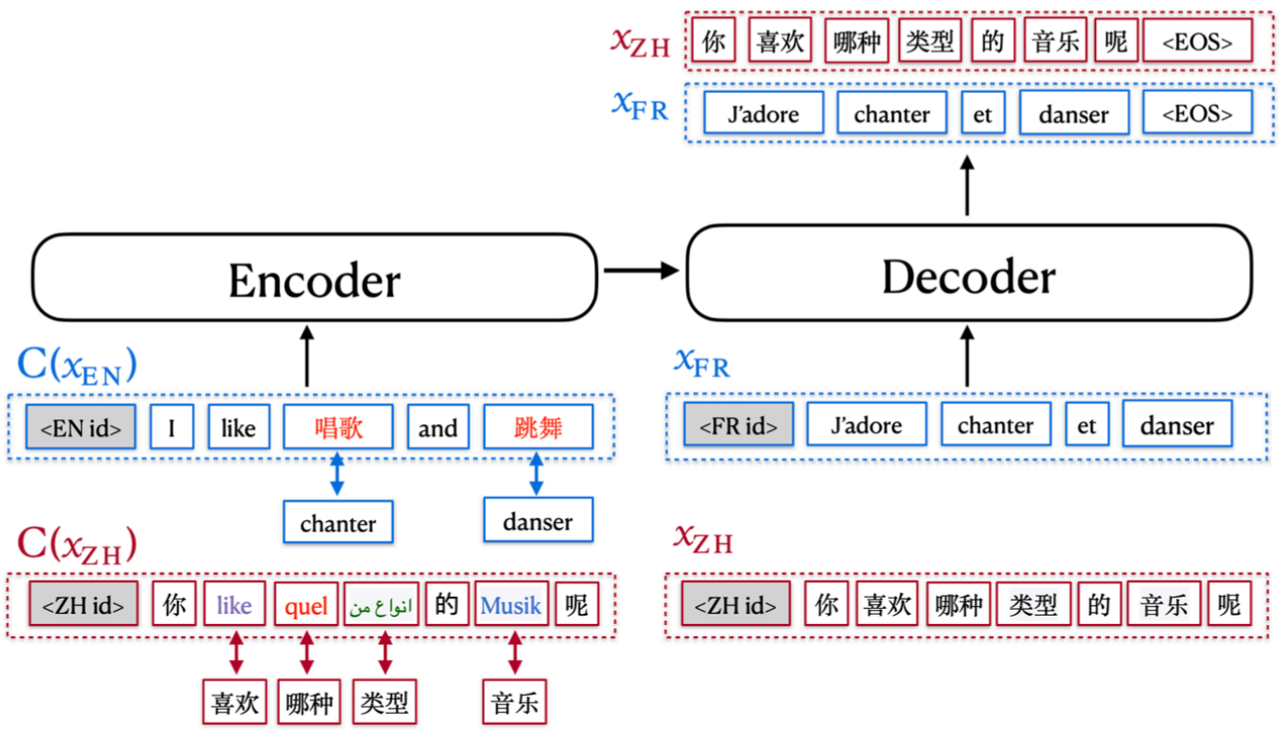


Figure 2: Aligned augmentation on both parallel and monolingual data by replacing words with the same meaning in synonym dictionaries. It either creates a pseudo-parallel example $(C(\mathbf{x}^i), \mathbf{x}^j)$ or a pseudo self-parallel example $(C(\mathbf{x}^i), \mathbf{x}^i)$.

English-Centric Supervised

- mCOLT improves over a strong baseline m-Transformer, which is on par with the strong mBART bilingual models.

	En-Fr wmt14		En-Tr wmt17		En-Es wmt13		En-Ro wmt16		En-Fi wmt17		Avg	Δ
	→	←	→	←	→	←	→(*)	←	→	←		
Transformer(**)	41.4	-	9.5	12.2	35.0	-	34.3	36.8	20.2	21.8	-	
mBART25(**)	41.1	-	17.8	22.5	33.3	-	37.7	38.8	22.4	28.5	-	
m-Transformer	42.0	38.1	18.8	23.1	32.8	33.7	35.9	37.7	20.0	28.2	31.11	
mCOLT	**43.7**	**39.4**	**21.9**	**23.9**	**34.1**	**34.3**	**38.0**	**38.8**	**23.0**	**28.6**	**32.57**	**+1.46**

Table 1: mCOLT outperforms m-Transformer in **supervised** translation directions. Consistent BLEU gains are observed in 20 directions (See Appendix) and in this table we pick the representative ones. We report tokenized BLEU above. Different from our work, their final BLEU scores are obtained by fine-tuning on a single direction. (*) Note that for En→Ro direction, we follow the previous setting to calculate BLEU score after removing Romanian dialects. (**) BLEU scores for Transformer and mBART are cited from (Liu et al., 2020)

English-Centric Unsupervised

- mCOLT obtains significant improvements without explicitly introducing supervision signals for these directions.

	En-Nl iwslt2014		En-Pt opus-100		En-Pl wmt20		Avg	Δ
	→	←	→	←	→	←		
m-Transformer	1.3	7.0	3.7	10.7	0.6	3.2	4.42	
mCOLT	**10.1**	**28.5**	**18.4**	**30.5**	**6.7**	**17.1**	**18.55**	**+14.13**

Table 2: mCOLT outperforms m-Transformer in **unsupervised** translation directions by a large margin. We report tokenized BLEU above.

Non-English Directions — Zero-Shot

- mCOLT narrows the gap with pivot-based models, in line with the intuition that bridging the representation gap of different languages can improve the zero-shot translation.

	Ar		Zh		Nl(*)	
	X→Ar	Ar→X	X→Zh	Zh→X	X→Nl	Nl→X
Pivot	5.5	17.0	28.5	16.4	2.2	6.0
m-Transformer	3.7	5.6	6.7	4.1	2.3	6.3
mCOLT	**5.8**	**17.2**	**28.6**	**14.2**	**5.5**	**7.1**

	Fr		De		Ru		Avg of all
	X→Fr	Fr→X	X→De	De→X	X→Ru	Ru→X	
Pivot	26.1	22.3	14.4	14.2	16.6	19.9	15.56
m-Transformer	7.7	4.8	4.2	4.8	5.7	4.8	5.05
mCOLT	**24.5**	**20.2**	**11.9**	**14.5**	**14.8**	**18.8**	**15.17**

Table 3: **Zero-Shot:** We report de-tokenized BLEU using sacreBLEU in OPUS-100. We observe consistent BLEU gains in zero-shot directions on different evaluation sets, see Appendix for more details. mCOLT further improves the quality. We also list BLEU of pivot-based model (X→En then En→Y using m-Transformer) as a reference, mCOLT only lags behind Pivot by -0.39 BLEU. (*) Note that Dutch(Nl) is not included in PC32.

Analysis: Ablation Study

- Contrastive learning is crucial for the improvement of zero-shot translations.
- Incorporating monolingual data helps mCOLT learn a better representation space.

	model	CTL	AA	MC24	Supervised	Unsupervised	Zero-shot
①	m-Transformer				28.65	4.42	5.05
②	mCOLT w/o AA	✓			28.79	4.75	13.55
③	mCOLT w/o MC24&CTL		✓		29.82	5.40	4.91
④	mCOLT w/o MC24	✓	✓		29.96	5.80	14.60
⑤	mCOLT	✓	✓	✓	**30.02**	**18.55**	**15.00**

Table 4: Summary of average BLEU of mCOLT w/o AA and mCOLT in different scenarios. We report averaged tokenized BLEU. For supervised translation, we report the average of 20 directions; for zero-shot translation, we report the average of 30 directions of OPUS-100. mCOLT w/o AA only adopts contrastive learning on the basis of m-Transformer. mCOLT w/o MC24&CTL excludes MC24 and contrastive loss from mCOLT. mCOLT w/o MC24 excludes MC24 from mCOLT.

Analysis: Similarity Search

- Cross-lingual retrieval as a quantitative indicator of cross-lingual alignment.

Lang	Fr	De	Zh	Ro	Cs	Tr	Ru	NL	PL	Pt
m-Transformer	91.7	96.8	87.0	90.6	84.8	91.1	89.1	25.6	6.3	37.3
mCOLT w/o AA	91.7	97.3	89.9	91.4	86.1	92.4	90.4	35.7	14.3	46.5
mCOLT	**93.0**	**98.0**	**90.7**	**91.9**	**89.3**	**92.4**	**92.3**	**60.3**	**28.1**	**58.6**

Table 5: **English-Centric:** Sentence retrieval top-1 accuracy on Tatoeba evaluation set. The reported accuracy is the average of En→X and X→En accuracy.

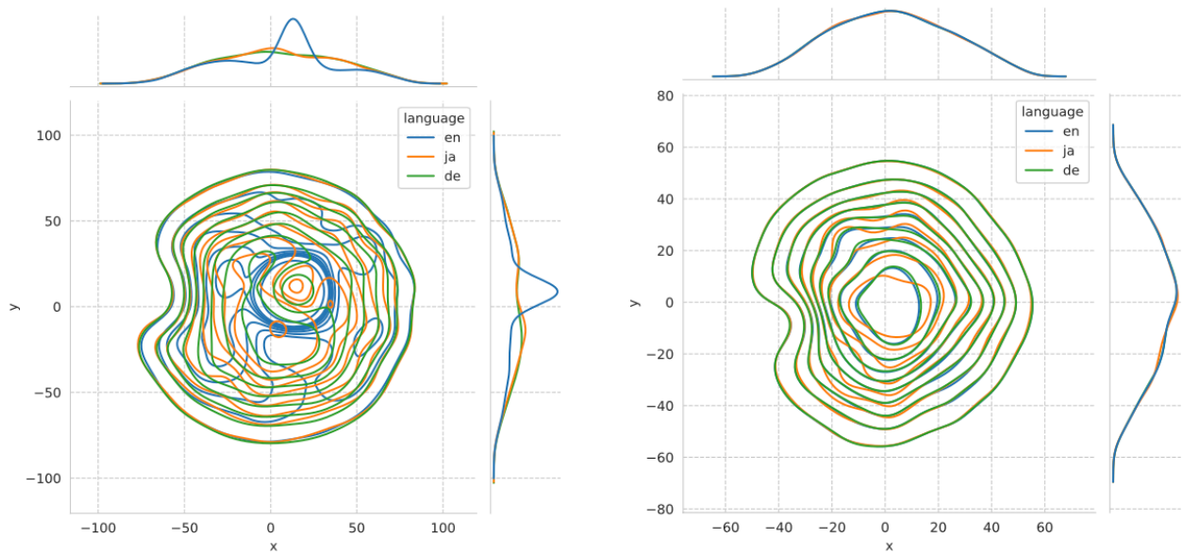
	Top1 Acc	Δ
m-Transformer	79.8	-
mCOLT w/o AA	84.4	+4.8
mCOLT	**89.6**	+9.8

Table 6: **Non-English:** The averaged sentence similarity search top-1 accuracy on Ted-M testset. m-Transformer < mCOLT w/o AA < mCOLT, which is consistent with the results in English-centric scenario.

mCOLT narrows the representation gap across languages.

Analysis: Visualization

- Bivariate kernel density estimation suggests that the sentence representations are drawn closer after applying mCOLT.



(a) m-Transformer (b) mCOLT

Figure 4: Bivariate kernel density estimation plots of representations after using T-SNE dimensionality reduction to 2 dimension. The blue line is English, the orange line is Japanese and the green line is German. This figure illustrates that the sentence representations are drawn closer after applying mCOLT

Summary

- Monolingual data is a rich resource to be exploited for multilingual NMT.
- Exploring the alignment signals to learn universal representations is important for multilingual NMT models.