CS11-737 Multilingual NLP

Multilingual Neural Machine Translation Pre-training and Joint Training Strategies

Lei Li

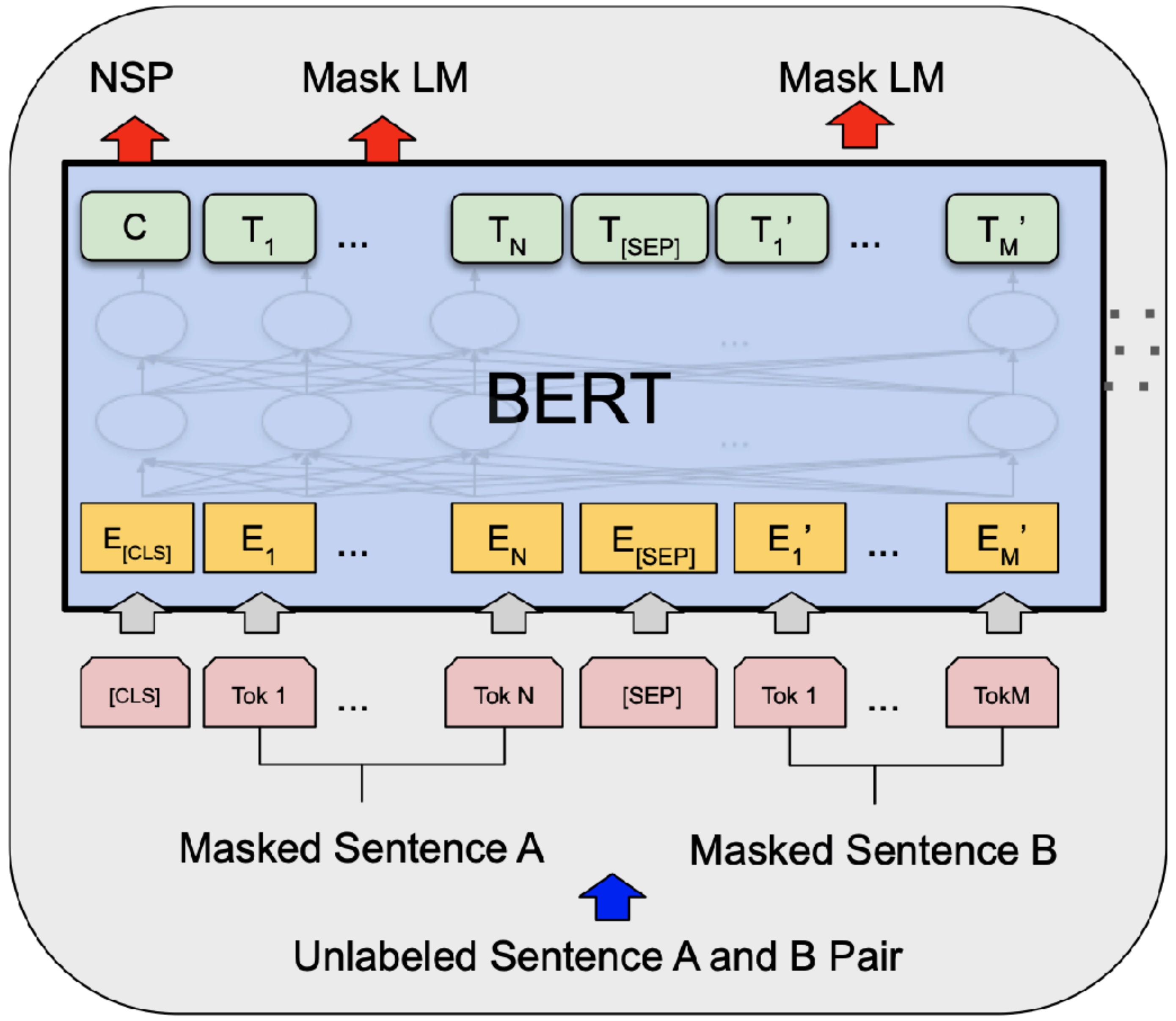
https://lileicc.github.io/course/11737mnlp23fa/

Carnegie Mellon University
Language Technologies Institute

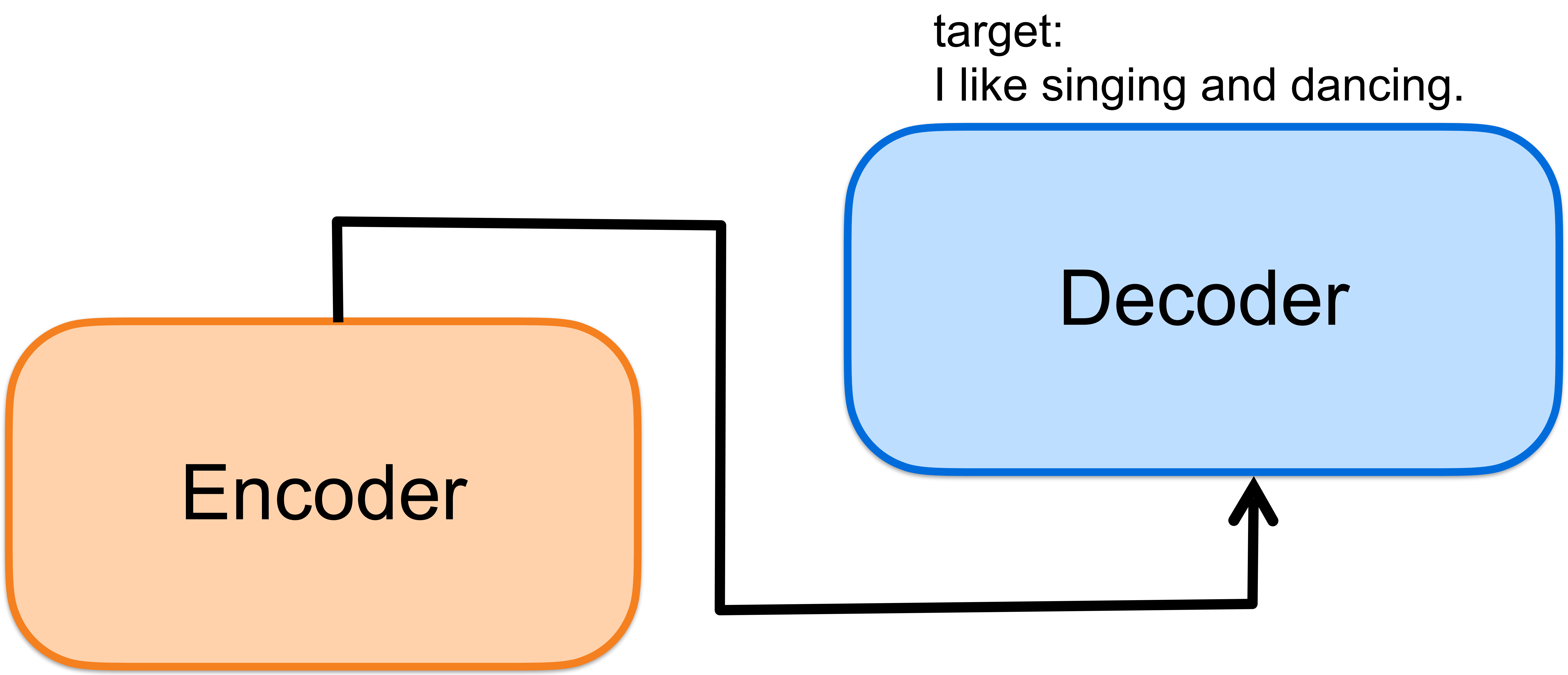
Sequence-to-sequence Pre-training

Mismatch between Pre-trained LM and MT

- BERT/GPT pre-training objective is different from MT



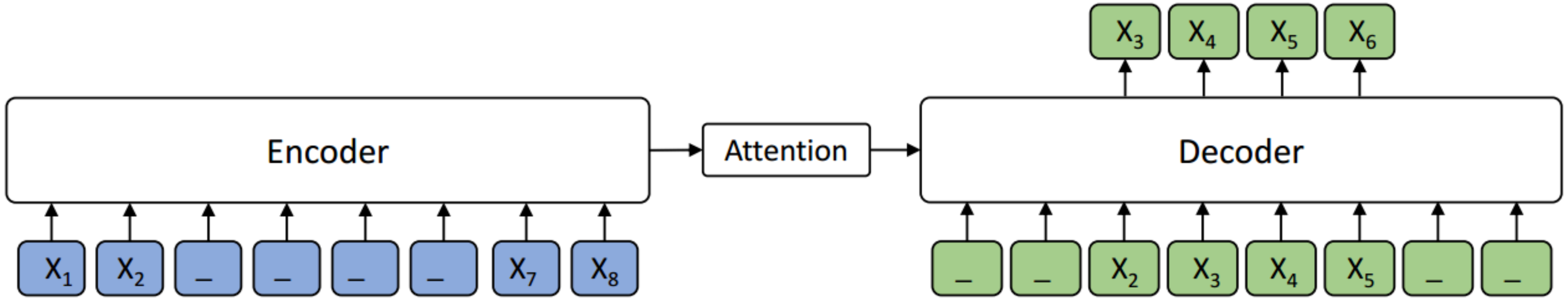
Pre-training



Source: 我喜欢唱歌和跳舞。

MASS: Pre-train for Sequence to Sequence Generation

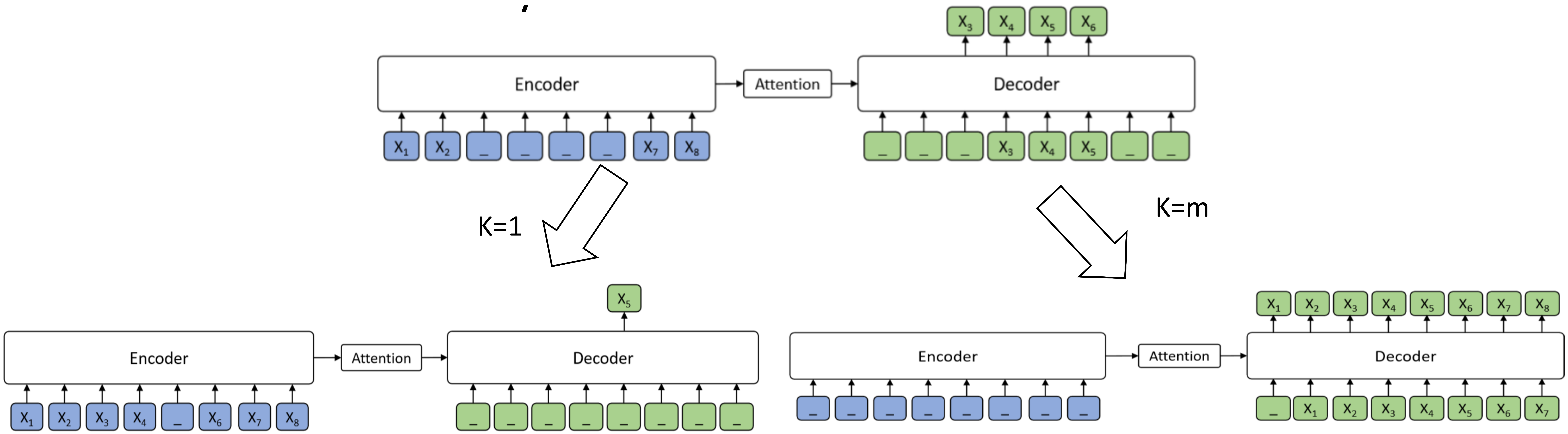
- MASS is carefully designed to jointly pre-train the encoder and decoder



- Mask k consecutive tokens (segment)
  - Force the decoder to attend on the source representations, i.e., encoder-decoder attention
  - Develop the decoder with the ability of language modeling

MASS: Pre-train for Sequence to Sequence Generation, [Song et al ICML 2019]

MASS vs. BERT/GPT

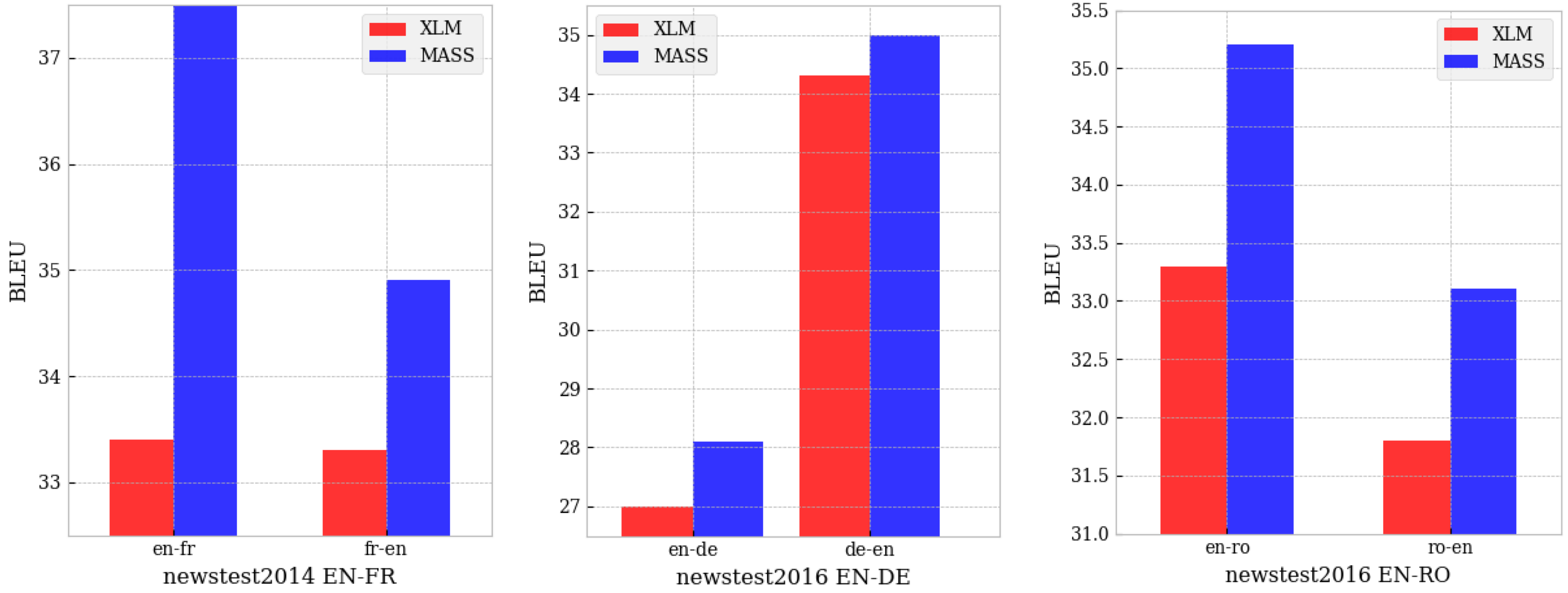


Length	Probability	Model
$k = 1$	$P(x^u \| x^{\setminus u}; \theta)$	masked LM in BERT
$k \in [1, m]$	$P(x^{u:v} \| x^{\setminus u:v}; \theta)$	MASS

Length	Probability	Model
$k = m$	$P(x^{1:m} \| x^{\setminus 1:m}; \theta)$	standard LM in GPT
$k \in [1, m]$	$P(x^{u:v} \| x^{\setminus u:v}; \theta)$	MASS

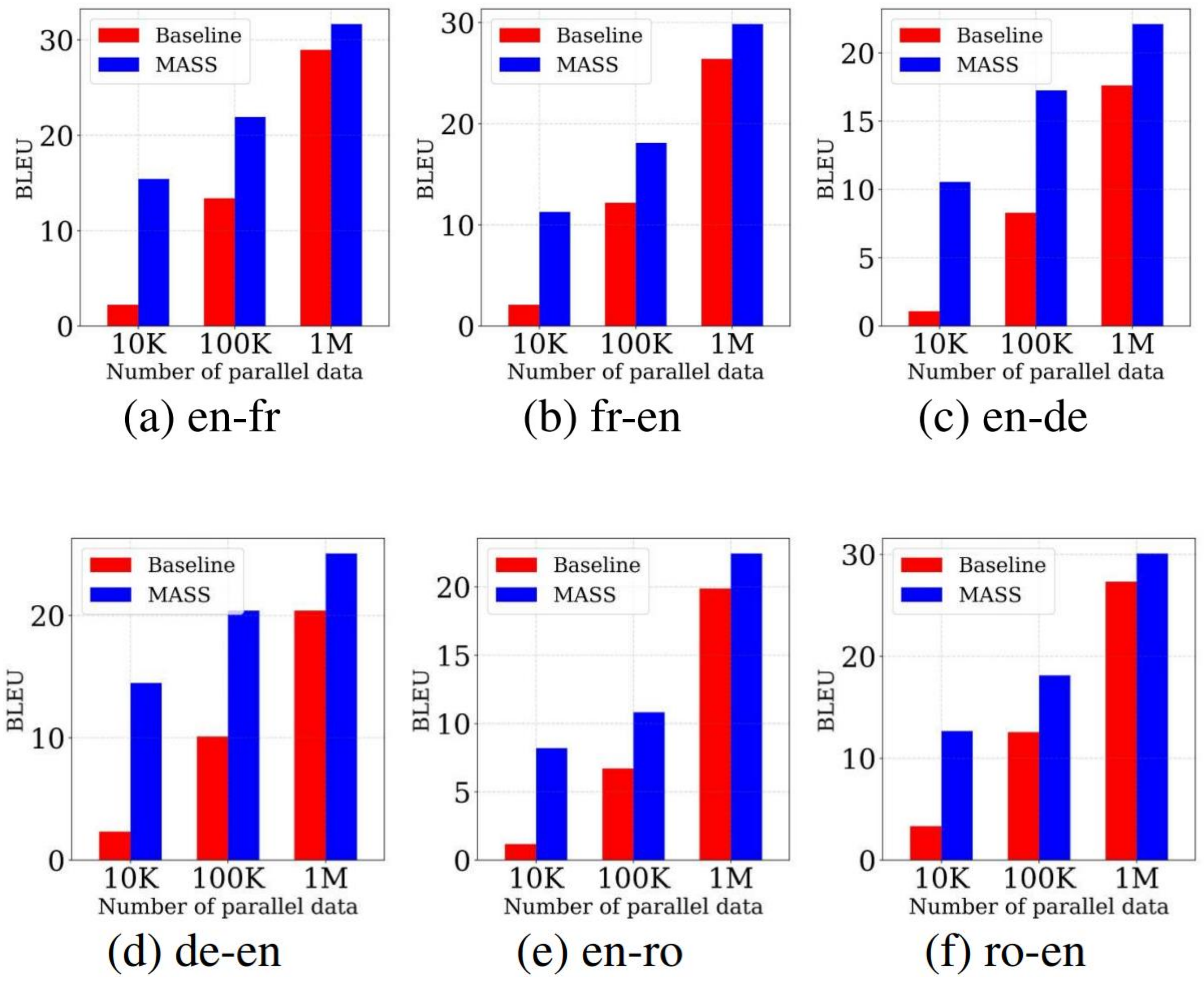
MASS: Pre-train for Sequence to Sequence Generation, [Song et al ICML 2019]

Unsupervised NMT



MASS: Pre-train for Sequence to Sequence Generation, [Song et al ICML 2019]

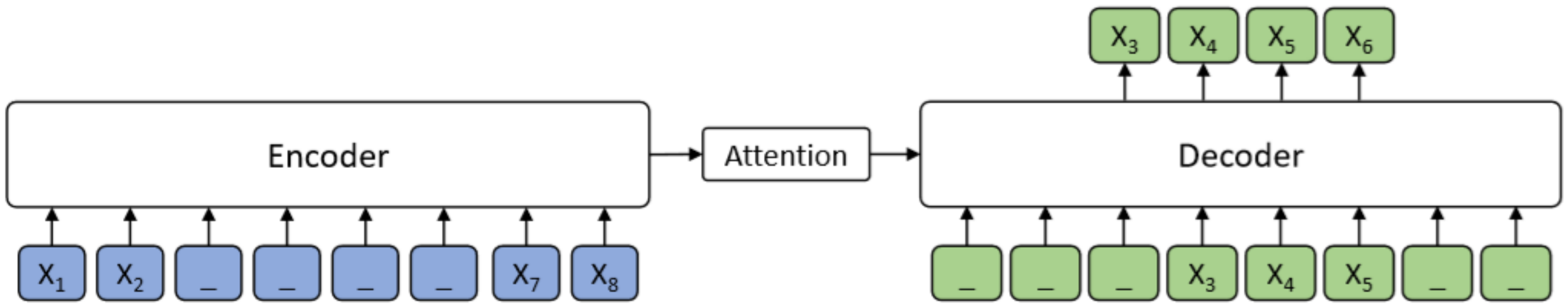
Low-resource NMT



MASS: Pre-train for Sequence to Sequence Generation, [Song et al ICML 2019]

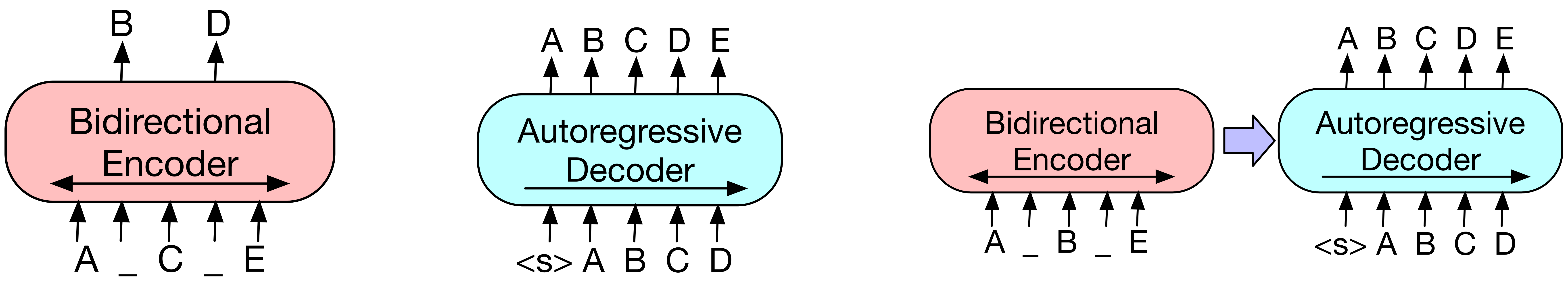
Summary

- Advantages
  - Unified sequence-to-sequence pretraining which jointly pretrains encoder, decoder and cross attention
  - Achieves improvements on zero-shot / unsupervised NMT
- Limitions
  - No evidence on rich resource NMT
  - Pre-training objective inconsistent with NMT, e.g. monolingual v.s. multilingual



MASS: Pre-train for Sequence to Sequence Generation, [Song et al ICML 2019]

BART: Denoising Sequence-to-Sequence Pre-training

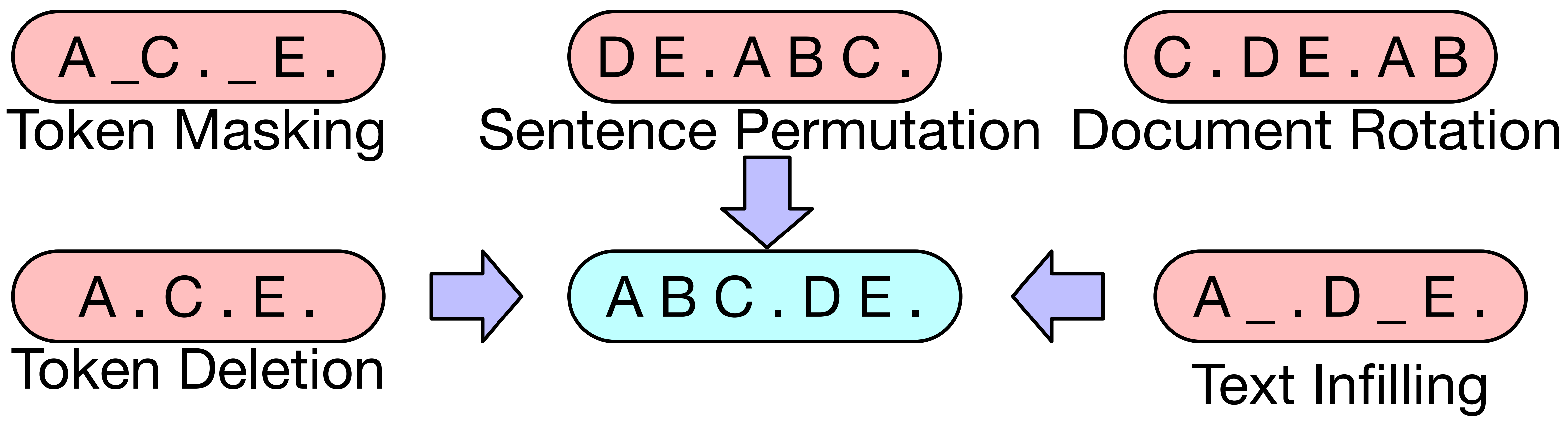


A schema comparison with BERT, GPT and BART.

- Standard sequence-to-sequence Transformer architecture
- Trained by corrupting documents and then optimizing a reconstruction loss
- Allows to apply *any* type of document corruption.

BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, [Lewis et al ACL 2020]

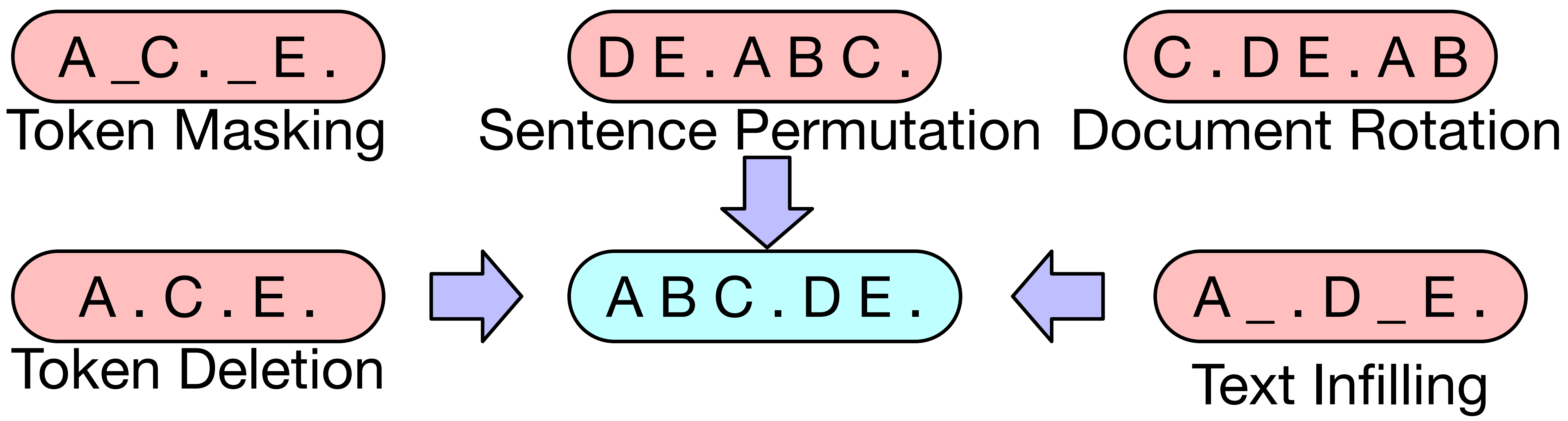
Noising the input



- Token masking: Random tokens are sampled and replaced with [MASK]
- Token deletion: Random tokens are deleted from the input.
- Text infilling: A number of span are sampled. Each span is replaced with [MASK]. 0-length span corresponding the insertion of [MASK].
- Sentence permutation: Sentences are shuffled with random order.
- Document Rotation: A token is chosen uniformly at random, and the document is rotated so that it begins with that token.

BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, [Lewis et al ACL 2020]

Noising the input

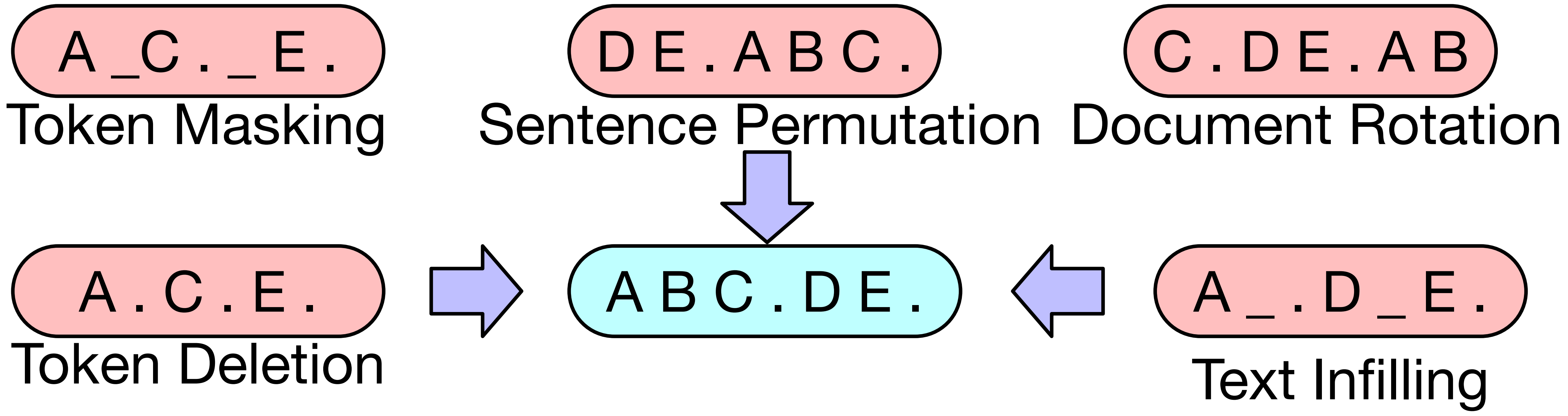


- Token masking: Random tokens are sampled and replaced with [MASK]
- Token deletion: Random tokens are deleted from the input.
- Text infilling: A number of span are sampled. Each span is replaced with [MASK]. 0-length span corresponding the insertion of [MASK].
- Sentence permutation: Sentences are shuffled with random order.
- Document Rotation: A token is chosen uniformly at random, and the document is rotated so that it begins with that token.

.

BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, [Lewis et al ACL 2020]

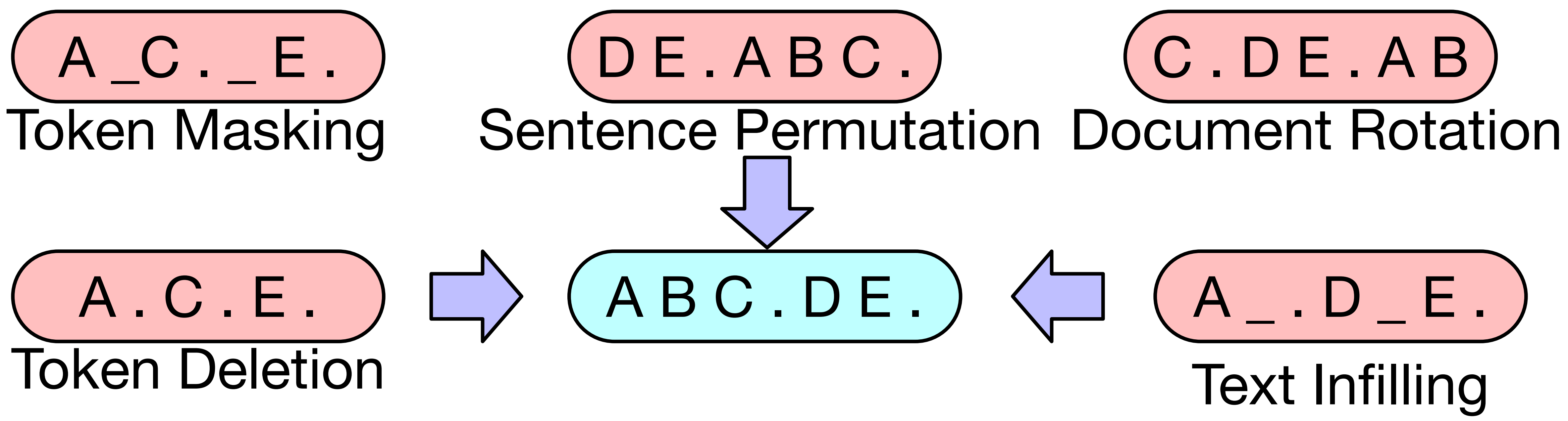
Noising the input



- Token masking: Random tokens are sampled and replaced with [MASK]
- Token deletion: Random tokens are deleted from the input.
- Text infilling: A number of span are sampled. Each span is replaced with [MASK]. 0-length span corresponding the insertion of [MASK].
- Sentence permutation: Sentences are shuffled with random order.
- Document Rotation: A token is chosen uniformly at random, and the document is rotated so that it begins with that token.

BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, [Lewis et al ACL 2020]

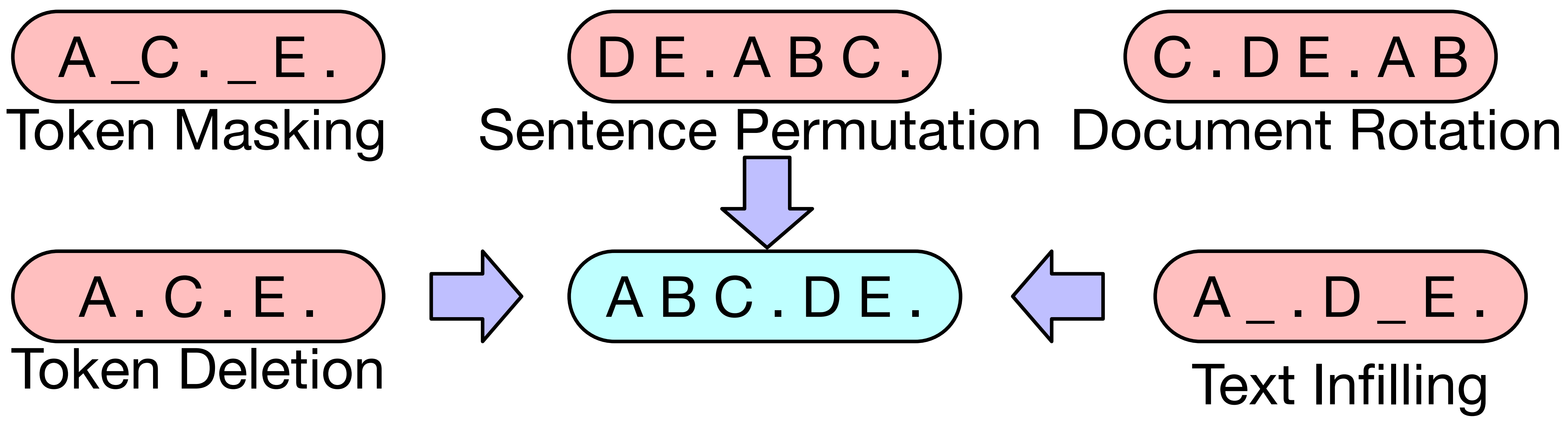
Noising the input



- Token masking: Random tokens are sampled and replaced with [MASK]
- Token deletion: Random tokens are deleted from the input.
- Text infilling: A number of span are sampled. Each span is replaced with [MASK]. 0-length span corresponding the insertion of [MASK].
- Sentence permutation: Sentences are shuffled with random order.
- Document Rotation: A token is chosen uniformly at random, and the document is rotated so that it begins with that token.

BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, [Lewis et al ACL 2020]

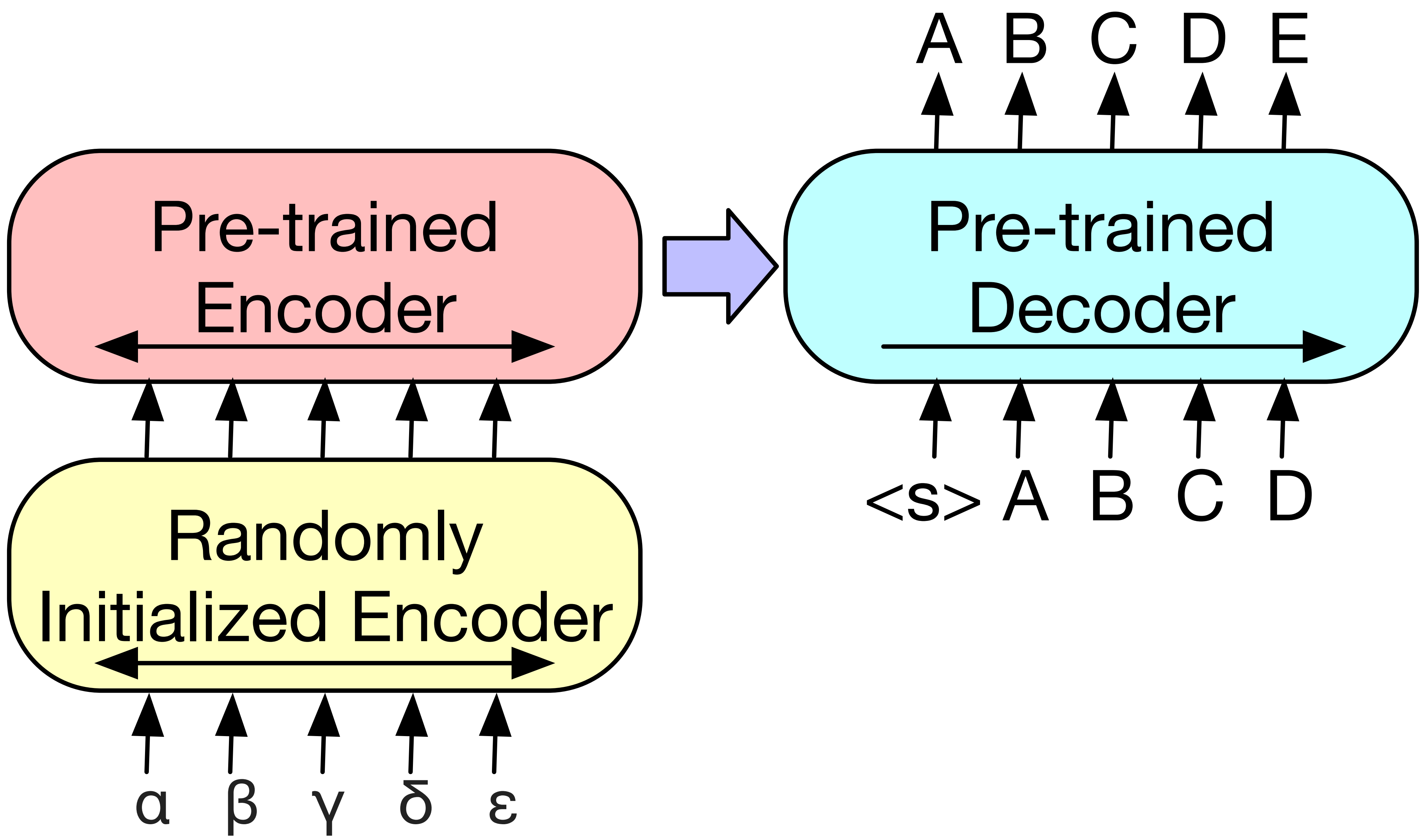
Noising the input



- Token masking: Random tokens are sampled and replaced with [MASK]
- Token deletion: Random tokens are deleted from the input.
- Text infilling: A number of span are sampled. Each span is replaced with [MASK]. 0-length span corresponding the insertion of [MASK].
- Sentence permutation: Sentences are shuffled with random order.
- Document Rotation: A token is chosen uniformly at random, and the document is rotated so that it begins with that token.

BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, [Lewis et al ACL 2020]

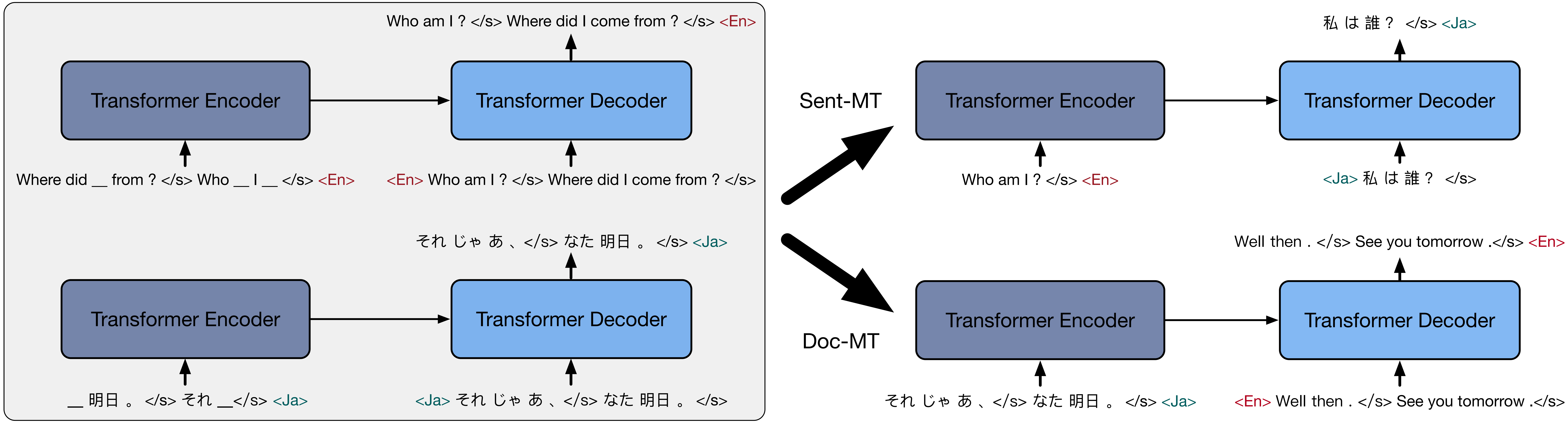
Fine-Tune on Neural Machine Translation



- Replace BART's encoder embedding layer with a new randomly initialized encoder
- The new encoder uses a separate vocabulary from the original BART mode
- First, freeze BART parameters and only update the randomly initialized source encoder. Then, jointly tuning with a few steps.

BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, [Lewis et al ACL 2020]

mBART: Multilingual Denoising Pre-training for Neural Machine Translation



Multilingual Denoising Pre-Training (mBART) | Fine-tuning on Machine Translation

- Multilingual denoising pre-training (25 languages)
  - Sentence permutation
  - Word-span masking
- Fine-tuning on MT with special language id

Multilingual Denoising Pre-training for Neural Machine Translation [Liu et al., TACL 2020]

Dataset

- Data: CC25 corpus
  - CC25 includes 25 languages from different families and with varied amounts of text from Common Crawl (CC)
  - Rebalanced the corpus by up/down-sampling text

$$\lambda_i = \frac{1}{p_i} \cdot \frac{p_i^\alpha}{\sum_i p_i^\alpha},$$

  - Sentence Piece which includes 25,000 subwords
  - Noisy function follows BART

Code	Language	Tokens/M	Size/GB
En	English	55608	300.8
Ru	Russian	23408	278.0
Vi	Vietnamese	24757	137.3
Ja	Japanese	530 (*)	69.3
De	German	10297	66.6
Ro	Romanian	10354	61.4
Fr	French	9780	56.8
Fi	Finnish	6730	54.3
Ko	Korean	5644	54.2
Es	Spanish	9374	53.3
Zh	Chinese (Sim)	259 (*)	46.9
It	Italian	4983	30.2
Nl	Dutch	5025	29.3
Ar	Arabic	2869	28.0
Tr	Turkish	2736	20.9
Hi	Hindi	1715	20.2
Cs	Czech	2498	16.3
Lt	Lithuanian	1835	13.7
Lv	Latvian	1198	8.8
Kk	Kazakh	476	6.4
Et	Estonian	843	6.1
Ne	Nepali	237	3.8
Si	Sinhala	243	3.6
Gu	Gujarati	140	1.9
My	Burmese	56	1.6

Multilingual Denoising Pre-training for Neural Machine Translation [Liu et al., TACL 2020]

mBART: Low-medium translation results

Languages	**En-Gu**		**En-Kk**		**En-Vi**		**En-Tr**		**En-Ja**		**En-Ko**	
Data Source	**WMT19**		**WMT19**		**IWSLT15**		**WMT17**		**IWSLT17**		**IWSLT17**	
Size	10K		91K		133K		207K		223K		230K	
Direction	←	→	←	→	←	→	←	→	←	→	←	→
Random	0.0	0.0	0.8	0.2	23.6	24.8	12.2	9.5	10.4	12.3	15.3	16.3
mBART25	**0.3**	**0.1**	**7.4**	**2.5**	**36.1**	**35.4**	**22.5**	**17.8**	**19.1**	**19.4**	**24.6**	**22.6**

Languages	**En-Nl**		**En-Ar**		**En-It**		**En-My**		**En-Ne**		**En-Ro**	
Data Source	**IWSLT17**		**IWSLT17**		**IWSLT17**		**WAT19**		**FLoRes**		**WMT16**	
Size	237K		250K		250K		259K		564K		608K	
Direction	←	→	←	→	←	→	←	→	←	→	←	→
Random	34.6	29.3	27.5	16.9	31.7	28.0	23.3	34.9	7.6	4.3	34.0	34.3
mBART25	**43.3**	**34.8**	**37.6**	**21.6**	**39.8**	**34.0**	**28.3**	**36.9**	**14.5**	**7.4**	**37.8**	**37.7**

Languages	**En-Si**		**En-Hi**		**En-Et**		**En-Lt**		**En-Fi**		**En-Lv**	
Data Source	**FLoRes**		**ITTB**		**WMT18**		**WMT19**		**WMT17**		**WMT17**	
Size	647K		1.56M		1.94M		2.11M		2.66M		4.50M	
Direction	←	→	←	→	←	→	←	→	←	→	←	→
Random	7.2	1.2	10.9	14.2	22.6	17.9	18.1	12.1	21.8	20.2	15.6	12.9
mBART25	**13.7**	**3.3**	**23.5**	**20.8**	**27.8**	**21.4**	**22.4**	**15.3**	**28.5**	**22.4**	**19.3**	**15.9**

Low resource: more than 6 BLEU. But fails in extremely low-resource setting

Multilingual Denoising Pre-training for Neural Machine Translation [Liu et al., TACL 2020]

mBART: Low-medium translation results

Languages	En-Gu		En-Kk		En-Vi		En-Tr		En-Ja		En-Ko	
Data Source	**WMT19**		**WMT19**		**IWSLT15**		**WMT17**		**IWSLT17**		**IWSLT17**	
Size	10K		91K		133K		207K		223K		230K	
Direction	←	→	←	→	←	→	←	→	←	→	←	→
Random	0.0	0.0	0.8	0.2	23.6	24.8	12.2	9.5	10.4	12.3	15.3	16.3
mBART25	**0.3**	**0.1**	**7.4**	**2.5**	**36.1**	**35.4**	**22.5**	**17.8**	**19.1**	**19.4**	**24.6**	**22.6**

Languages	En-Nl		En-Ar		En-It		En-My		En-Ne		En-Ro	
Data Source	**IWSLT17**		**IWSLT17**		**IWSLT17**		**WAT19**		**FLoRes**		**WMT16**	
Size	237K		250K		250K		259K		564K		608K	
Direction	←	→	←	→	←	→	←	→	←	→	←	→
Random	34.6	29.3	27.5	16.9	31.7	28.0	23.3	34.9	7.6	4.3	34.0	34.3
mBART25	**43.3**	**34.8**	**37.6**	**21.6**	**39.8**	**34.0**	**28.3**	**36.9**	**14.5**	**7.4**	**37.8**	**37.7**

Languages	En-Si		En-Hi		En-Et		En-Lt		En-Fi		En-Lv	
Data Source	**FLoRes**		**ITTB**		**WMT18**		**WMT19**		**WMT17**		**WMT17**	
Size	647K		1.56M		1.94M		2.11M		2.66M		4.50M	
Direction	←	→	←	→	←	→	←	→	←	→	←	→
Random	7.2	1.2	10.9	14.2	22.6	17.9	18.1	12.1	21.8	20.2	15.6	12.9
mBART25	**13.7**	**3.3**	**23.5**	**20.8**	**27.8**	**21.4**	**22.4**	**15.3**	**28.5**	**22.4**	**19.3**	**15.9**

Low resource: more than 6 BLEU. But fails in extremely low-resource setting

Medium resource: more than 3 BLEU

Multilingual Denoising Pre-training for Neural Machine Translation [Liu et al., TACL 2020]

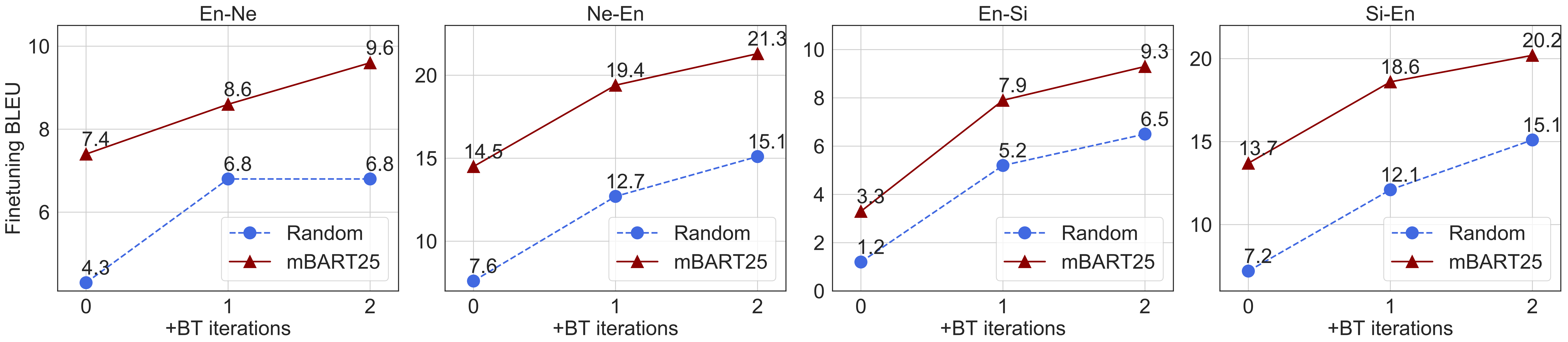
mBART: Rich-resource translation

Languages	Cs	Es	Zh	De	Ru	Fr
Size	11M	15M	25M	28M	29M	41M
Random	16.5	33.2	**35.0**	**30.9**	**31.5**	**41.4**
mBART25	**18.0**	**34.0**	33.3	30.5	31.3	41.0

- Pre-training slightly hurts performance when >25M parallel sentence are available.
- When a significant amount of bi-text data is given, supervised training are supposed to wash out the pre-trained weights completely.

Multilingual Denoising Pre-training for Neural Machine Translation [Liu et al., TACL 2020]

mBART: Pre-training complementary to BT



- Test on low resource FLoRes dataset [Guzmán et al., 2019]
- Use the same monolingual data to generate BT data
- Initializing the model with mBART25 pre-trained parameters improves BLEU scores at each iteration of back translation, resulting in new state-of-the-art results in all four translation directions

Is pre-training on multilingual better than on single language?

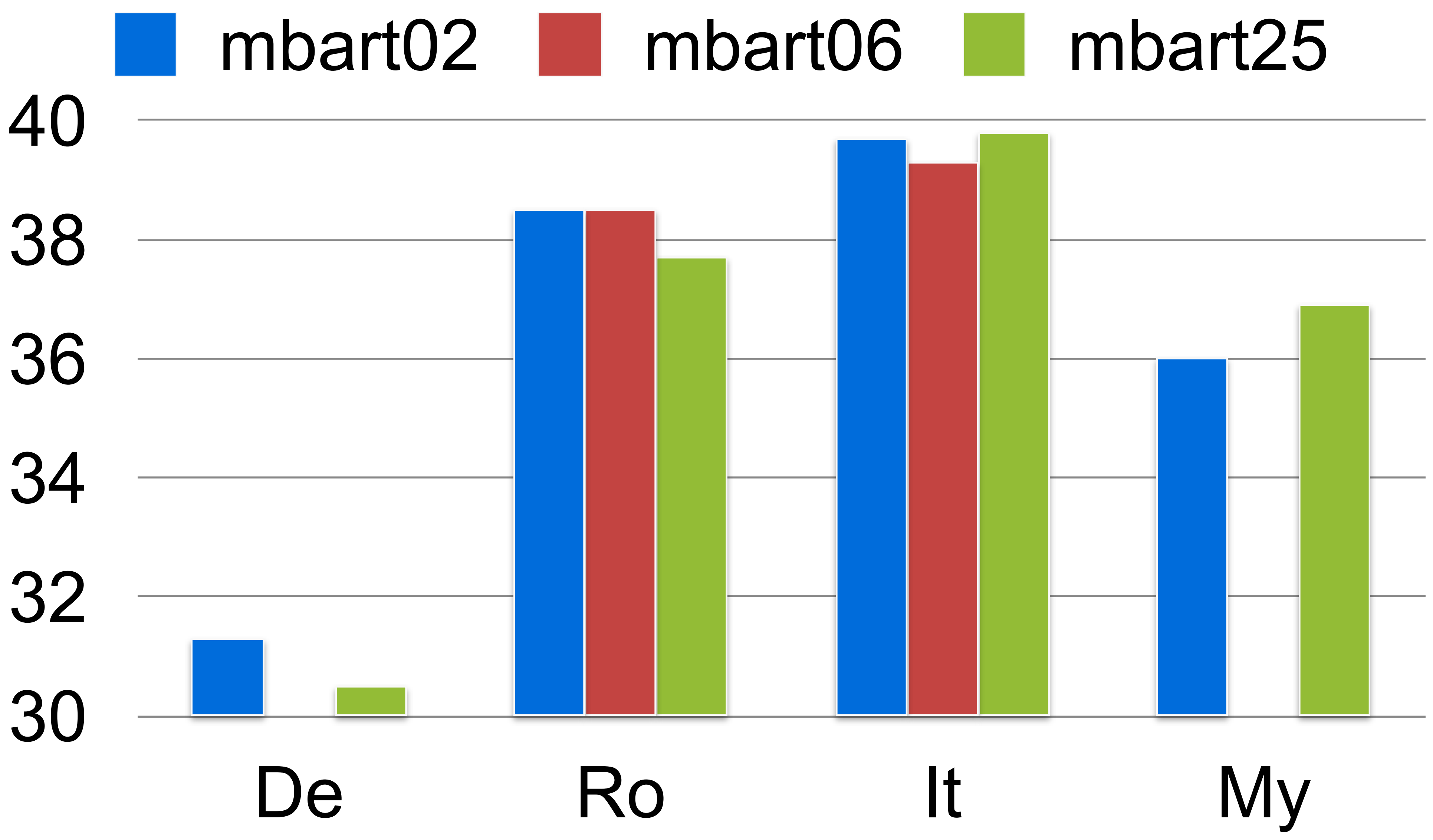
Pre-training		Fine-tuning		
Model	**Data**	**En→Ro**	**Ro→En**	**+BT**
Random	None	34.3	34.0	36.8
XLM (2019)	En Ro	-	35.6	38.5
MASS (2019)	En Ro	-	-	39.1
BART (2019)	En	-	-	38.0
XLM-R (2019)	CC100	35.6	35.8	-
BART-En	En	36.0	35.8	37.4
BART-Ro	Ro	37.6	36.8	38.1
mBART02	En Ro	**38.5**	**38.5**	**39.9**
mBART25	CC25	37.7	37.8	38.8

- BART model trained on the same En and Ro data only. Both have improvements over baselines, while worse than mBART results, indicating pre-training in a multilingual setting is essential.
- Combining BT leads to additional gains, resulting in a new state-of-the-art for Ro-En translation
- mBART02 is better than mBART25. The more seems not the better?

Multilingual Denoising Pre-training for Neural Machine Translation [Liu et al., TACL 2020]

How many languages should you pre-train on?

Languages	De	Ro	It	My	En
Size/GB	66.6	61.4	30.2	1.6	300.8
mBART02	**31.3**	**38.5**	39.7	36.5	
mBART06	-	**38.5**	39.3	-	
mBART25	30.5	37.7	**39.8**	**36.9**	



- Pretraining on more languages helps most when the target language monolingual data is limited
- When monolingual data is plentiful (De, Ro), pre-training on multiple languages slightly hurts the final results (<1 BLEU)

Multilingual Denoising Pre-training for Neural Machine Translation [Liu et al., TACL 2020]

Analysis: Generalization to unseen languages

	Monolingual	Nl-En	En-Nl	Ar-En	En-Ar	Nl-De	De-Nl
Random	None	34.6 (-8.7)	29.3 (-5.5)	27.5 (-10.1)	16.9 (-4.7)	21.3 (-6.4)	20.9 (-5.2)
mBART02	En Ro	41.4 (-2.9)	34.5 (-0.3)	34.9 (-2.7)	21.2 (-0.4)	26.1 (-1.6)	25.4 (-0.7)
mBART06	En Ro Cs It Fr Es	43.1 (-0.2)	34.6 (-0.2)	37.3 (-0.3)	21.1 (-0.5)	26.4 (-1.3)	25.3 (-0.8)
mBART25	All	**43.3**	**34.8**	**37.6**	**21.6**	**27.7**	**26.1**

Nl-De and Ar are not included in the pre-training corpus

- mBART can improve performance even with fine tuning for languages that did not appear in the pre-training corpora,
- Pre-training has language universal aspects, especially within the parameters learned at the Transformer layers.
- The more pre-trained languages the better

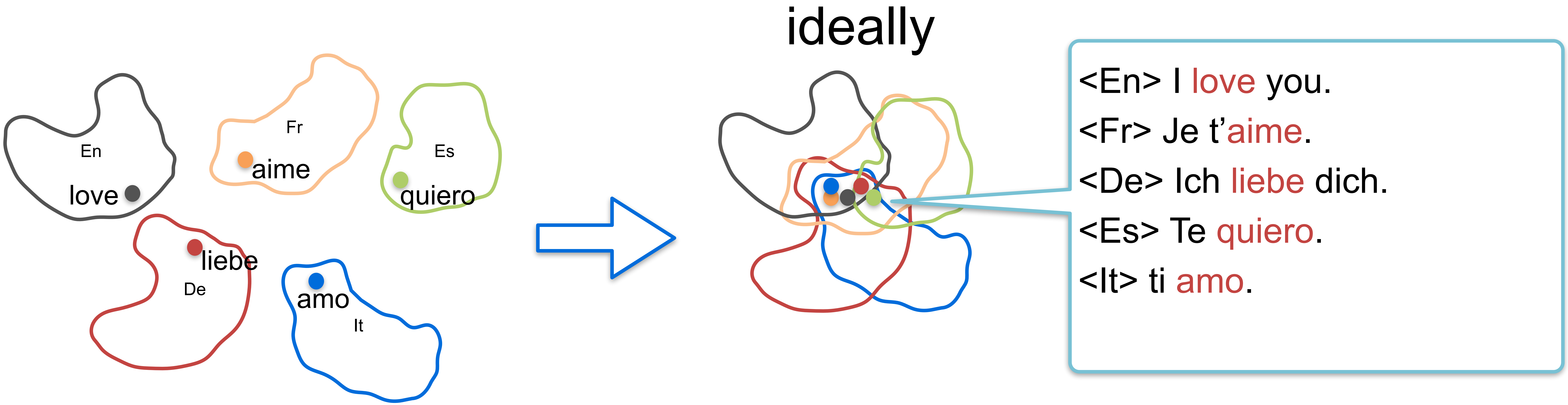
Multilingual Denoising Pre-training for Neural Machine Translation [Liu et al., TACL 2020]

Multilingual Training

How can we build a single unified Multilingual MT models with superior performance on all language directions?

Idea 1: Aligning Semantic Representations across Languages

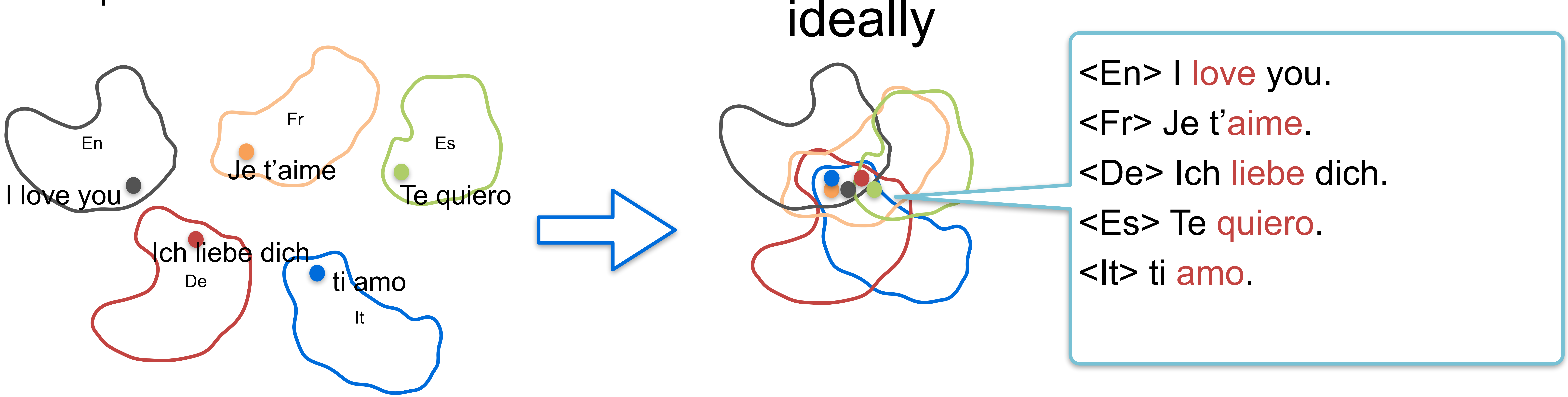
- Key idea:
  - Words in difference languages with the same meaning should have the same embedding, but the training objective does not necessarily encourage that!



Pre-training Multilingual Neural Machine Translation by Leveraging Alignment Information [Lin et al., EMNLP 2020]
Contrastive Learning for Many-to-many Multilingual Neural Machine Translation [Pan et al., ACL 2021]

Proposed mRASP: Aligning Semantic Representations across Languages

- Key idea:
  - Words in difference languages with the same meaning should have the same embedding
  - Parallel sentences in difference languages should have the same representation

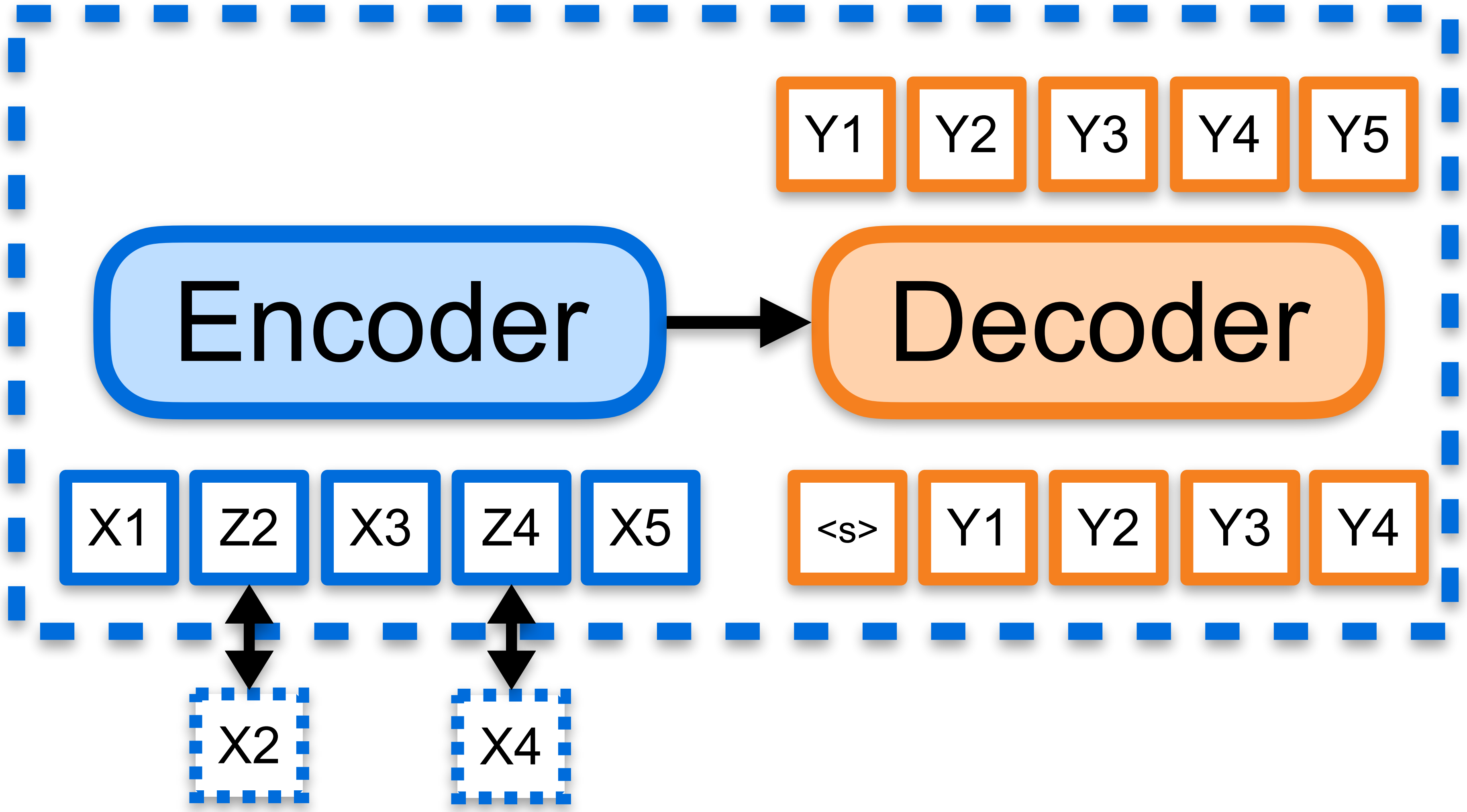


mRASP

Pre-training Multilingual Neural Machine Translation by Leveraging Alignment Information
[Lin, Pan, Wang, Qiu, Feng, Zhou, **Lei Li**, EMNLP2020]
Contrastive Learning for Many-to-many Multilingual Neural Machine Translation [Pan, Wu, Wang, **Lei Li**, ACL 2021]

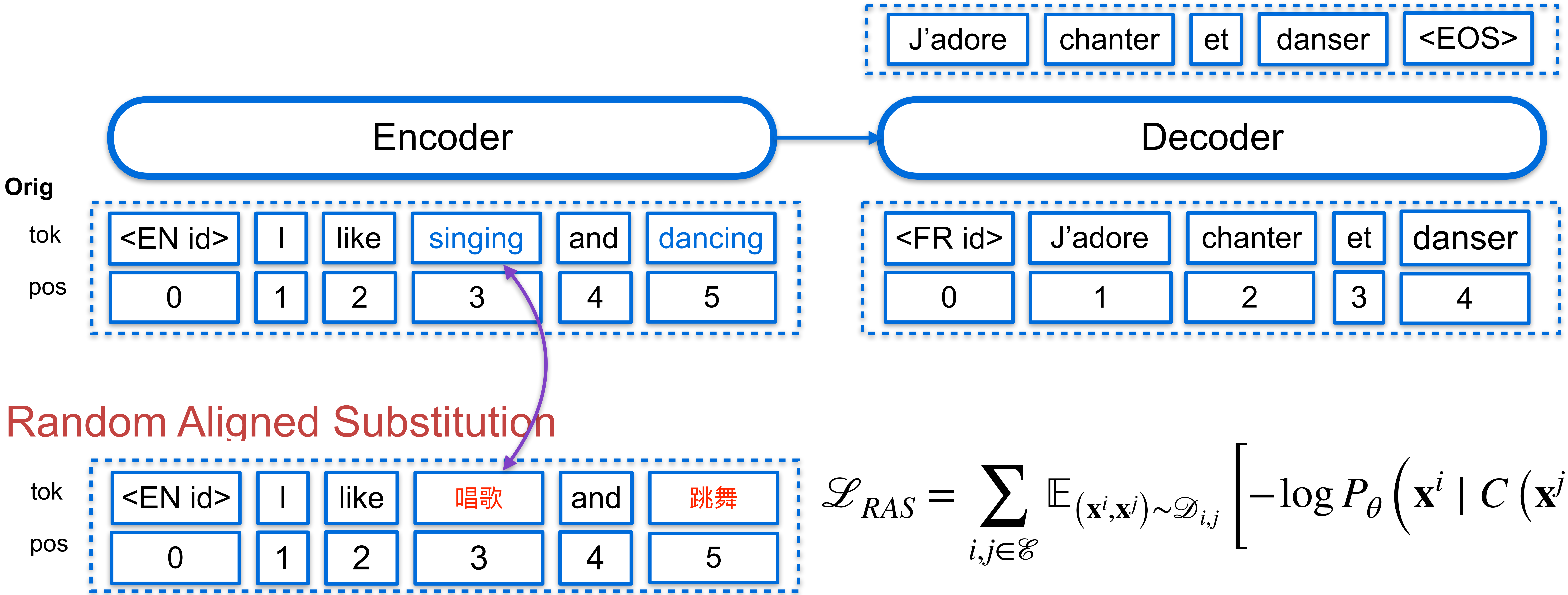
Aligning Semantic Representations across Languages

- **mRASP**: **m**ultilingual **R**andom **A**ligned **S**ubstitution **P**re-training
  - Multilingual Pre-training Approach
  - RAS: specially designed training method to align semantic embeddings



Pre-training Multilingual Neural Machine Translation by Leveraging Alignment Information [Lin et al., EMNLP 2020]

mRASP: Random Aligned Substitution



$$\mathscr{L}_{RAS} = \sum_{i,j\in\mathscr{E}} \mathbb{E}_{(\mathbf{x}^i,\mathbf{x}^j)\sim\mathscr{D}_{i,j}} \left[-\log P_\theta\left(\mathbf{x}^i \mid C\left(\mathbf{x}^j\right)\right)\right]$$

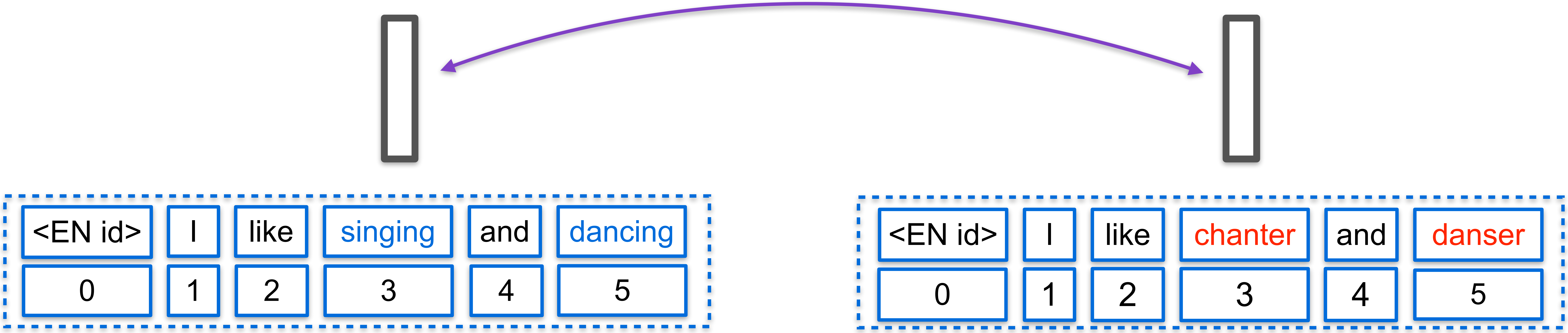
Randomly replace a source word to its synonym in different language.

Pre-training Multilingual Neural Machine Translation by Leveraging Alignment Information
[Lin, Pan, Wang, Qiu, Feng, Zhou, **Lei Li**, EMNLP2020]

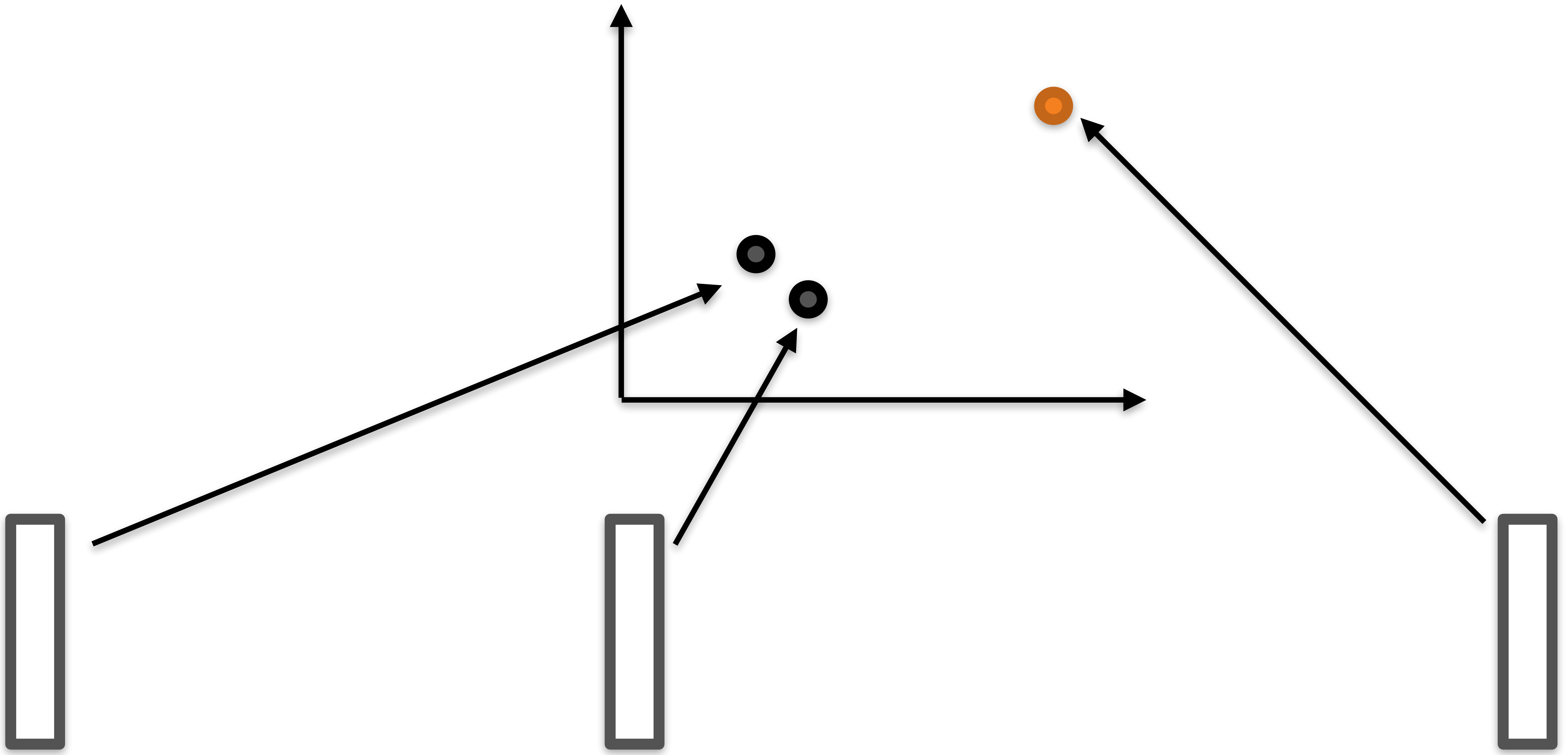
mRASP: Bringing Synonym Representations Closer

$$\mathcal{L}^{pre} = \sum_{i,j\in\mathcal{E}} \mathbb{E}_{(\mathbf{x}^i,\mathbf{x}^j)\sim\mathcal{D}_{i,j}}\left[-\log P_\theta\left(\mathbf{x}^i \mid C\left(\mathbf{x}^j\right)\right)\right]$$

training with translation loss to bring closer



Idea 2: Bring parallel sentence representations closer



<En> I love you. <Fr> Je t’aime. <En> It’s sunny.

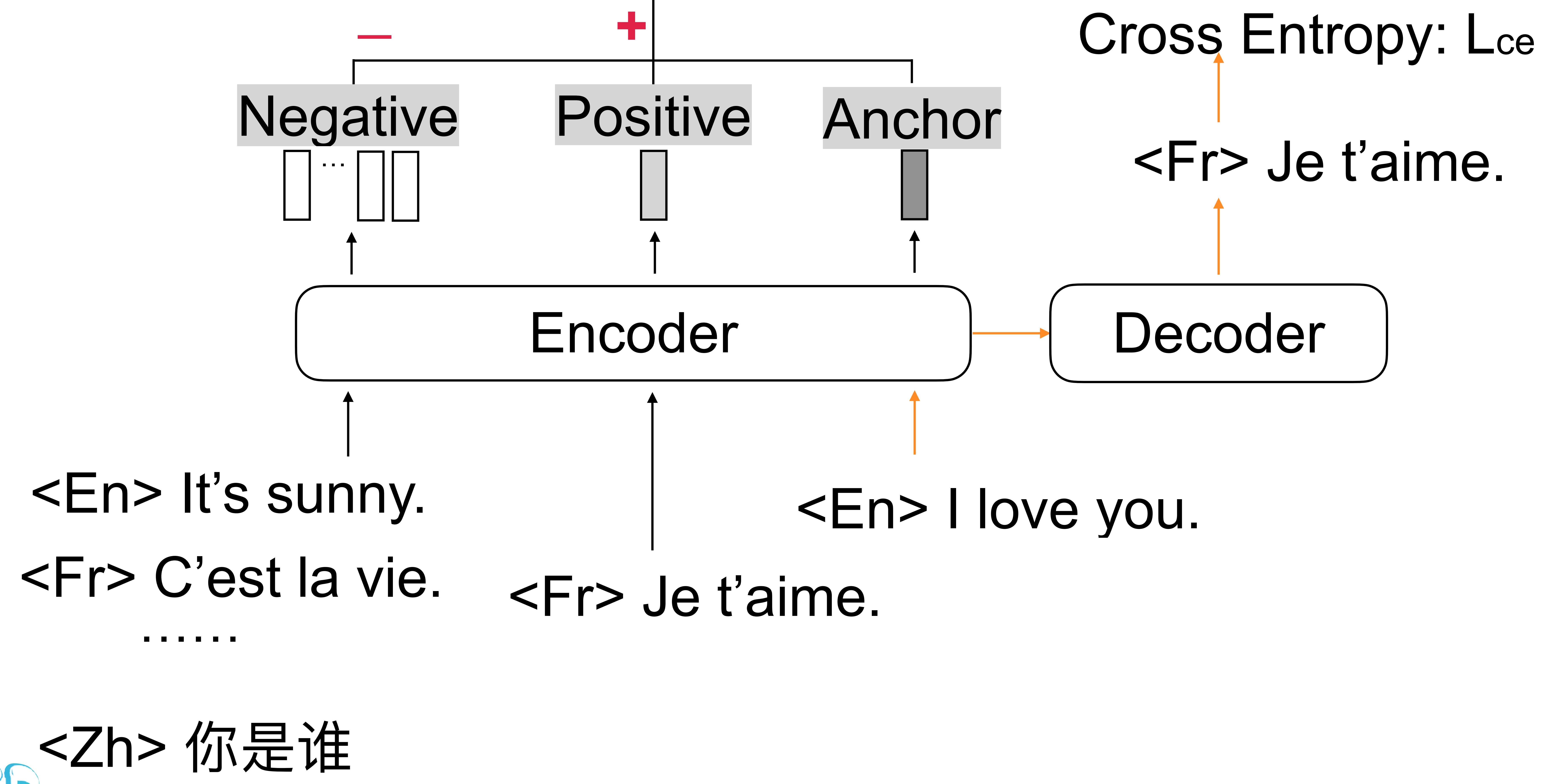
Contrastive Learning for Many-to-many Multilingual Neural Machine Translation [Pan, Wu, Wang, **Lei Li**, ACL 2021]

mRASP2: Contrastive Learning to Bring Sentence Representations Closer

Contrastive Loss: Lctr

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

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

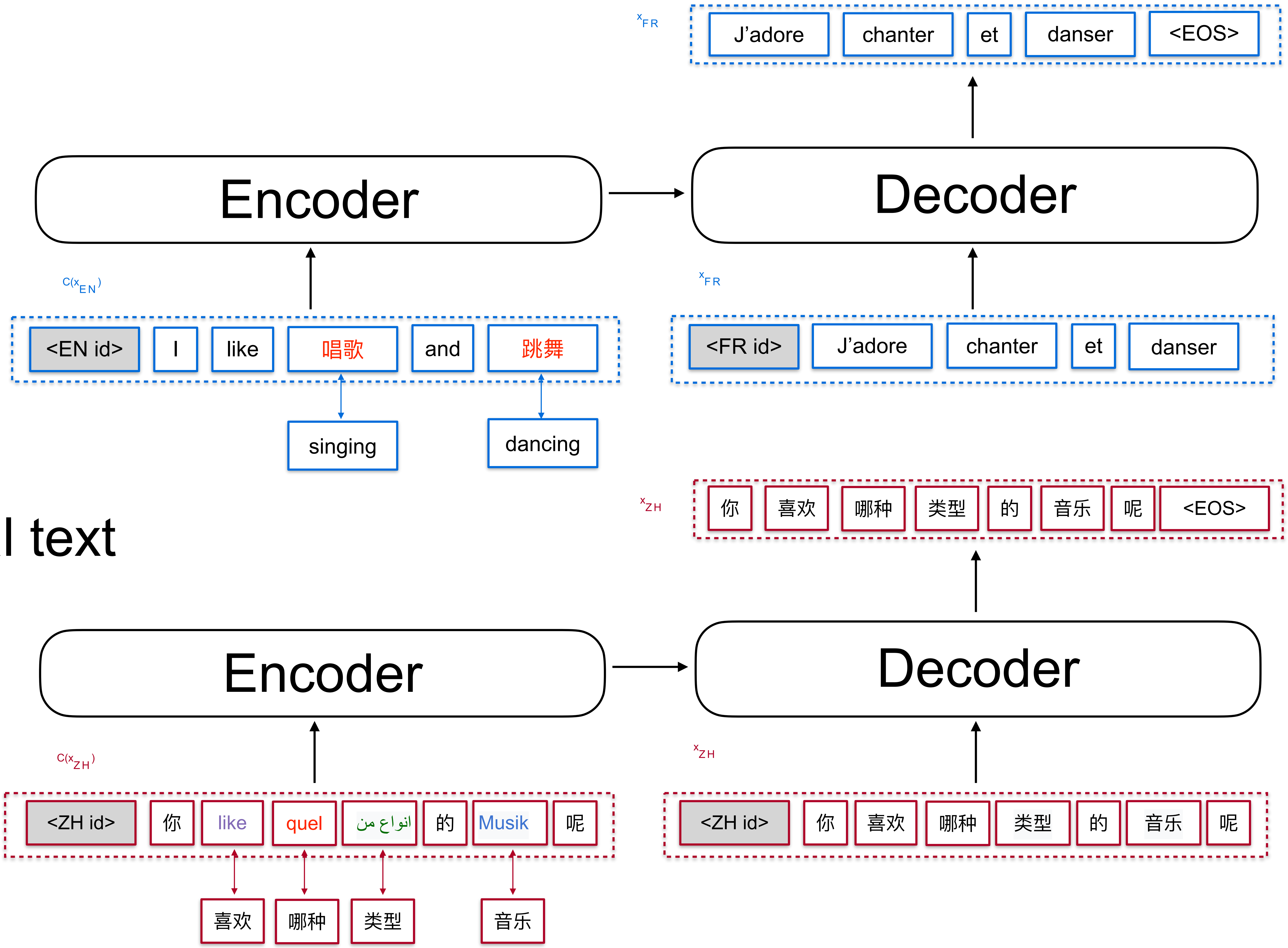


Contrastive Learning for Many-to-many Multilingual Neural Machine Translation [Pan, Wu, Wang, **Lei Li**, ACL 2021]

mRASP2: Integrating Monolingual Data in Unified Training

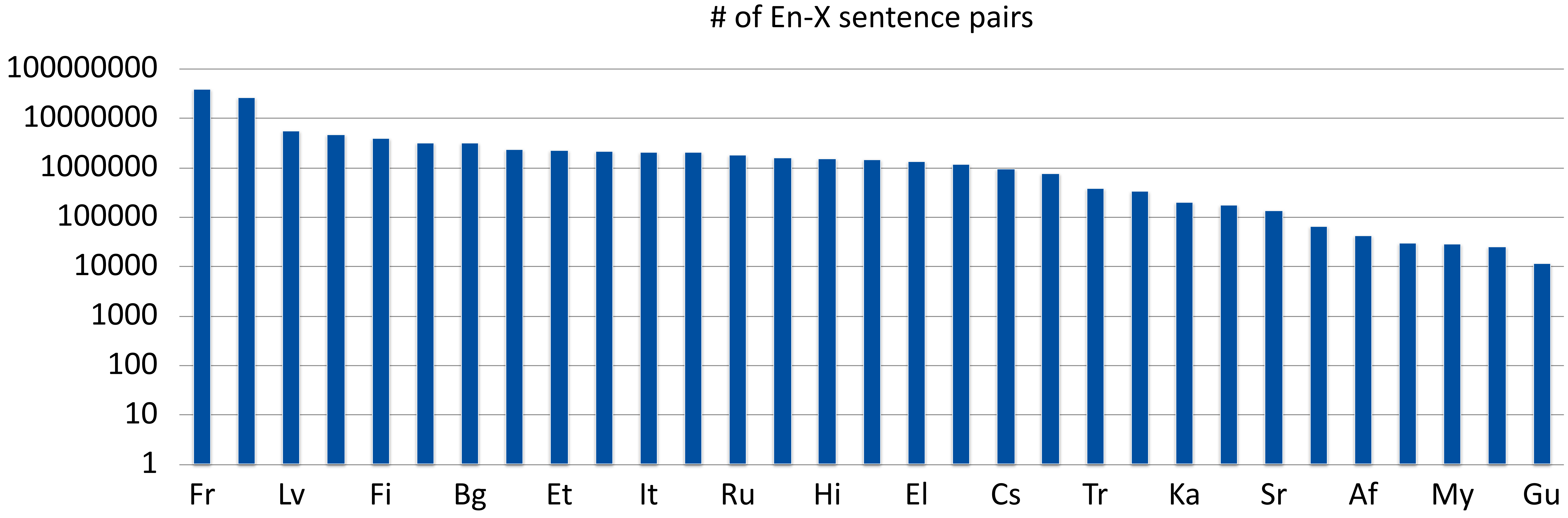
- Parallel text

- Monolingual text



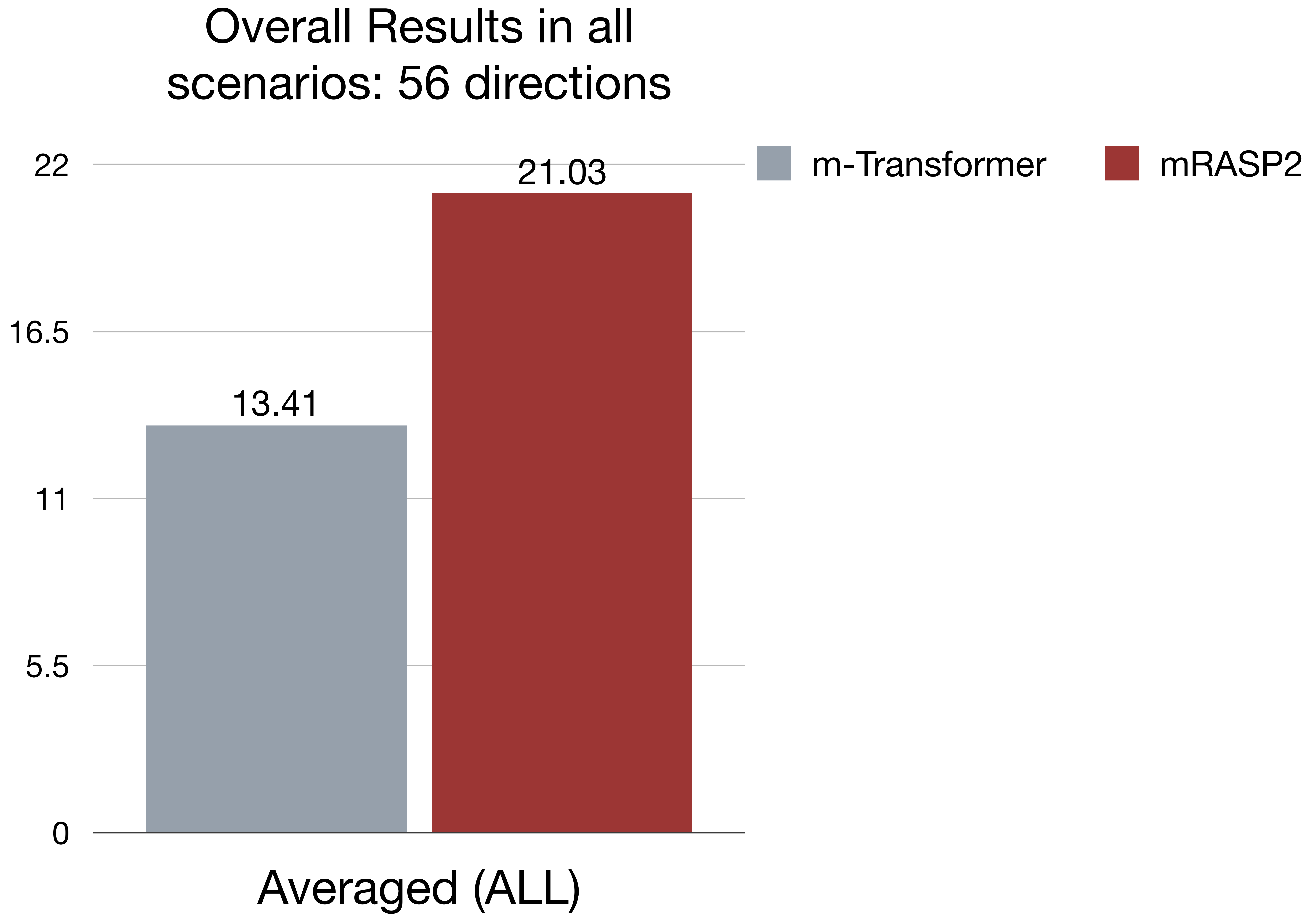
Training Data for mRASP

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

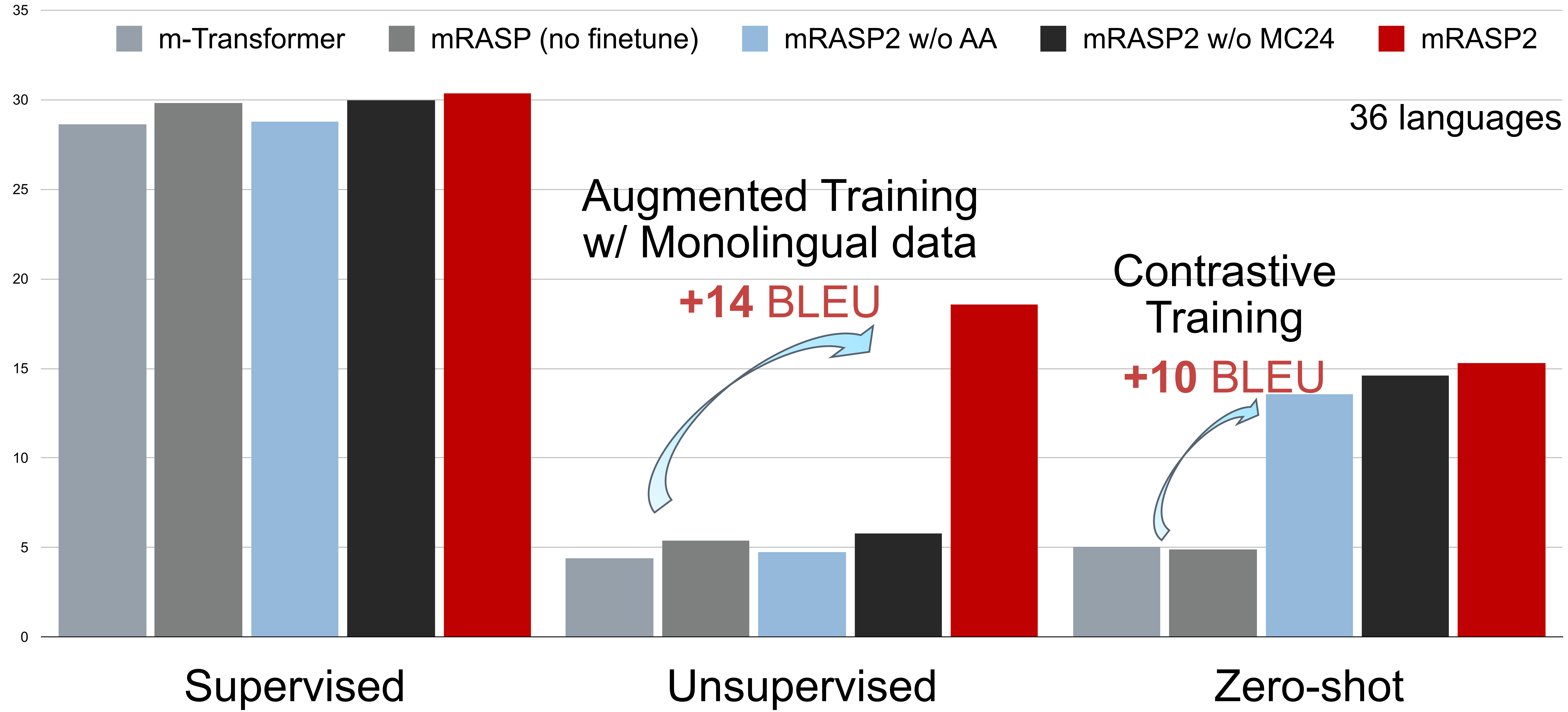


Pre-training Multilingual Neural Machine Translation by Leveraging Alignment Information [Lin et al., EMNLP 2020]

mRASP2: a single MNMT model (no fine-tuning)

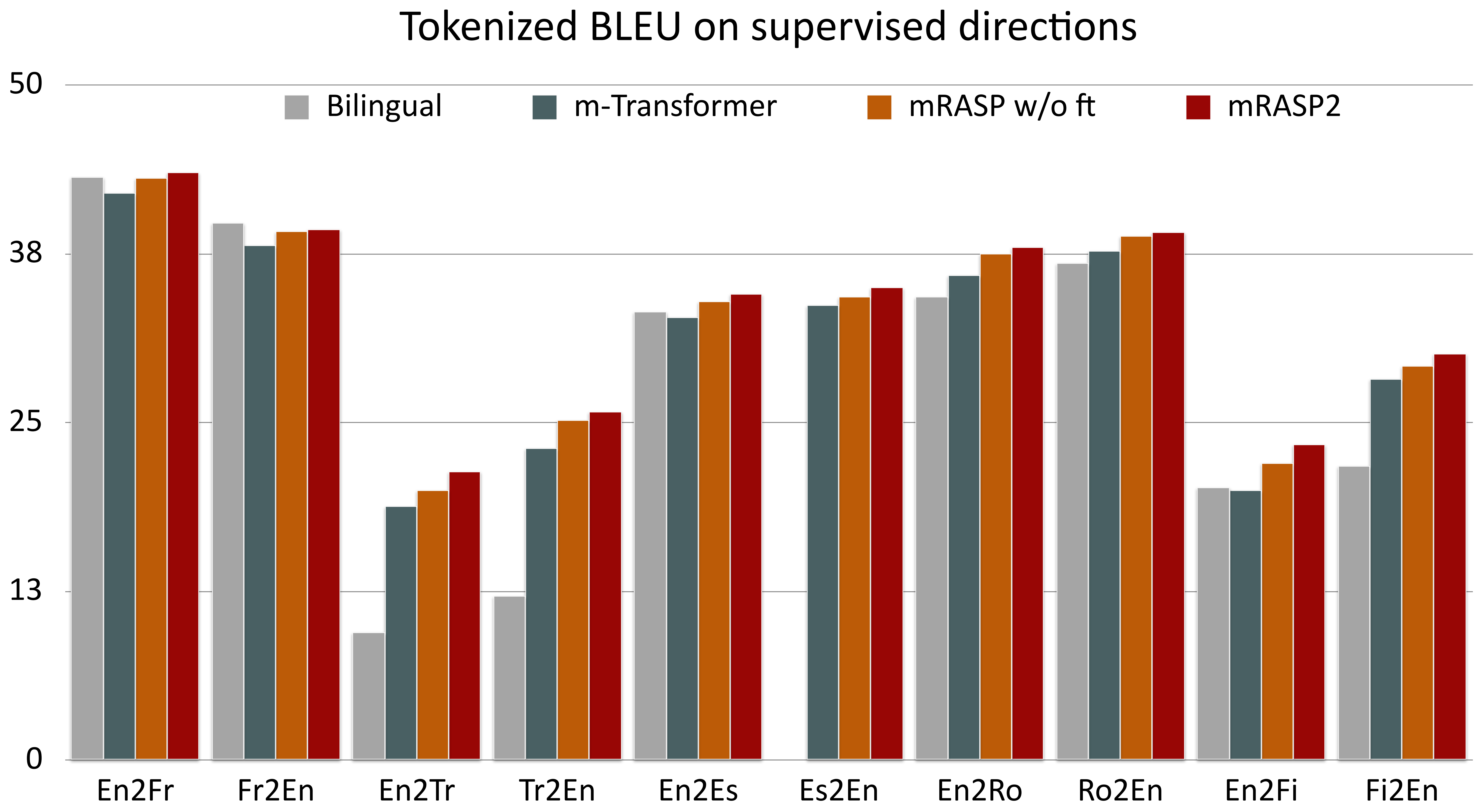


mRASP significantly improves Zero-shot and Unsupervised Translation

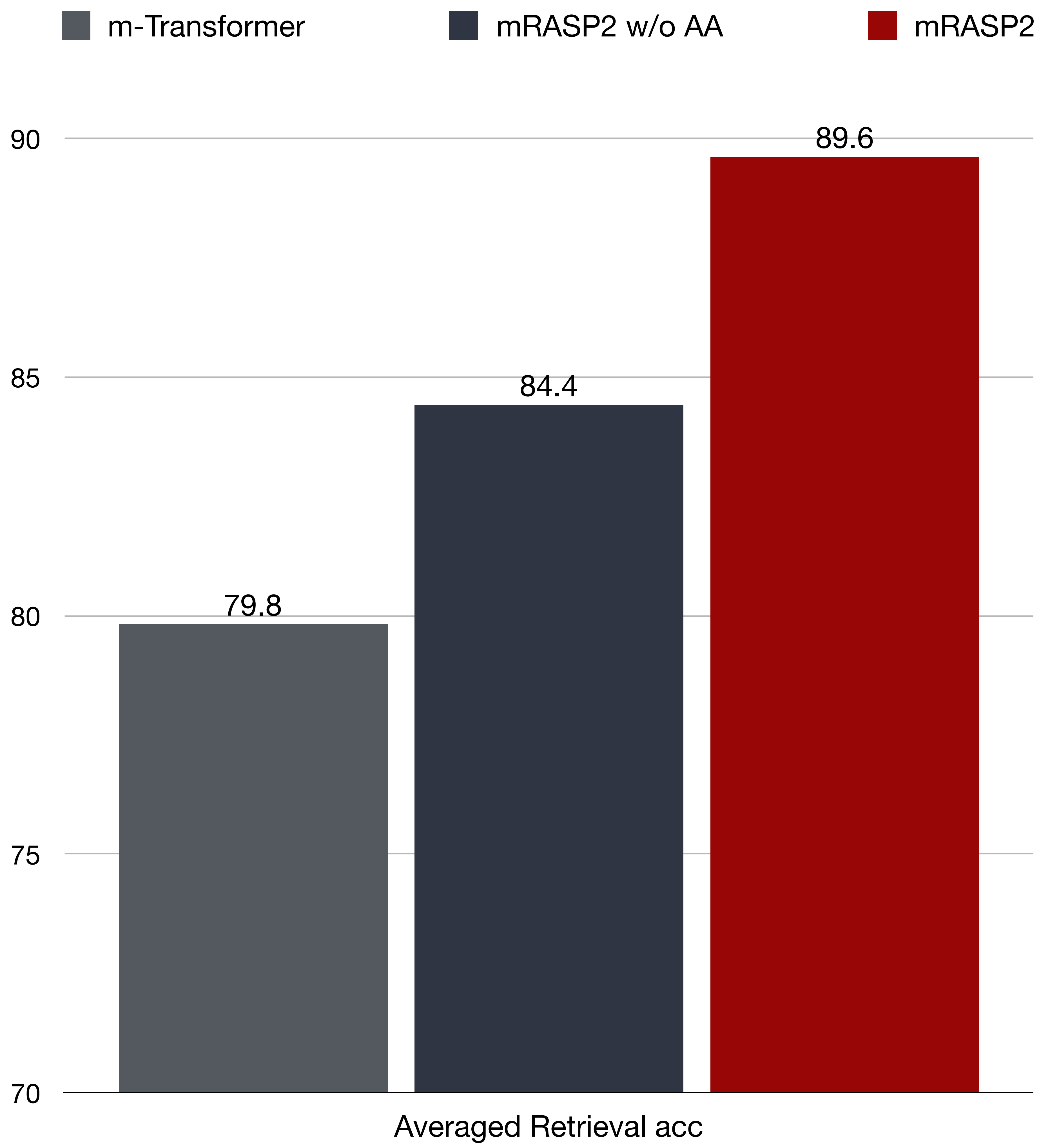


Pre-training Multilingual Neural Machine Translation by Leveraging Alignment Information
[Lin, Pan, Wang, Qiu, Feng, Zhou, **Lei Li**, EMNLP2020]
Contrastive Learning for Many-to-many Multilingual Neural Machine Translation [Pan, Wu, Wang, **Lei Li**, ACL 2021]

mRASP2: Comparable or Better Performance on Supervised Directions



Better Semantic Alignment: Sentence Retrieval

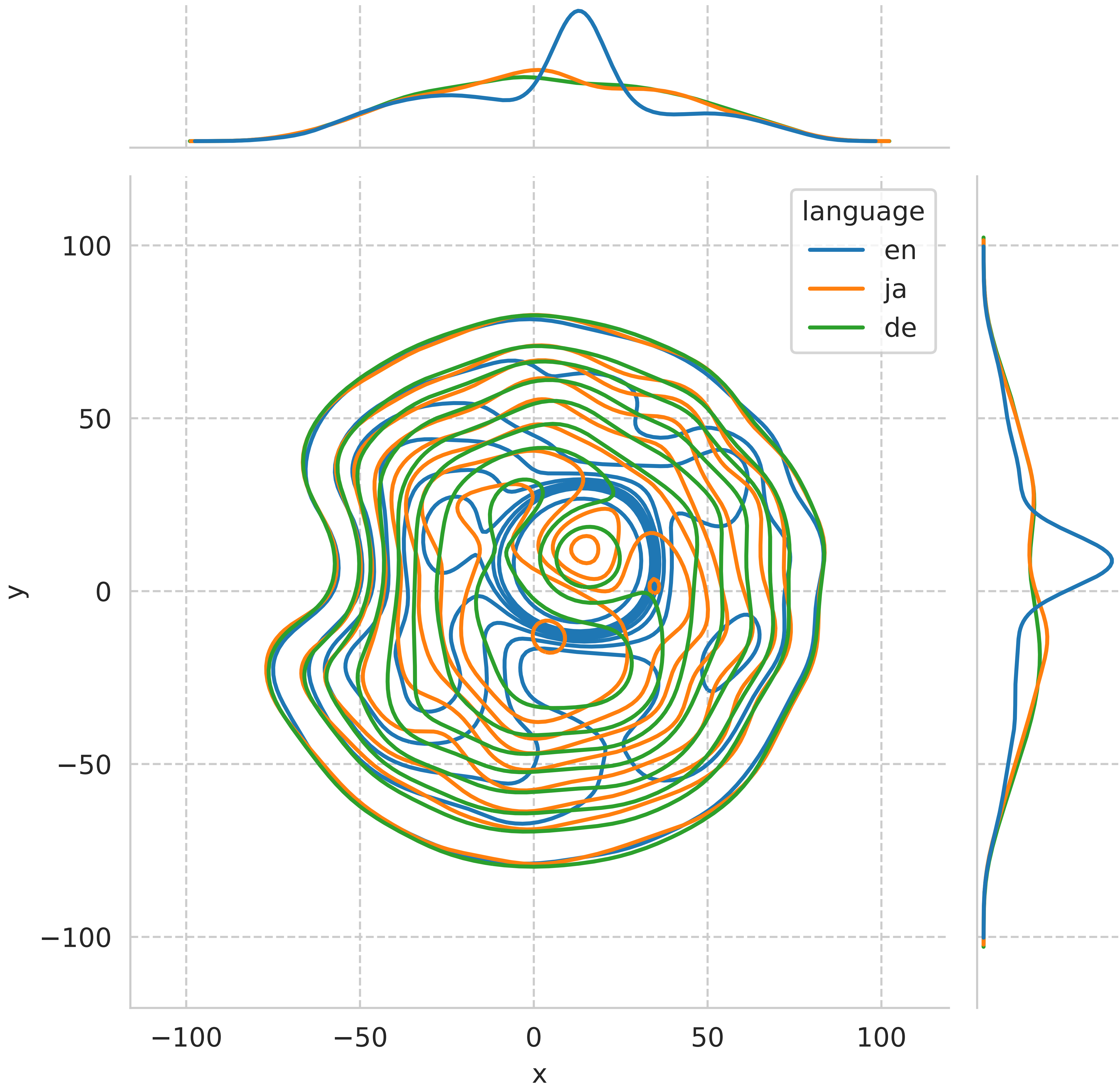


15-way parallel test set(Ted-M): 2284 samples

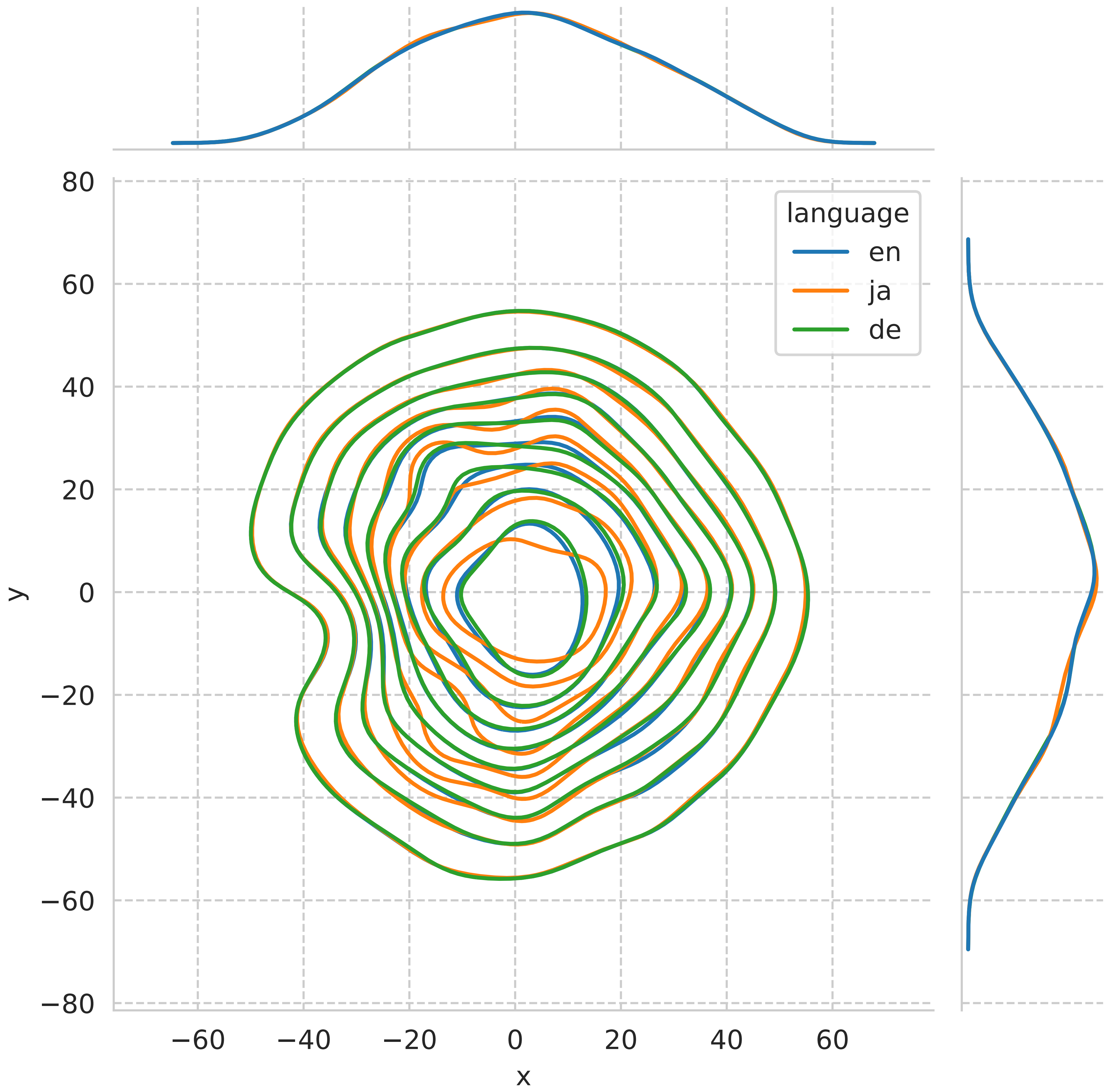
Contrastive Learning and Randomly Aligned Substitution both contribute to the improvement on sentence retrieval

mRASP Produces Better Semantic Alignment

m-Transformer



mRASP



Contrastive Learning for Many-to-many Multilingual Neural Machine Translation [Pan, Wu, Wang, **Lei Li**, ACL 2021]

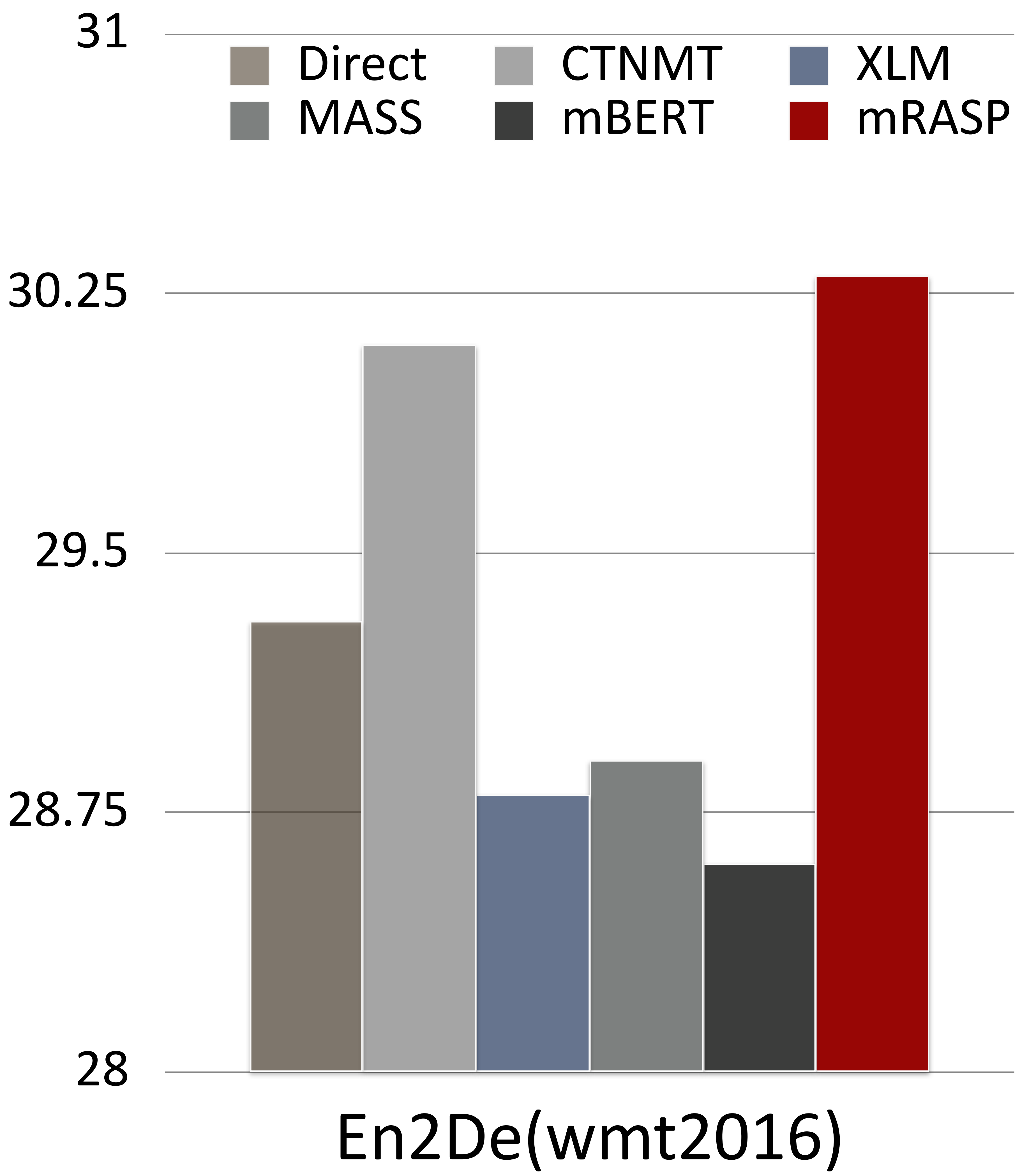
mRASP Fine-tuning

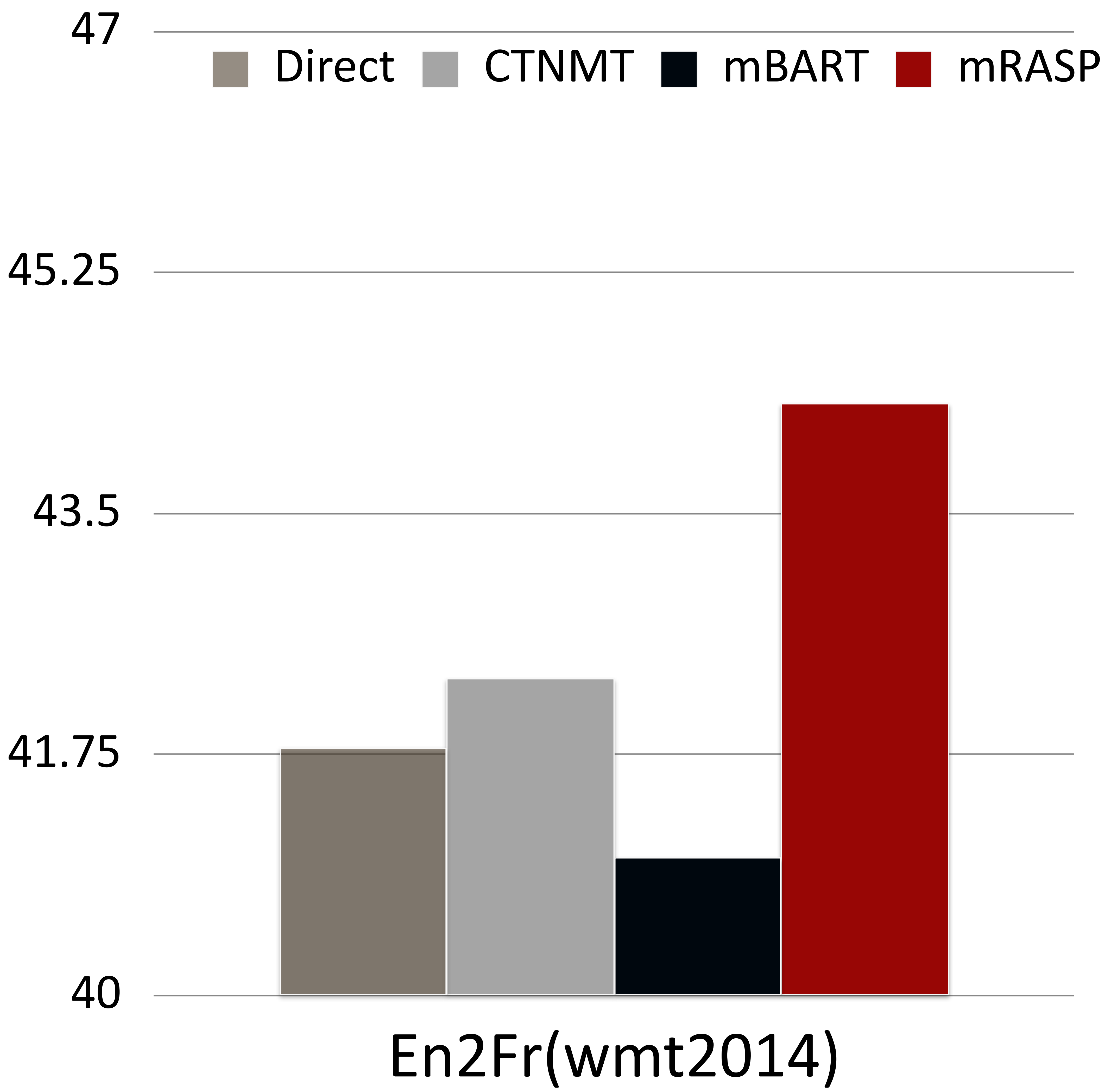
- Further fine-tuning based on mRASP model
- Fine-tuning Dataset
- Indigenous Corpus: included in pre-training phase
    - Extremely low resource (<100K) (Be, My, etc.)
    - Low resource(>100k and <1M) (He, Tr, etc.)
    - Medium resource (>1M and <10M) (De, Et, etc.)
    - Rich resource (>10M) (Zh, Fr, etc.)

Pre-training Multilingual Neural Machine Translation by Leveraging Alignment Information [Lin et al., EMNLP 2020]

mRASP Fine-tunes better: Rich resource works

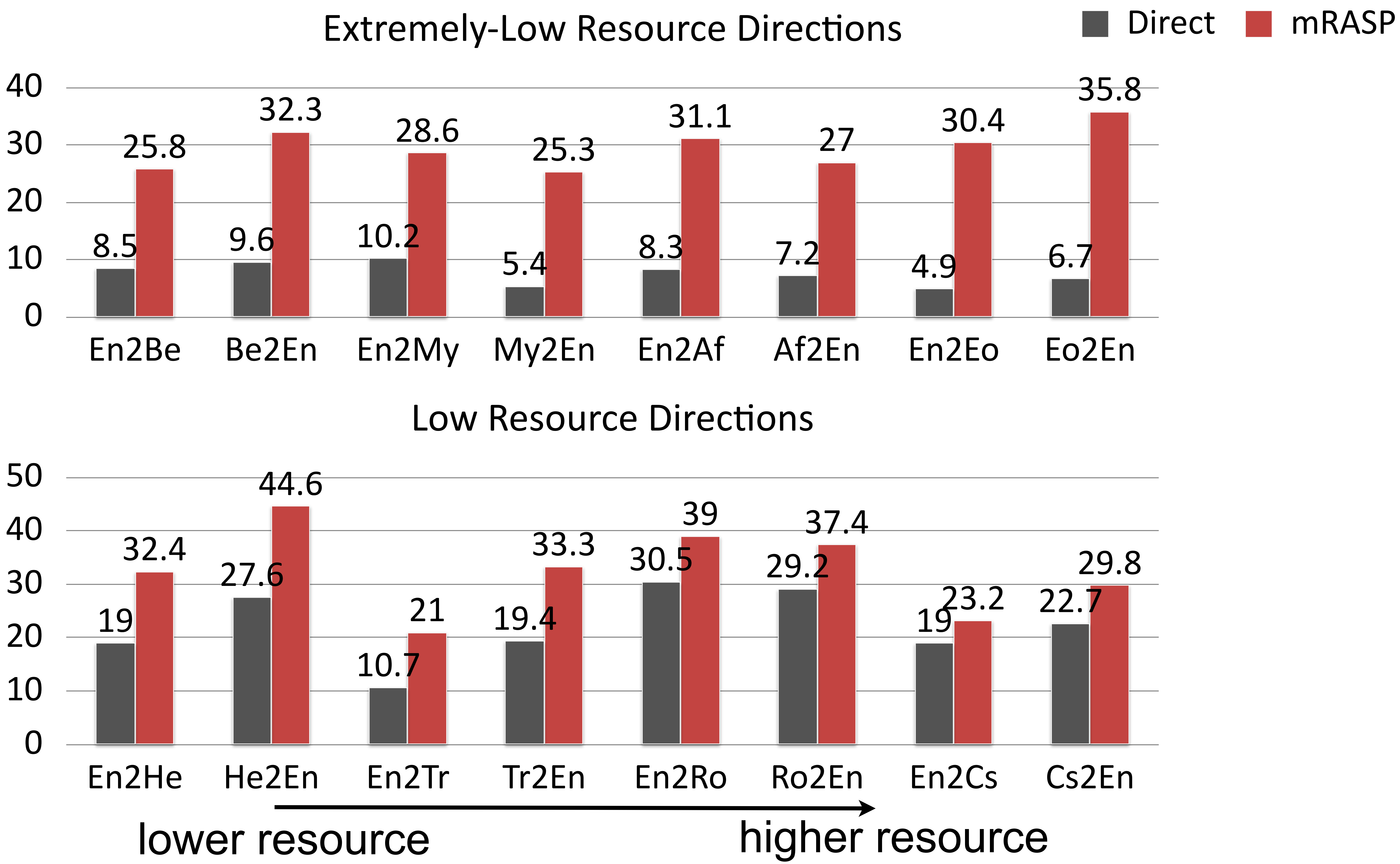
- En->Fr +1.1BLEU.





Pre-training Multilingual Neural Machine Translation by Leveraging Alignment Information [Lin et al., EMNLP 2020]

mRASP: Low resource works



Pre-training Multilingual Neural Machine Translation by Leveraging Alignment Information [Lin et al., EMNLP 2020]

mRASP: Unseen languages

- mRASP generalizes on all exotic scenarios.

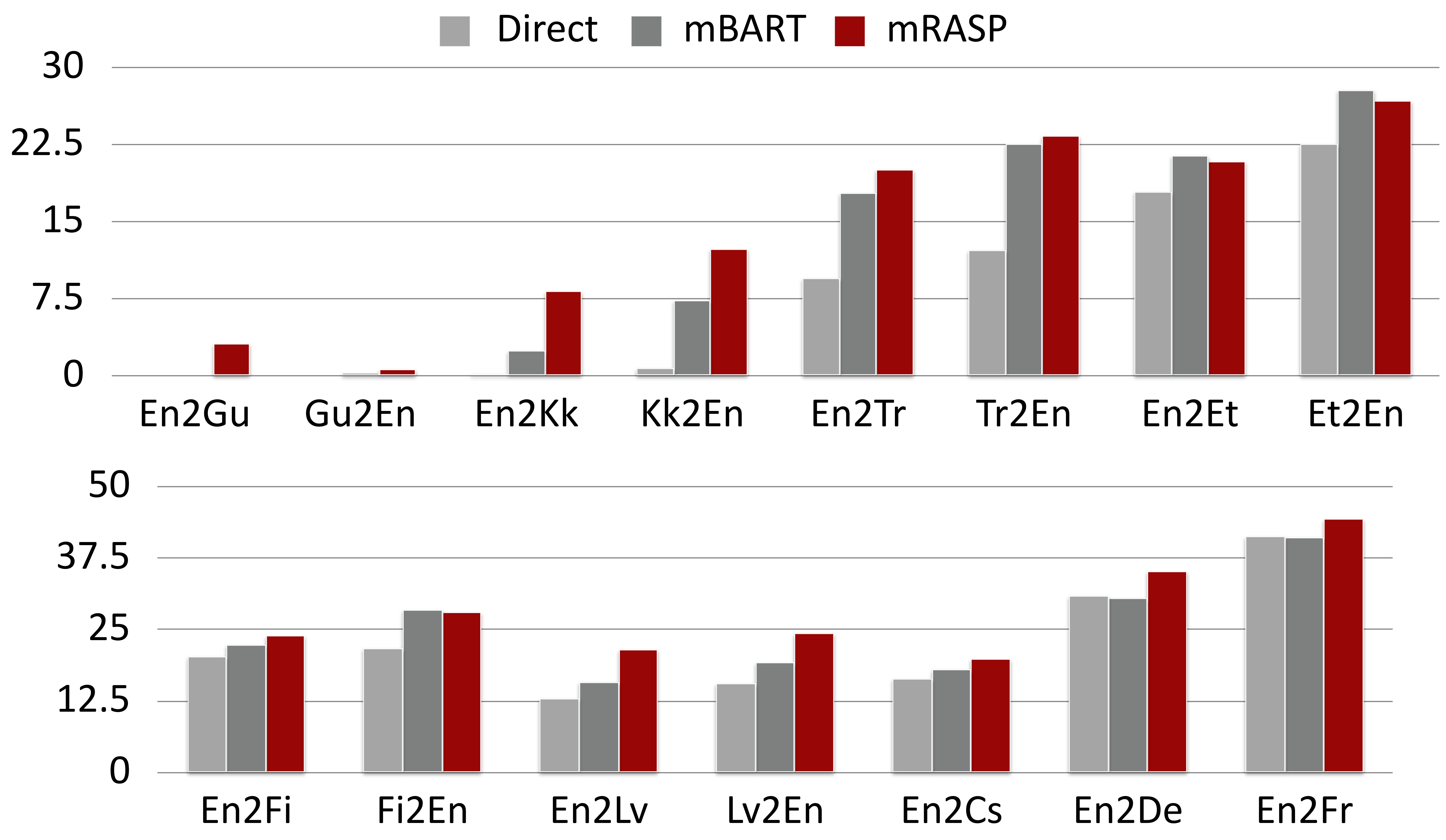
		Fr-Zh(20K)		De-Fr(9M)	
		—>	<—	—>	<—
Exotic Pair	Direct	0.7	3	23.5	21.2
	mRASP	25.8	26.7	29.9	23.4
		Nl-Pt(12K)		Da-El(1.2M)	
		—>	<—	—>	<—
Exotic Full	Direct	0.0	0.0	14.1	16.9
	mRASP	14.1	13.2	17.6	19.9
		En-Mr(11K)		En-Gl(1.2M)	
		—>	<—	—>	<—
Exotic Source/Target	Direct	6.4	6.8	8.9	12.8
	mRASP	22.7	22.9	32.1	38.1
		En-Eu(726k)		En-Sl(2M)	
		—>	<—	—>	<—
	Direct	7.1	10.9	24.2	28.2
	mRASP	19.1	28.4	27.6	29.5

12k: **Direct** not work **VS** **mRASP** achieves 10+ BLEU!!

Pre-training Multilingual Neural Machine Translation by Leveraging Alignment Information [Lin et al., EMNLP 2020]

mRASP: Compare with other methods

- mRASP outperforms mBART for all but two language pairs.



Pre-training Multilingual Neural Machine Translation by Leveraging Alignment Information [Lin et al., EMNLP 2020]

Summary

- Pre-training for NMT
  - sequence to sequence training objective
  - MASS: masked prediction using seq2seq
  - mBart: Recover original sentence from noised ones in multiple languages.
- Multilingual joint training
  - mRASP & mRASP2:
    - augmenting data with randomly substitute of words from bilingual lexicon
    - monolingual reconstruction
    - contrastive learning

Discussion

- What strategies for training multilingual NMT