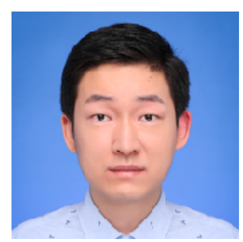
The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)

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

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li






ByteDance AI Lab
字节跳动人工智能实验室

Outline

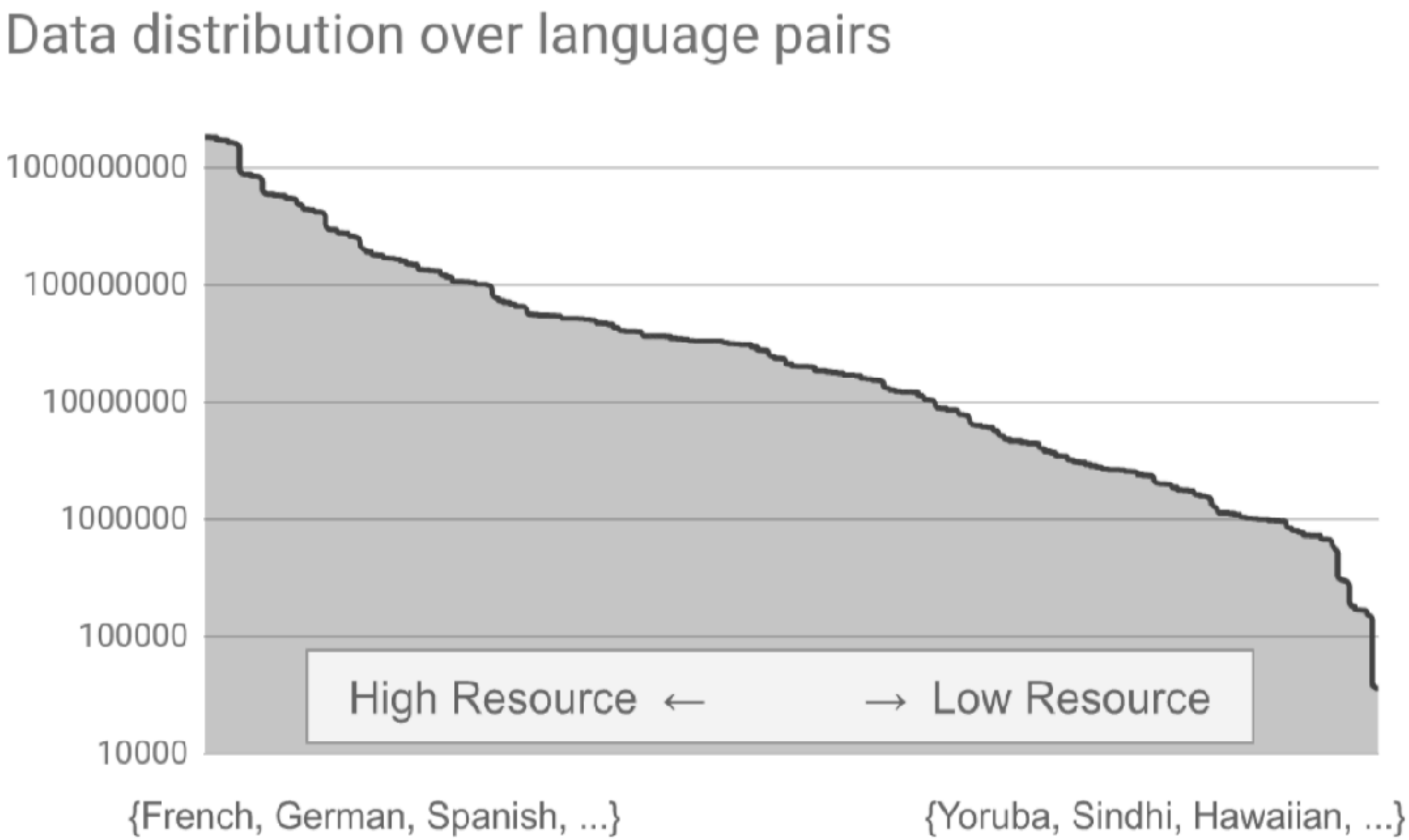
- Motivation and Goal
- mRASP2 Methodology
- Experiments and Analysis
  - Supervised / Unsupervised / Zero-shot
  - Better alignment
- Summary and Take-away

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?

Why Training Multilingual MT Jointly?

- Data scarcity for low/zero resource languages.

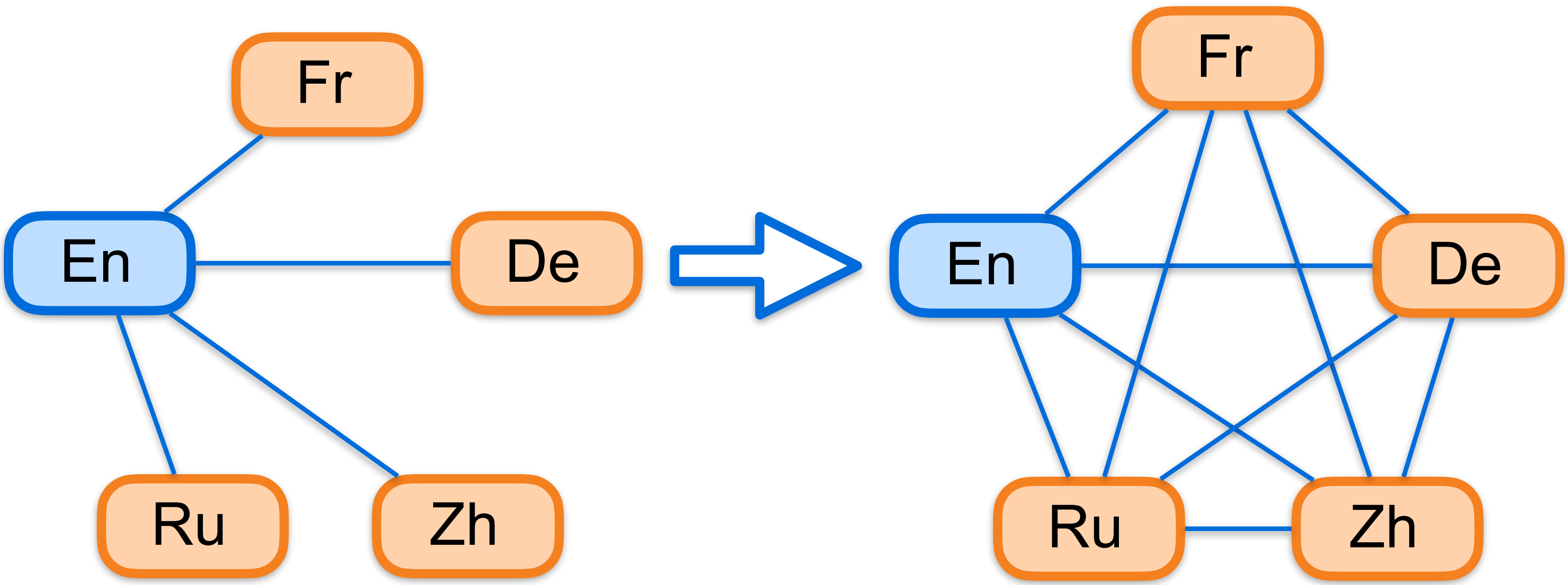


Arivazhagan et al. 2019

Many-to-many Multilingual NMT

Training only w/
En-X Corpus

Many-to-many MNMT



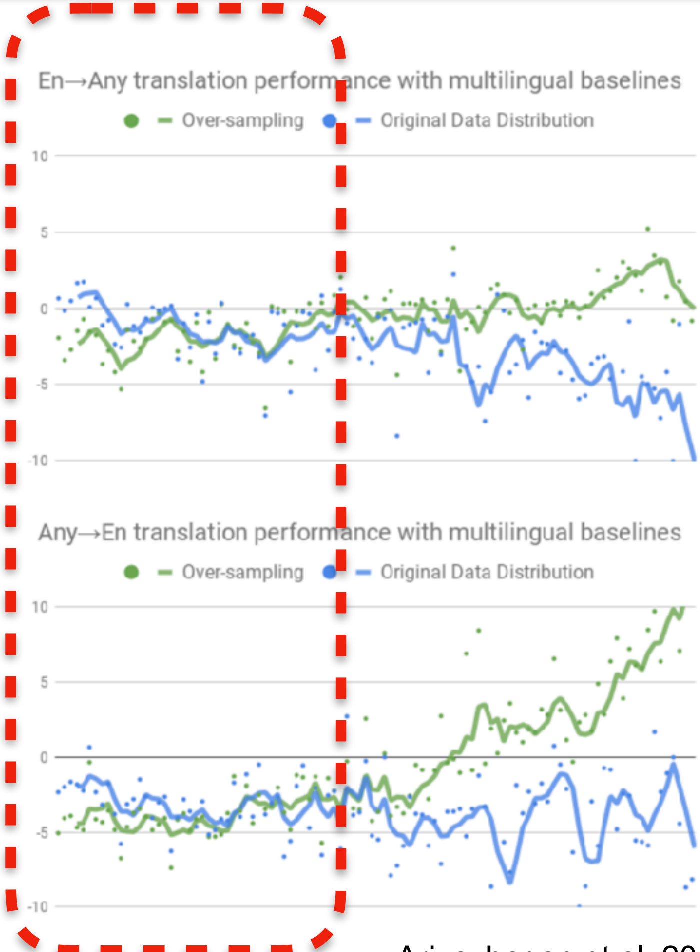
Existing Multilingual NMT(1)

Supervised	✓	En-Zh, En-Fr, En-De
Unsupervised	✗	Fr-Zh, Fr-De, De-Zh
Zero-shot	↓	En-Pt (Assume only have monolingual data of Pt)

Severe degradation on zero-shot translation

- M Johnson, 2017
- N Arivazhagan, 2019

Existing Multilingual NMT(2)



Arivazhagan et al. 2019

Degradation on high-resource directions

Existing Multilingual NMT(3)

Parallel ✓

Monolingual ✗

Only use parallel data

Parallel Monolingual

We want ......

Supervised ✓

Unsupervised ✓

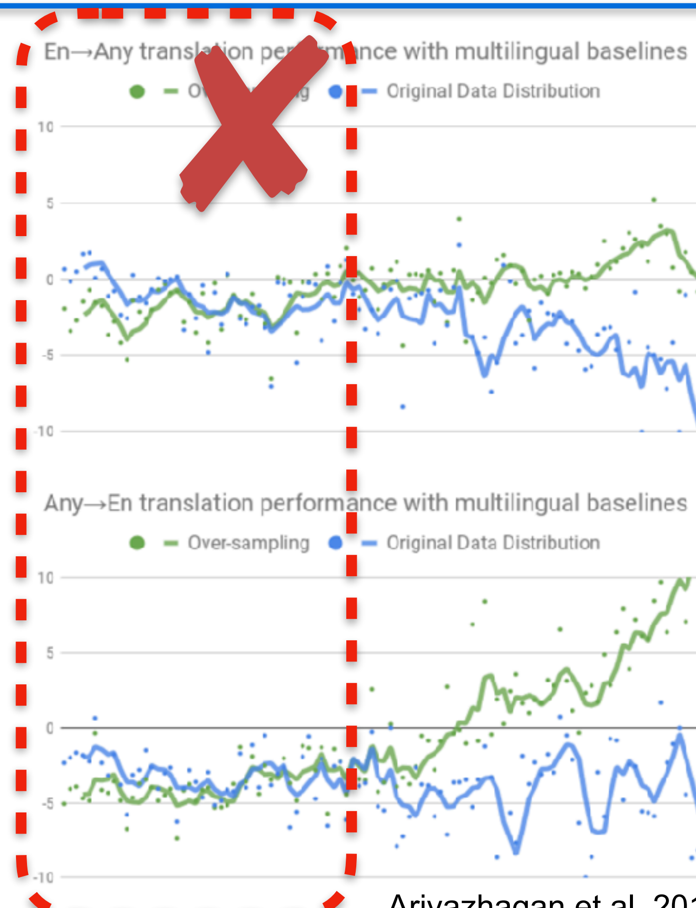
Zero-shot ✓

Enabling unsupervised / zero-shot translation

Parallel ✓

Monolingual ✓

Leveraging both parallel & monolingual data



Arivazhagan et al. 2019

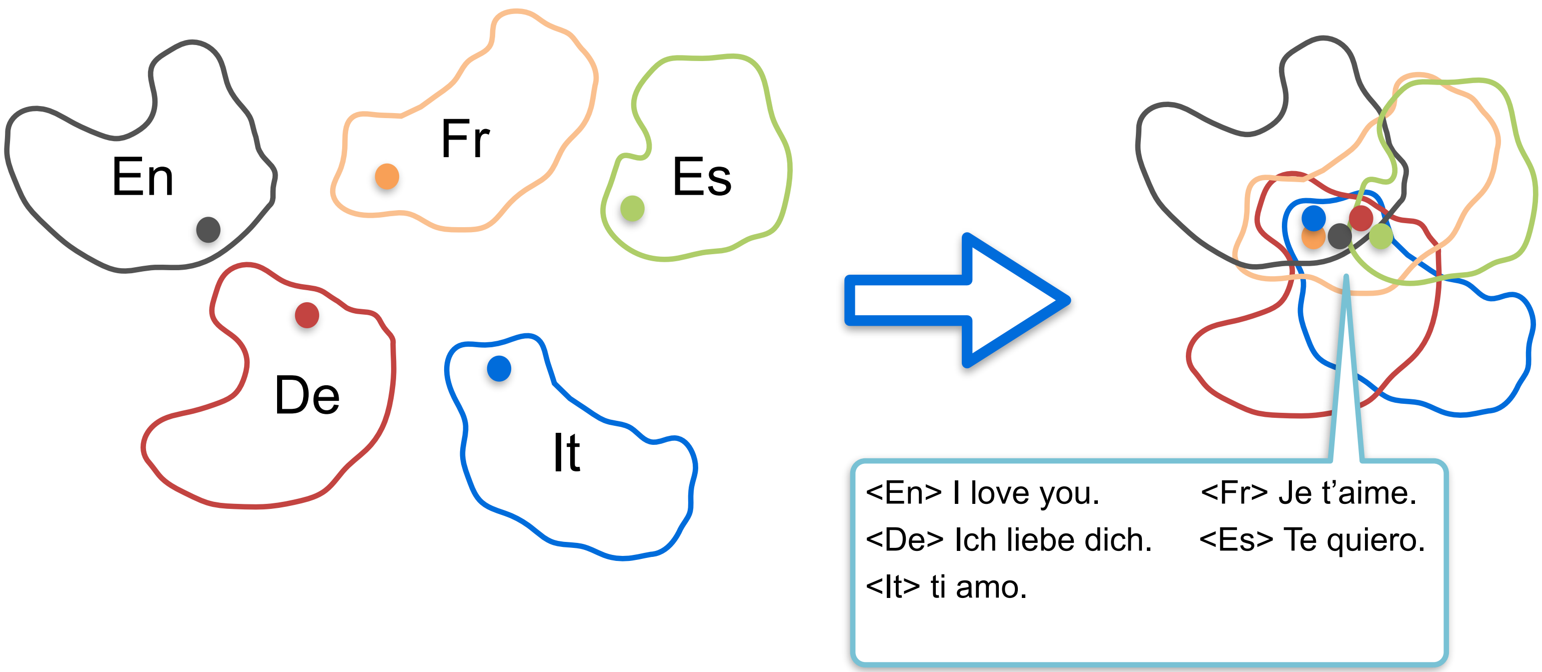
Comparable / better performance on high-resource directions

Goal of mRASP2

- Build a universal NMT model that is both
  - A unified multilingual NMT model that support complete many-to-many translation.
  - A ready-to-use model from which we can derive any NMT model for specific translation direction

Intuition of mRASP2: Bring Representation Closer

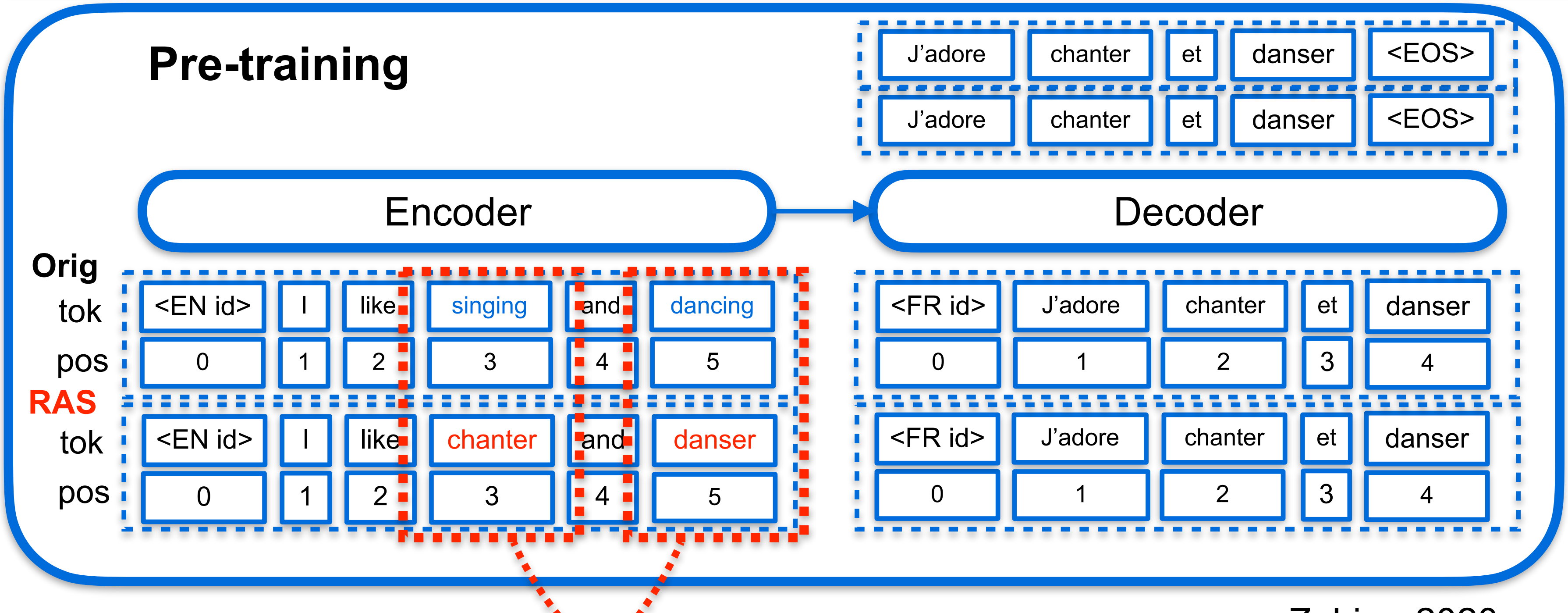
- Sentences with the same semantics across different languages should have similar representations.



Outline

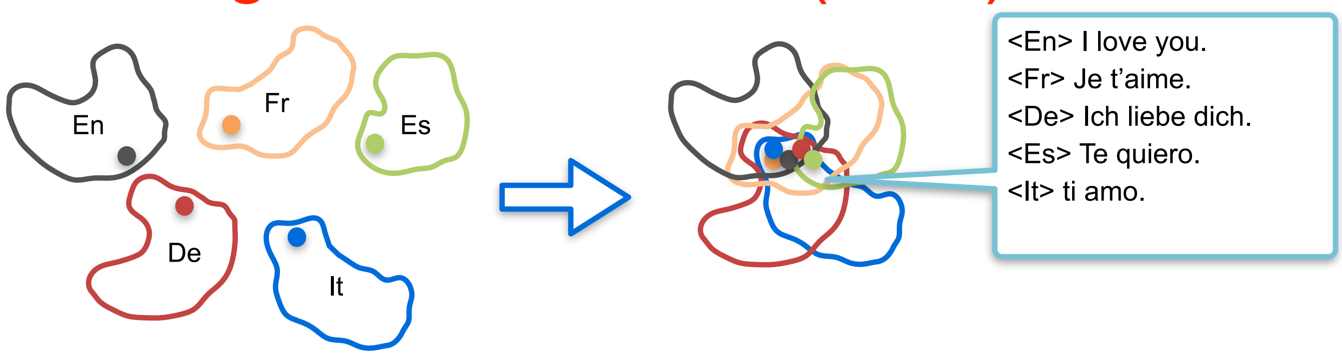
- Motivation and Goal
- mRASP2 Methodology
- Experiments and Analysis
  - Supervised / Unsupervised / Zero-shot
  - Better alignment
- Summary and Take-away

mRASP



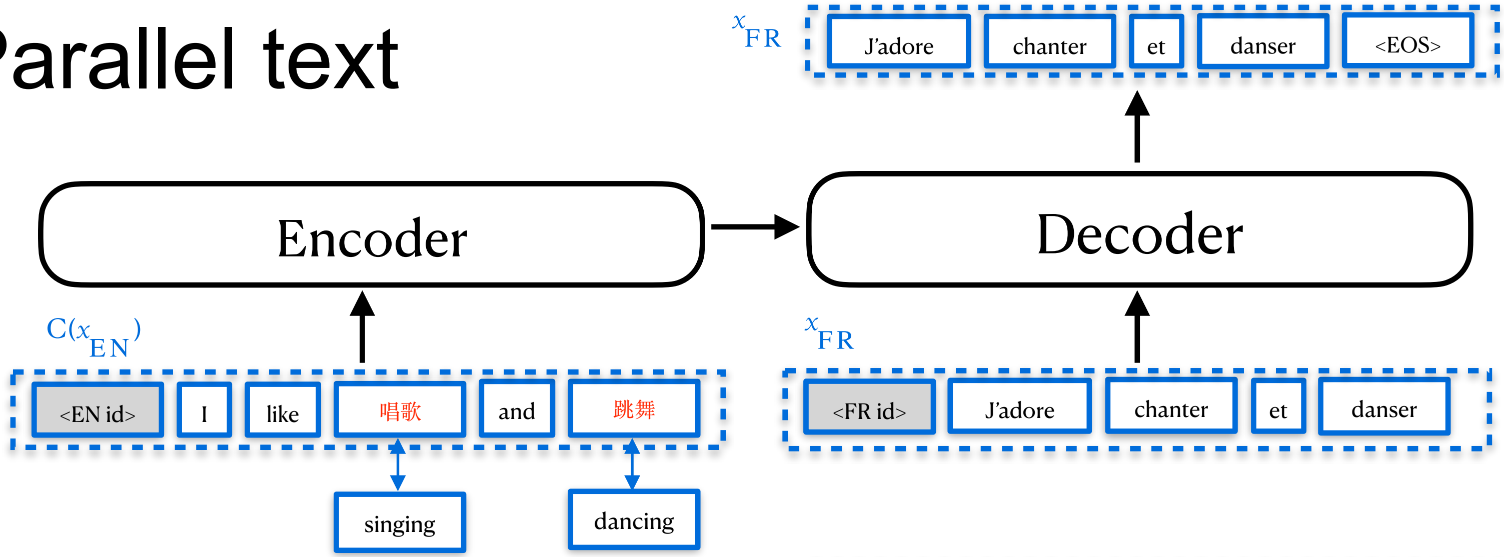
Z Lin · 2020

Random Aligned Substitution(RAS)

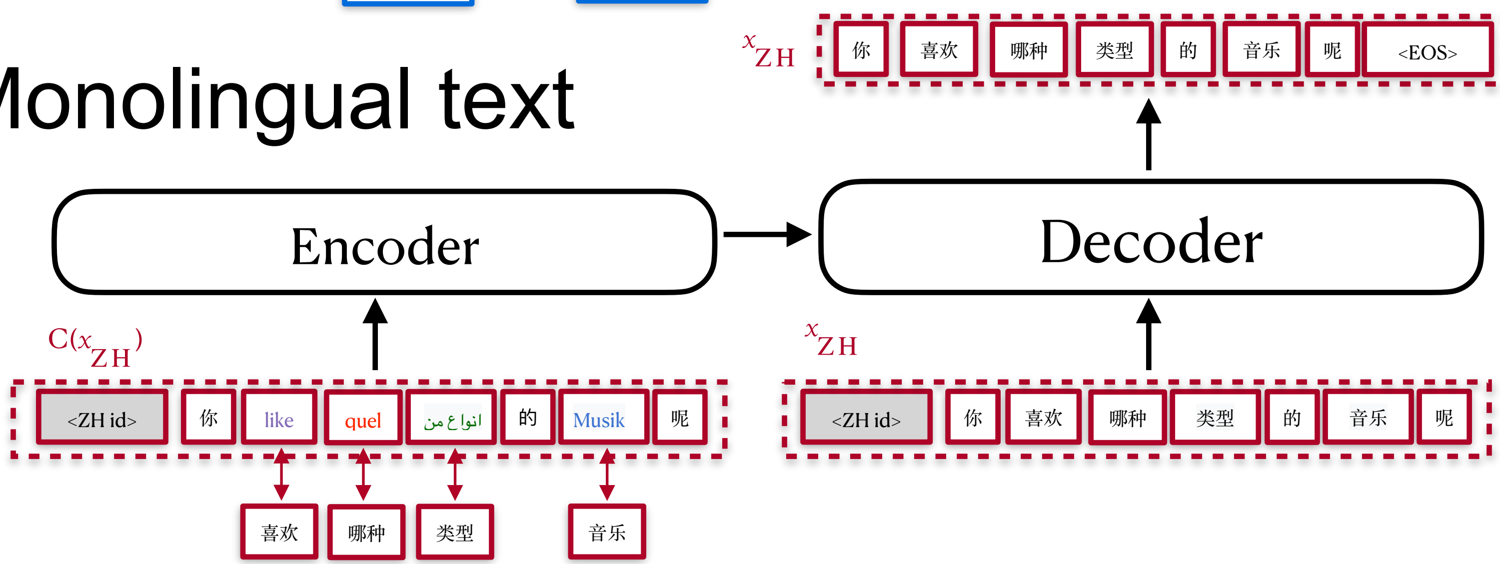


Seq2seq Training with Aligned Augmentation

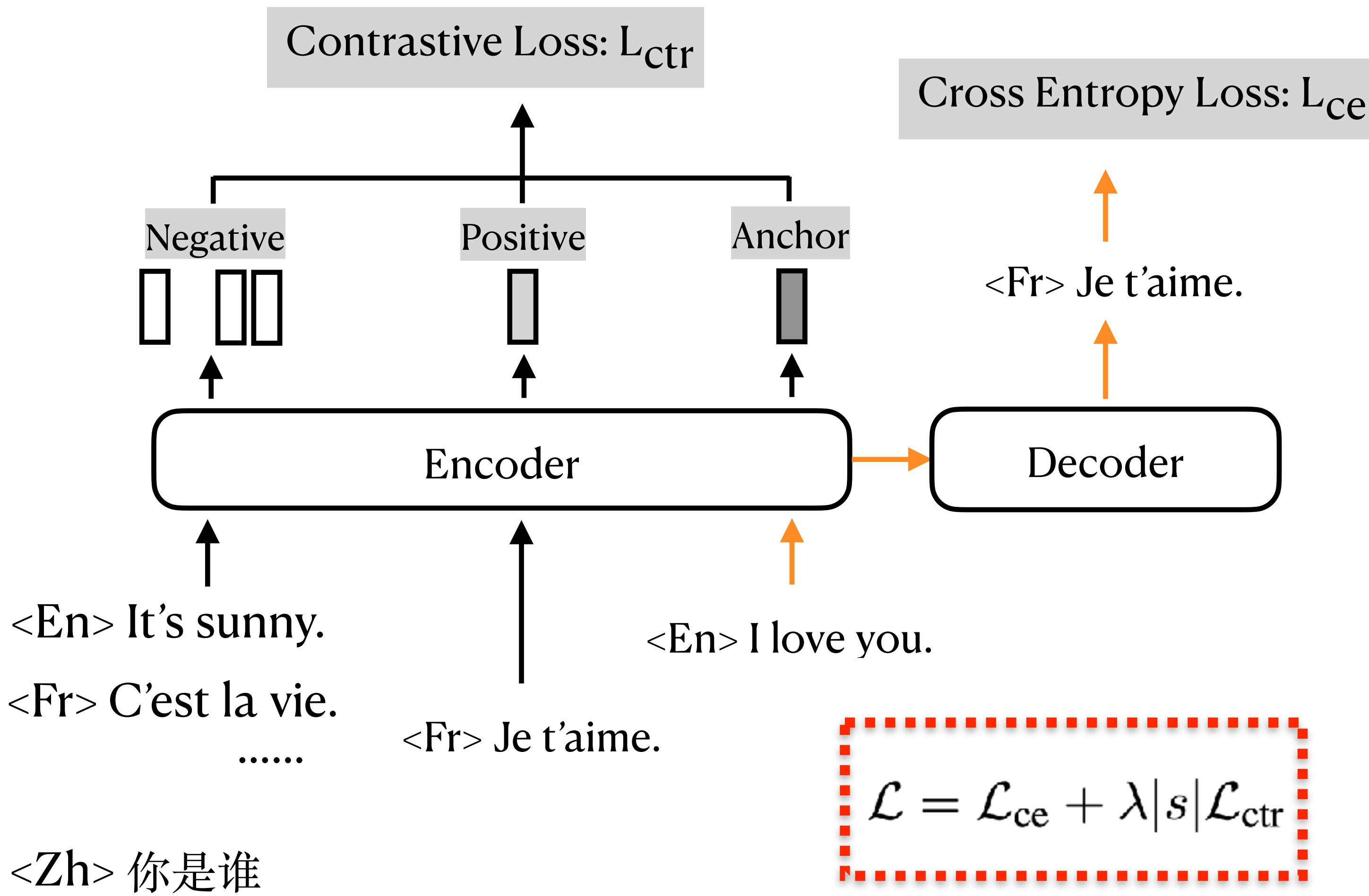
- Parallel text



- Monolingual text

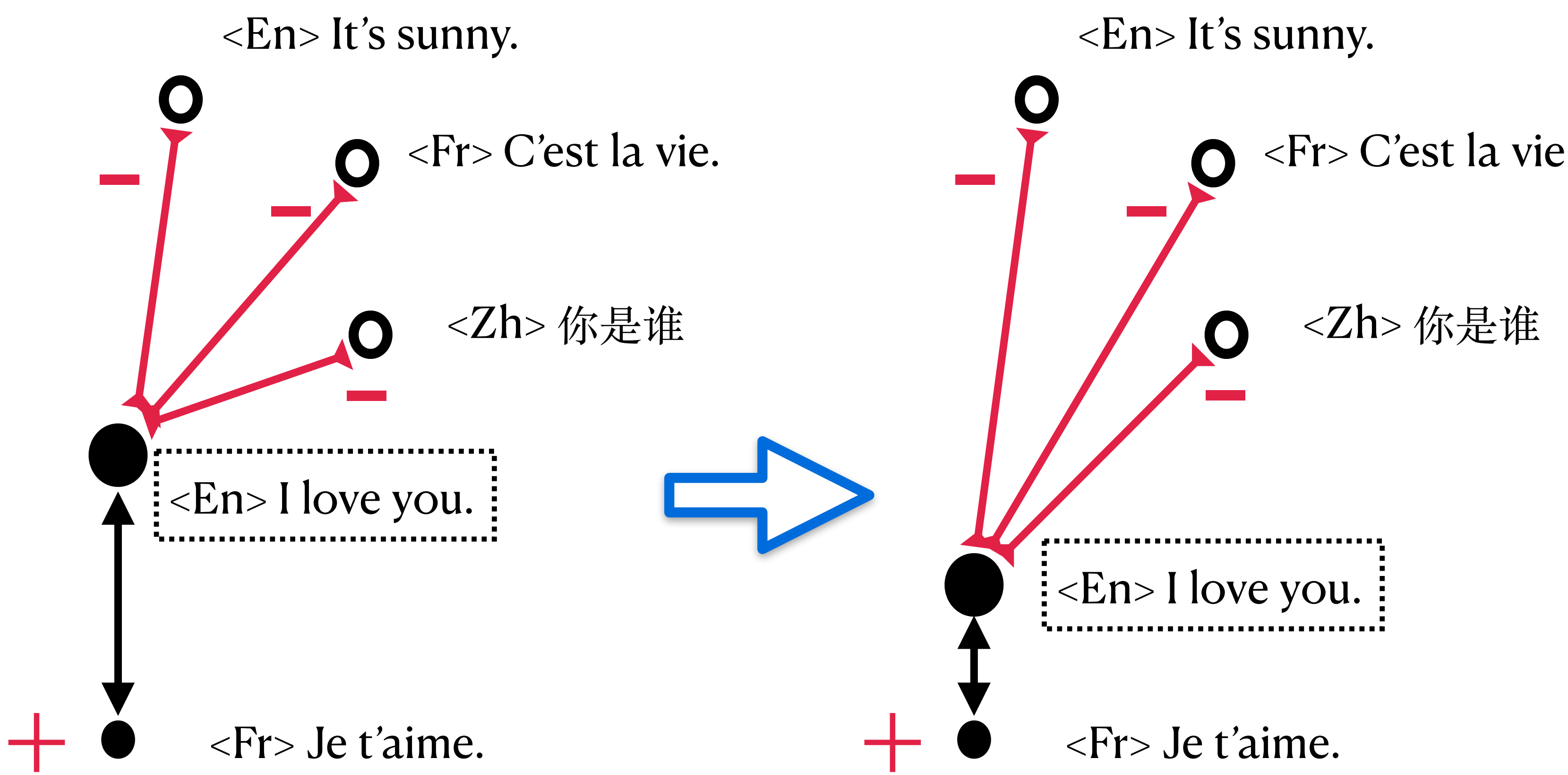


mRASP2 Training



$$\mathcal{L} = \mathcal{L}_{ce} + \lambda |s| \mathcal{L}_{ctr}$$

Contrastive Learning



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

$$\boxed{\mathcal{L} = \mathcal{L}_{\text{ce}} + \lambda |s| \mathcal{L}_{\text{ctr}}}$$

Outline

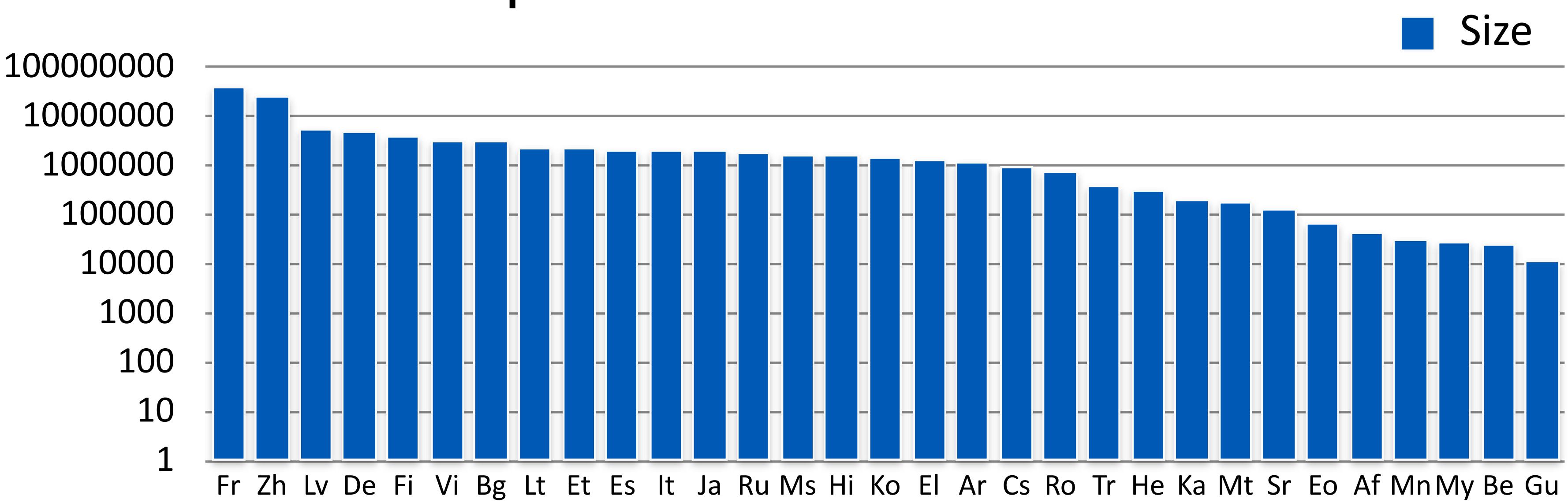
- Motivation and Goal
- mRASP Methodology
- Experiments and Analysis
  - Supervised / Unsupervised / Zero-shot
  - Better alignment
- Summary and Take-away

Two Main Questions

- Does mRASP2 work on supervised / unsupervised / zero-shot scenarios?
- Why mRASP2 works?

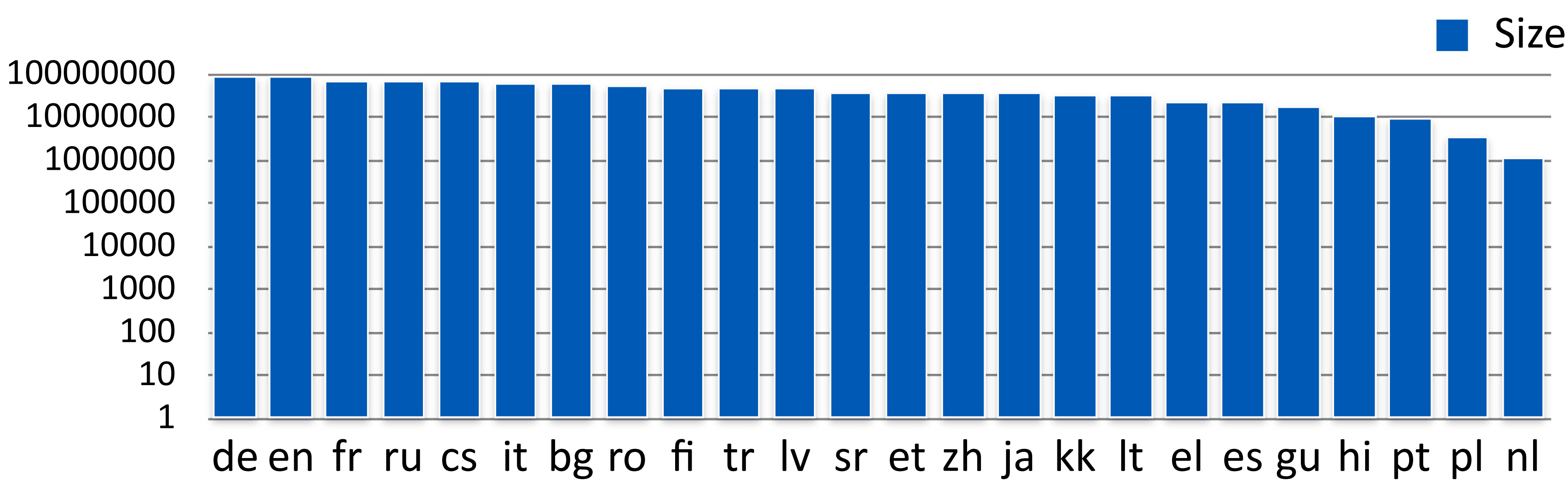
Datasets

- Parallel Dataset: **PC32** (32 language pairs)
  - 32 English-centric language pairs, resulting in 64 directed translation pairs in total
  - Contains a total size of 110.4M public parallel sentence pairs

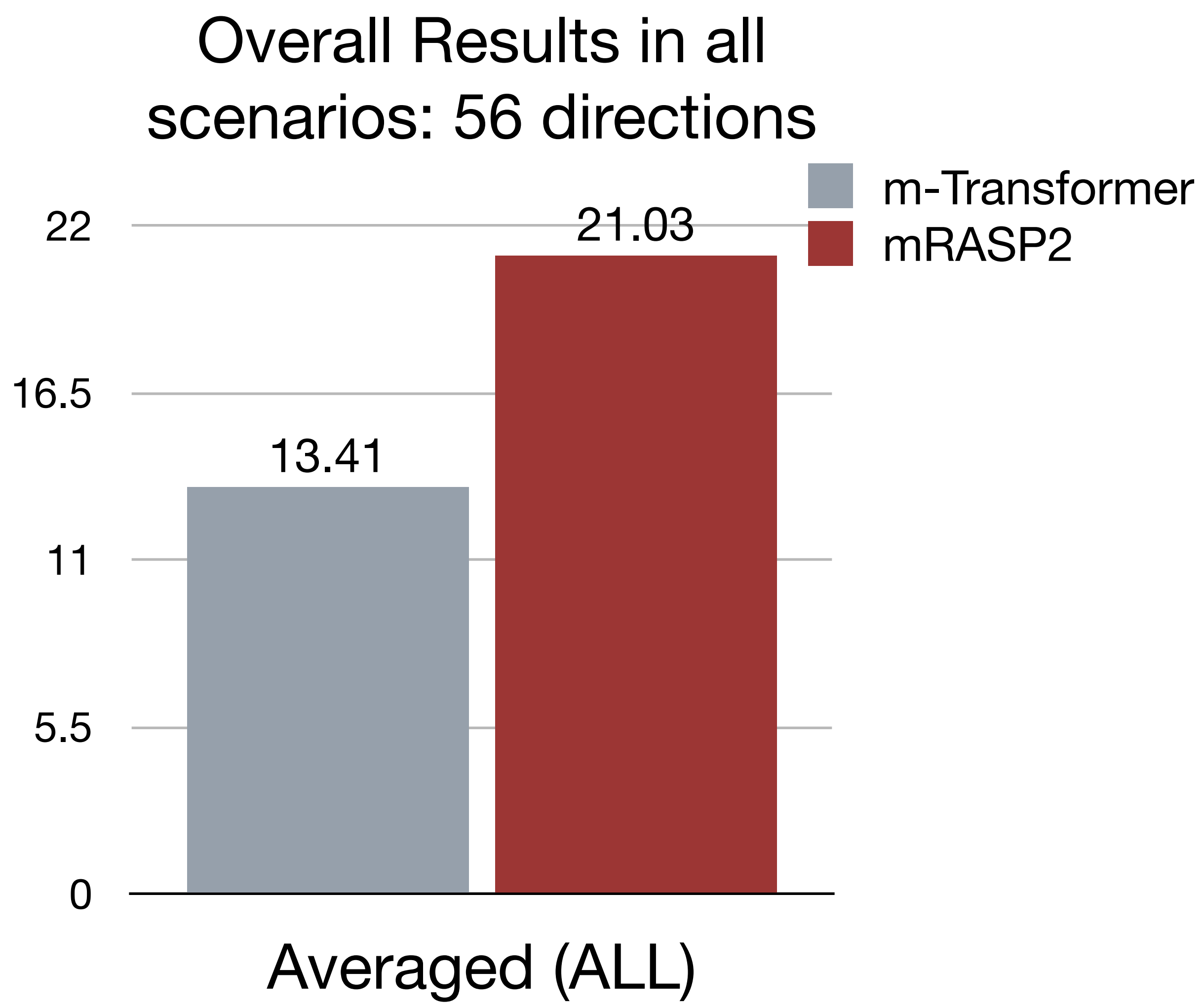


Datasets

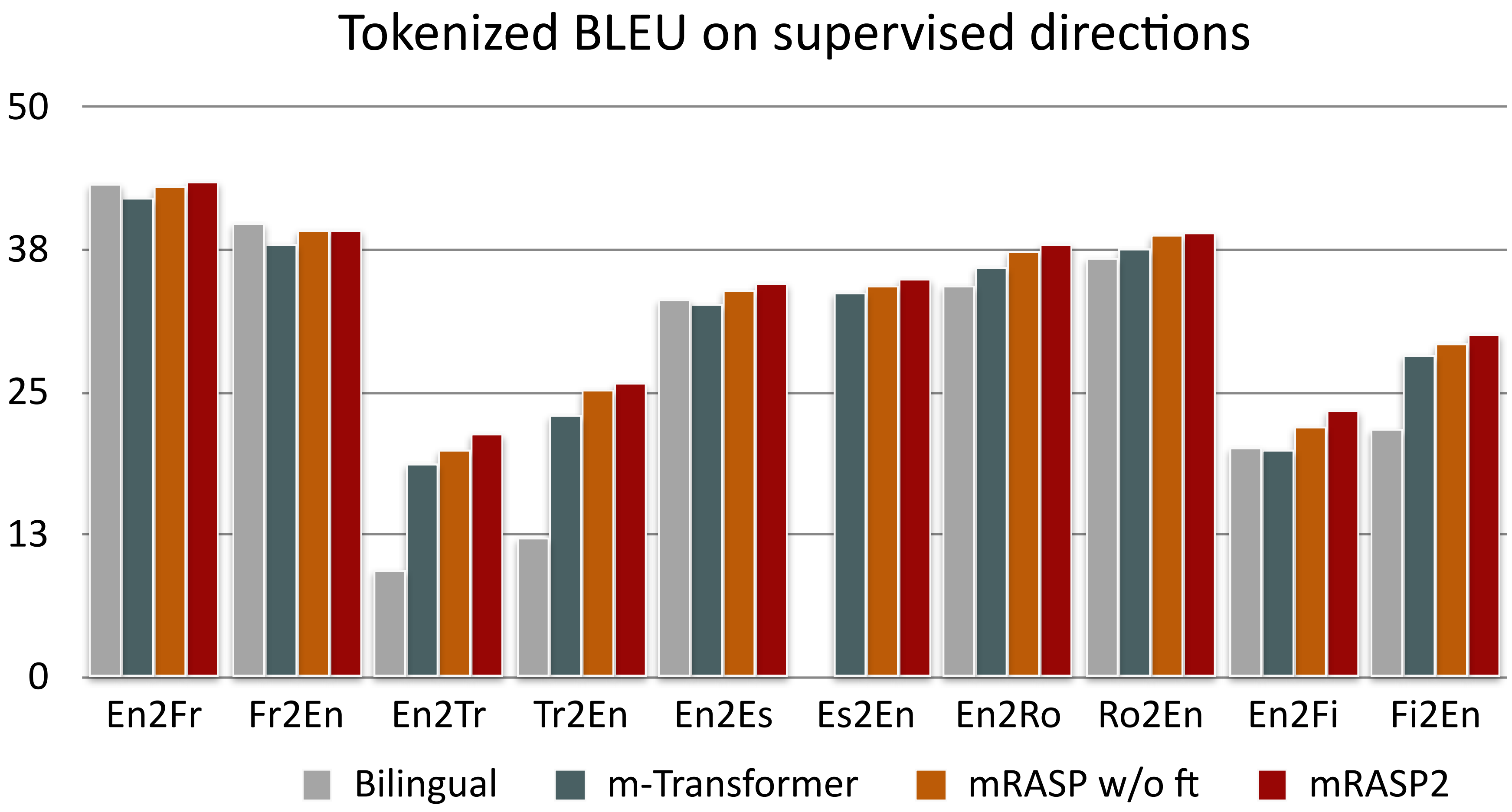
- Monolingual Dataset: **MC24** (24 languages)
  - 21 languages that also appear in **PC32**
  - 3 additional languages: Nl, Pl, Pt
  - Temperature sampling: T=5



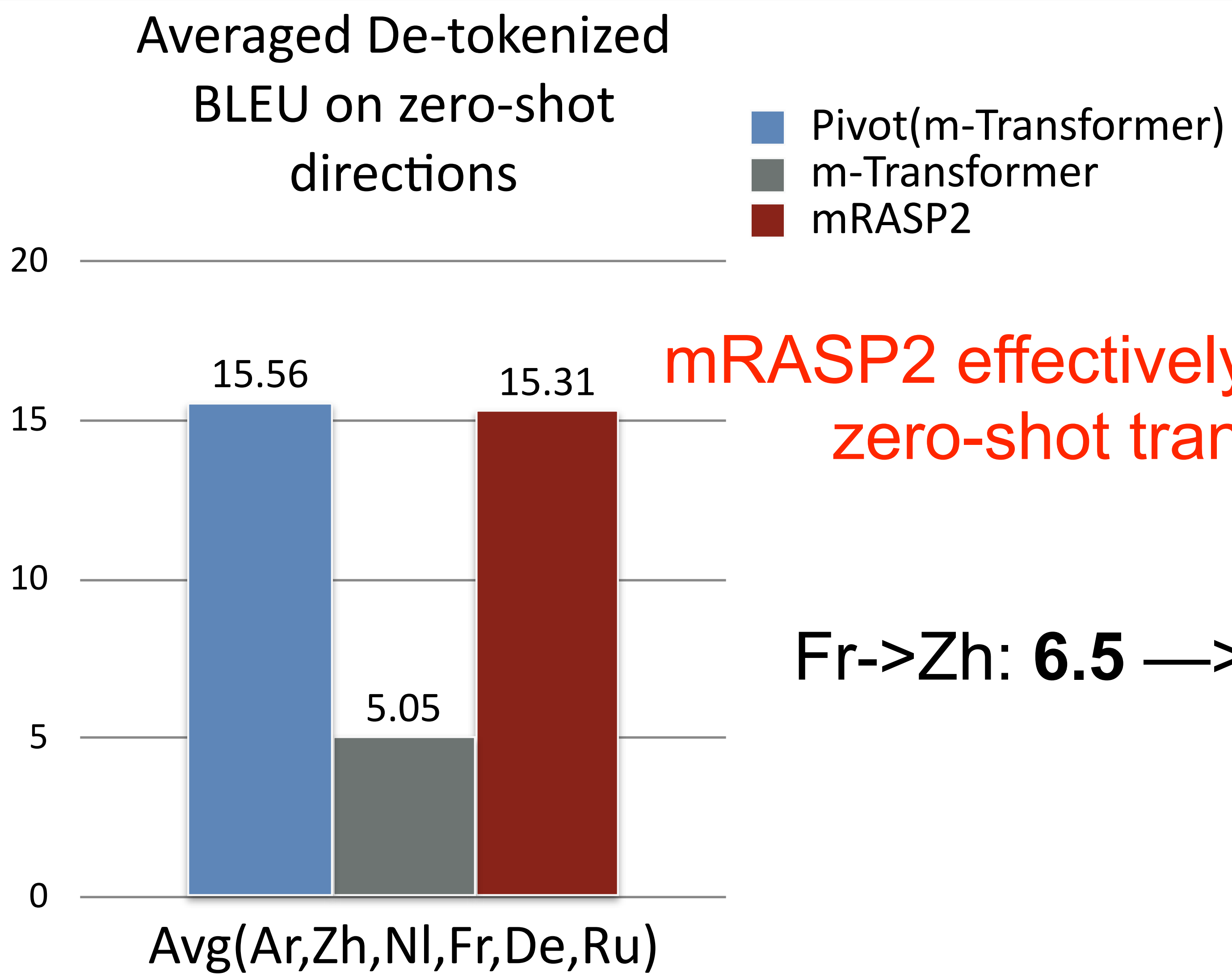
Overall Results



Comparable or Better Performance on Supervised Directions



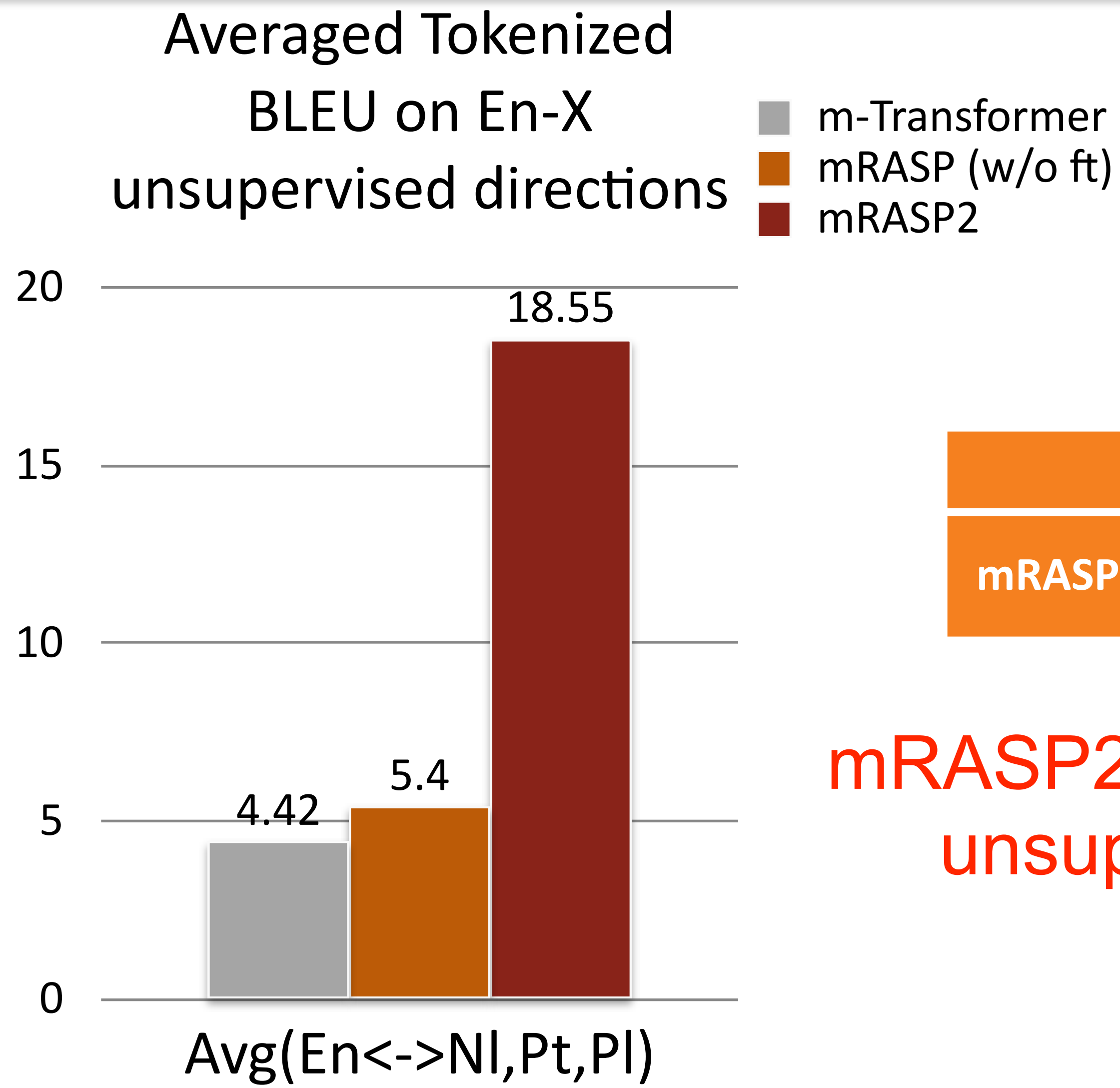
Effectiveness on Zero-shot Directions



mRASP2 effectively improves zero-shot translation

Fr->Zh: **6.5** —> **42.3**

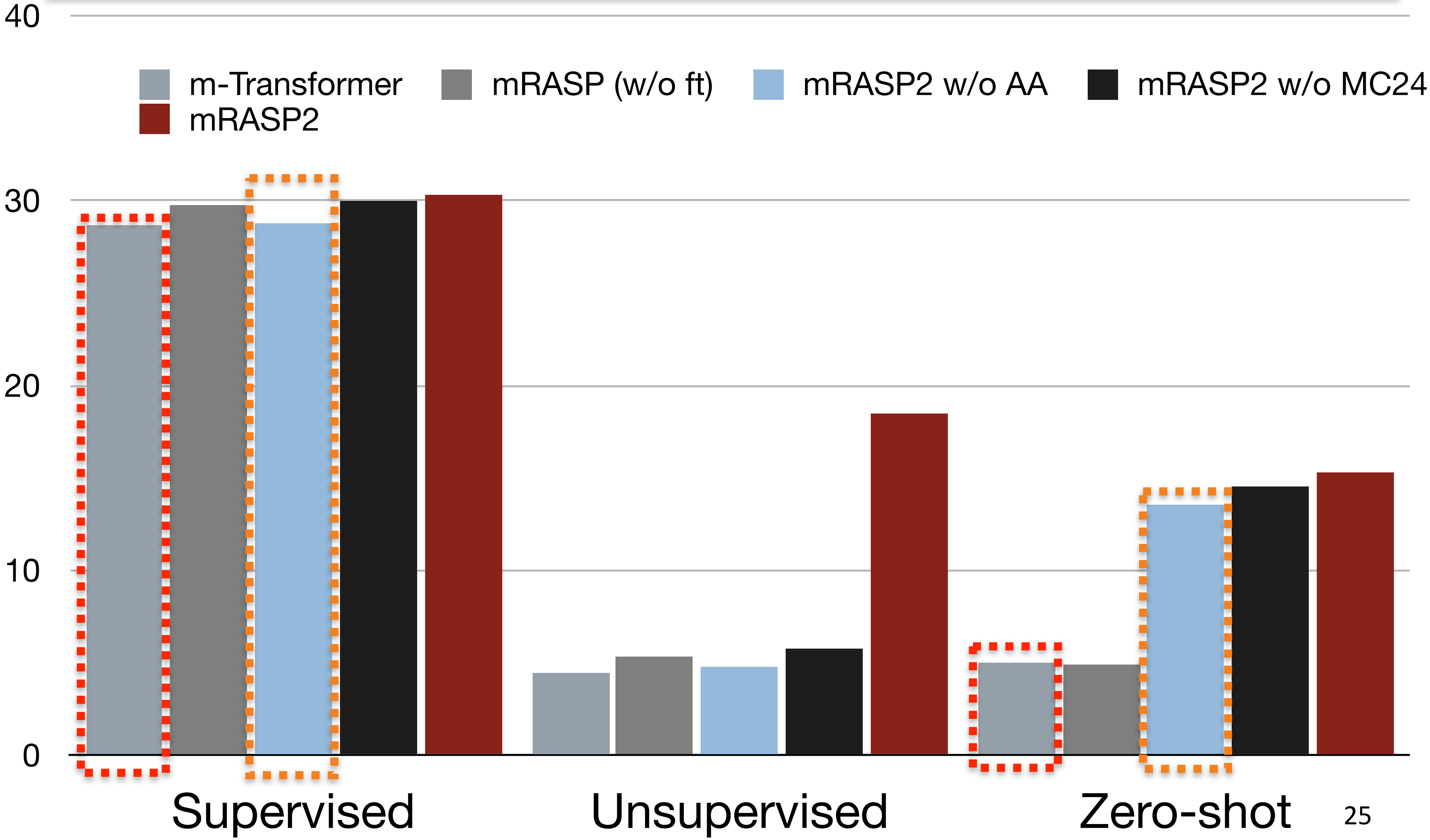
Effectiveness on Unsupervised Directions



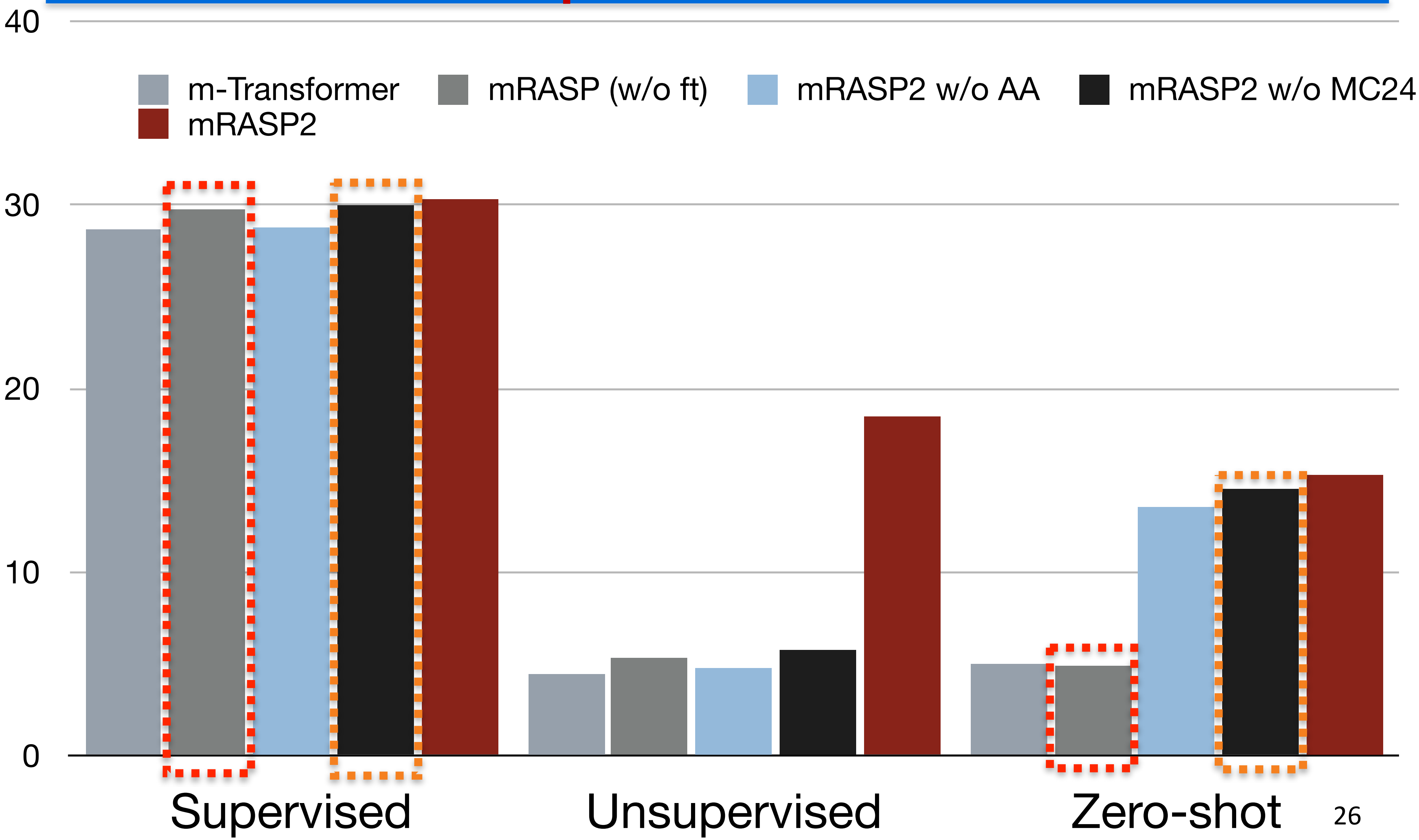
	Nl->Pt	Pt->Nl
mRASP2	9.3	8.3

mRASP2 also works on fully unsupervised directions

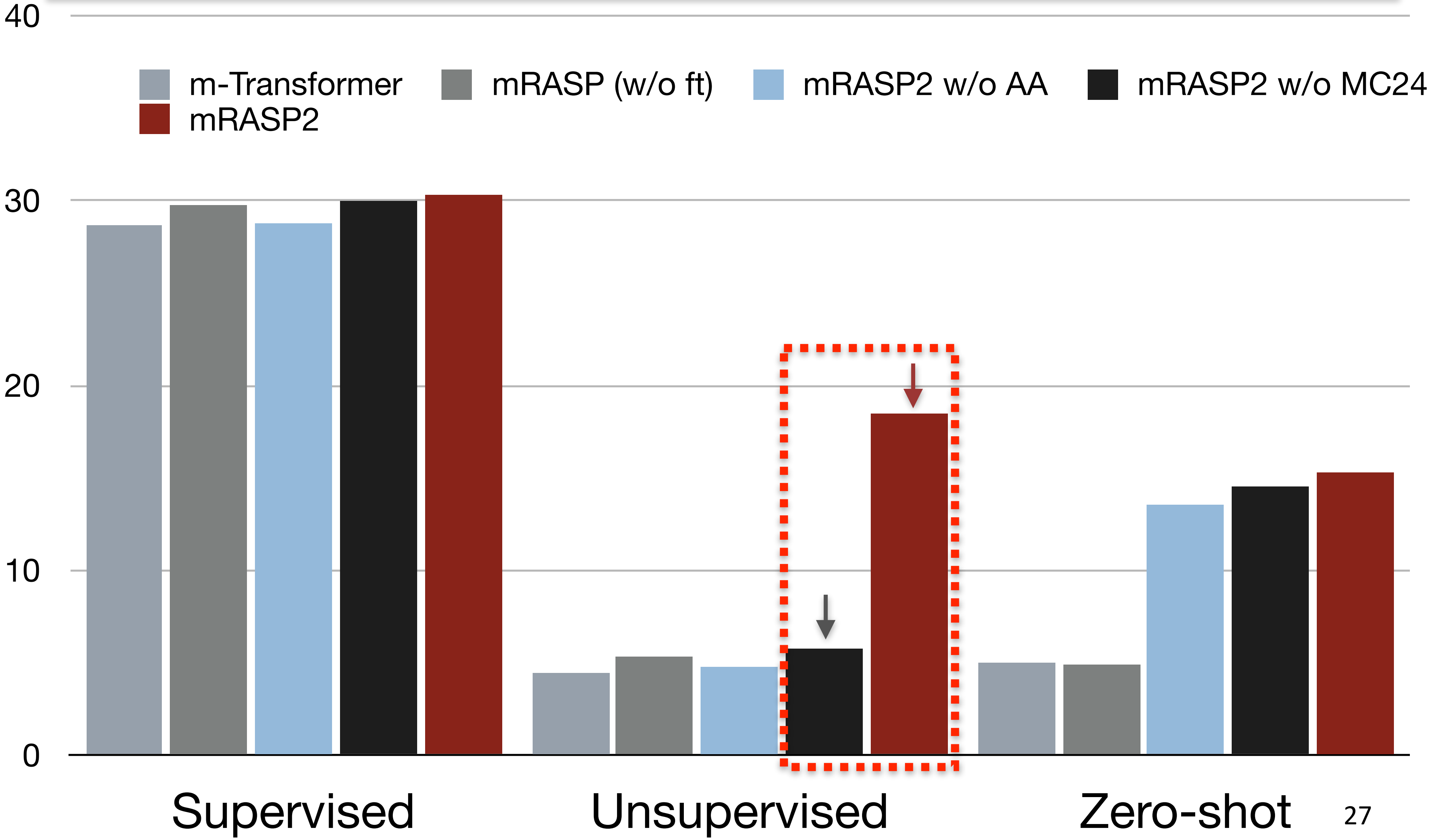
Contrastive Learning effectively improves zero-shot translation without hurting supervised translation performance



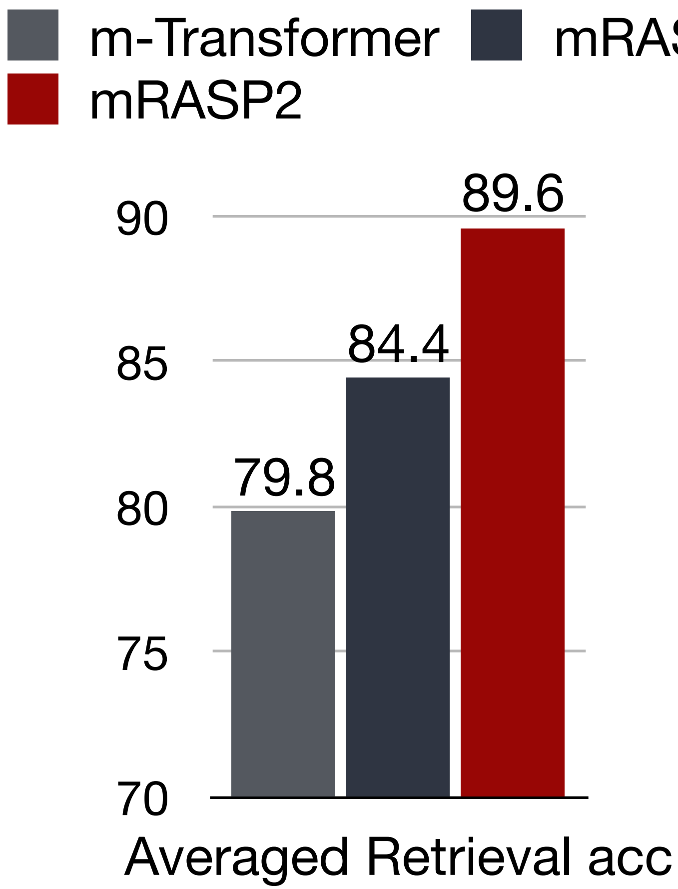
Contrastive Learning effectively improves zero-shot translation without hurting supervised translation performance



Monolingual Corpus mainly contributes to unsupervised translation



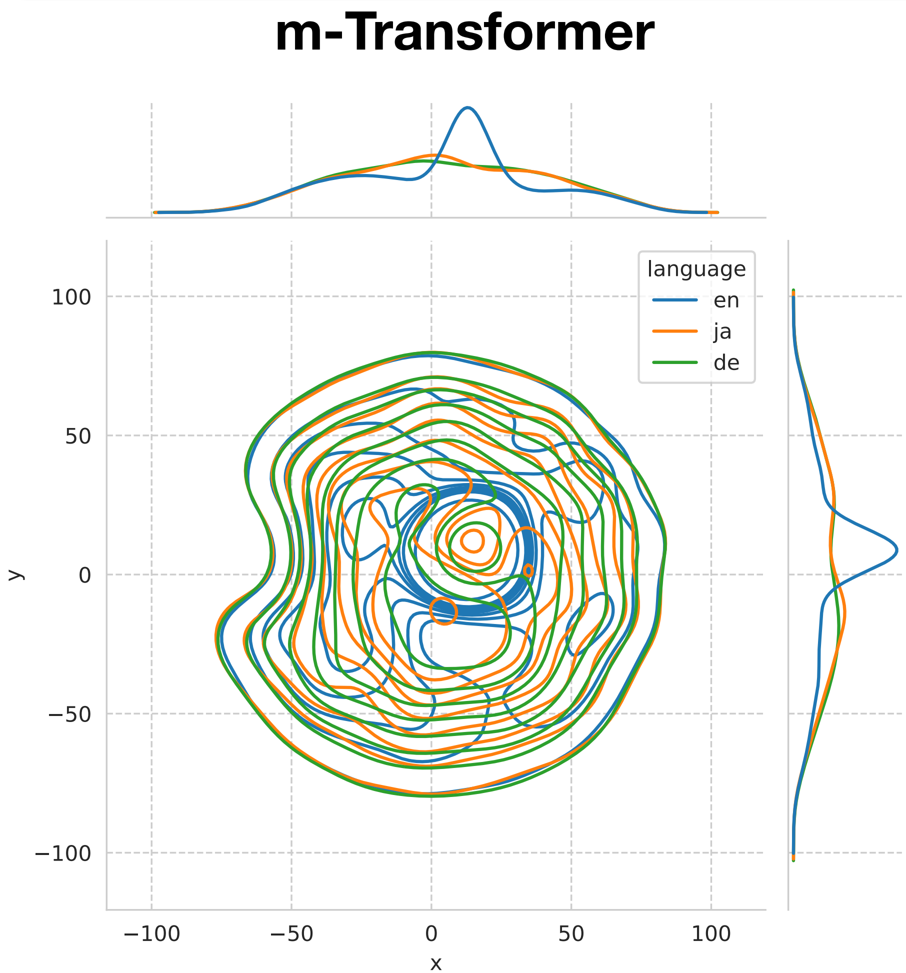
Better Semantic Alignment Across Languages: Improved Sentence Retrieval

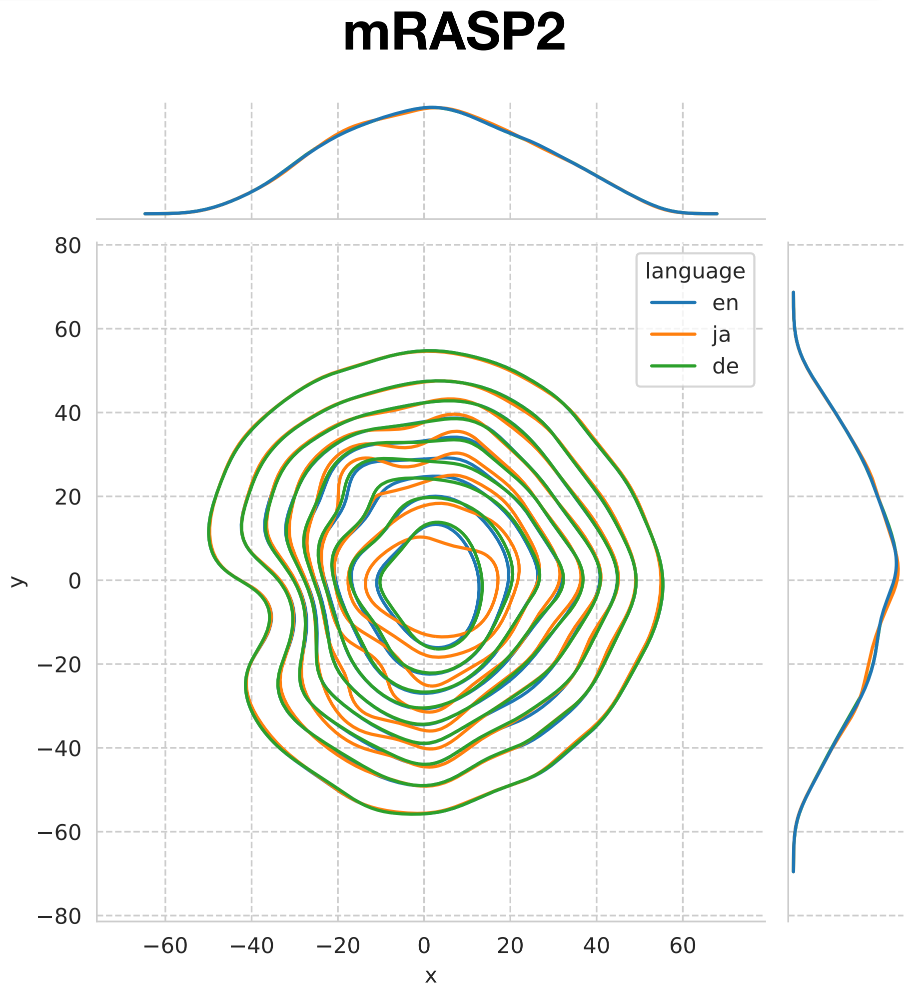


15-way parallel test set(Ted-M): 2284 samples

Contrastive Learning and Aligned Augmentation both contribute to the improvement on sentence retrieval

Better Semantic Alignment: Visualization of Sentence Repr





Better Alignment of En, Ja, De Representations !!

Outline

- Motivation and Goal
- mRASP2 Methodology
- Experiments and Analysis
  - Supervised / Unsupervised / Zero-shot
  - Better alignment
- Summary and Take-away

Summary

- We propose mRASP2
  - A universal Multilingual MT model
  - Leverages monolingual data along with parallel data in a unified framework
  - Bridges the representation gap of utterances in different languages with the same semantics.

Take Home Messages

- Closer representation —> Improved multilingual MT performance
- Leverage both parallel and monolingual corpora !!
- Contrastive Learning and Aligned Augmentation are effective in bridging representation

Thanks!

Paper Video Blog

- Code and models available at:
  - https://github.com/PANXiao1994/mRASP2
- Also MT in ACL21:
  - Green vocabulary learning: VOLT [Xu et al. 2021]
  - Language-specific subnets for MNMT: LaSS [Lin et al. 2021]
  - Language Tag Matters [Wu et al. 2021]
  - Glancing Transformer [Qian et al. 2021]
- Other tools:
  - Transformer fast training and inference: https://github.com/bytedance/lightseq
  - Speech & MT toolkit: https://github.com/bytedance/neurst