291K
Deep Learning for Machine Translation
Pre-training for NMT

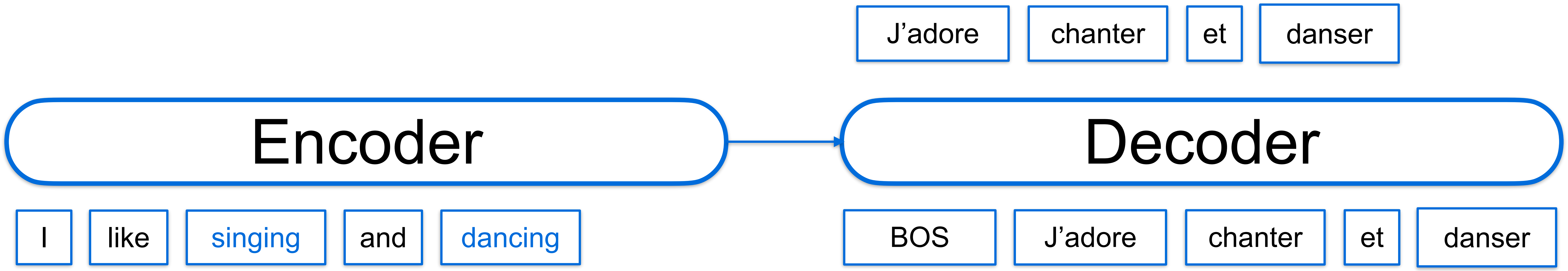
Lei Li
UCSB
11/10/2021
11/15/2021

Outline

- Monolingual Sequence-to-sequence pre-training
  - MASS: Masked seq-to-seq pretraining
  - BART
- Multilingual fused pre-training
  - Cross-lingual Language Model Pre-training [NeurIPS, 2019]
  - Alternating Language Modeling Pre-training [AAAI, 2020]
  - XLM-T: Cross-lingual Transformer Encoders
- Multilingual sequence to sequence pre-training
  - mBART [TACL, 2020]
  - mRASP & mRASP2 [EMNLP, 2020] [ACL, 2021]
  - LaSS: Learning language-specific sub-network via pre-training & fine-tuning [ACL, 2021]

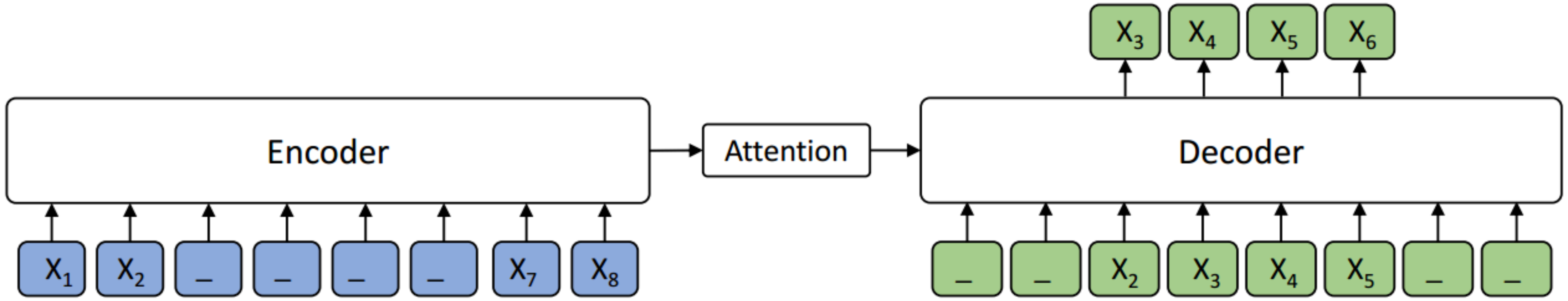
Sequence-to-sequence Pre-training

Sequence-to-sequence learning for MT



MASS: Pre-train for Sequence to Sequence Generation

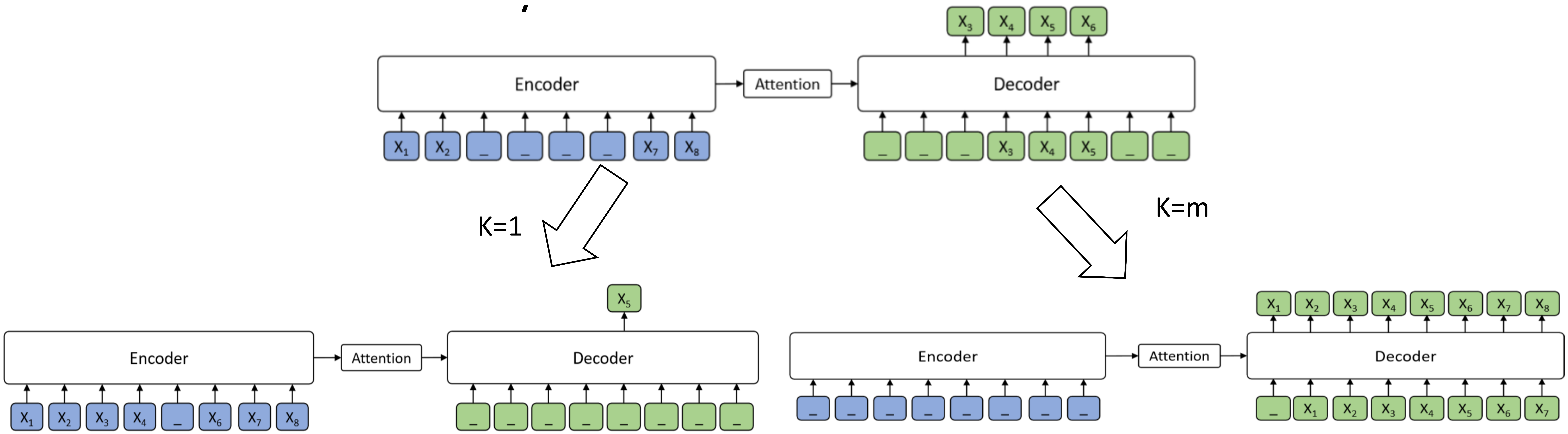
- MASS is carefully designed to jointly pre-train the encoder and decoder



- 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

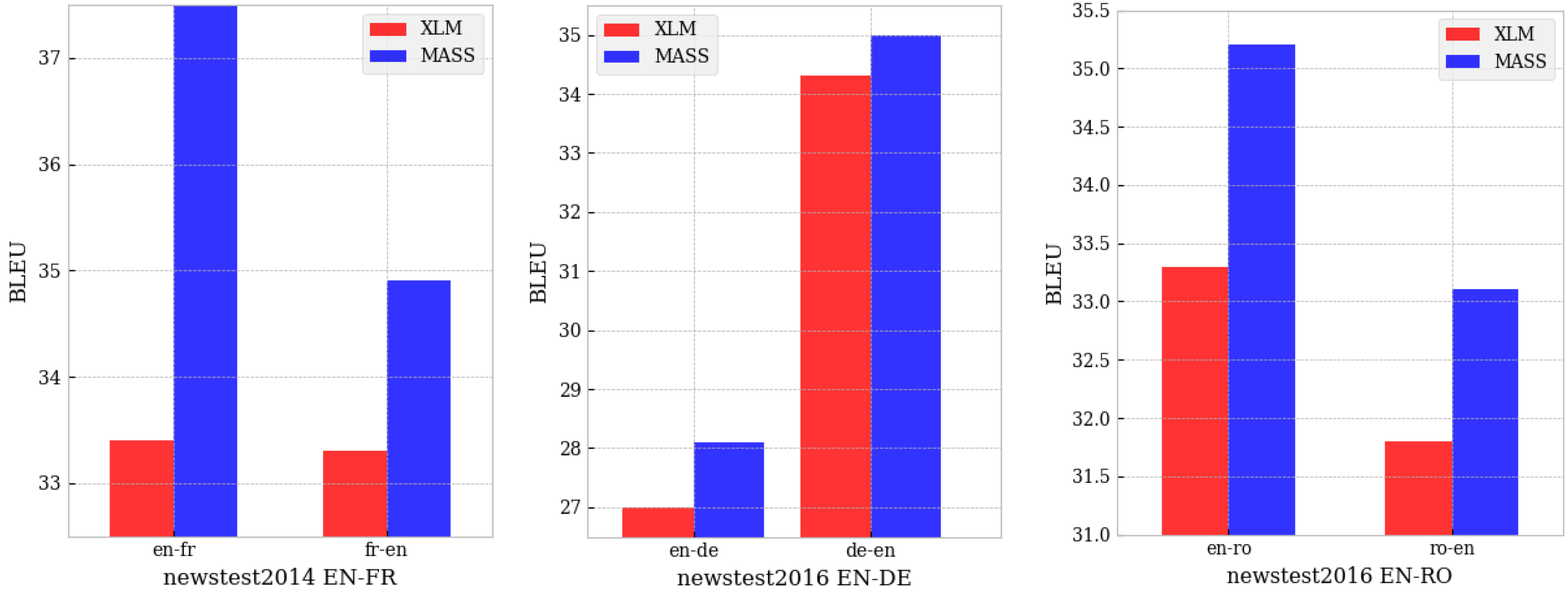


Length	Probability	Model
$k = 1$	$P(x^u \mid x^{\setminus u}; \theta)$	masked LM in BERT
$k \in [1, m]$	$P(x^{u:v} \mid x^{\setminus u:v}; \theta)$	MASS

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

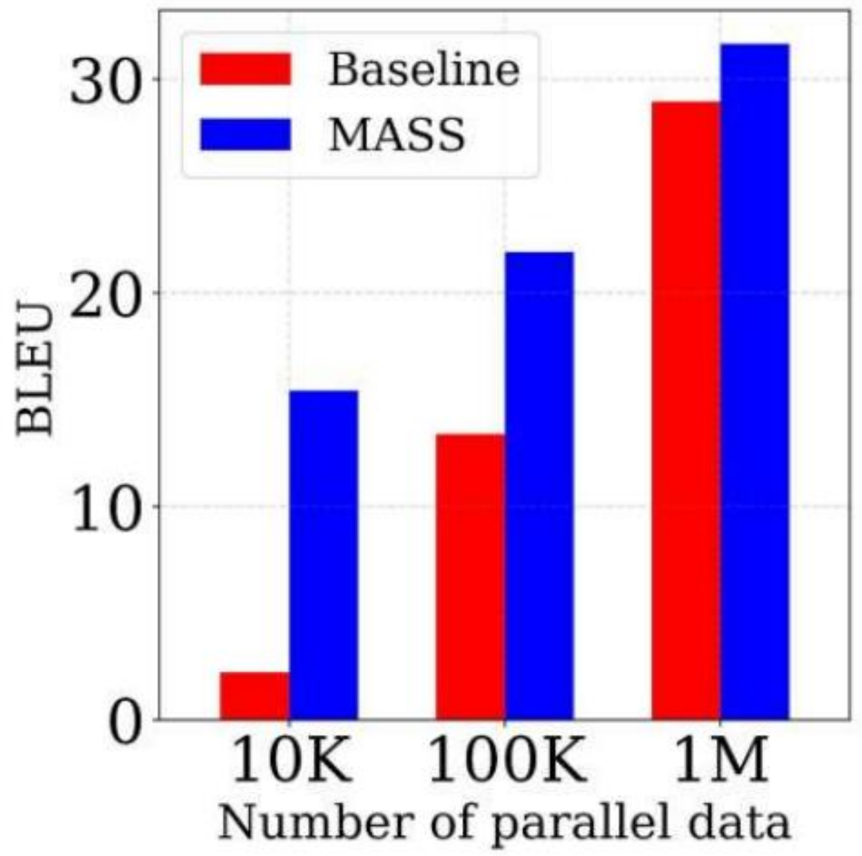
MASS: Pre-train for Sequence to Sequence Generation, [Song et al ICML 2019]

Unsupervised NMT

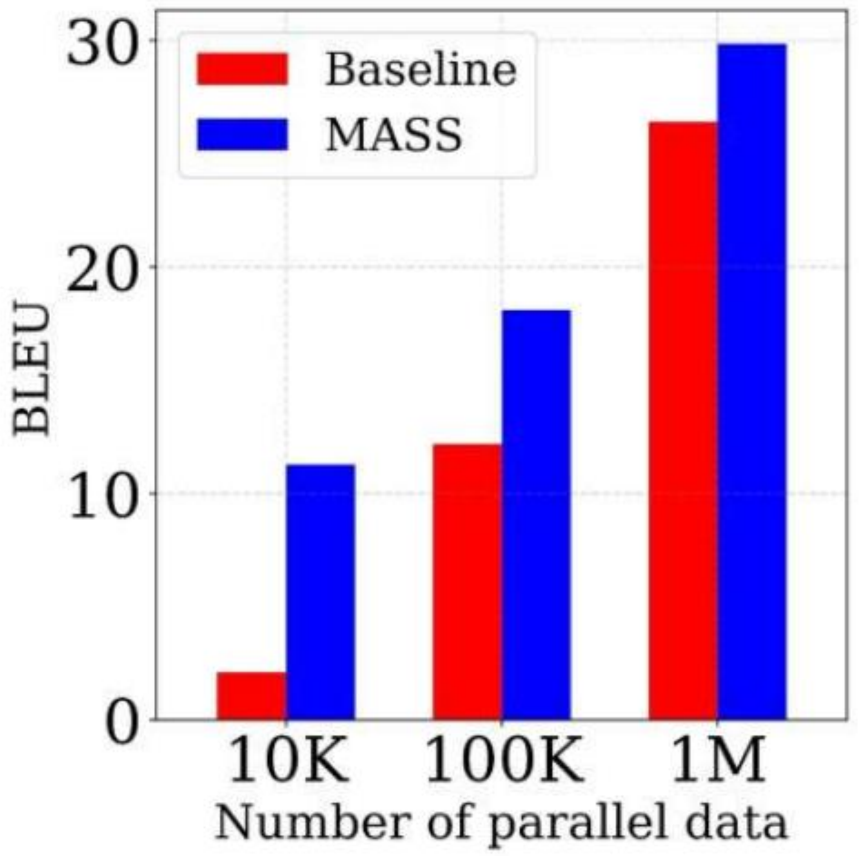


MASS: Pre-train for Sequence to Sequence Generation, [Song et al ICML 2019]

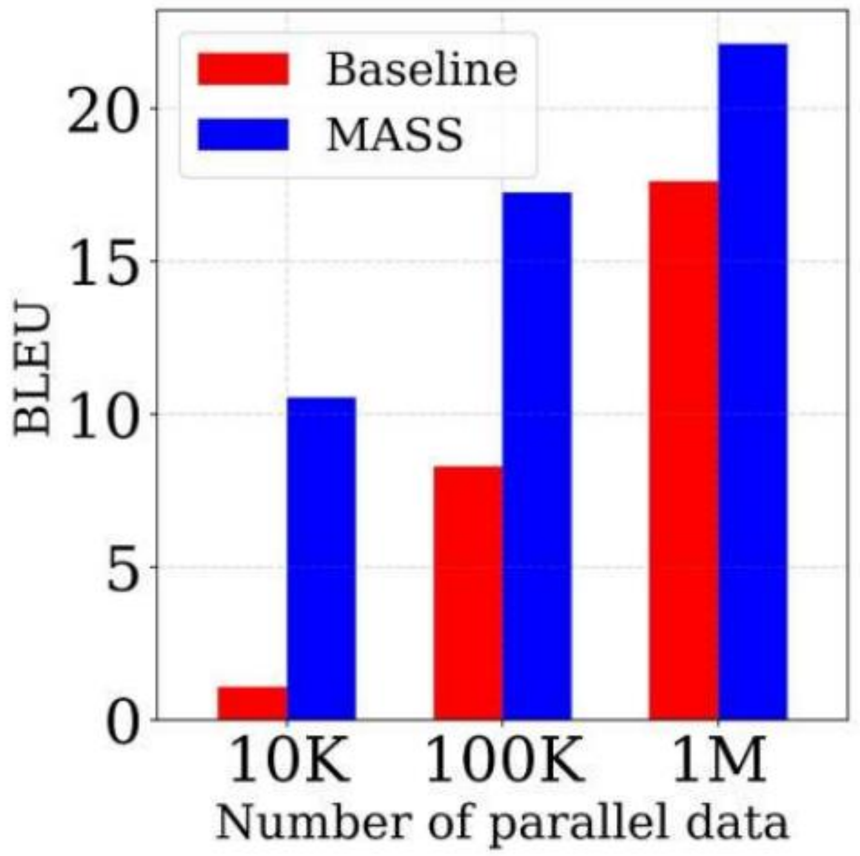
Low-resource NMT



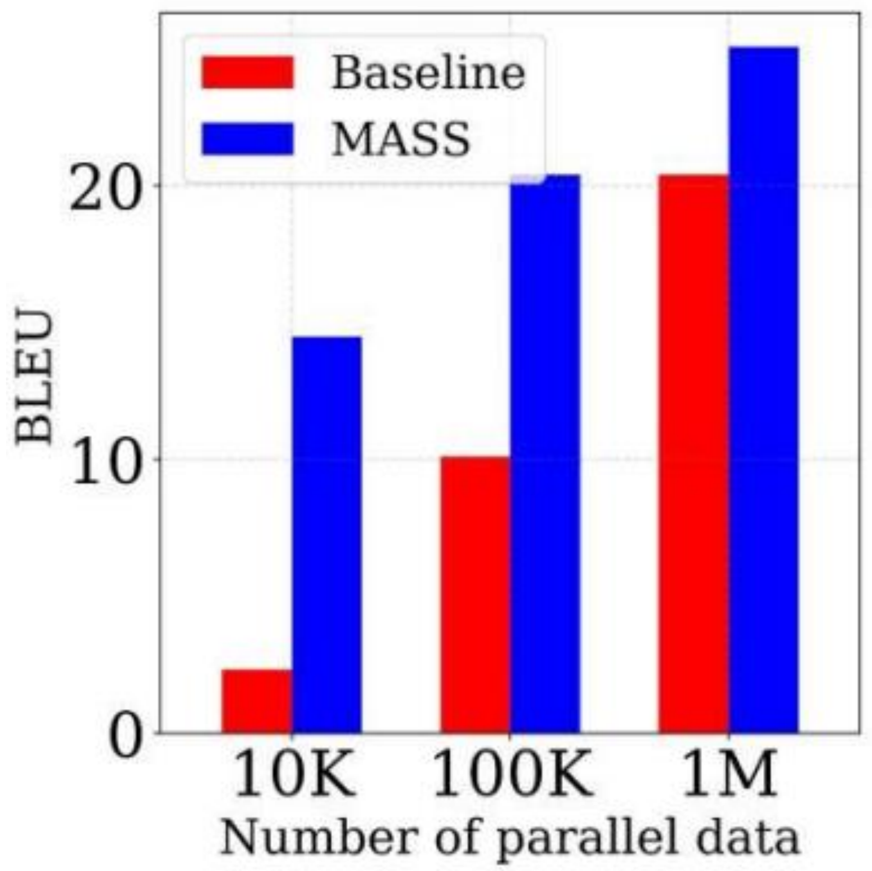
(a) en-fr



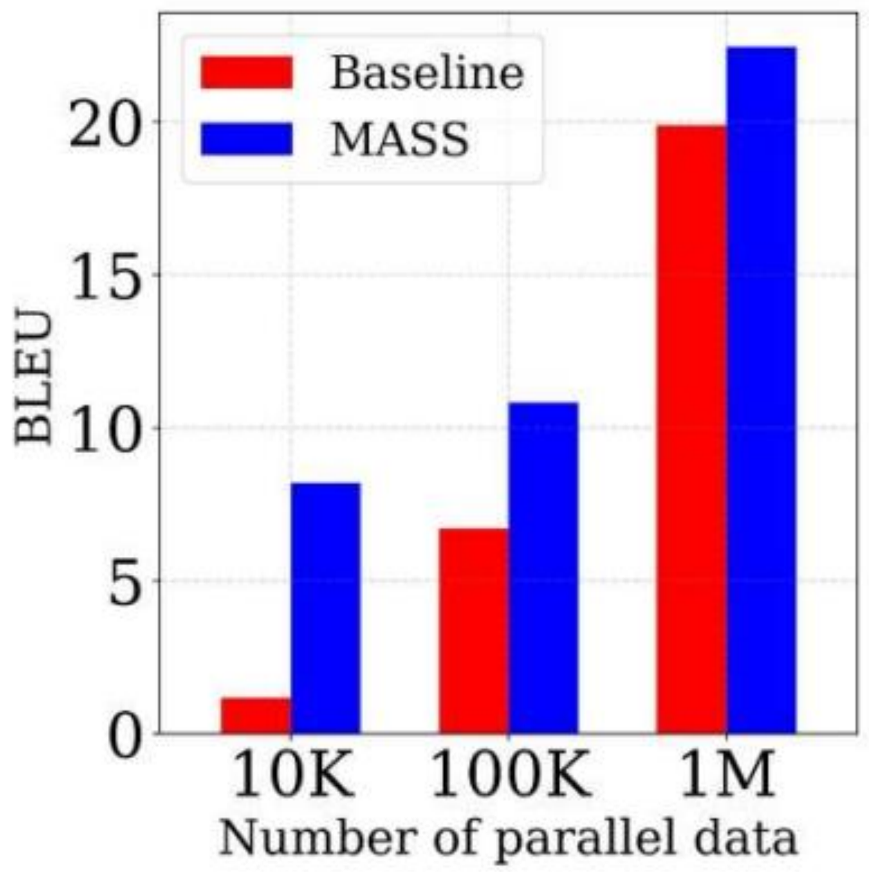
(b) fr-en



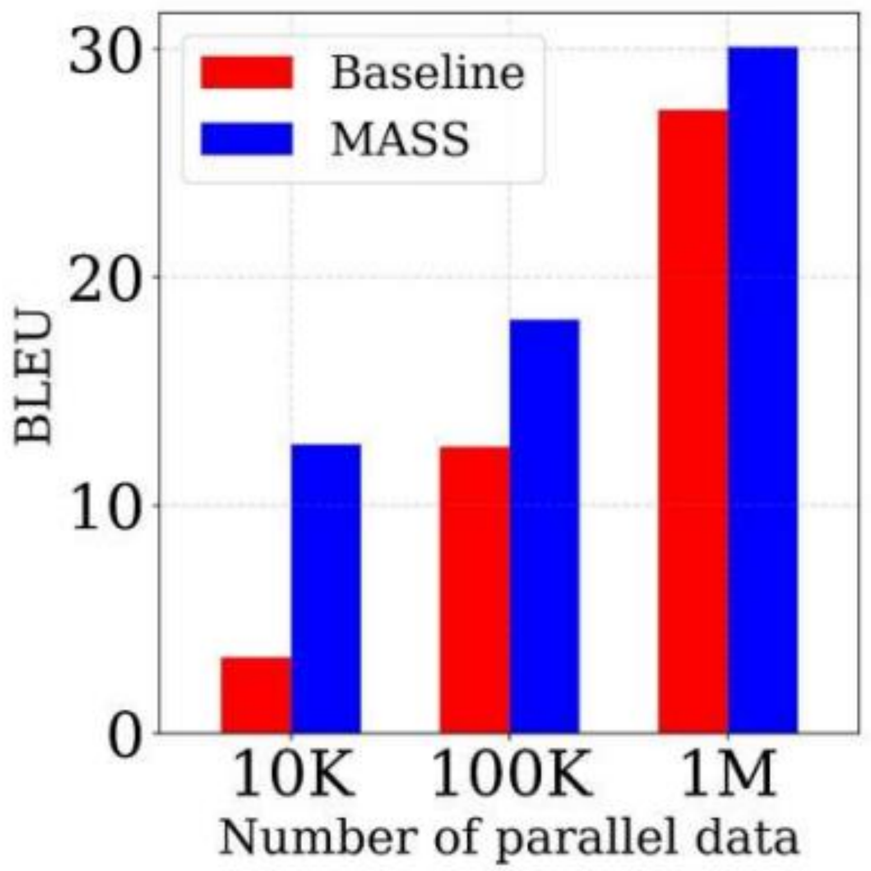
(c) en-de



(d) de-en



(e) en-ro

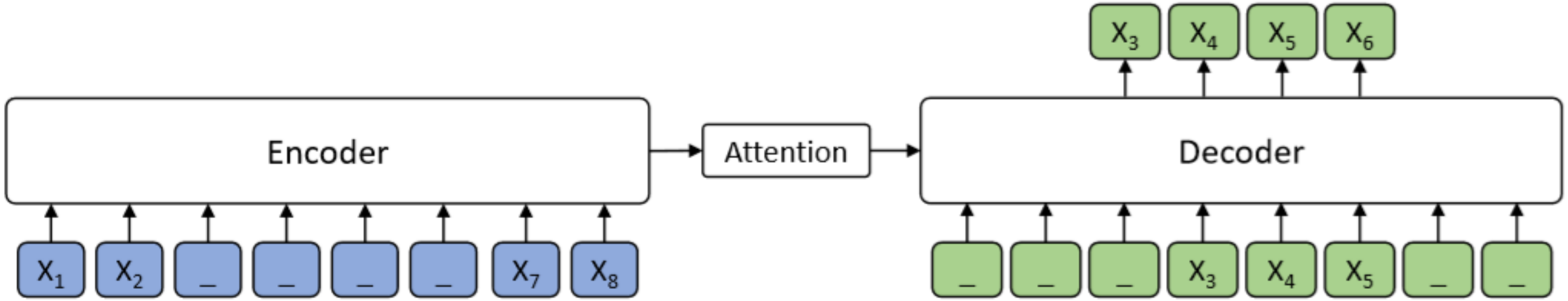


(f) ro-en

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

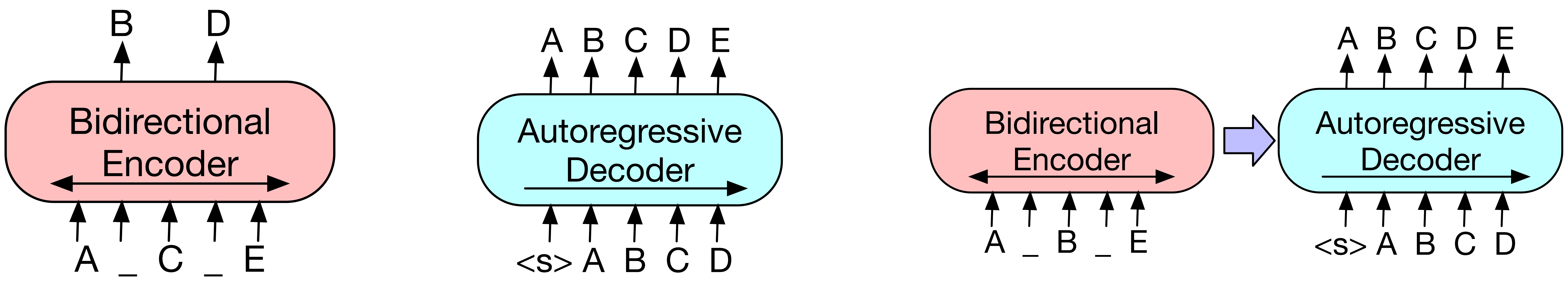
Summary

- 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 for Natural Language Generation, Translation, and Comprehension

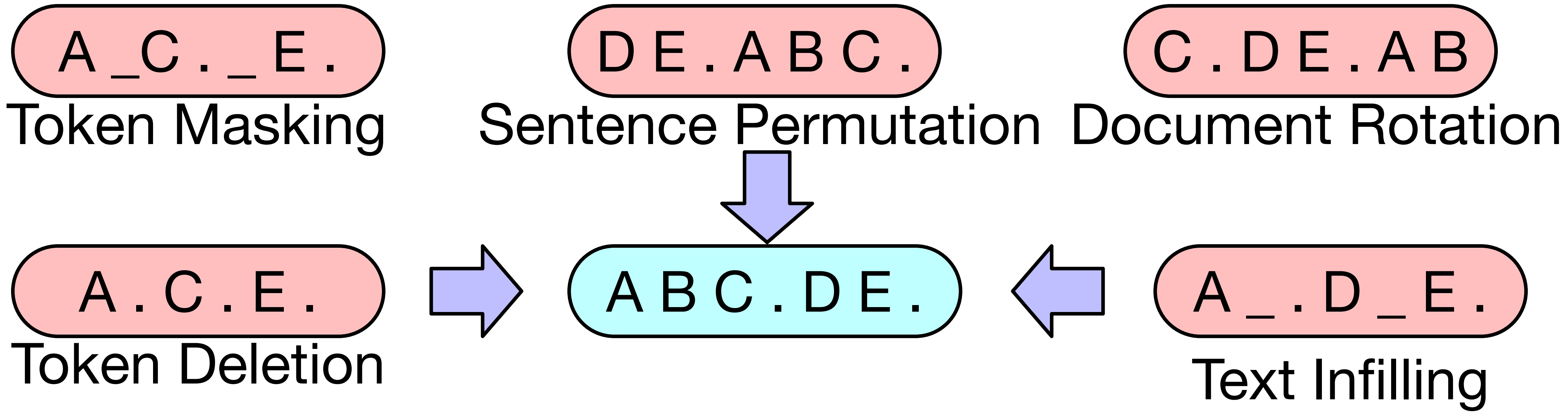


A schema comparison with BERT, GPT and BART.

- 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]

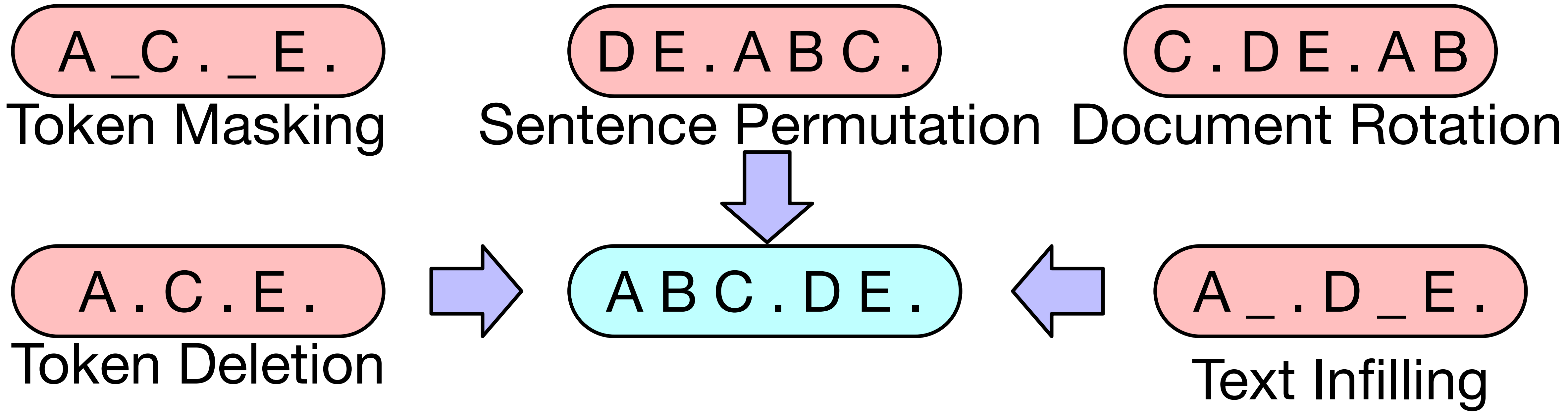
Noising the input



- 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]

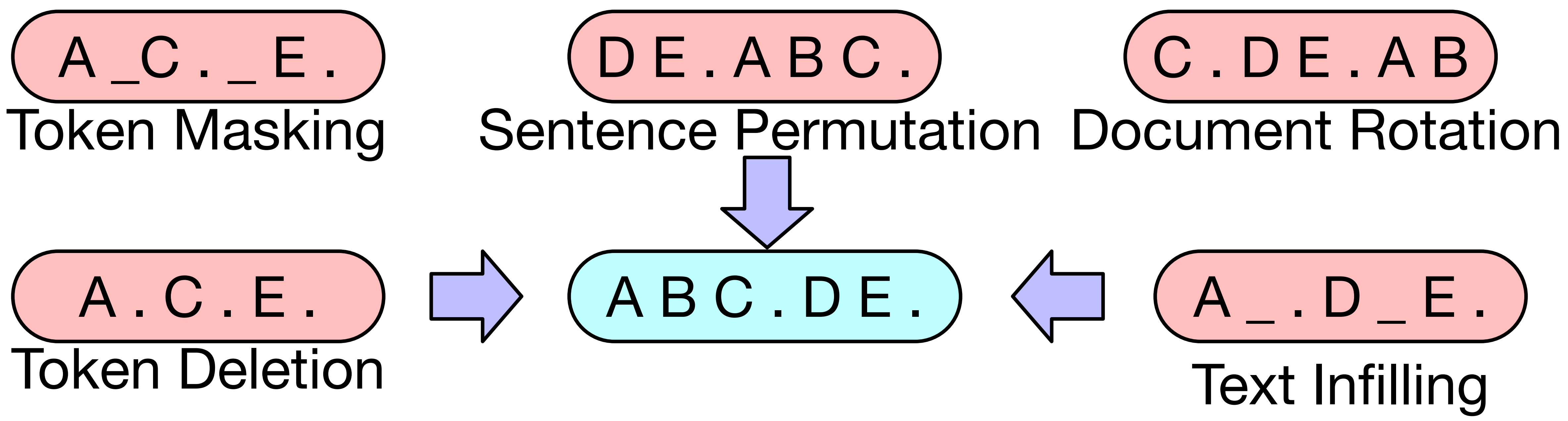
Noising the input



- 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]

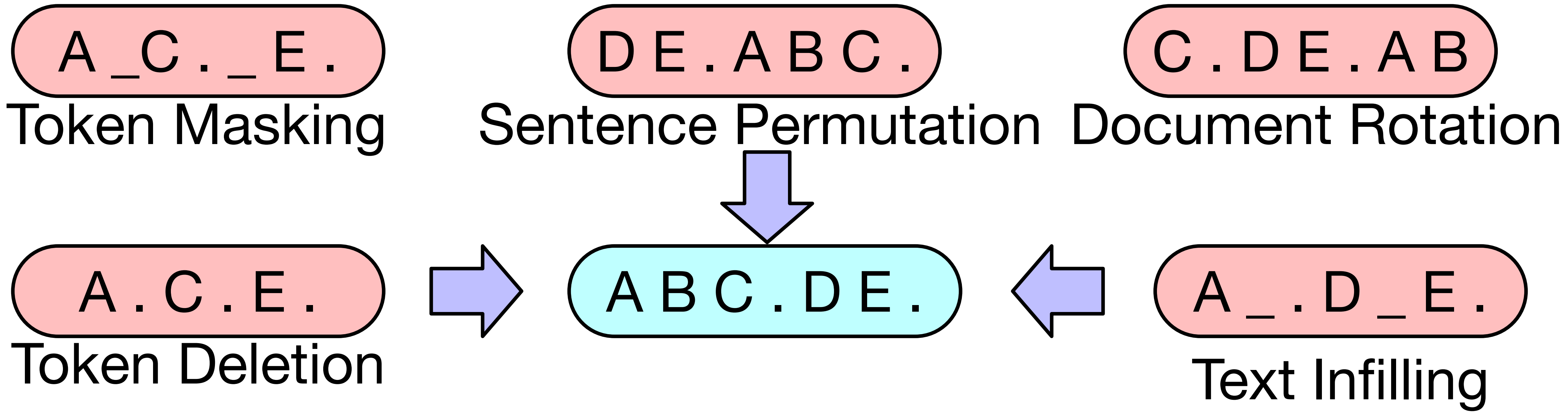
Noising the input



- 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]

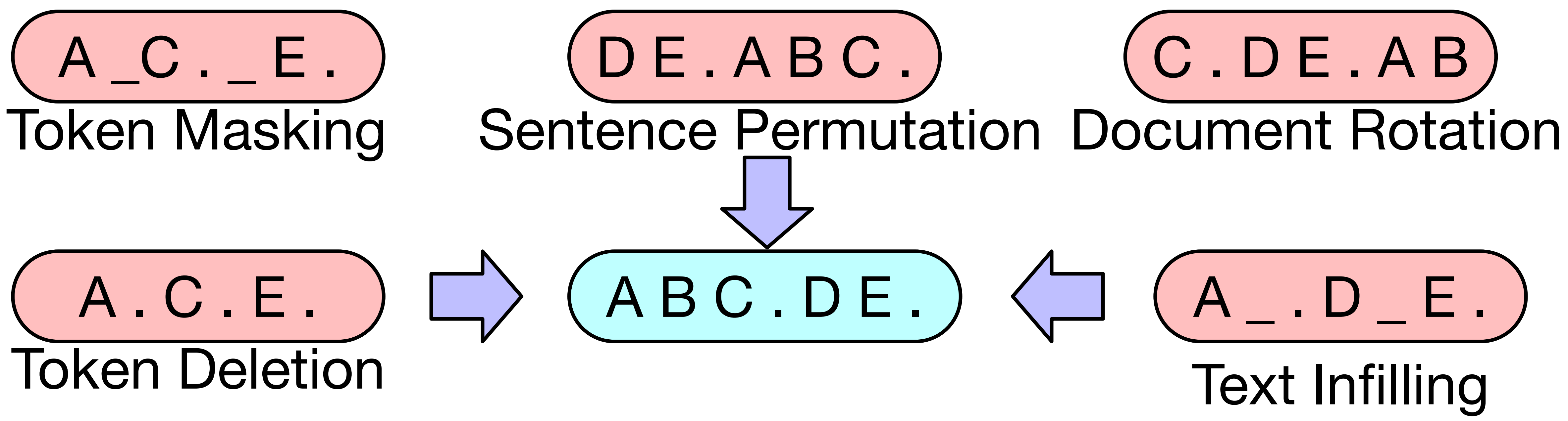
Noising the input



- 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]

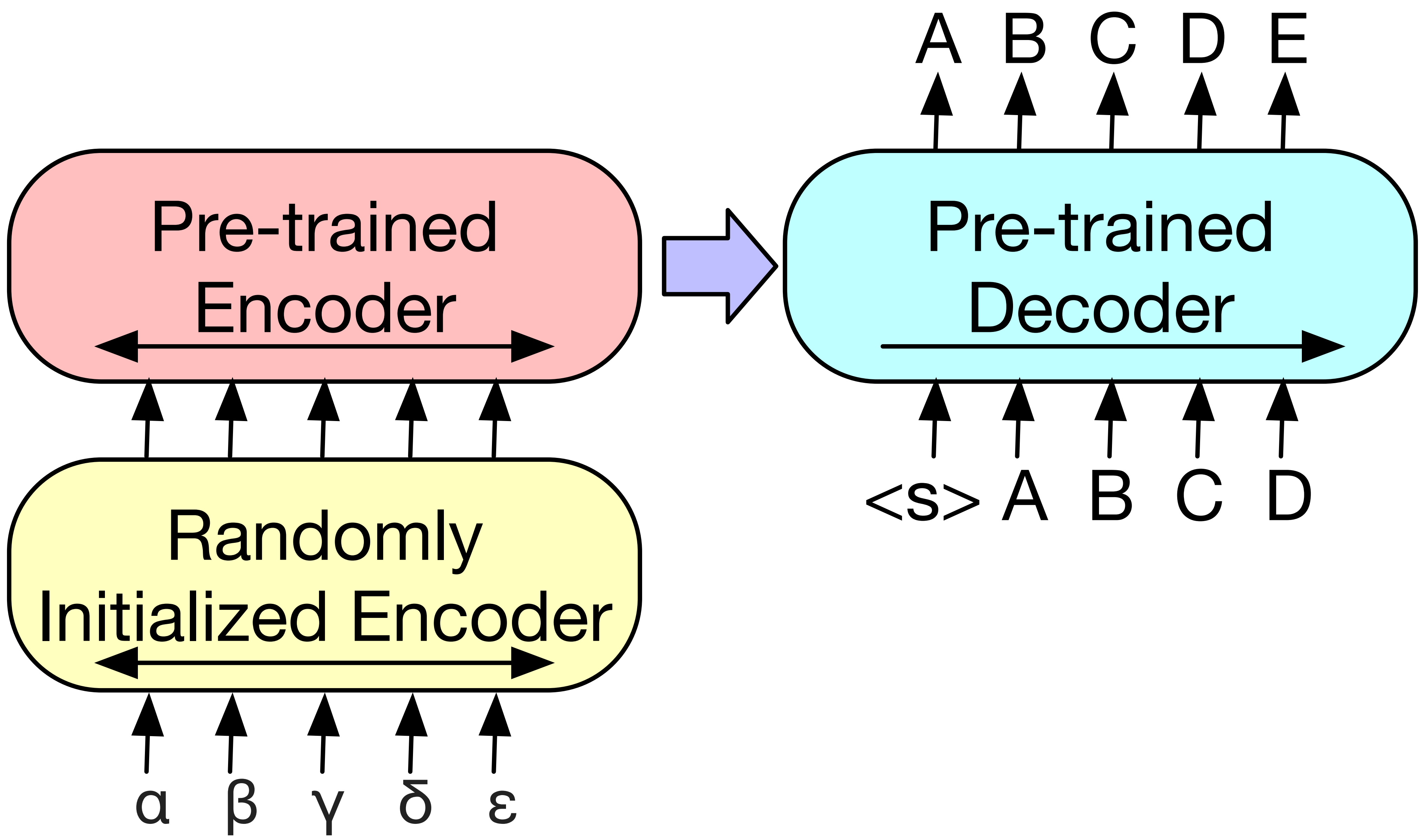
Noising the input



- 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]

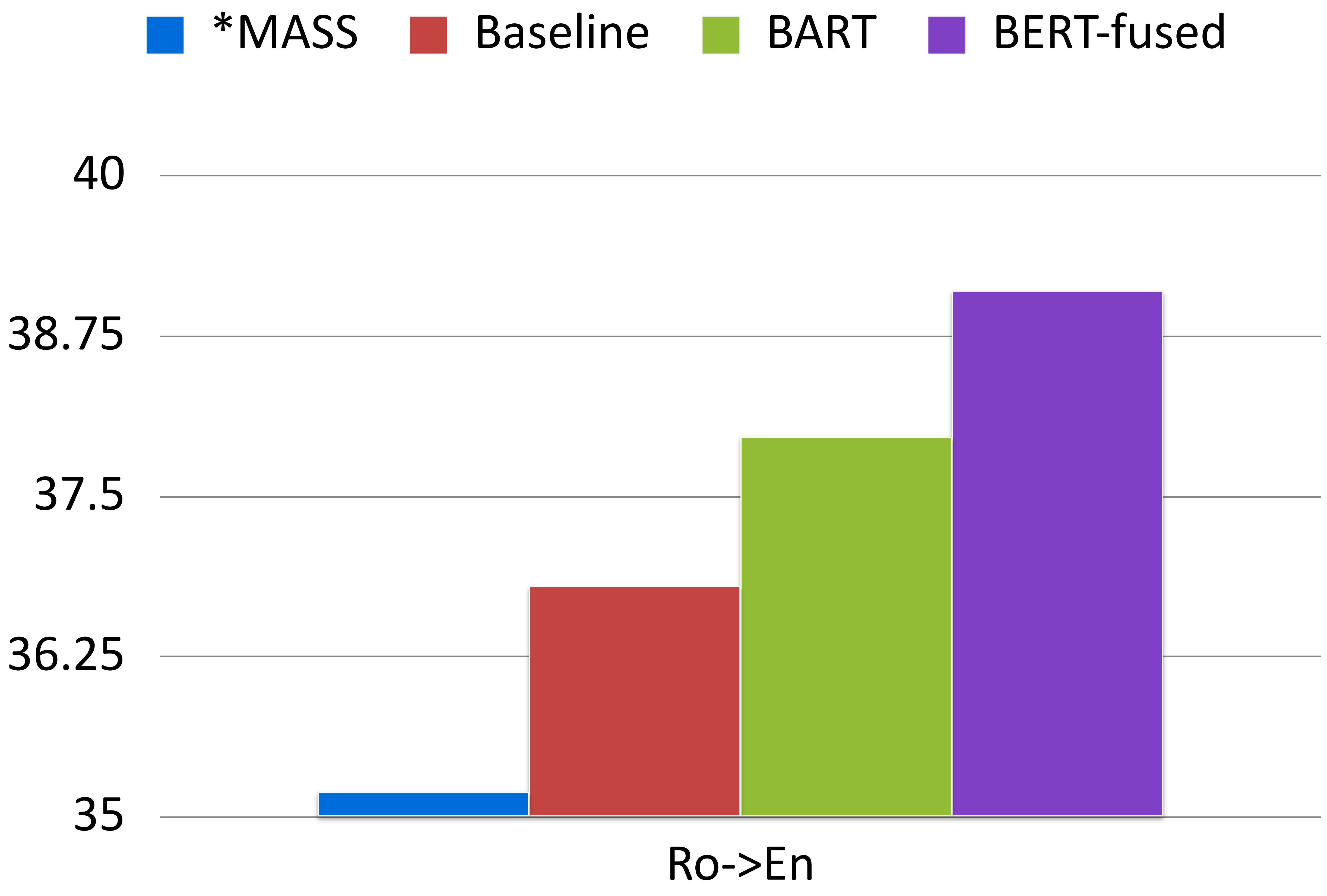
Fine-Tune on Neural Machine Translation



- 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]

Results on NMT



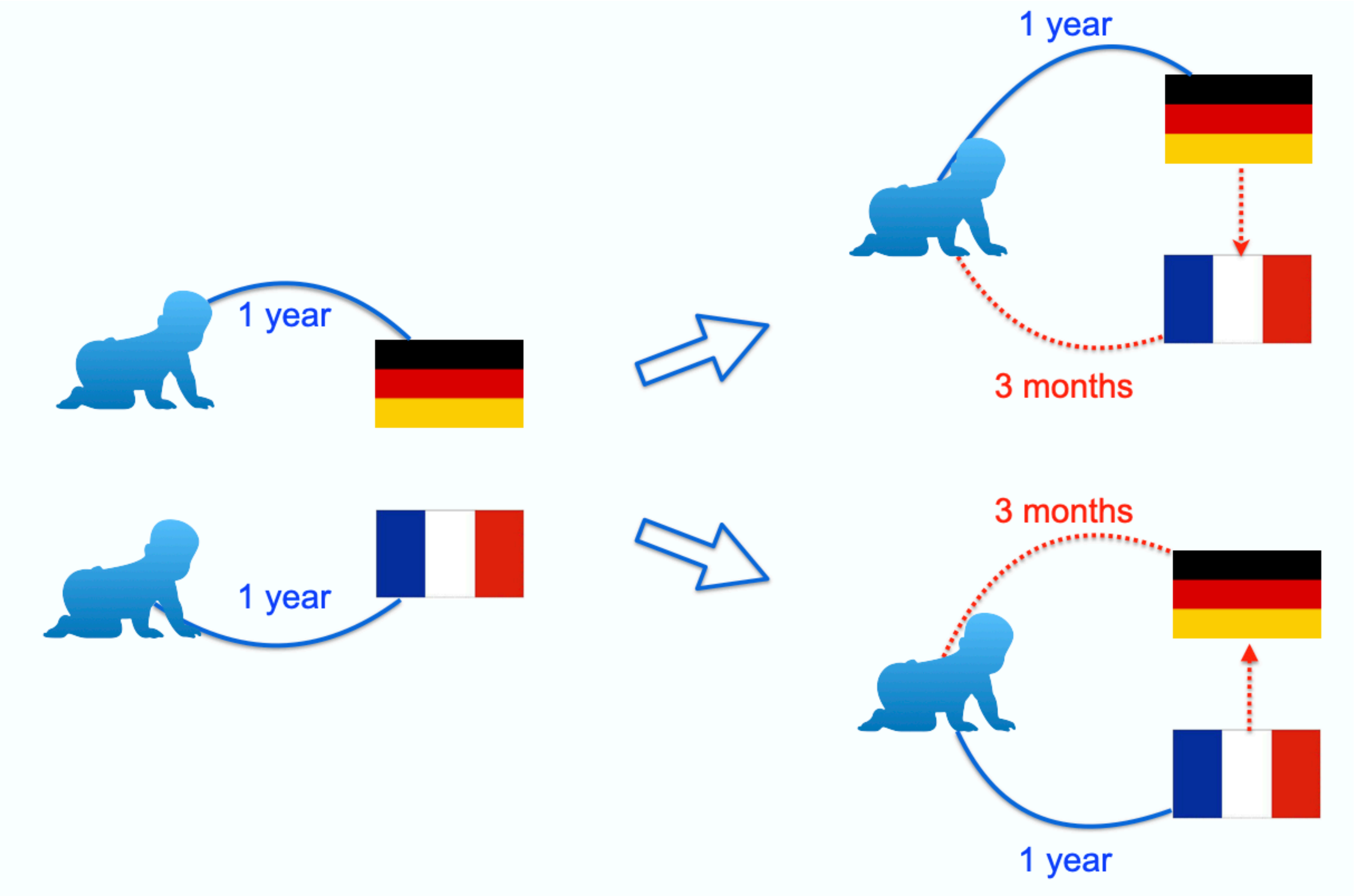
- Results on IWSLT 2016 En->Ro augmented with back-translation data
- 6 layer of additional transformer encoder to encoding Romania representation.
- *MASS reports unsupervised results

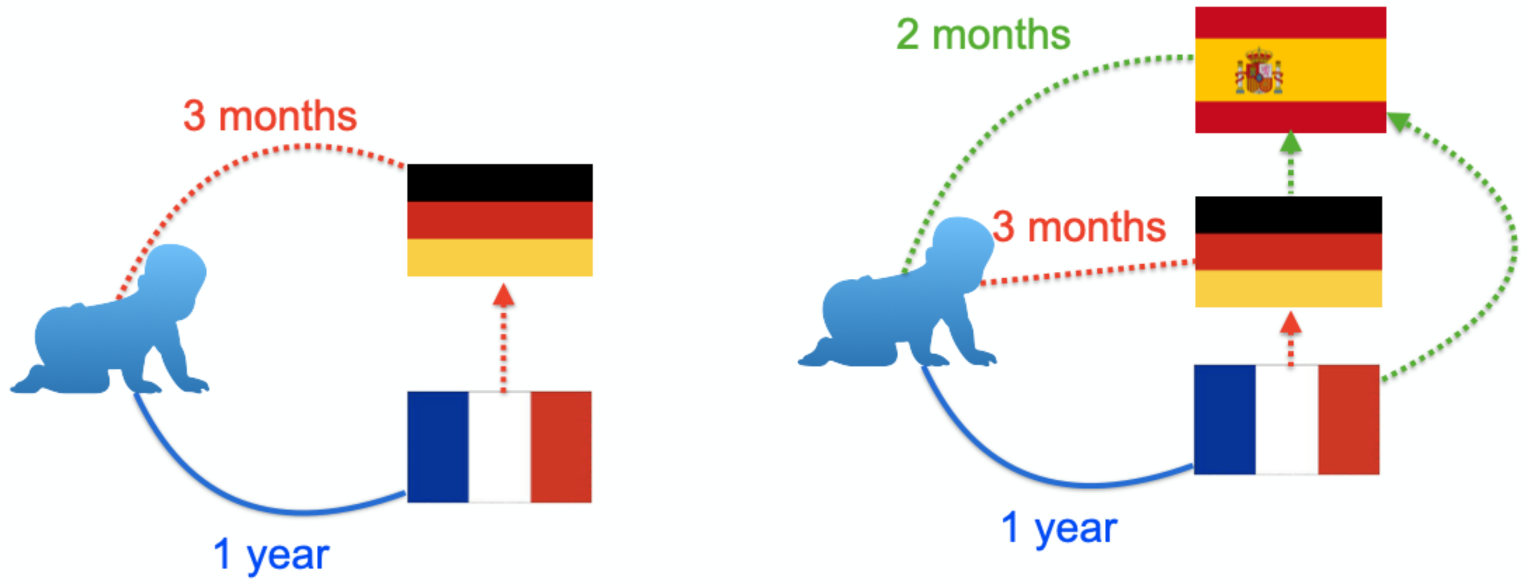
BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, [Lewis et al ACL 2020]

Multilingual Fused Pretraining

Multi-lingual Pre-training for NMT

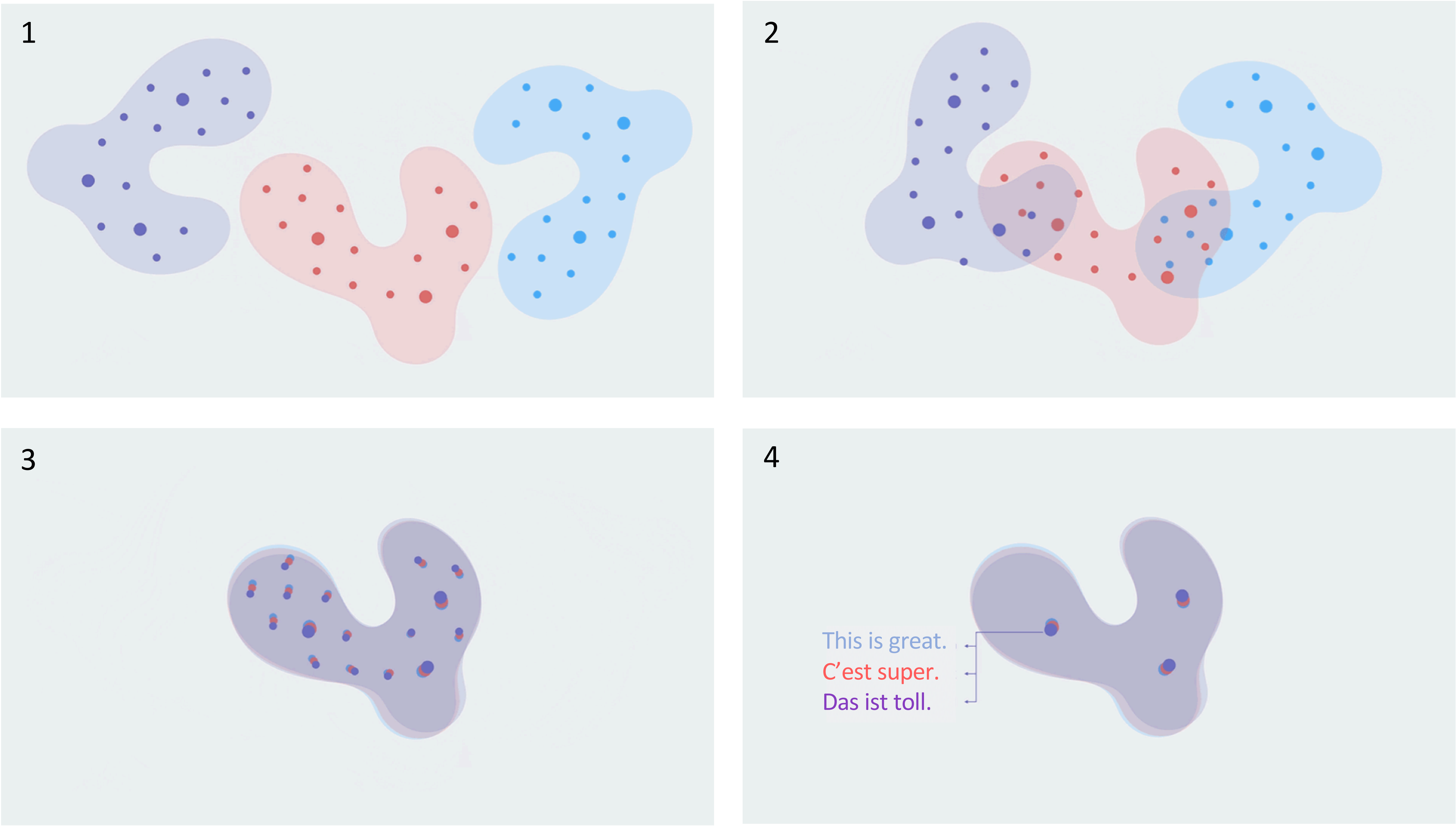
- Data scarcity for low/zero resource languages.
- <u>Transfer knowledge</u> between languages.





Cross-lingual Language Model Pretraining

Learning cross-lingual representation

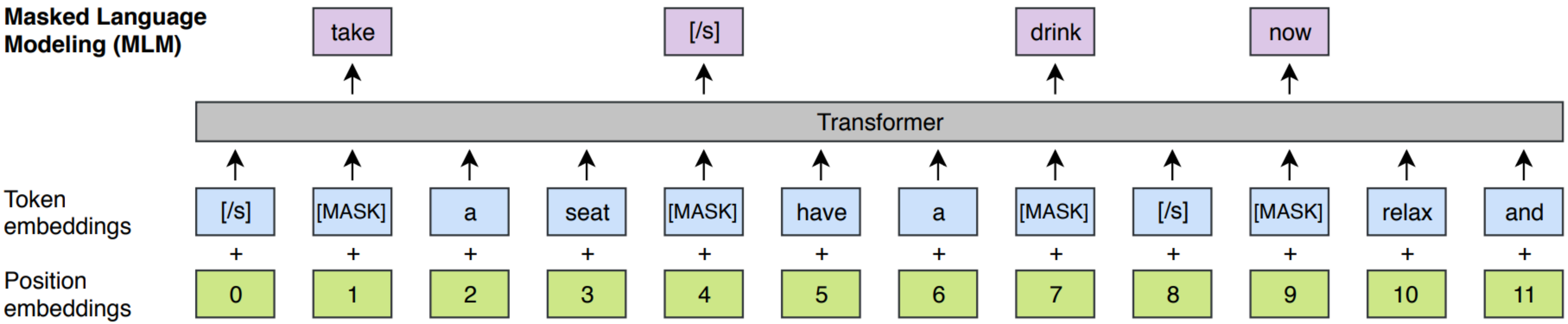


Cross-lingual Language Model Pre-training, [Conneau et al NeurIPS 2019]

Multiple masked language model (MLM)

Similar to BERT, but in many languages…

Multilingual representations emerge from a single model trained on many languages



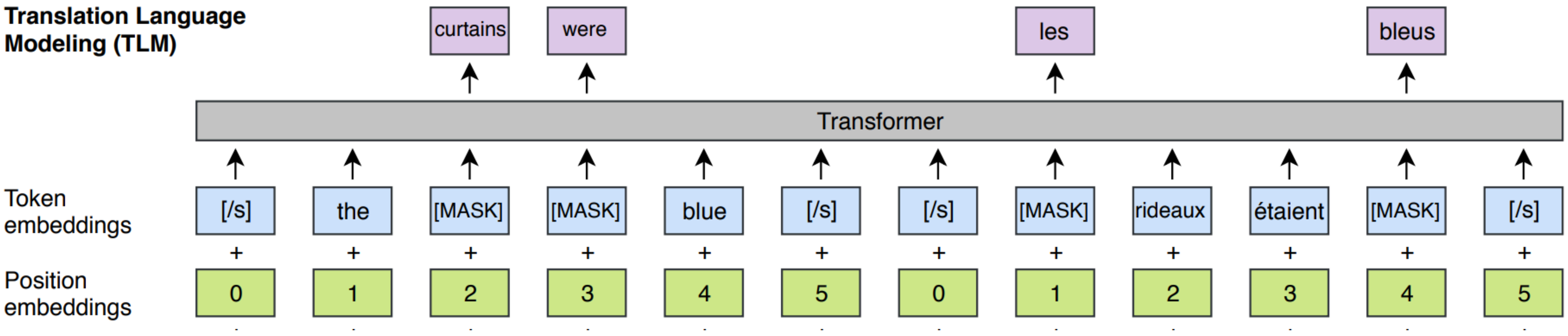
Multilingual Masked language modeling pretraining

Cross-lingual Language Model Pre-training, [Conneau et al NeurIPS 2019]

Translation language model (TLM)

MLM is unsupervised, but TLM leverages parallel data…

Encourage the model to learn cross-lingual context when predicting



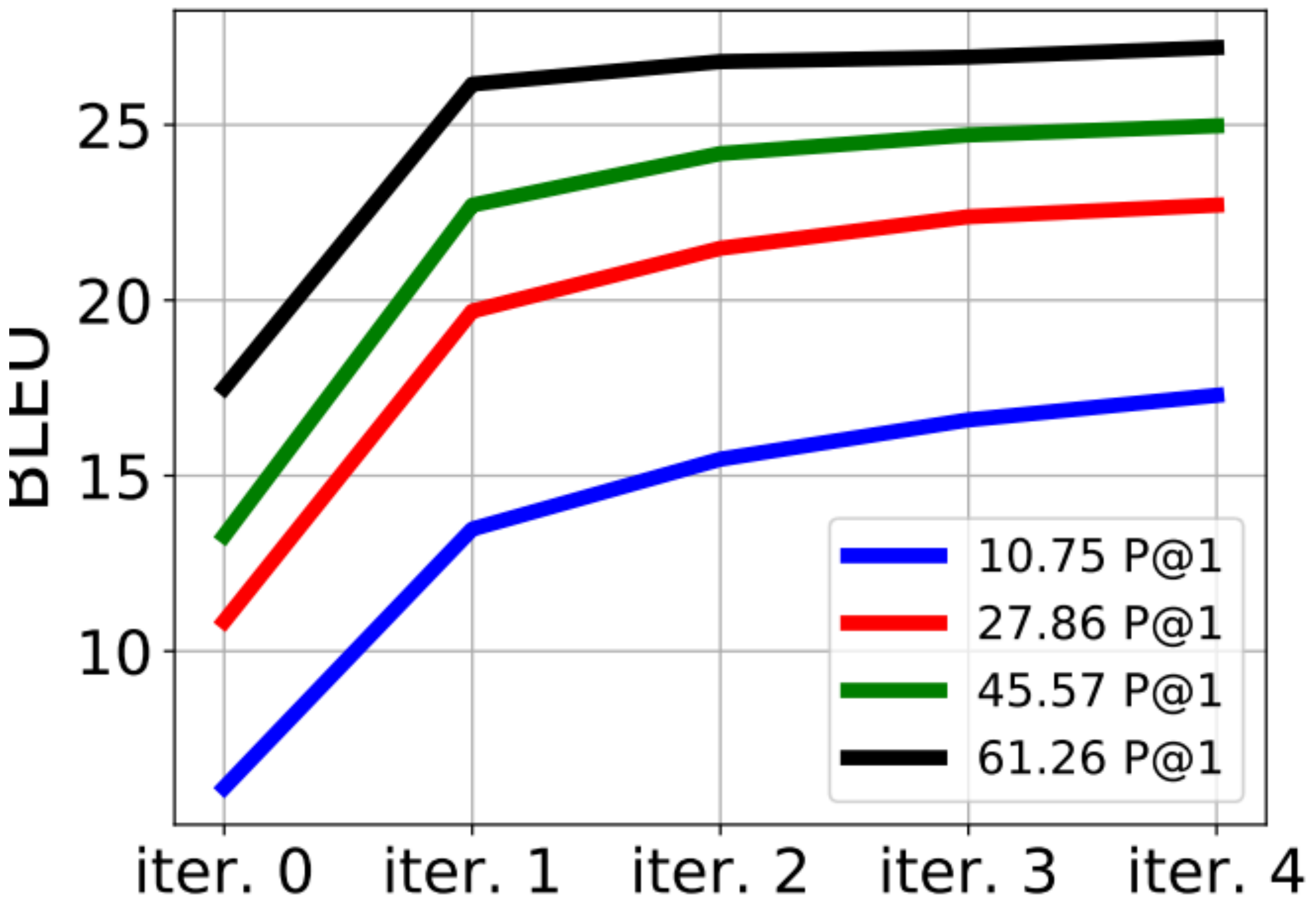
Translation language modeling (TLM) pretraining

Cross-lingual Language Model Pre-training, [Conneau et al NeurIPS 2019]

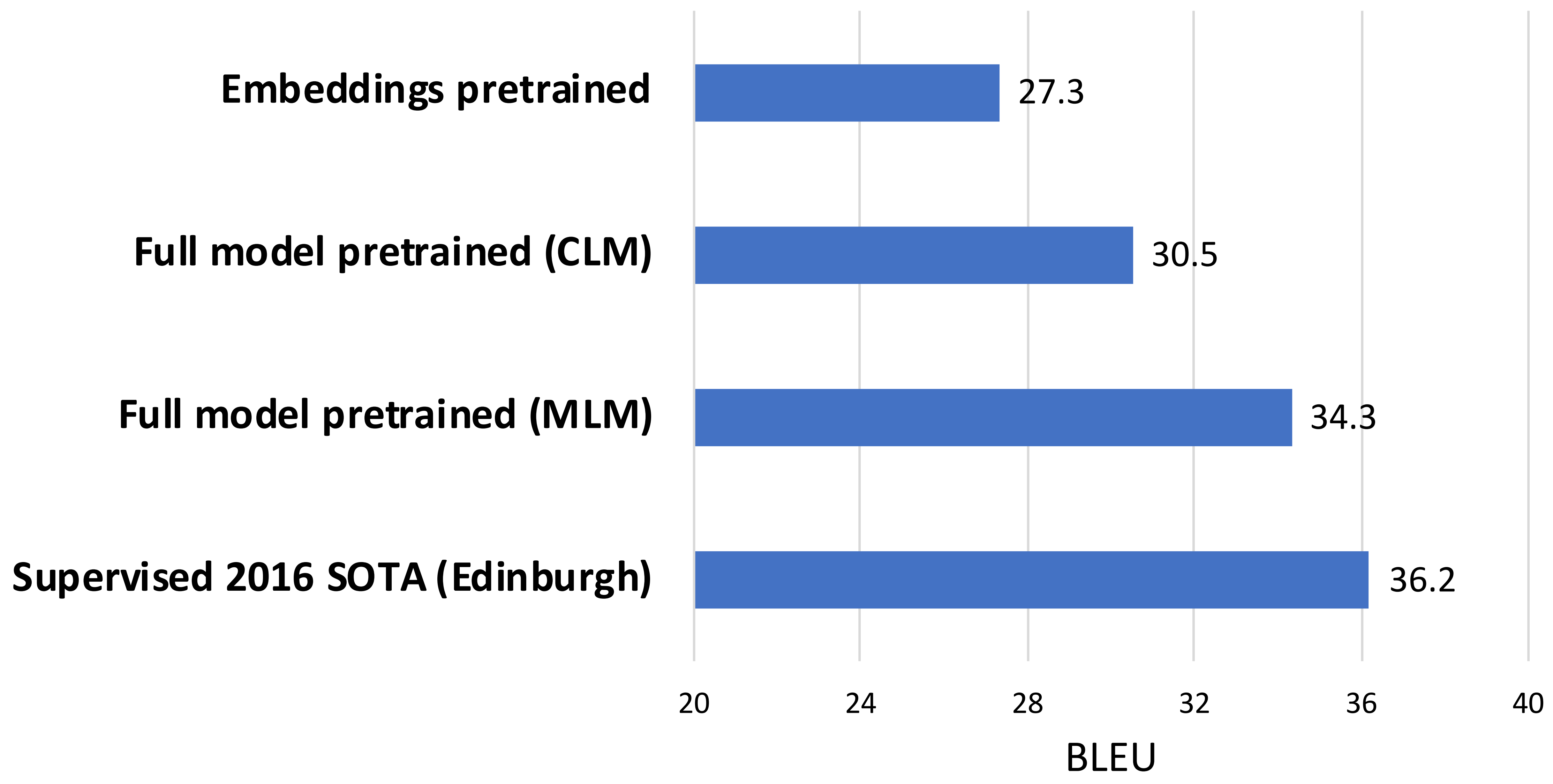
Results on Unsupervised Machine Translation

Initialization is key in unsupervised MT to bootstrap the iterative BT process

Embedding layer initialization
is essential for neural unsupervised MT (*)



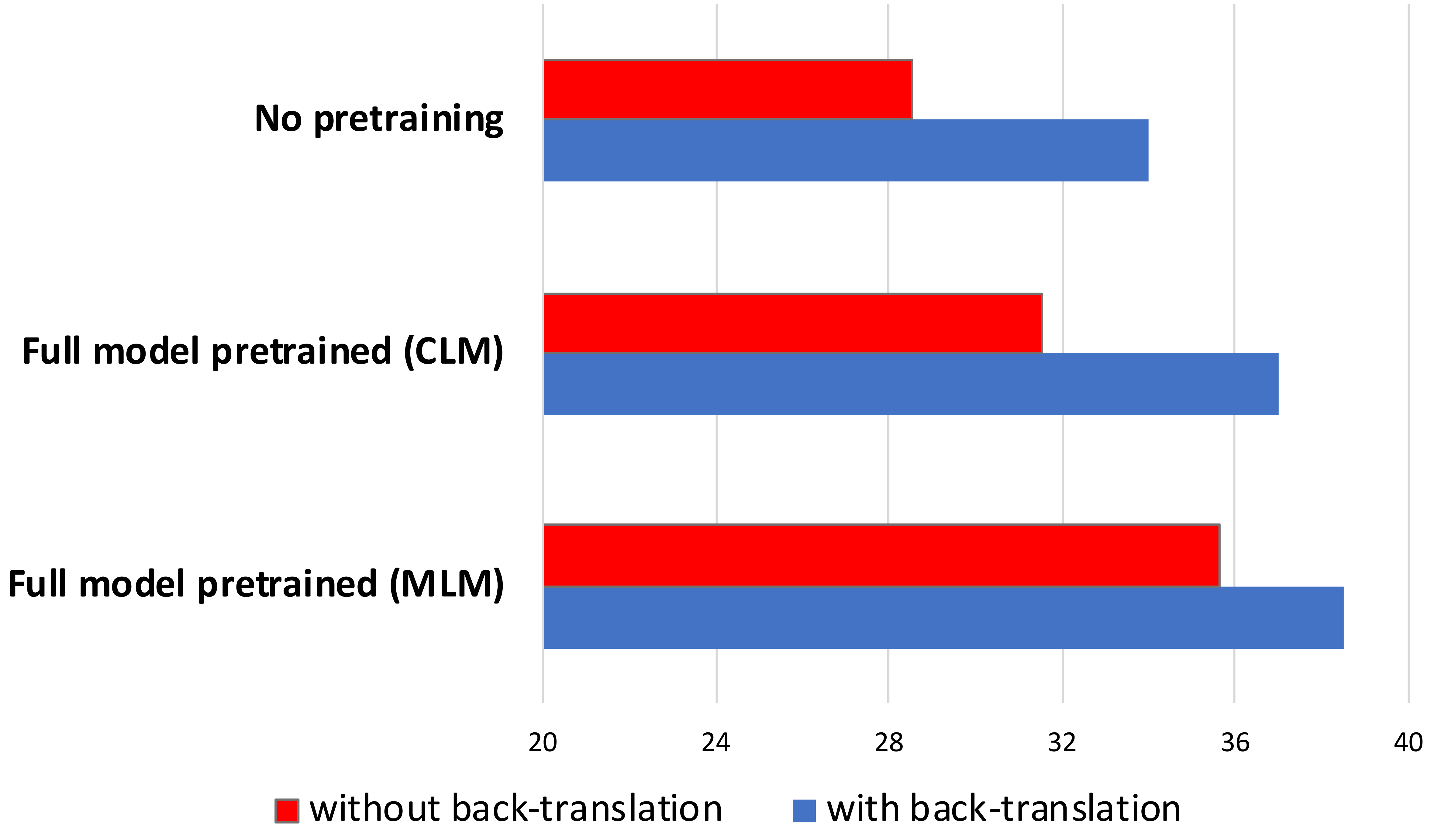
Full Transformer model initialization
significantly improves performance (+7 BLEU)



Cross-lingual Language Model Pre-training, [Conneau et al NeurIPS 2019]

Results on supervised machine translation

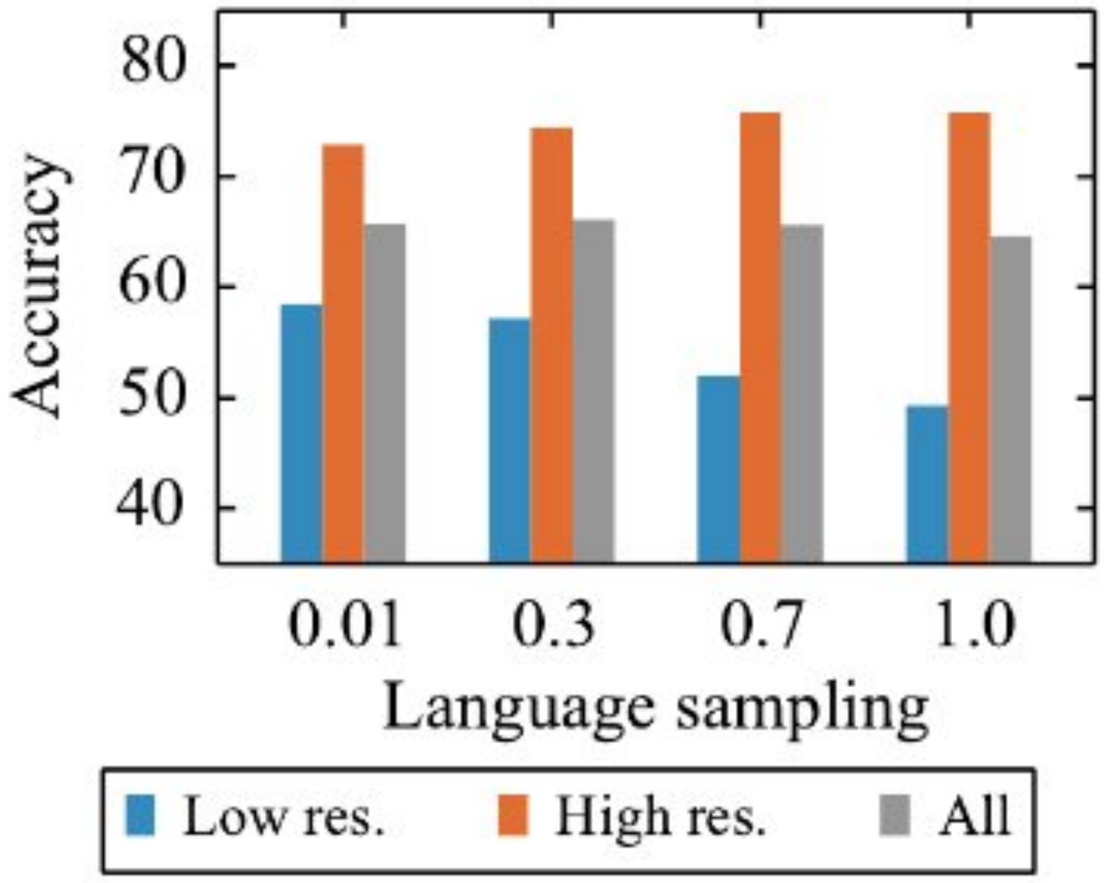
- Pre-training is important for translation
  - Pre-training both encoder and decoder improves
  - MLM is better than CLM
  - Back translation + Pre-training achieve the best



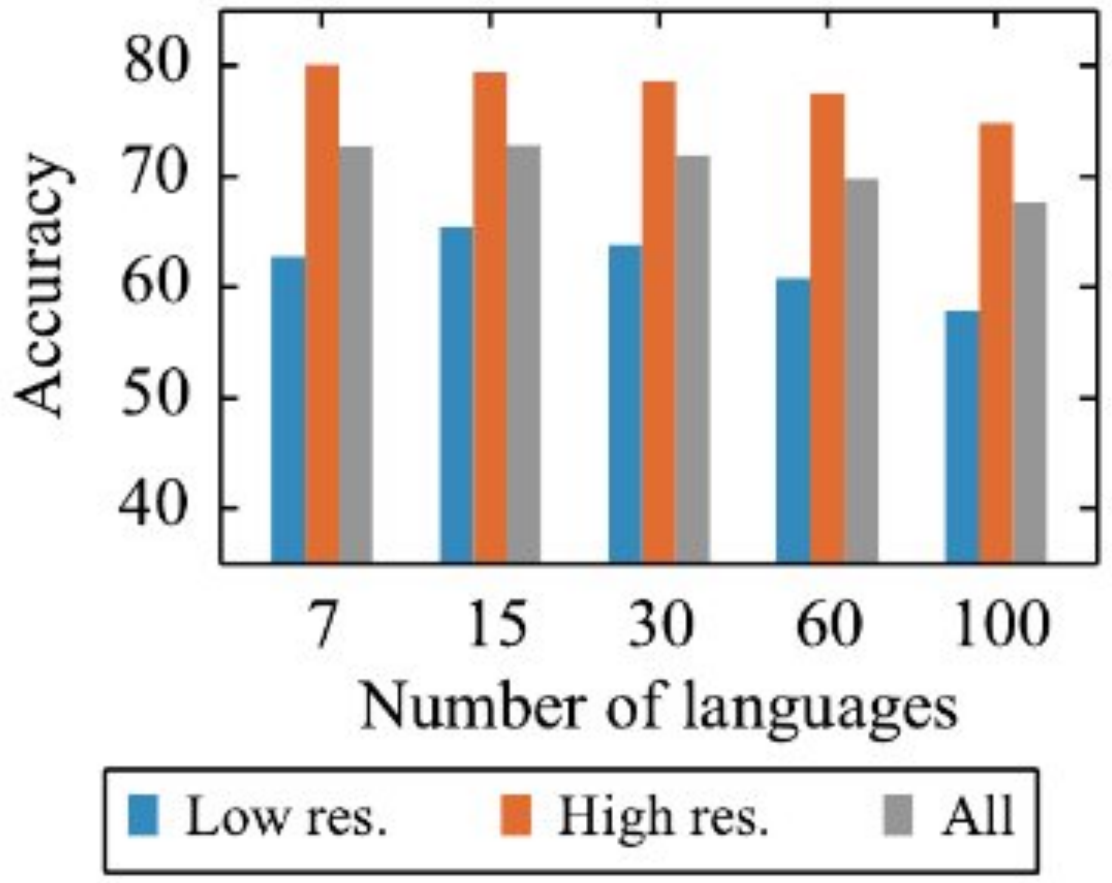
Cross-lingual Language Model Pre-training, [Conneau et al NeurIPS 2019]

Ablation study

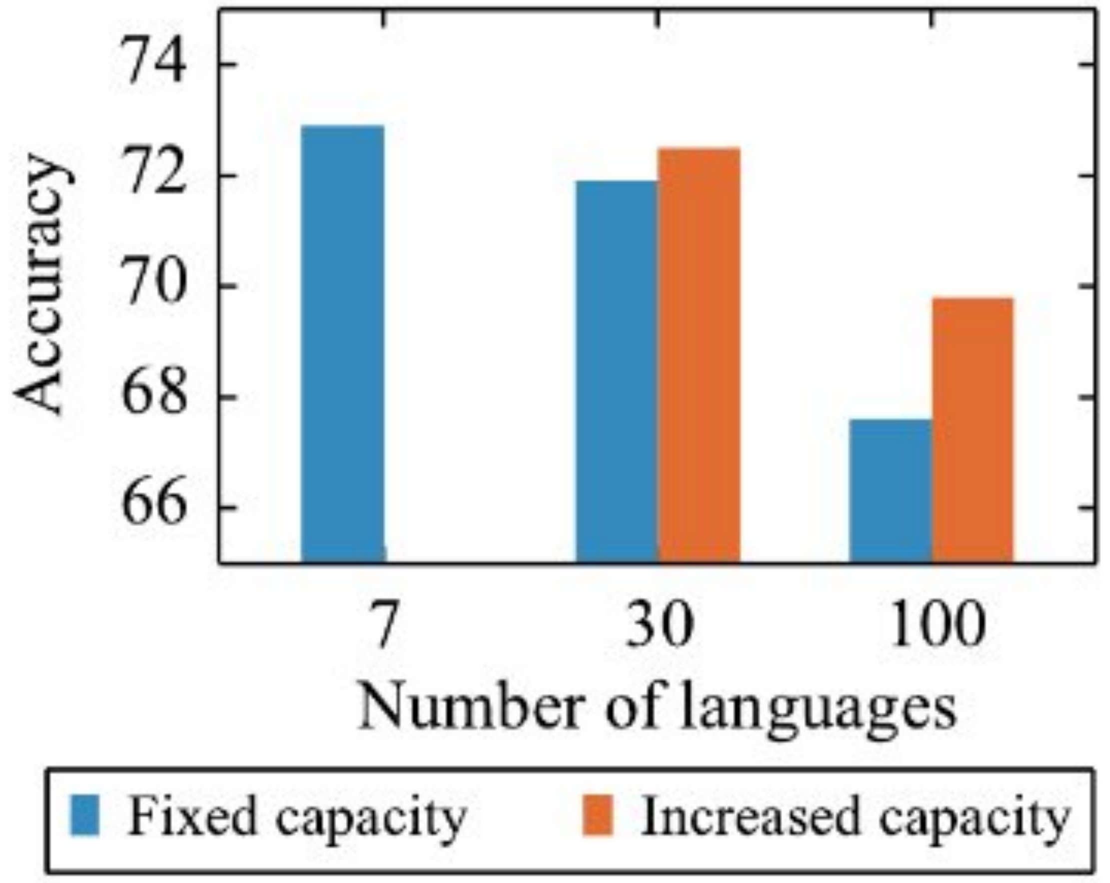
- Adding more languages improves performance on low-resource languages due to positive knowledge transfer
- Sampling batches more often in some languages improves performance in these languages but decrease performance in other languages (capacity allocation problem)



High-res/low-res trade-off



The transfer-interference trade-off



The curse of multilinguality

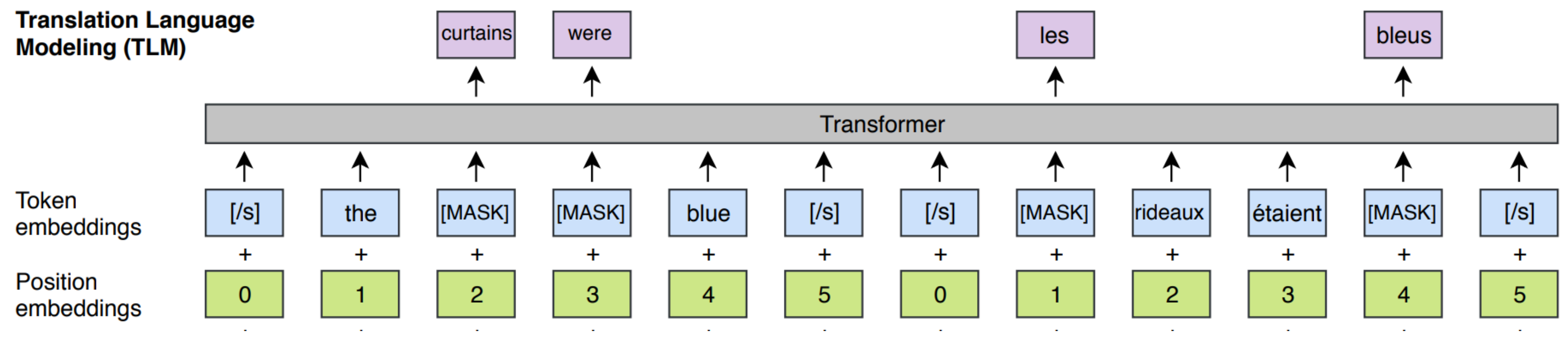
Cross-lingual Language Model Pre-training, [Conneau et al NeurIPS 2019]

Summary

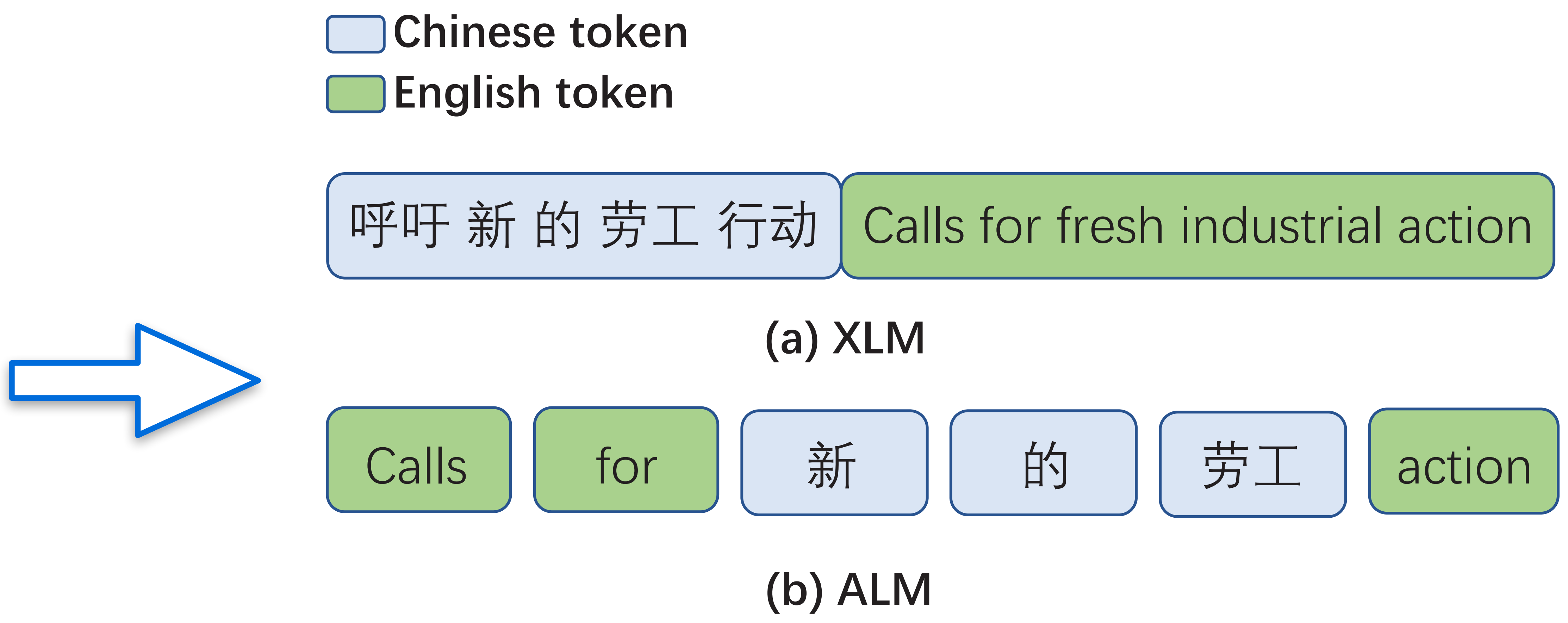
- Cross-lingual language model pre-training is very effective for NMT
- Pre-training reduces the gap between unsupervised and supervised MT
- Encourage knowledge transfer across languages is promising

Cross-lingual Language Model Pre-training, [Conneau et al NeurIPS 2019]

Alternating Language Modeling for Cross-Lingual Pre-Training



Translation language modeling (TLM) pretraining

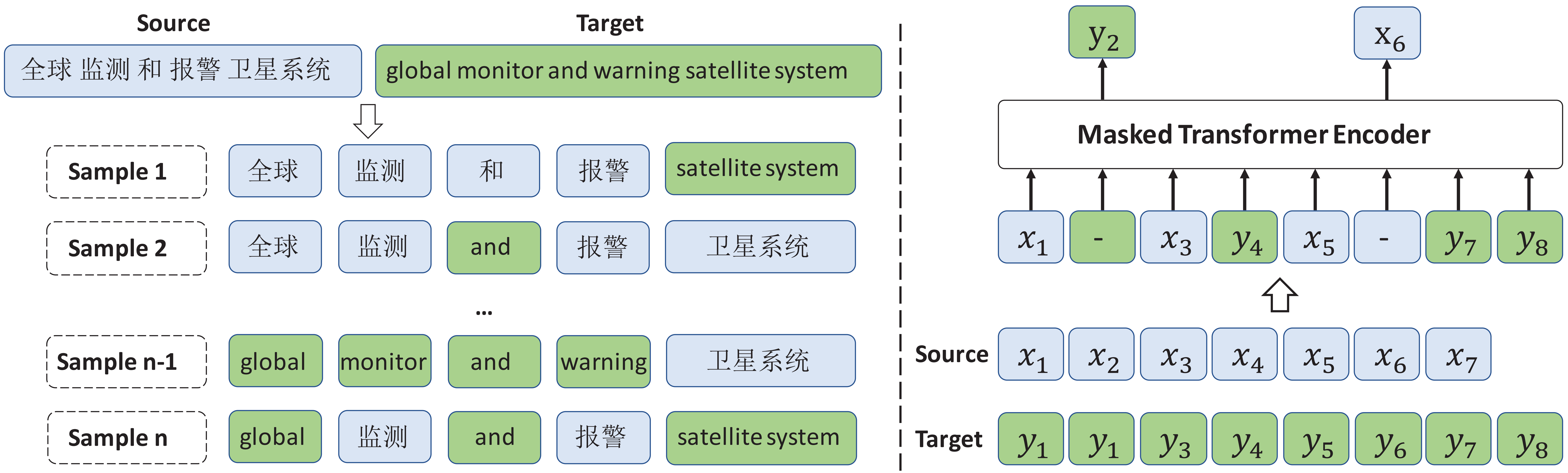


Sentence level mixing

Token level mixing

- ALM extend TLM in a sentence, which alternately predicts words of different languages
- ALM can capture the rich cross-lingual context of words and phrases

Overview of ALM pre-training



Training details

- Dataset
  - Original parallel data to generate 20 times code-switched sentences
  - Separately obtain the alternating language sentences of source language and target language, which are 40 times than original data
  - Totally, 1.5 billion code-switched sentences are used for pre-training
- Model
  - Transformer big
  - Reload the parameters of ALT for both encoder and decoder. The cross-lingual attention parameters are randomly initialized.

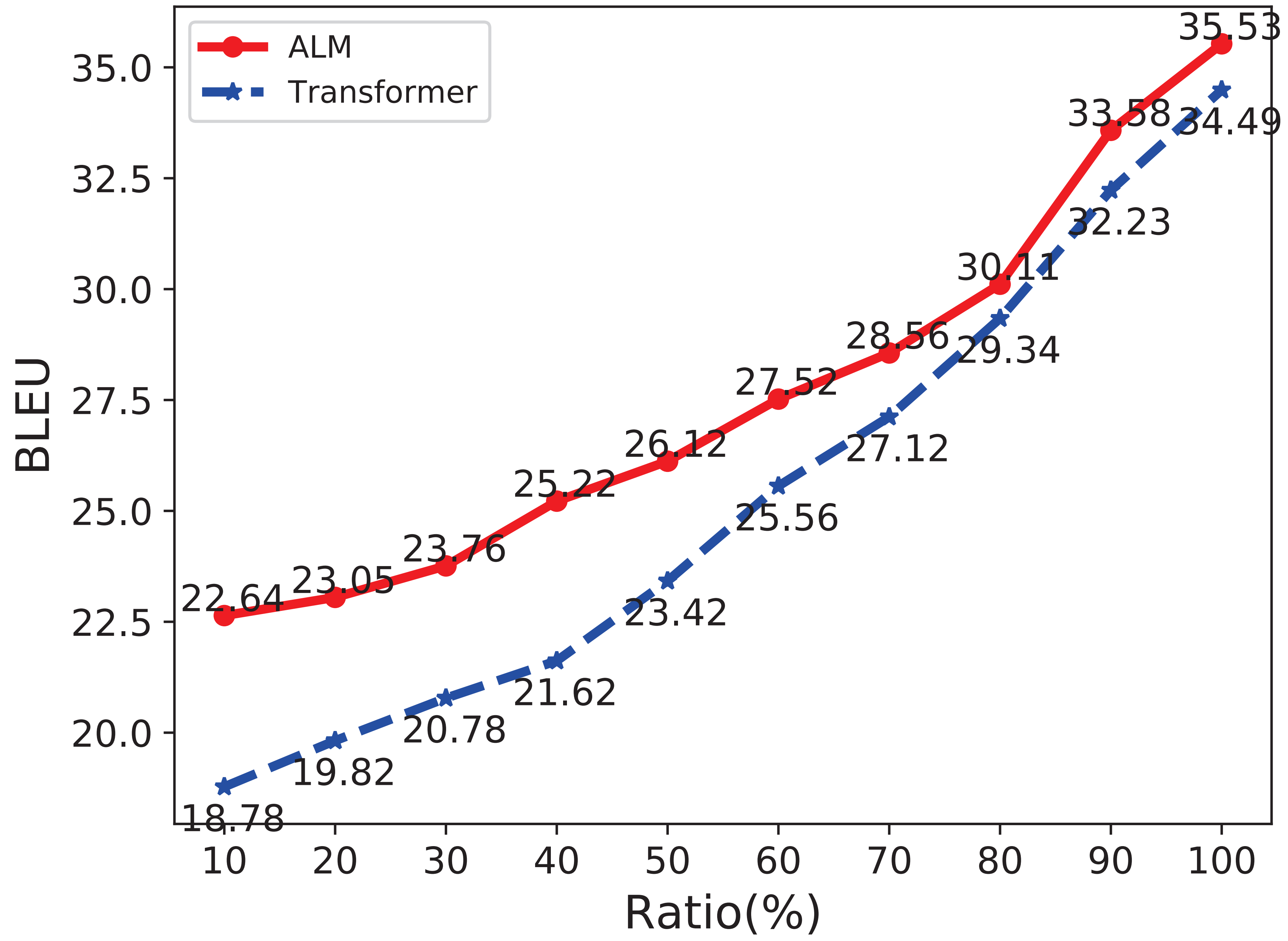
Results

En $\rightarrow$ De	BLEU(%)
Transformer (Vaswani et al. 2017)	28.40
ConvS2S (Gehring et al. 2017)	25.16
Weighted Transformer (Ahmed, Keskar, and Socher 2017)	28.90
Layer-wise Transformer (He et al. 2018)	29.01
RNMT+ (Chen et al. 2018)	28.50
mBERT (Devlin et al. 2019)	28.64
MASS (Song et al. 2019)	28.92
XLM (Lample and Conneau 2019)	28.88
ALM (this work)	**29.22**

De $\rightarrow$ En	BLEU(%)
Transformer (Vaswani et al. 2017)	34.49
LightConv (Wu et al. 2019)	34.80
DynamicConv (Wu et al. 2019)	35.20
Advsoft (Wang, Gong, and Liu 2019)	35.18
Layer-wise Transformer (He et al. 2018)	35.07
mBERT (Devlin et al. 2019)	34.82
MASS (Song et al. 2019)	35.14
XLM (Lample and Conneau 2019)	35.22
ALM (this work)	**35.53**

- mBERT: extends the BERT model to different languages
- XLM: the most related work. The results are implemented with released code.
- Mass: set the fragment length k as 50% of the total number of masked tokens in the sentence.

Results

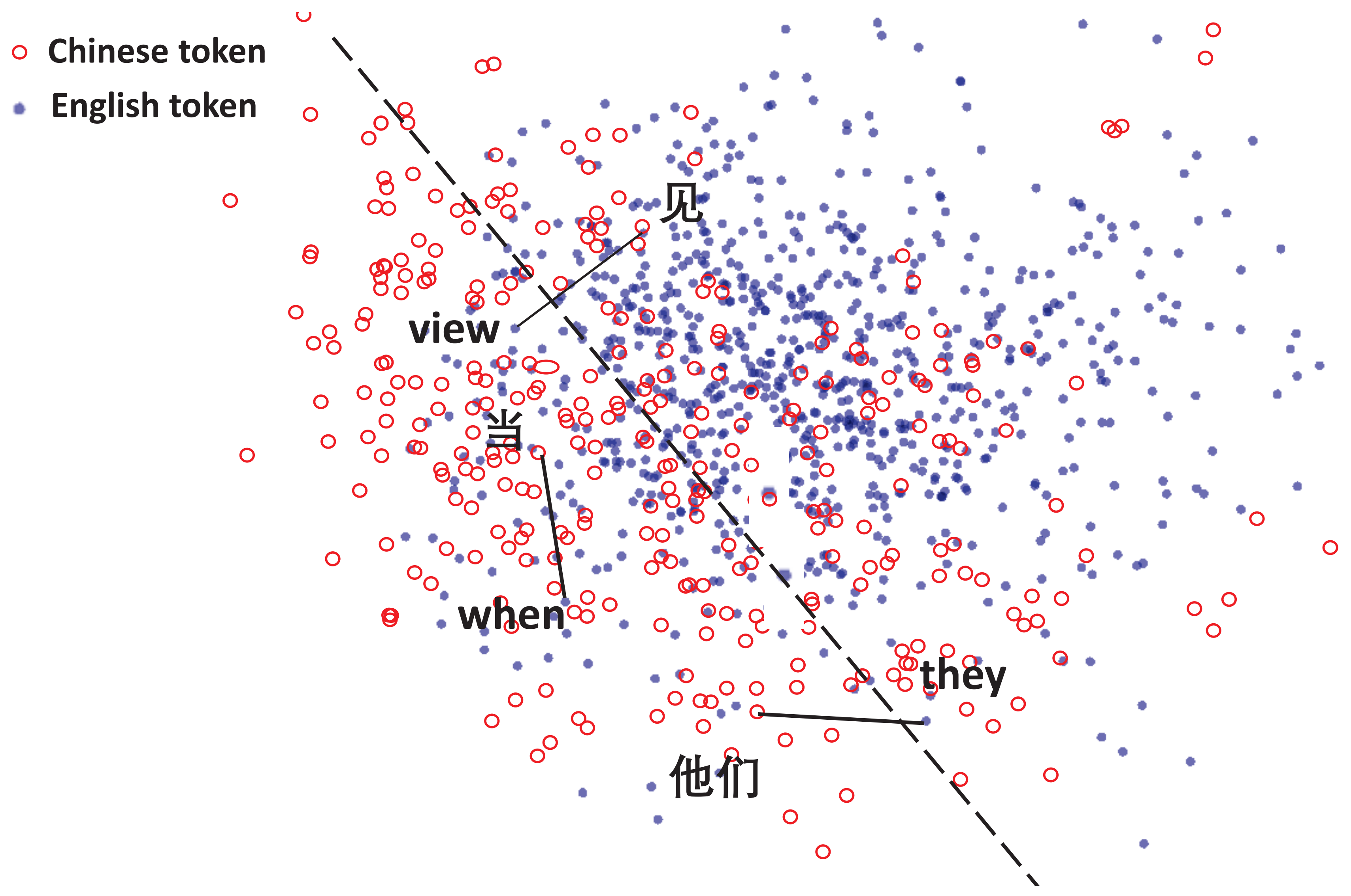


- Randomly shuffle the full parallel training set in the task of IWSLT14 German- to-English translation dataset. Then, extract the random K% samples as the fine-tuned parallel data
- Not surprise, the improvements of ALM is larger for low resource NMT

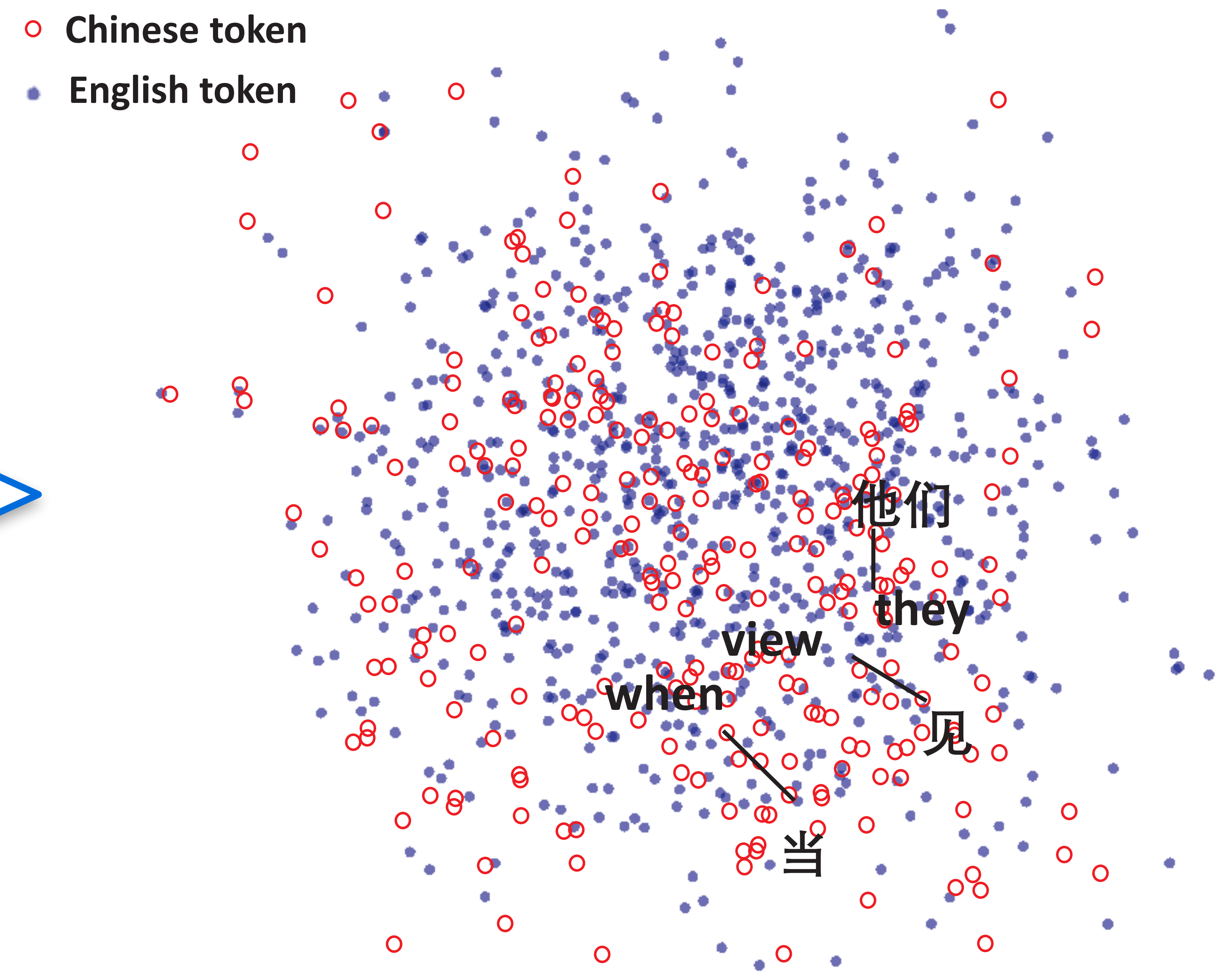
Alternating Language Modeling for Cross-Lingual Pre-Training [Yang et al AAAI 2020]

Visualization of word embedding

Mixing Chinese words and English words can draw the distribution of source language and target language in a same space



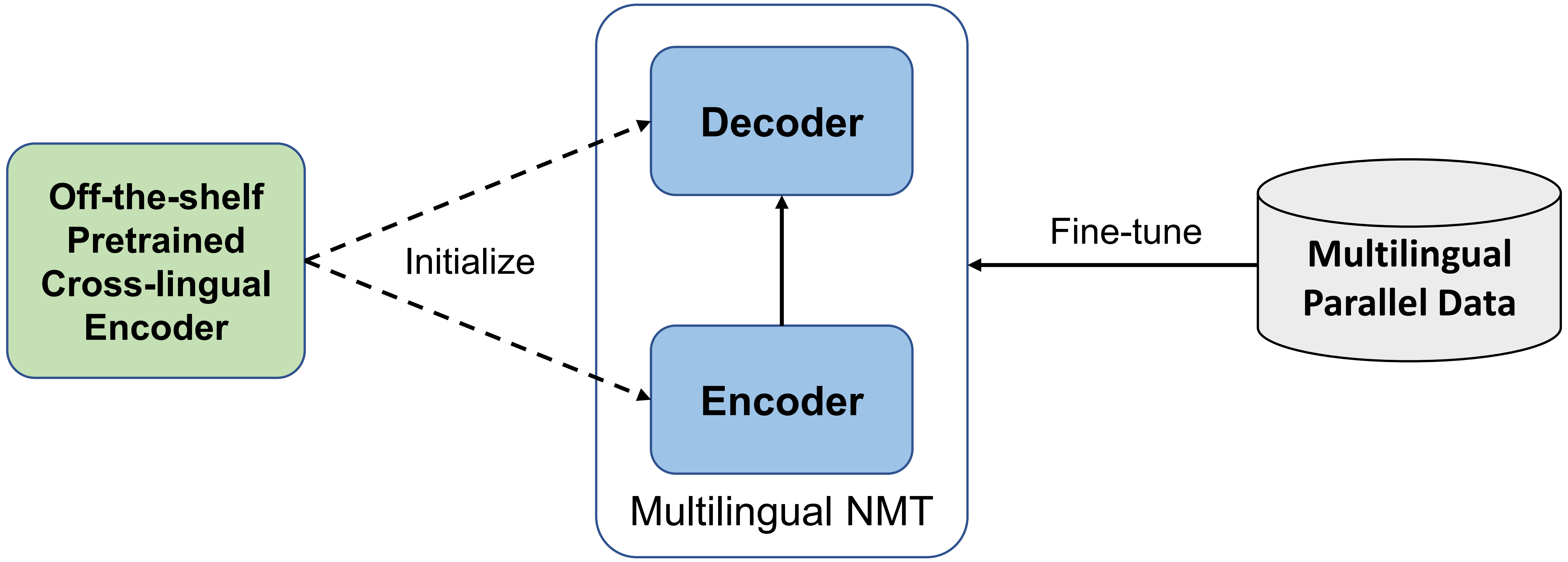
(a) Transformer



(b) ALM

Alternating Language Modeling for Cross-Lingual Pre-Training [Yang et al AAAI 2020]

XLM-T: Scaling up Multilingual Machine Translation with Pretrained Cross-lingual Transformer Encoders



- Initialize MT encoder and decoder with pre-trained cross-lingual encoders
- Fine-tune the model on multilingual parallel data

XLM-T: Scaling up Multilingual Machine Translation with Pretrained Cross-lingual Transformer Encoders [Ma et al, 2020]

XLM-T: Scaling up Multilingual Machine Translation with Pretrained Cross-lingual Transformer Encoders

X → En	Fr	Cs	De	Fi	Lv	Et	Ro	Hi	Tr	Gu	Avg
Train on Original Parallel Data (Bitext)											
Bilingual NMT	36.2	28.5	40.2	19.2	17.5	19.7	29.8	14.1	15.1	9.3	23.0
Many-to-One	34.8	29.0	40.1	21.2	20.4	26.2	34.8	22.8	23.8	19.2	27.2
XLM-T	35.9	30.5	41.6	22.5	21.4	28.4	36.6	24.6	25.6	20.4	**28.8**
Many-to-Many	35.9	29.2	40.0	21.1	20.4	26.3	35.5	23.6	24.3	20.6	27.7
XLM-T	35.5	30.0	40.8	22.1	21.5	27.8	36.5	25.3	25.0	20.6	**28.5**
*Train on Original Parallel Data and **Back-Translation** Data (Bitext+BT)*											
(Wang et al., 2020)	35.3	31.9	45.4	23.8	22.4	30.5	39.1	28.7	27.6	23.5	30.8
Many-to-One	35.9	32.6	44.1	24.9	23.1	31.5	39.7	28.2	27.8	23.1	31.1
XLM-T	36.0	33.1	44.8	25.4	23.9	32.7	39.8	30.1	28.8	23.6	**31.8**
(Wang et al., 2020)	35.3	31.2	43.7	23.1	21.5	29.5	38.1	27.5	26.2	23.4	30.0
Many-to-Many	35.7	31.9	43.7	24.2	23.2	30.4	39.1	28.3	27.4	23.8	30.8
XLM-T	36.1	32.6	44.3	25.4	23.8	32.0	40.3	29.5	28.7	24.2	**31.7**

- The multilingual models achieve much better performance on the low-resource languages and are worse on the high-resource languages
- XLM-T achieves significant improvements over the multilingual baseline across all 10 languages
- In the back-translation setting, XLM-T can further improve this strong baseline

XLM-T: Scaling up Multilingual Machine Translation with Pretrained Cross-lingual Transformer Encoders [Ma et al, 2020]

XLM-T: Scaling up Multilingual Machine Translation with Pretrained Cross-lingual Transformer Encoders

X → En	Fr	Cs	De	Fi	Lv	Et	Ro	Hi	Tr	Gu	Avg
Train on Original Parallel Data (Bitext)											
Bilingual NMT	36.2	28.5	40.2	19.2	17.5	19.7	29.8	14.1	15.1	9.3	23.0
Many-to-One	34.8	29.0	40.1	21.2	20.4	26.2	34.8	22.8	23.8	19.2	27.2
XLM-T	35.9	30.5	41.6	22.5	21.4	28.4	36.6	24.6	25.6	20.4	**28.8**
Many-to-Many	35.9	29.2	40.0	21.1	20.4	26.3	35.5	23.6	24.3	20.6	27.7
XLM-T	35.5	30.0	40.8	22.1	21.5	27.8	36.5	25.3	25.0	20.6	**28.5**
*Train on Original Parallel Data and **Back-Translation** Data (Bitext+BT)*											
(Wang et al., 2020)	35.3	31.9	45.4	23.8	22.4	30.5	39.1	28.7	27.6	23.5	30.8
Many-to-One	35.9	32.6	44.1	24.9	23.1	31.5	39.7	28.2	27.8	23.1	31.1
XLM-T	36.0	33.1	44.8	25.4	23.9	32.7	39.8	30.1	28.8	23.6	**31.8**
(Wang et al., 2020)	35.3	31.2	43.7	23.1	21.5	29.5	38.1	27.5	26.2	23.4	30.0
Many-to-Many	35.7	31.9	43.7	24.2	23.2	30.4	39.1	28.3	27.4	23.8	30.8
XLM-T	36.1	32.6	44.3	25.4	23.8	32.0	40.3	29.5	28.7	24.2	**31.7**

- The multilingual models achieve much better performance on the low-resource languages and are worse on the high-resource languages
- XLM-T achieves significant improvements over the multilingual baseline across all 10 languages
- In the back-translation setting, XLM-T can further improve this strong baseline

XLM-T: Scaling up Multilingual Machine Translation with Pretrained Cross-lingual Transformer Encoders

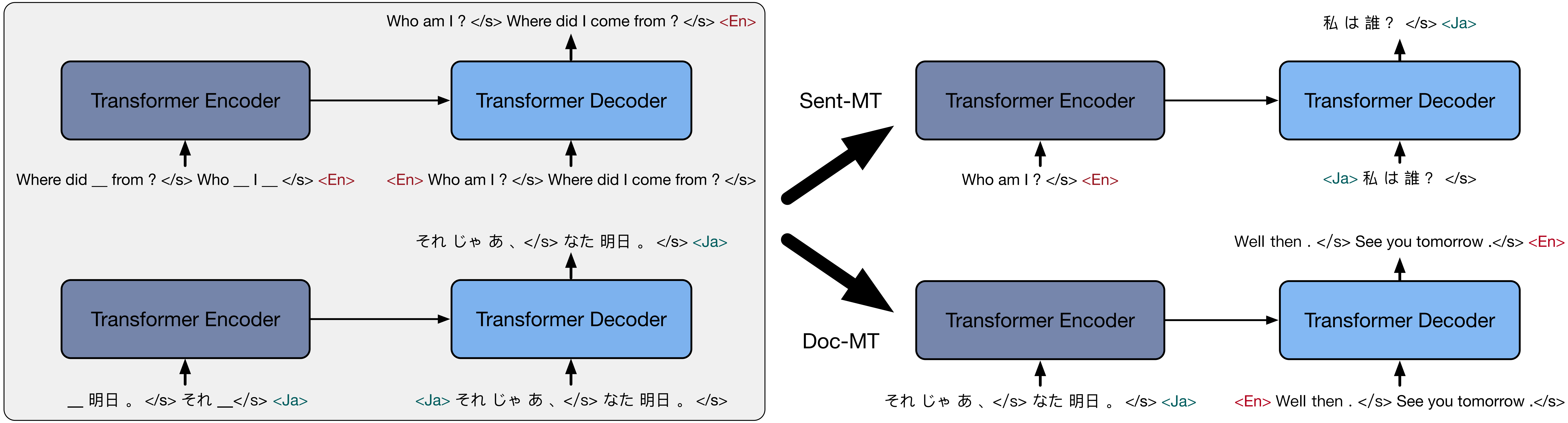
En → X	Fr	Cs	De	Fi	Lv	Et	Ro	Hi	Tr	Gu	Avg
Train on Original Parallel Data (Bitext)											
Bilingual NMT	36.3	22.3	40.2	15.2	16.5	15.0	23.0	12.2	13.3	7.9	20.2
One-to-Many	34.2	20.9	40.0	15.0	18.1	20.9	26.0	14.5	17.3	13.2	22.0
XLM-T	34.8	21.4	39.9	15.4	18.7	20.9	26.6	15.8	17.4	15.0	**22.6**
Many-to-Many	34.2	21.0	39.4	15.2	18.6	20.4	26.1	15.1	17.2	13.1	22.0
XLM-T	34.2	21.4	39.7	15.3	18.9	20.6	26.5	15.6	17.5	14.5	**22.4**
*Train on Original Parallel Data and **Back-Translation** Data (Bitext+BT)*											
(Wang et al., 2020)	36.1	23.6	42.0	17.7	22.4	24.0	29.8	19.8	19.4	17.8	25.3
One-to-Many	36.8	23.6	42.9	18.3	23.3	24.2	29.5	20.2	19.4	13.2	25.1
XLM-T	37.3	24.2	43.6	18.1	23.7	24.2	29.7	20.1	20.2	13.7	**25.5**
(Wang et al., 2020)	35.8	22.4	41.2	16.9	21.7	23.2	29.7	19.2	18.7	16.0	24.5
Many-to-Many	35.9	22.9	42.2	17.5	22.5	23.4	28.9	19.8	19.1	14.5	24.7
XLM-T	36.6	23.9	42.4	18.4	22.9	24.2	29.3	20.1	19.8	12.8	**25.0**

- Generally, the improvements are smaller than X → En
- The multilingual part of En → X is at the decoder side, which XLM-R is not an expert in.

PART 3: Multilingual Pre-training for NMT

- Multilingual fused pre-training
  - Cross-lingual Language Model Pre-training [NeurIPS, 2019]
  - Alternating Language Modeling Pre-training [AAAI, 2020]
  - XLM-T: Cross-lingual Transformer Encoders
- Multilingual sequence to sequence pre-training
  - mBART [TACL, 2020]
  - CSP [EMNLP, 2020]
  - mRASP & mRASP2 [EMNLP, 2020] [ACL, 2021]
  - LaSS: Learning language-specific sub-network via pre-training & fine-tuning [ACL, 2021]

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]

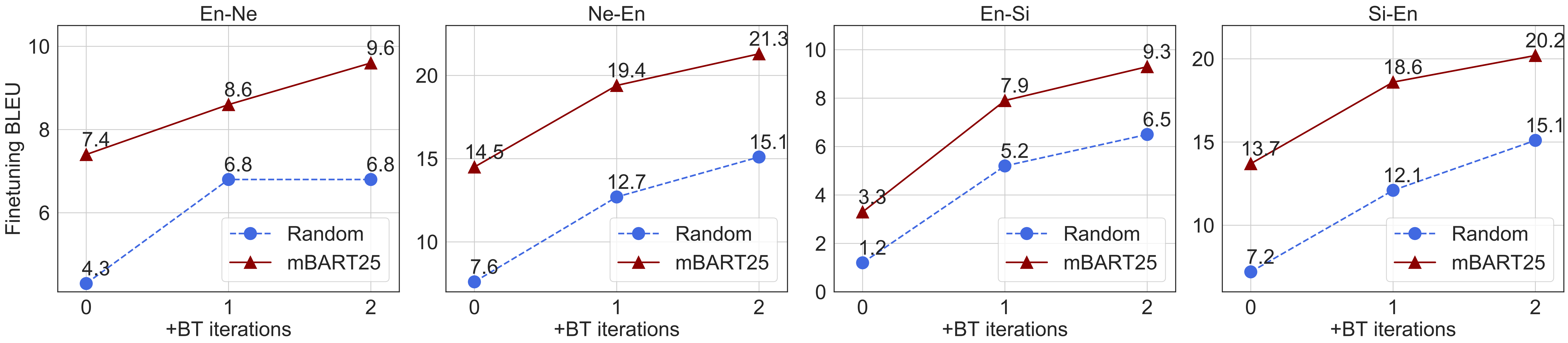
mBART: Rich-resource translation

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?

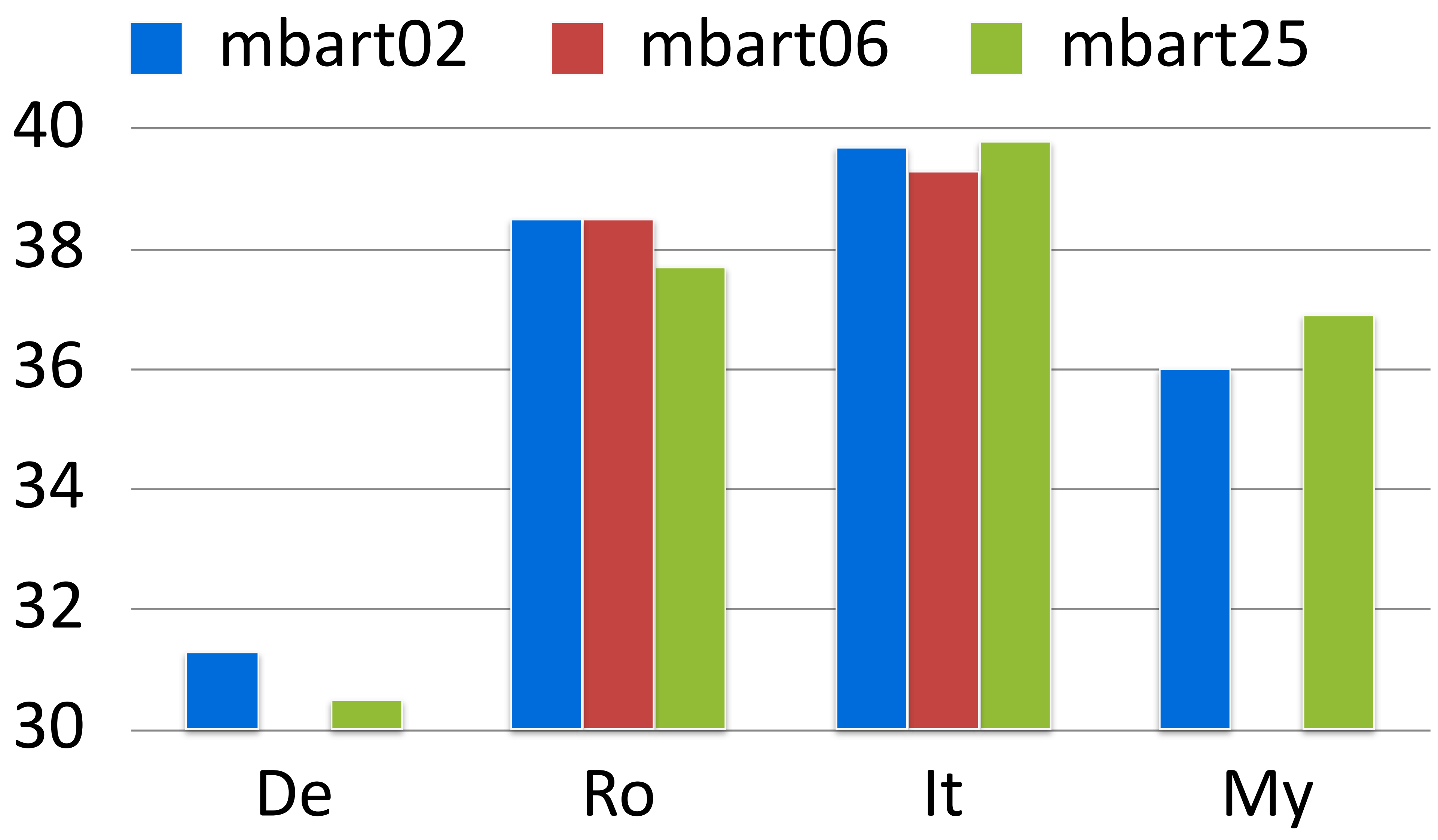
Pre-training		Fine-tuning		
Model	**Data**	**En→Ro**	**Ro→En**	**+BT**
Random	None	34.3	34.0	36.8
XLM (2019)	En Ro	-	35.6	38.5
MASS (2019)	En Ro	-	-	39.1
BART (2019)	En	-	-	38.0
XLM-R (2019)	CC100	35.6	35.8	-
BART-En	En	36.0	35.8	37.4
BART-Ro	Ro	37.6	36.8	38.1
mBART02	En Ro	**38.5**	**38.5**	**39.9**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

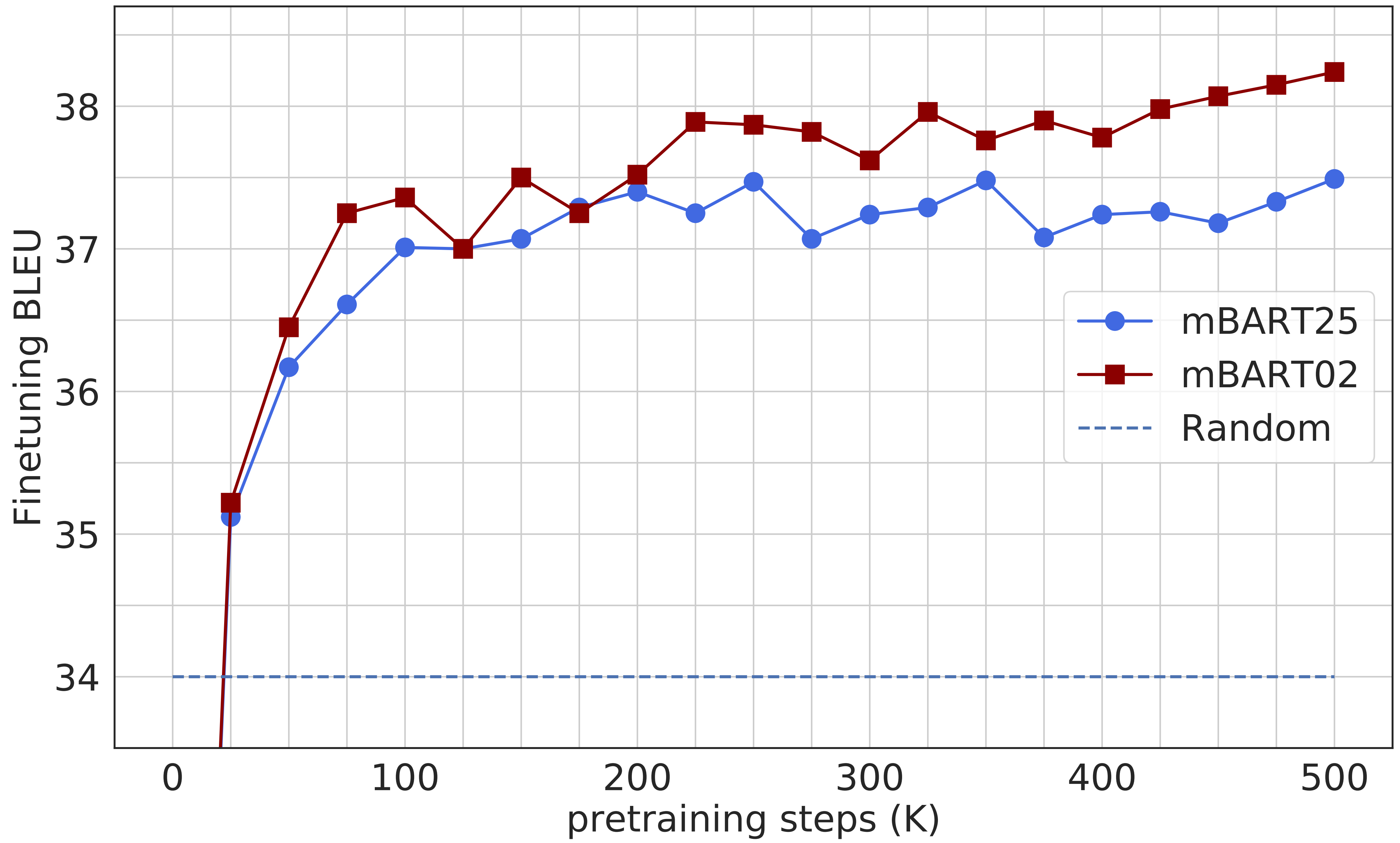
Languages	De	Ro	It	My	En
Size/GB	66.6	61.4	30.2	1.6	300.8
mBART02	**31.3**	**38.5**	39.7	36.5	
mBART06	-	**38.5**	39.3	-	
mBART25	30.5	37.7	**39.8**	**36.9**	



- 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]

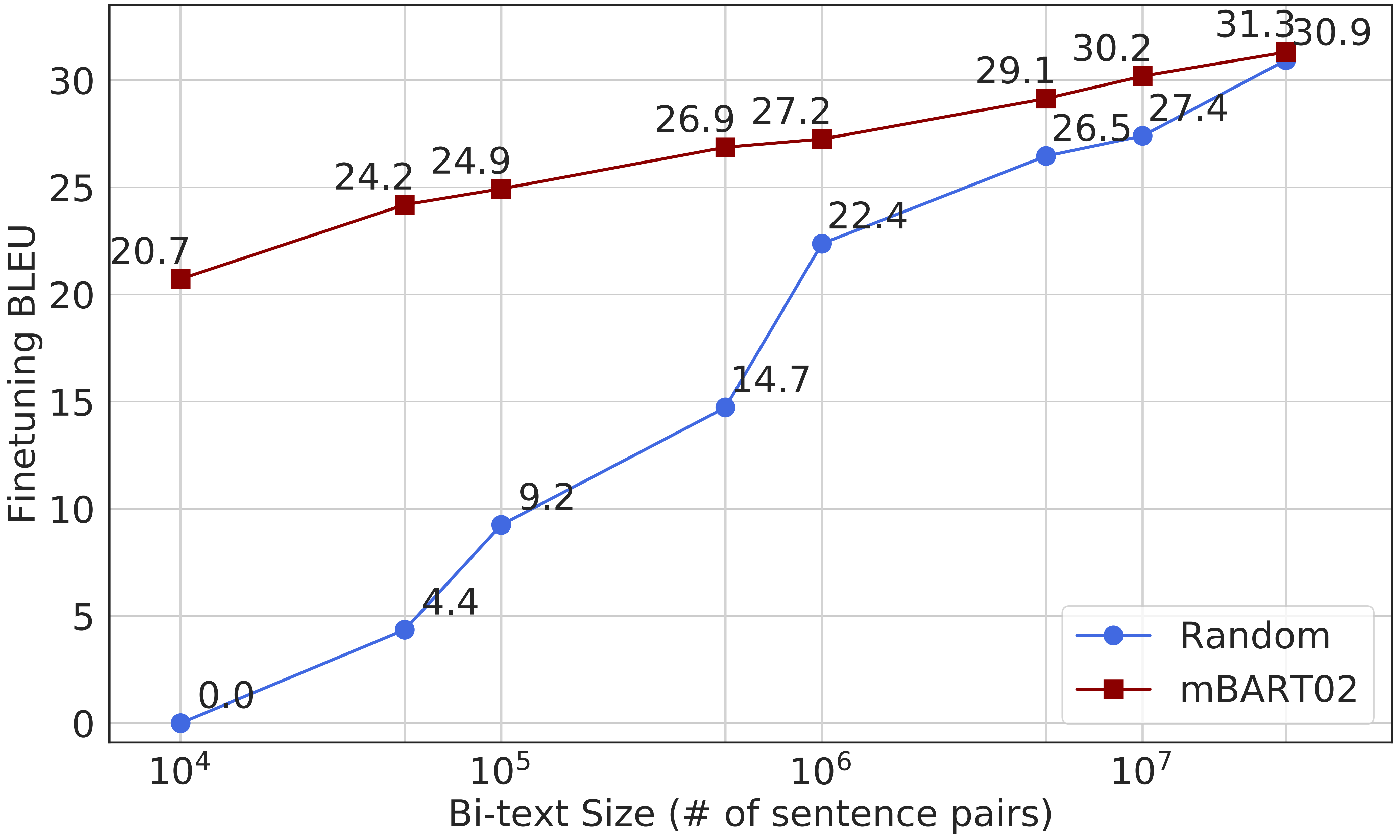
Analysis: Pre-training steps matters



- Without any pre-training, the model overfits and performs much worse than the baseline
- After just 25K steps (5% of training), both models outperform the best baseline.
- The models keep improving by over 3 BLEU for the rest of steps and have not fully converged after 500K steps.
- The more the better

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

Analysis: Perform better on low resource



- The pre-trained model is able to achieve over 20 BLEU with only 10K training examples, while the baseline system scores 0.
- Unsurprisingly, mBART consistently outperforms the baseline models, but the gap reduces with increasing amounts of bi-text, especially after 10M sentence pairs

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

Analysis: Generalization to unseen languages

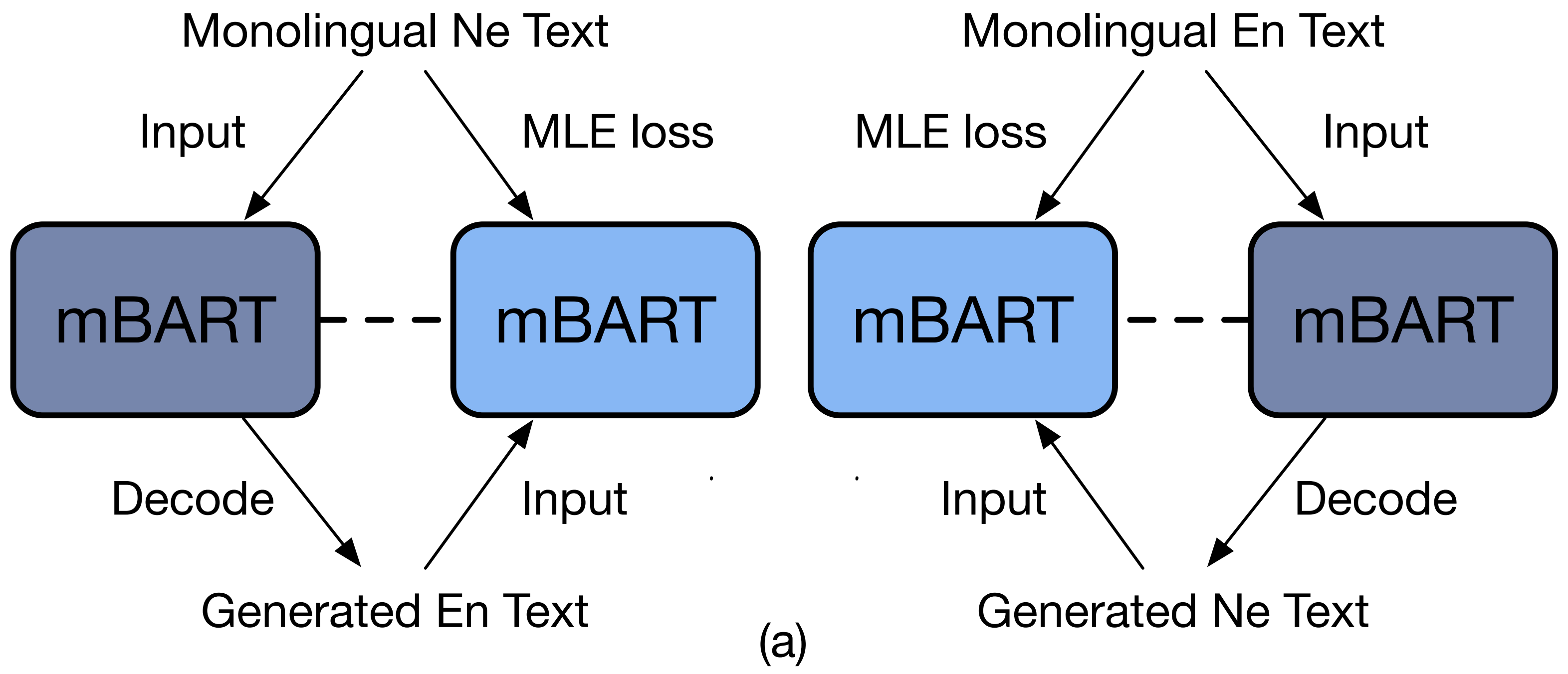
	Monolingual	Nl-En	En-Nl	Ar-En	En-Ar	Nl-De	De-Nl
Random	None	34.6 (-8.7)	29.3 (-5.5)	27.5 (-10.1)	16.9 (-4.7)	21.3 (-6.4)	20.9 (-5.2)
mBART02	En Ro	41.4 (-2.9)	34.5 (-0.3)	34.9 (-2.7)	21.2 (-0.4)	26.1 (-1.6)	25.4 (-0.7)
mBART06	En Ro Cs It Fr Es	43.1 (-0.2)	34.6 (-0.2)	37.3 (-0.3)	21.1 (-0.5)	26.4 (-1.3)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

Unsupervised Machine Translation



UNMT with back translation

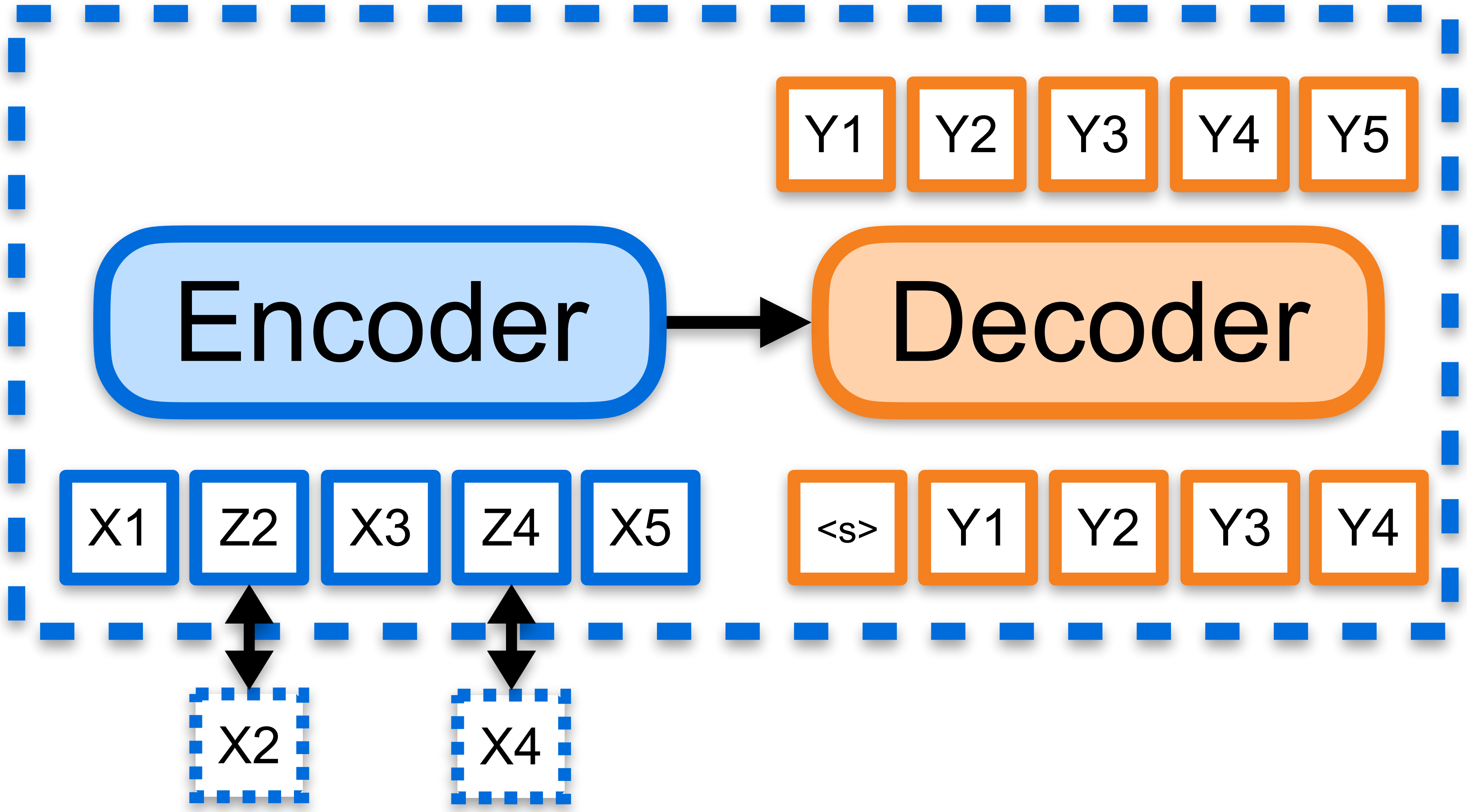
Model	Similar Pairs				Dissimilar Pairs			
	En-De		En-Ro		En-Ne		En-Si	
	←	→	←	→	←	→	←	→
Random	21.0	17.2	19.4	21.2	0.0	0.0	0.0	0.0
XLM (2019)	34.3	26.4	31.8	33.3	0.5	0.1	0.1	0.1
MASS (2019)	**35.2**	28.3	**33.1**	**35.2**	-	-	-	-
mBART	34.0	**29.8**	30.5	35.0	**10.0**	**4.4**	**8.2**	**3.9**

- Following the same procedure with UNMT, but initialize the translation model with the pre-trained mBART
- To avoid simply copying the source text, constrain mBART to only generating tokens in target language
- Achieve very competitive results

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

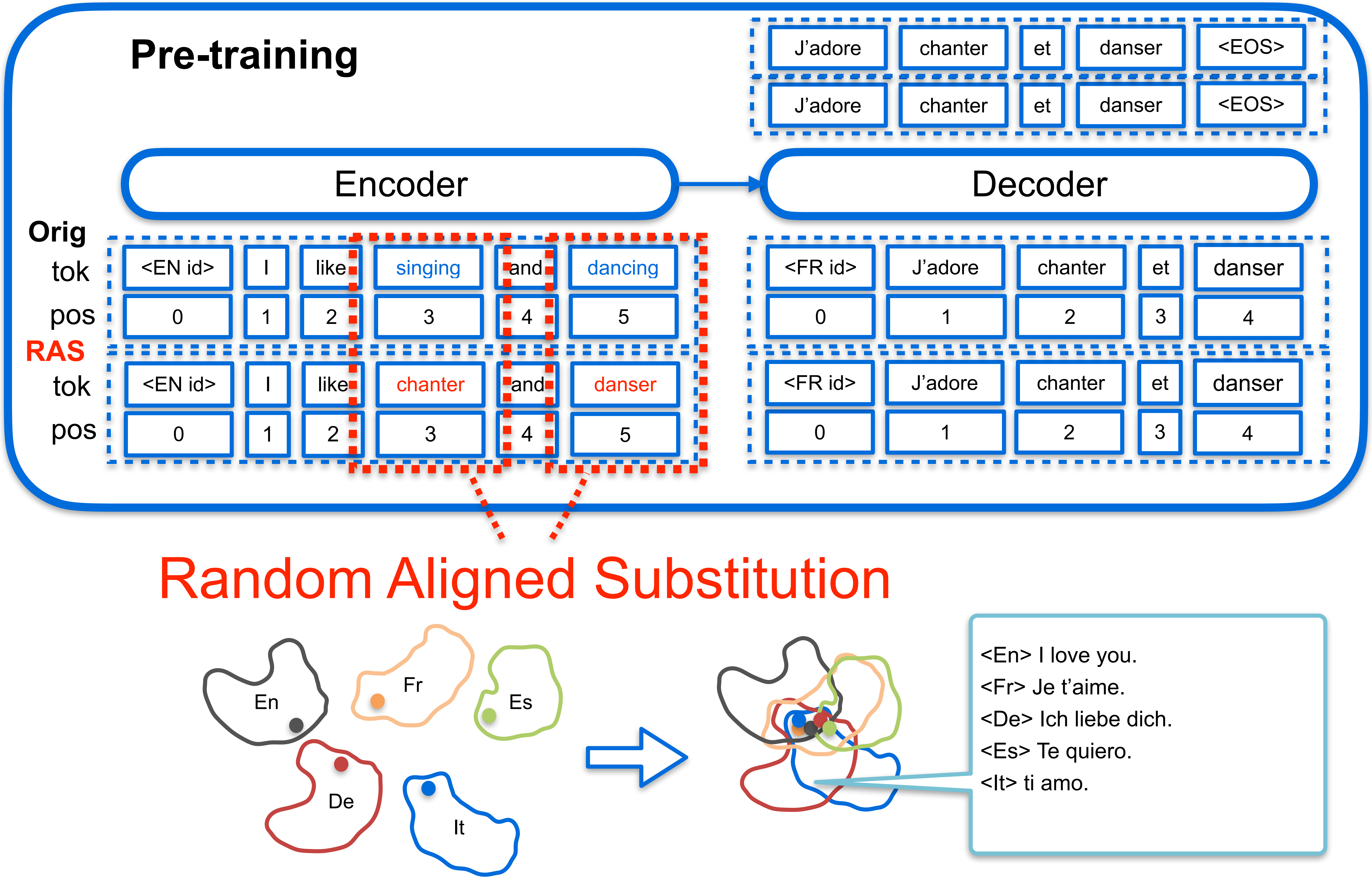
mRASP: multilingual Random Aligned Substitution Pre-training

- **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



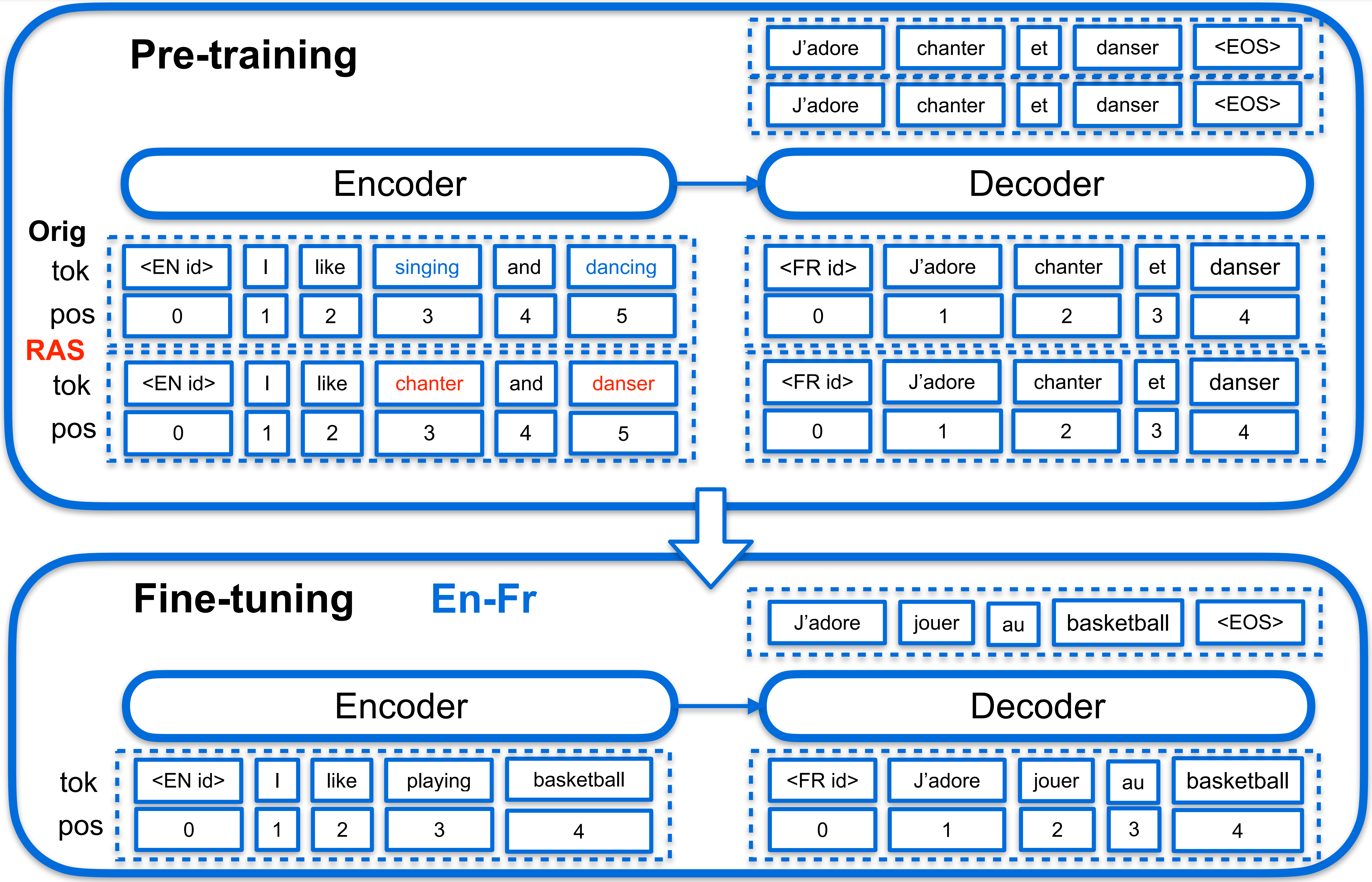
Pre-training Multilingual Neural Machine Translation by Leveraging Alignment Information [Lin et al., EMNLP 2020]

mRASP: Overview



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

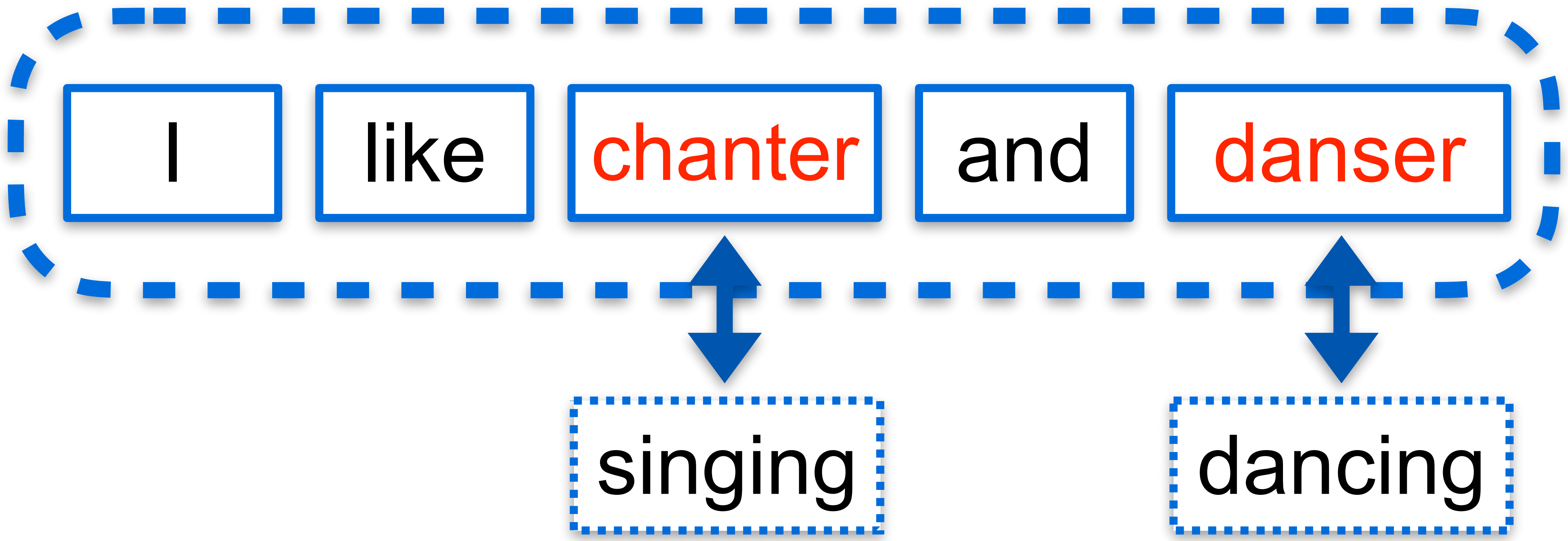
mRASP: Overview



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

mRASP: RAS method

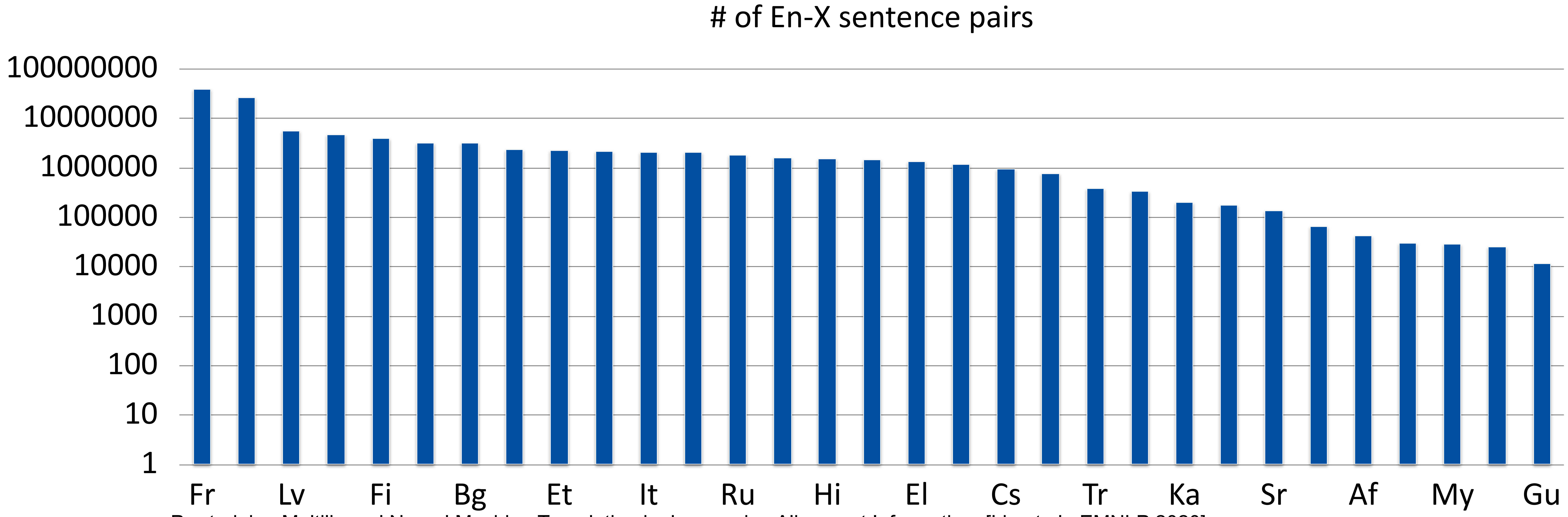
- **Random Aligned Substitution (RAS)**
  - Randomly replace a source word to its synonym in different language.
  - Draw the embedding space closer.



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

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



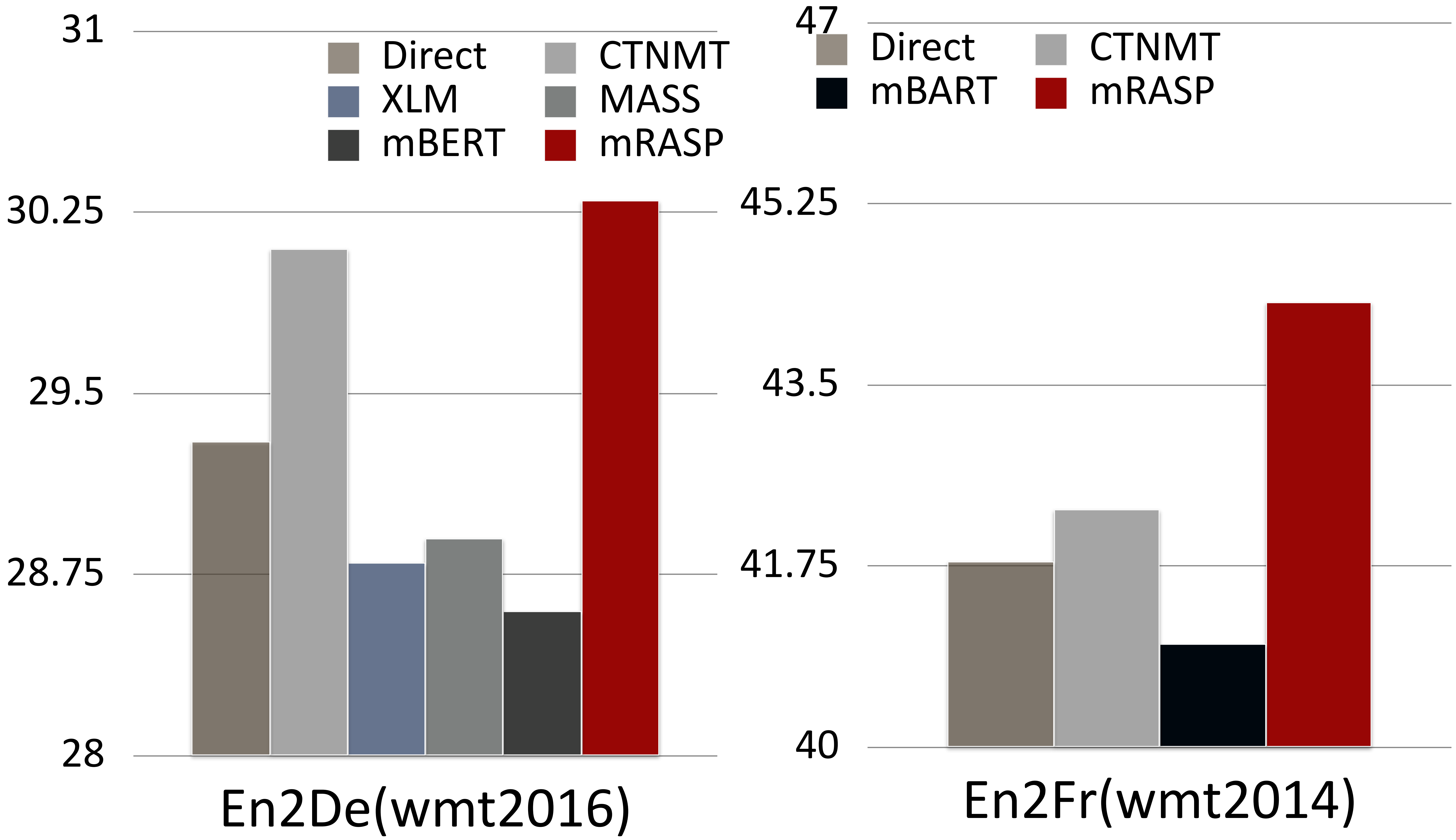
Pre-training Multilingual Neural Machine Translation by Leveraging Alignment Information [Lin et al., EMNLP 2020]

mRASP: Fine-tuning Dataset

- 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.)

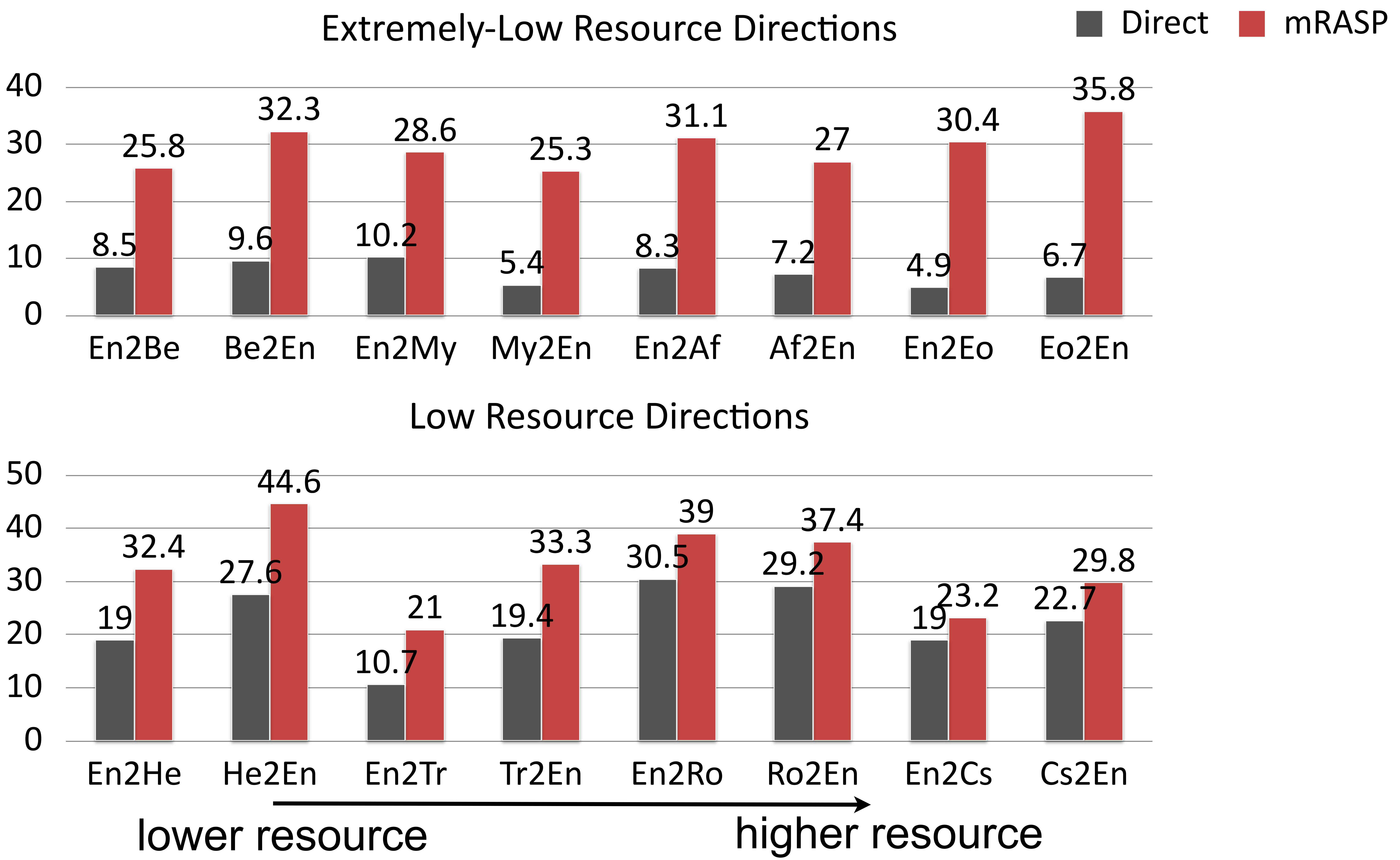
mRASP: Rich resource works

- Rich resource benchmarks can be further improved (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.

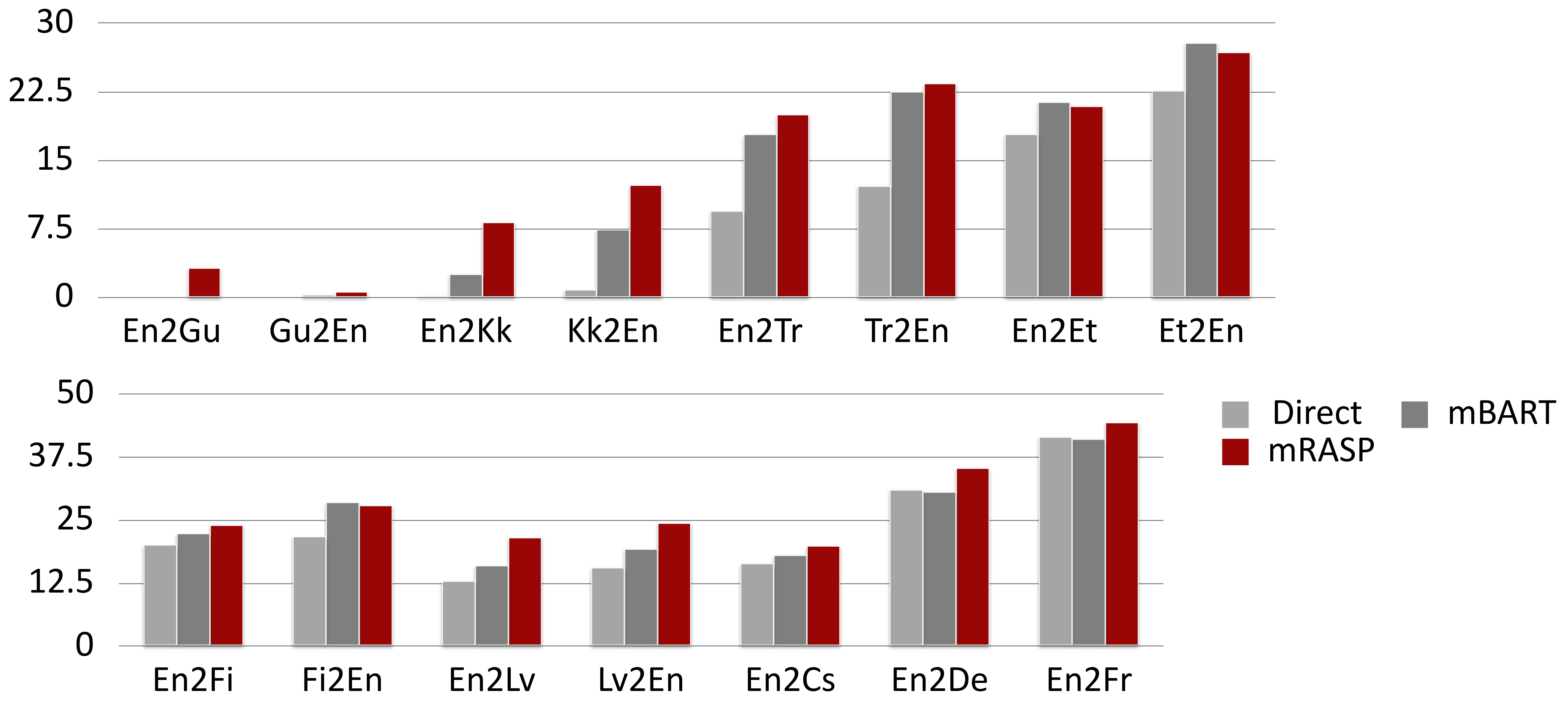
		Fr-Zh(20K)		De-Fr(9M)	
		—>	<—	—>	<—
Exotic Pair	Direct	0.7	3	23.5	21.2
	mRASP	25.8	26.7	29.9	23.4
		Nl-Pt(12K)		Da-El(1.2M)	
		—>	<—	—>	<—
Exotic Full	Direct	0.0	0.0	14.1	16.9
	mRASP	14.1	13.2	17.6	19.9
		En-Mr(11K)		En-Gl(1.2M)	
		—>	<—	—>	<—
Exotic Source/ Target	Direct	6.4	6.8	8.9	12.8
	mRASP	22.7	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]

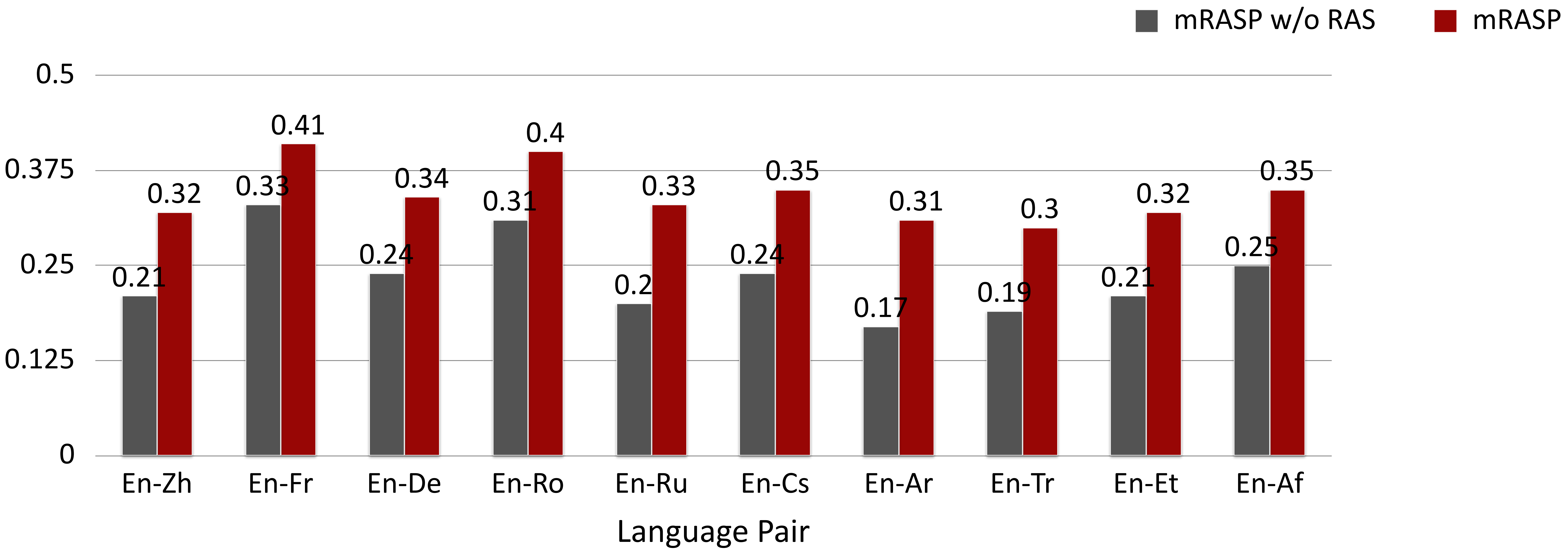
mRASP: Compare with other methods

- mRASP outperforms mBART for all but two language pairs.



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

mRASP: Makes multilingual embeddings more similar



RAS draws the embedding space of languages closer.

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

mRASP 2: Contrastive Learning for Many-to-many Multilingual Neural Machine Translation

Supervised

Unsupervised

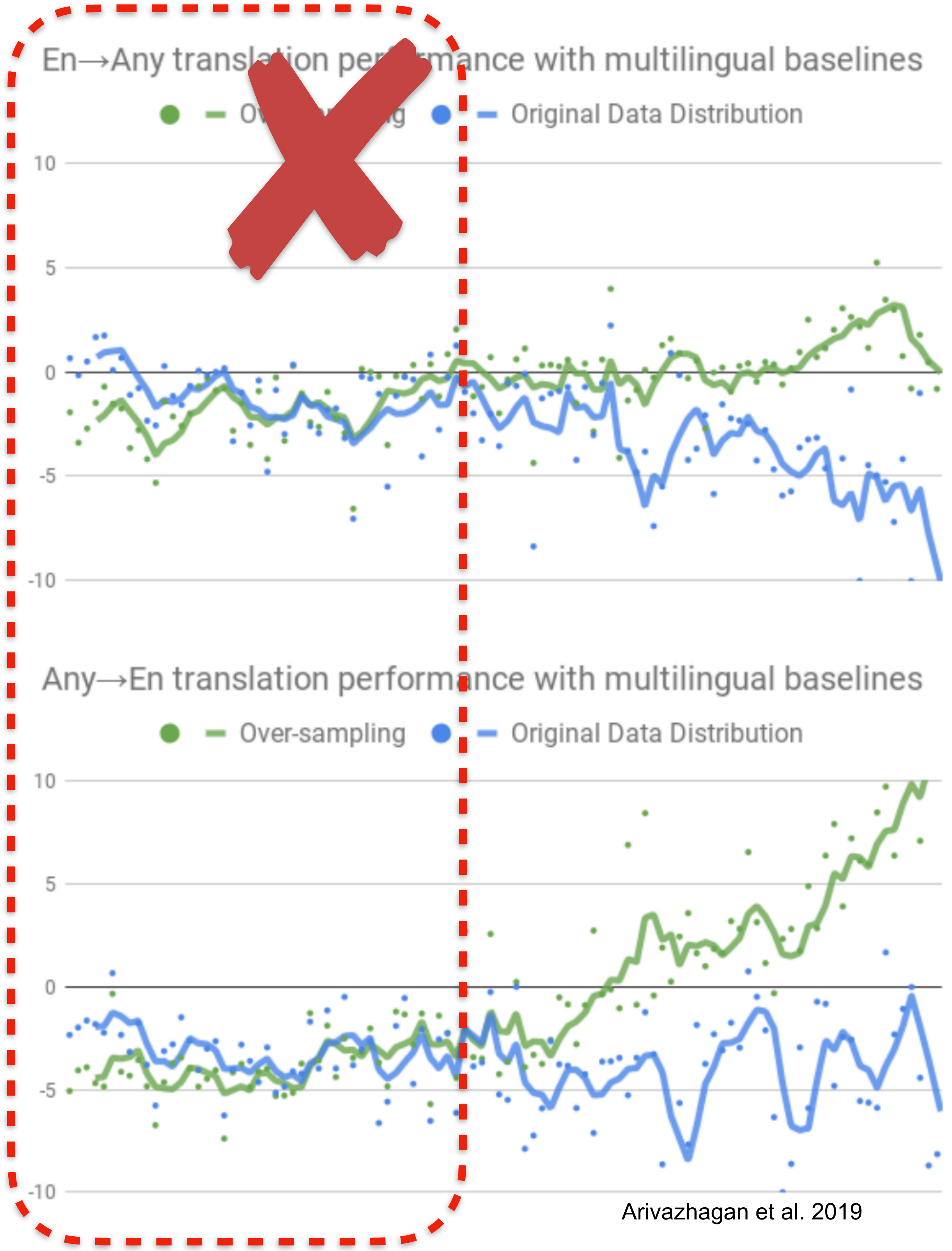
Zero-shot

Enabling unsupervised / zero-shot translation

Parallel

Monolingual

Leveraging both parallel & monolingual data

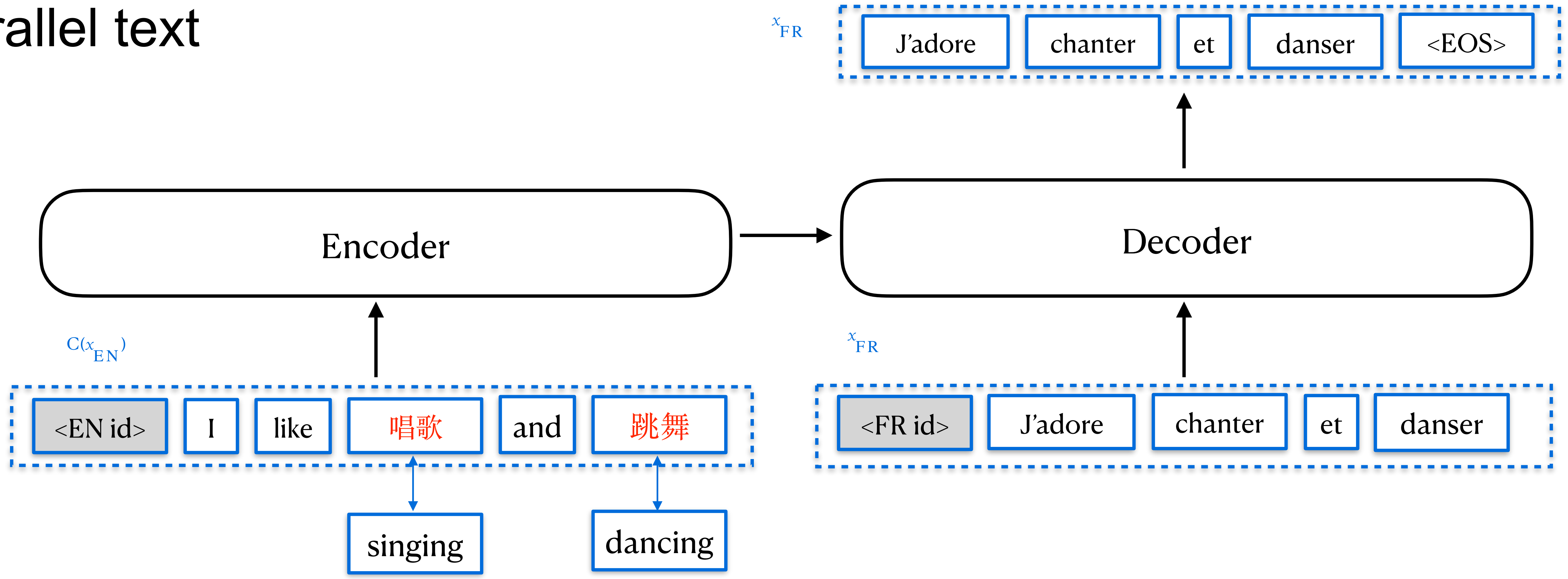


Comparable / better performance on high-resource directions

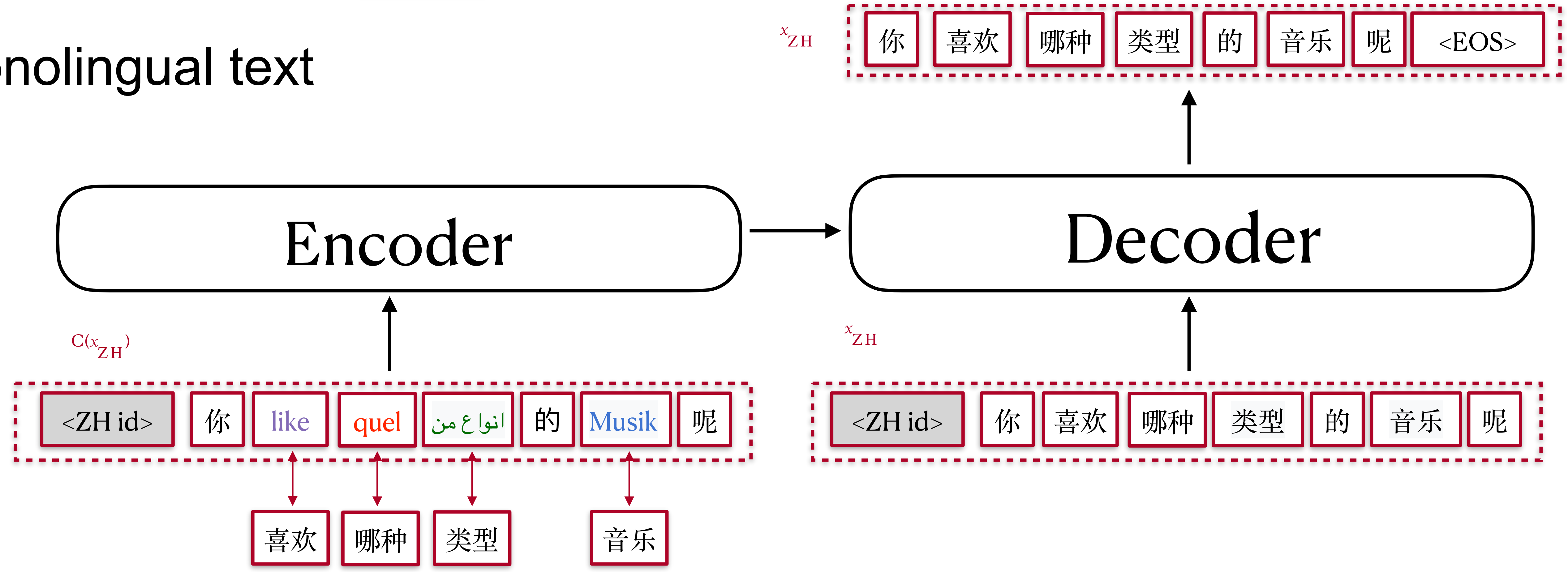
Contrastive Learning for Many-to-many Multilingual Neural Machine Translation [Pan et al., ACL 2021]

mRASP2 introduces monolingual data

- Parallel text

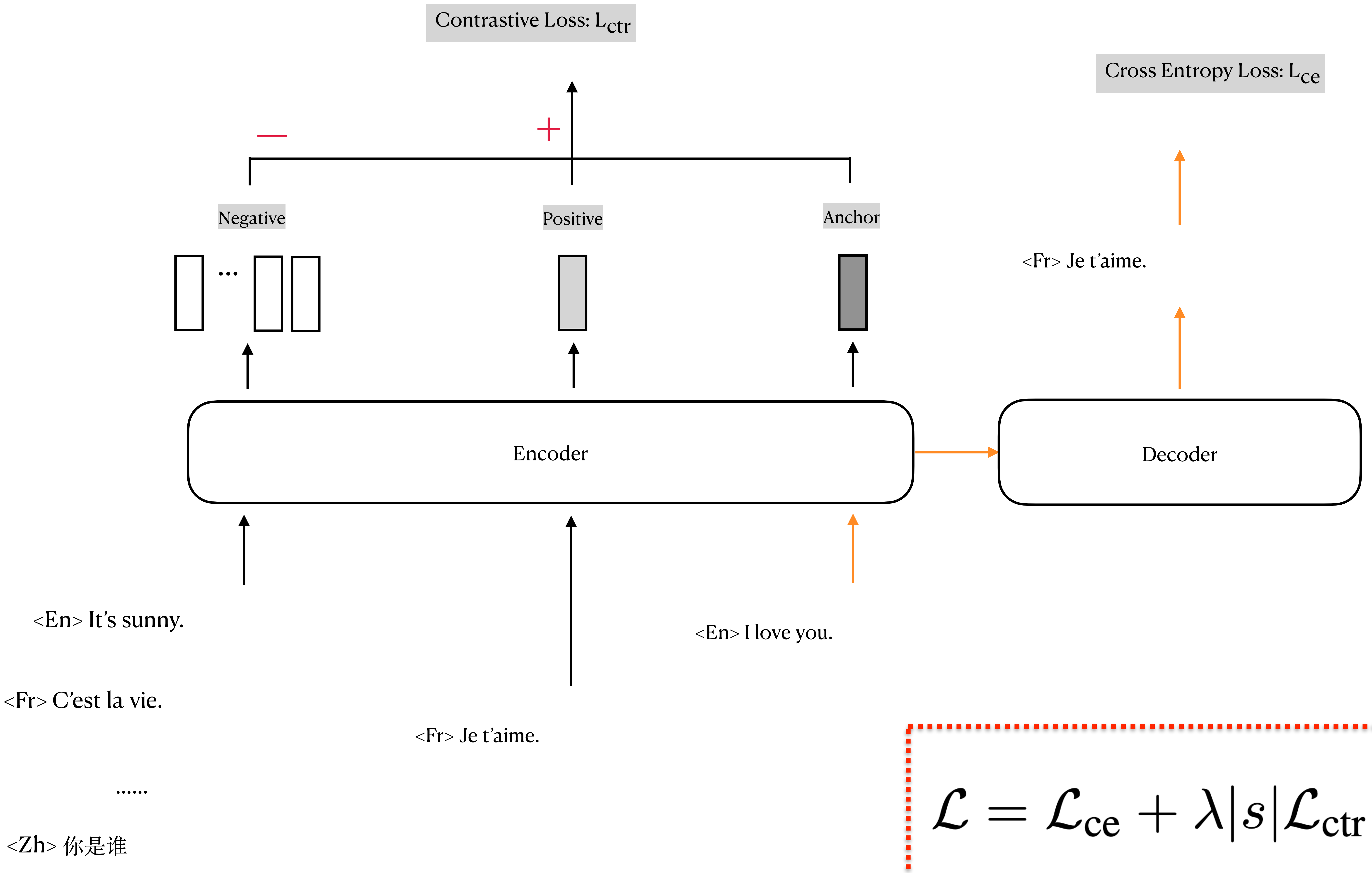


- Monolingual text



Contrastive Learning for Many-to-many Multilingual Neural Machine Translation [Pan et al., ACL 2021]

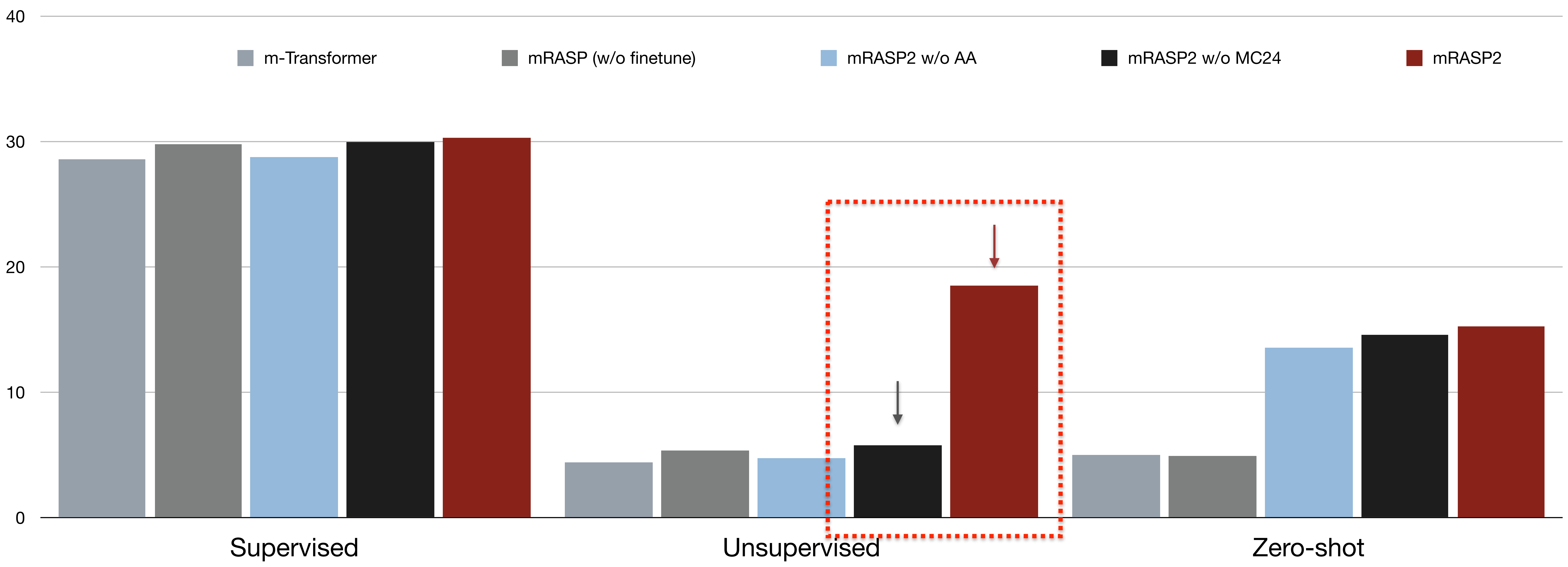
mRASP2 maps different languages in a same space



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

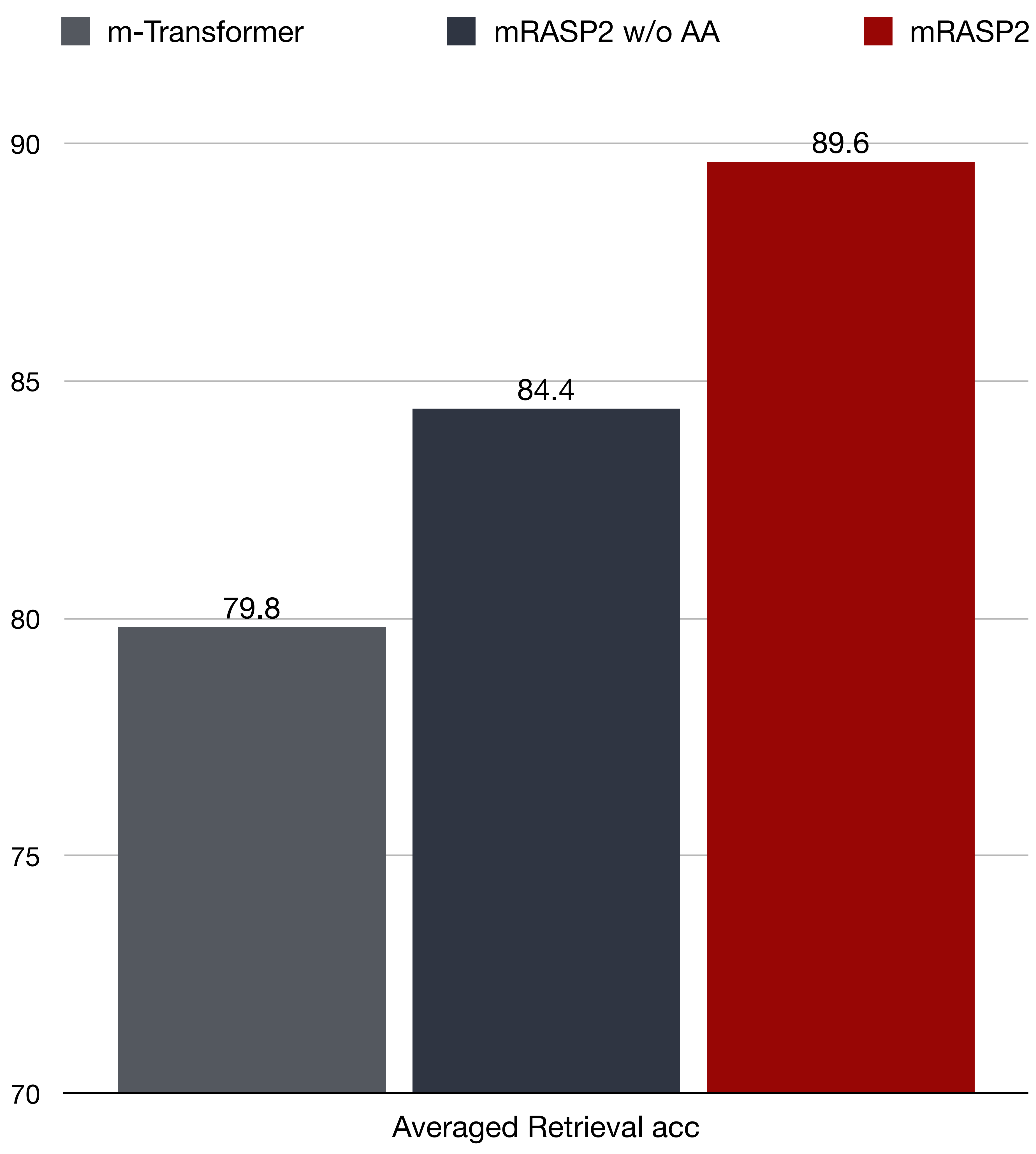
Contrastive Learning for Many-to-many Multilingual Neural Machine Translation [Pan et al., ACL 2021]

Experiments



Monolingual Corpus mainly contributes to unsupervised translation

Better Semantic Alignment: Sentence Retrieval

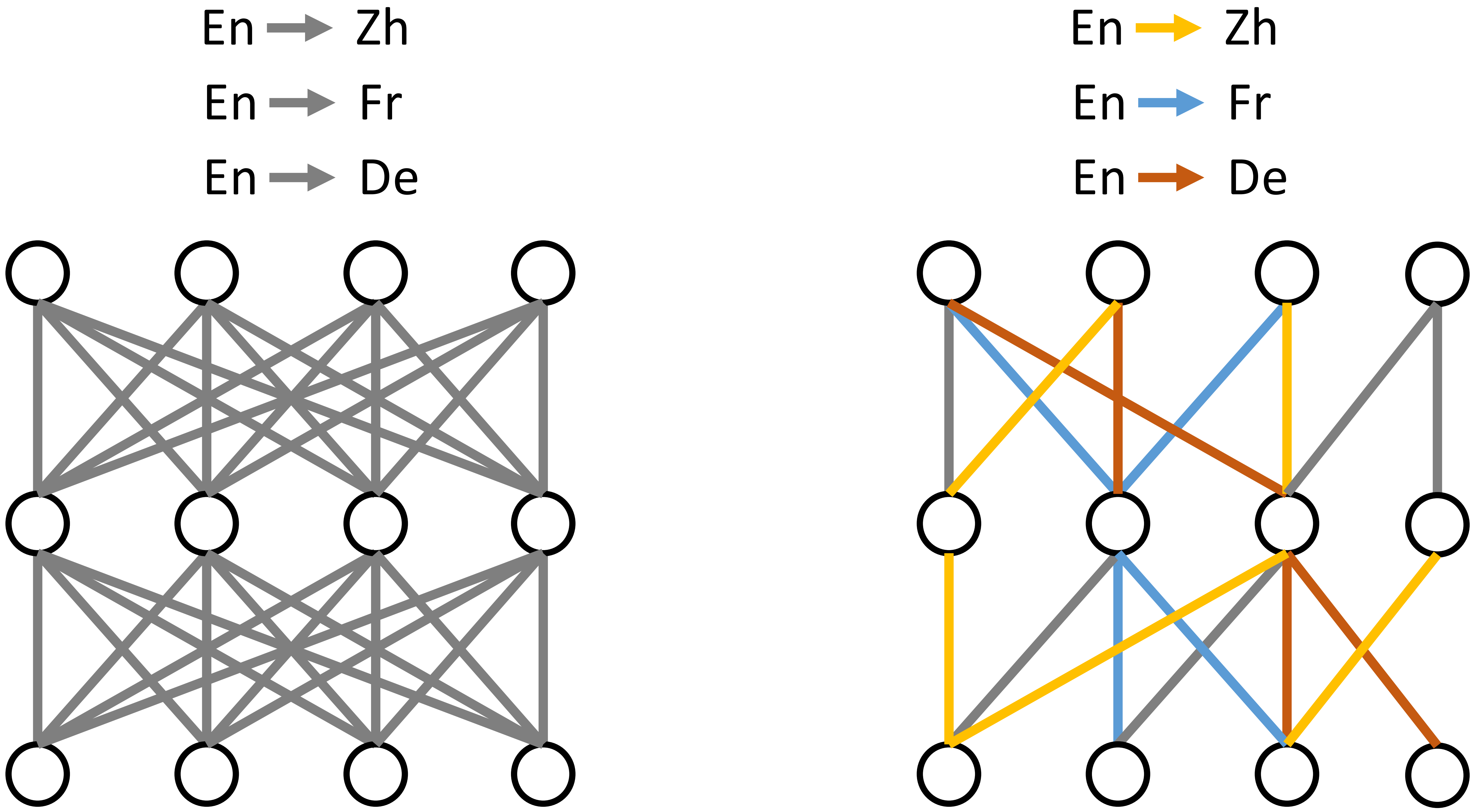


15-way parallel test set(Ted-M): 2284 samples

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

Learning Language Specific Sub-network for Multilingual Machine Translation

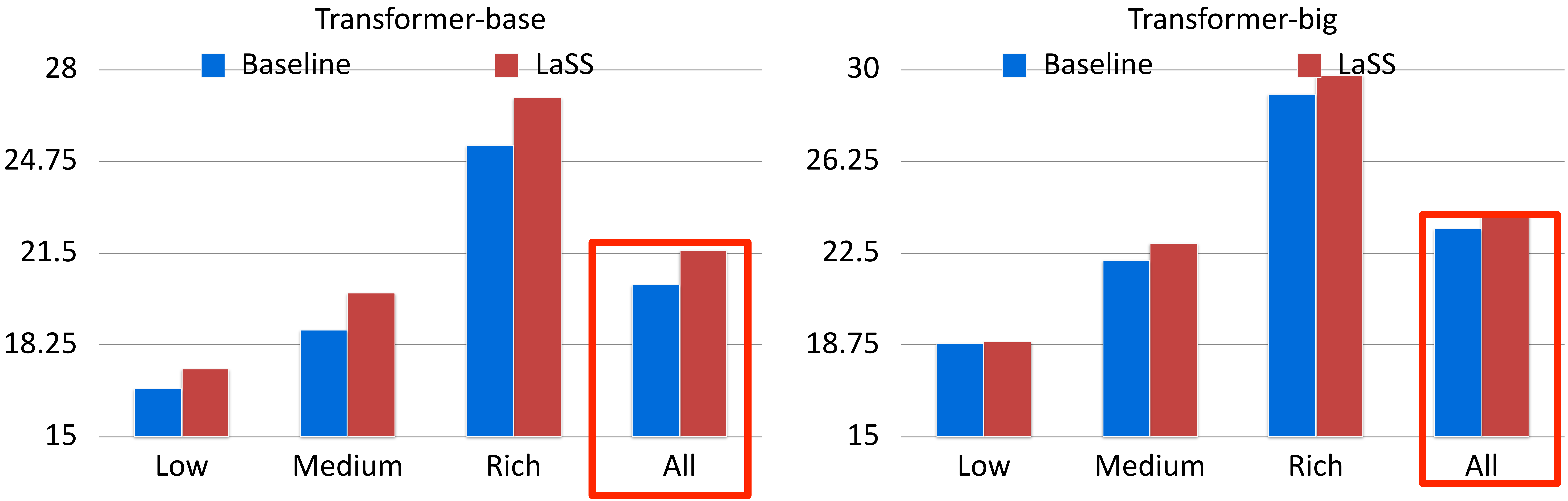
- LaSS accommodates one sub-network for each language pair.
  - Each language pair has shared parameters with some other language pairs and preserves its language-specific parameters
  - For fine-tuning, only updates the corresponding parameters



Learning Language Specific Sub-network for Multilingual Machine Translation [Lin et al., ACL 2021]

Efficacy in alleviating Parameter Interference

– WMT

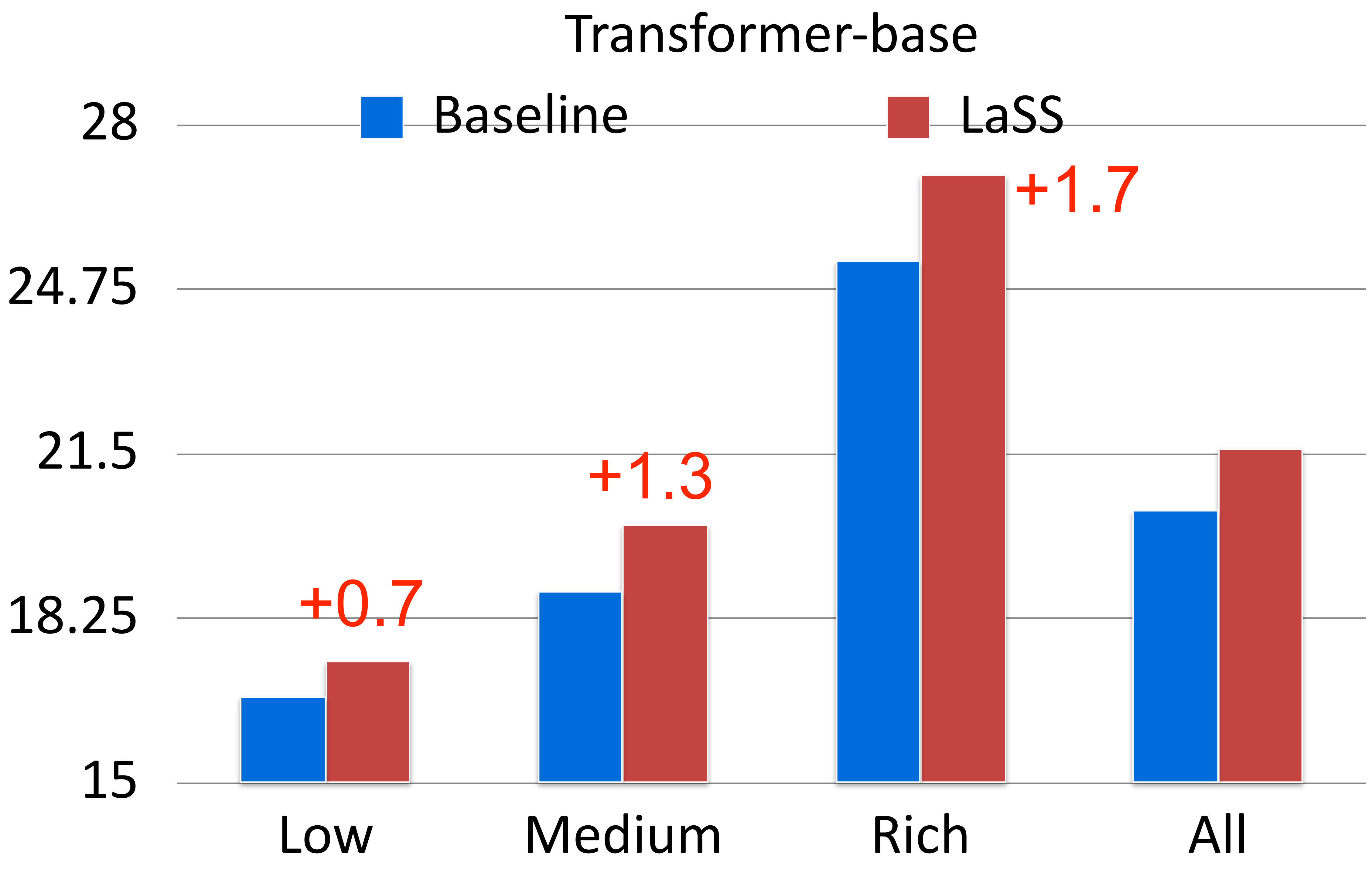


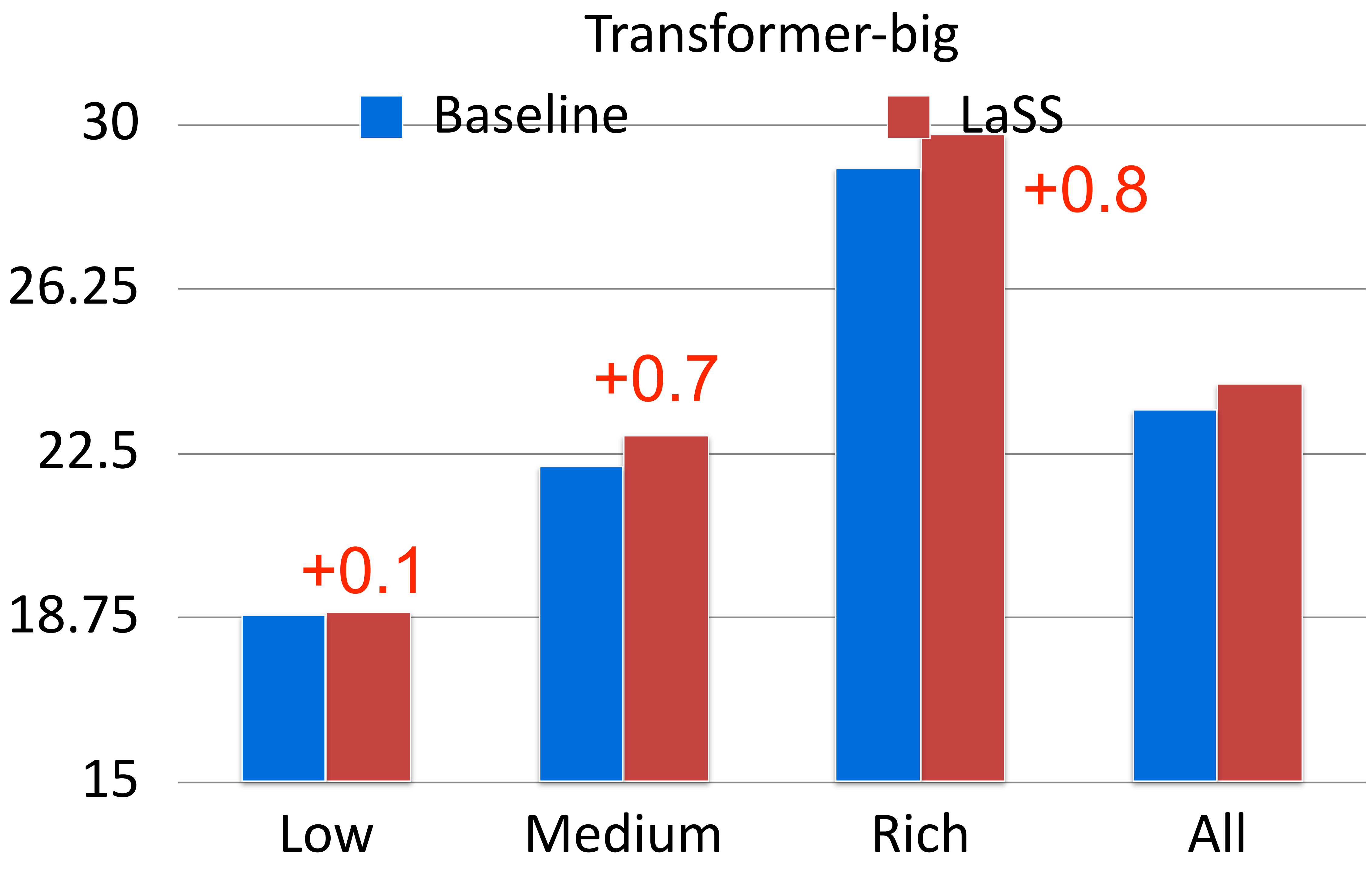
LaSS obtains consistent gains for both Transformer-base and Transformer-big

Learning Language Specific Sub-network for Multilingual Machine Translation [Lin et al., ACL 2021]

LaSS obtains more gains for rich resource

– WMT

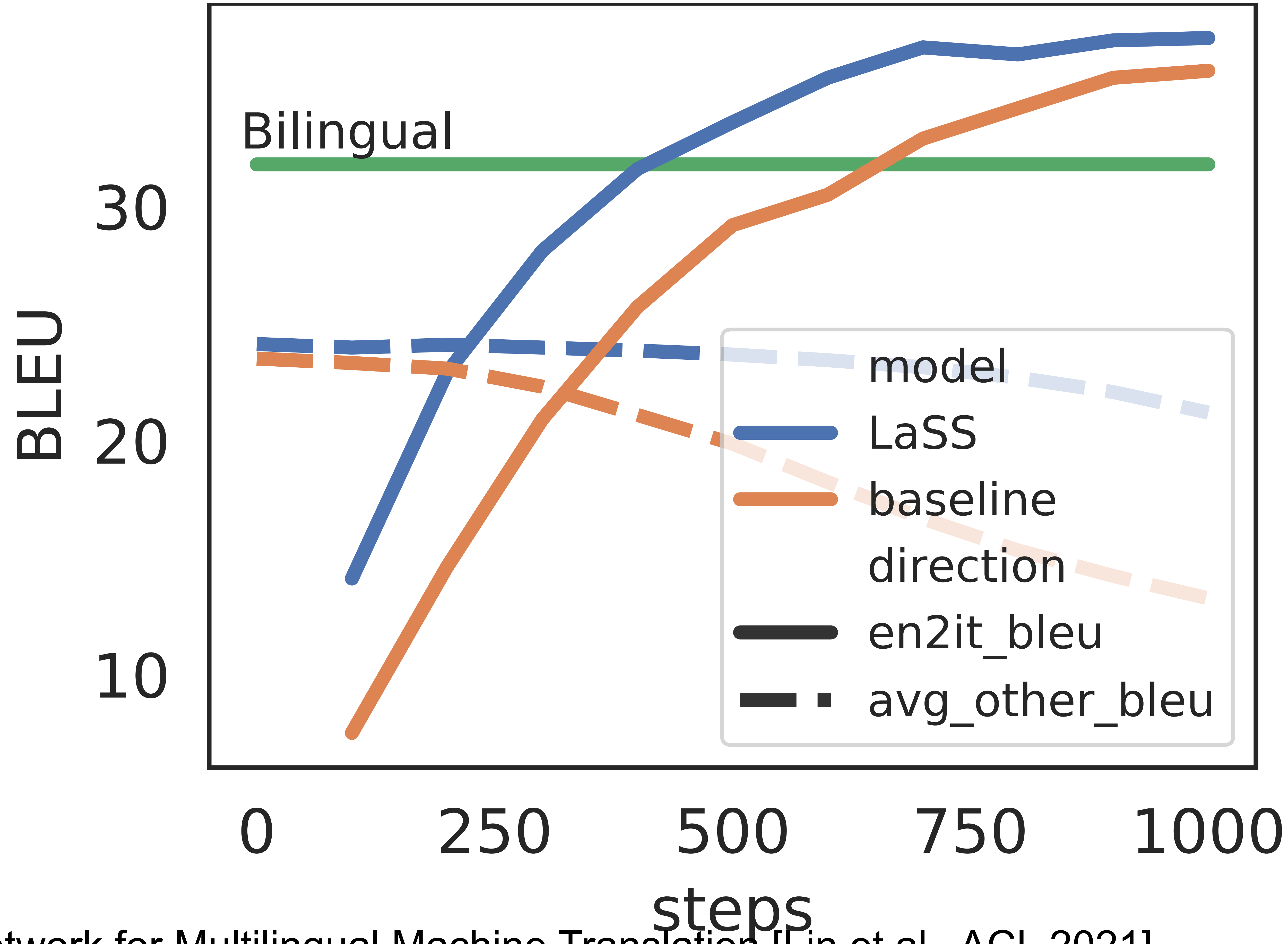




With the dataset scale increasing, the improvement becomes larger, since rich resource language pairs suffer more from parameter interference

Adaptation to New Language Pairs

- Distribute a new sub-network for new language pair and train the sub-network for fixed steps

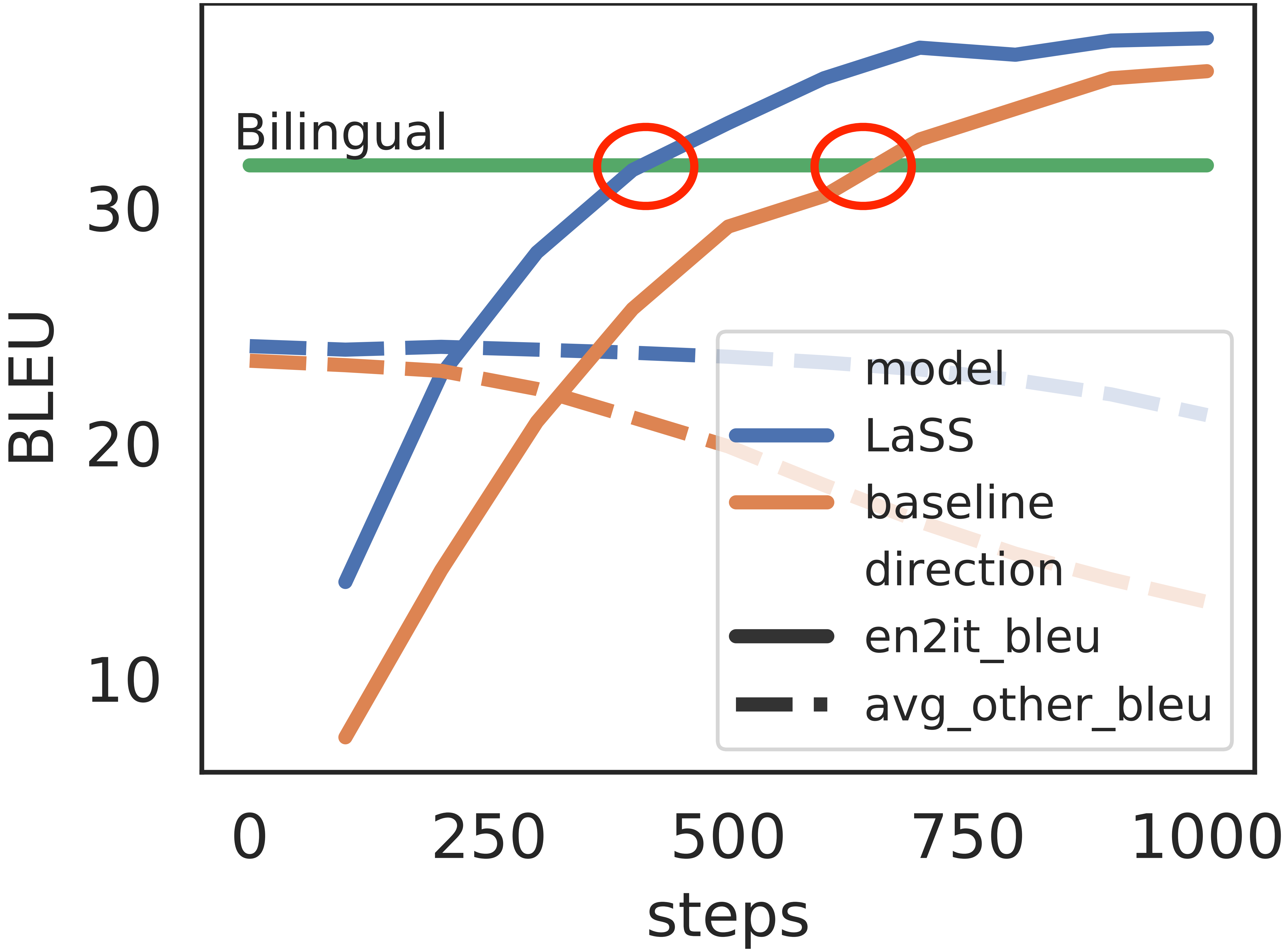


Learning Language Specific Sub-network for Multilingual Machine Translation [Lin et al., ACL 2021]

Adaptation to New Language Pairs

- Distribute a new sub-network for new language pair and train the sub-network for fixed steps

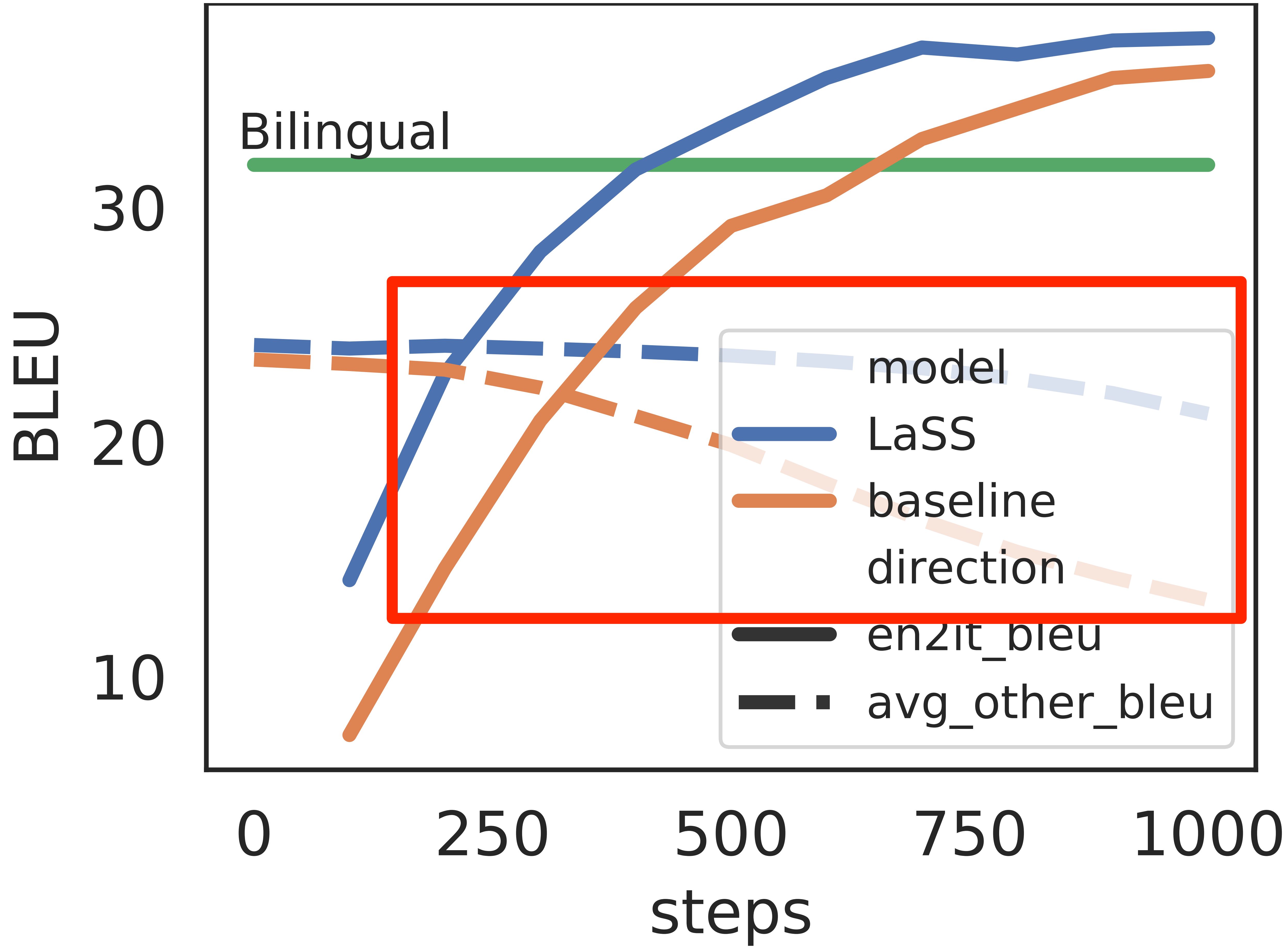
LaSS reaches the bilingual model performance with fewer steps.



Adaptation to New Language Pairs

- Distribute a new sub-network for new language pair and train the sub-network for fixed steps

LaSS hardly drops on existing language pairs

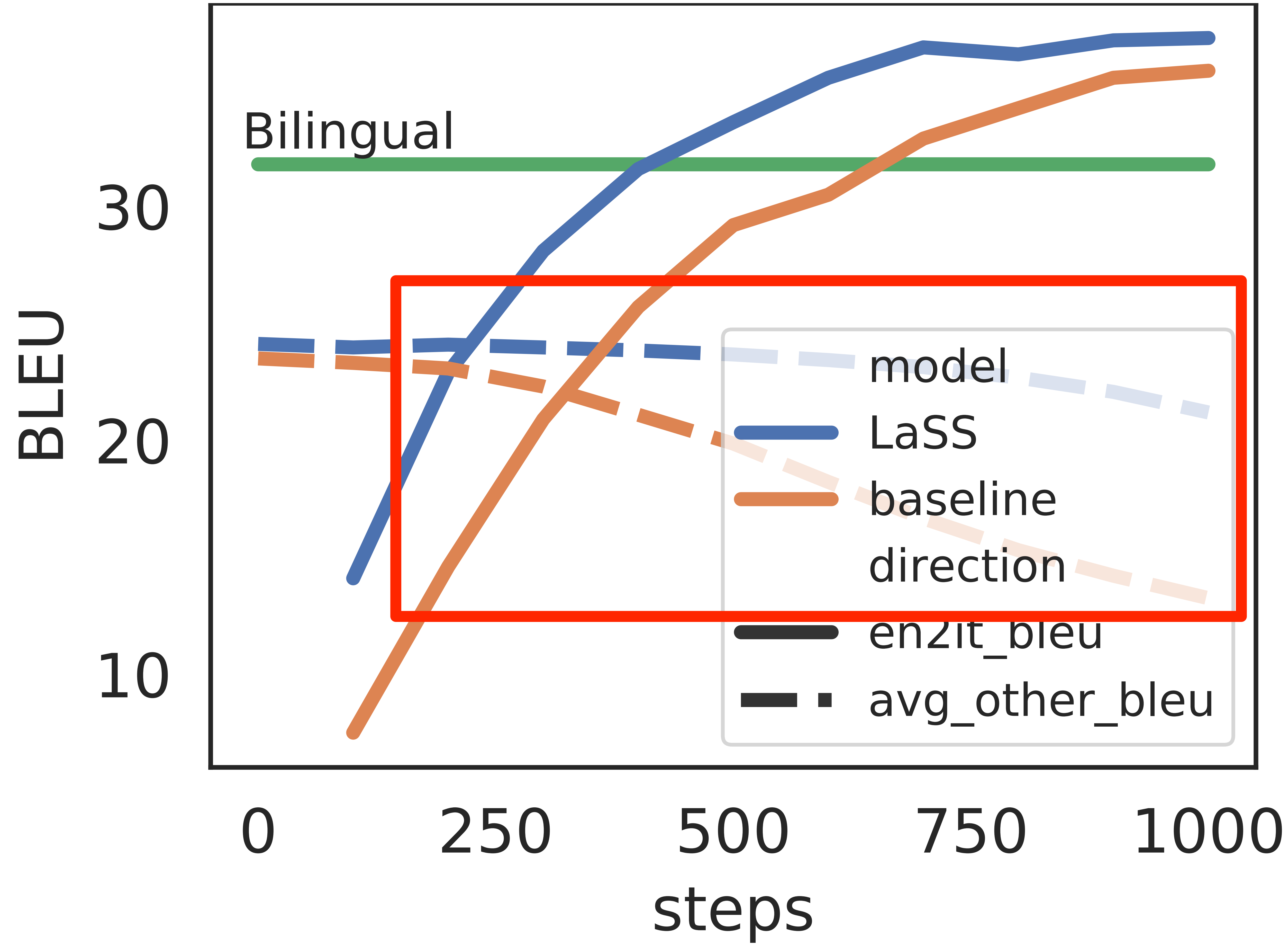


Adaptation to New Language Pairs

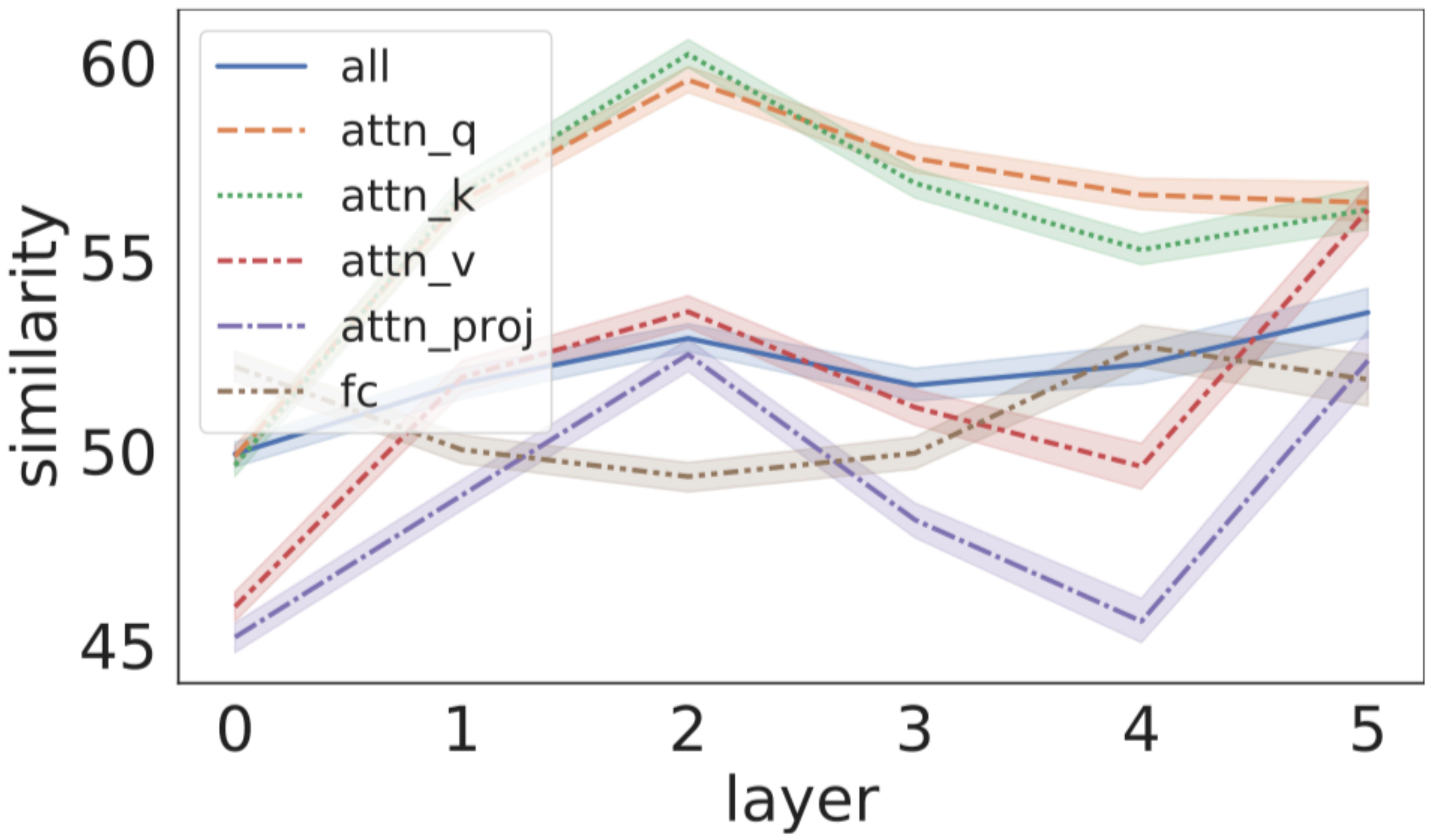
- Distribute a new sub-network for new language pair and train the sub-network for fixed steps

easy adaptation is attributed to the language specific sub-network

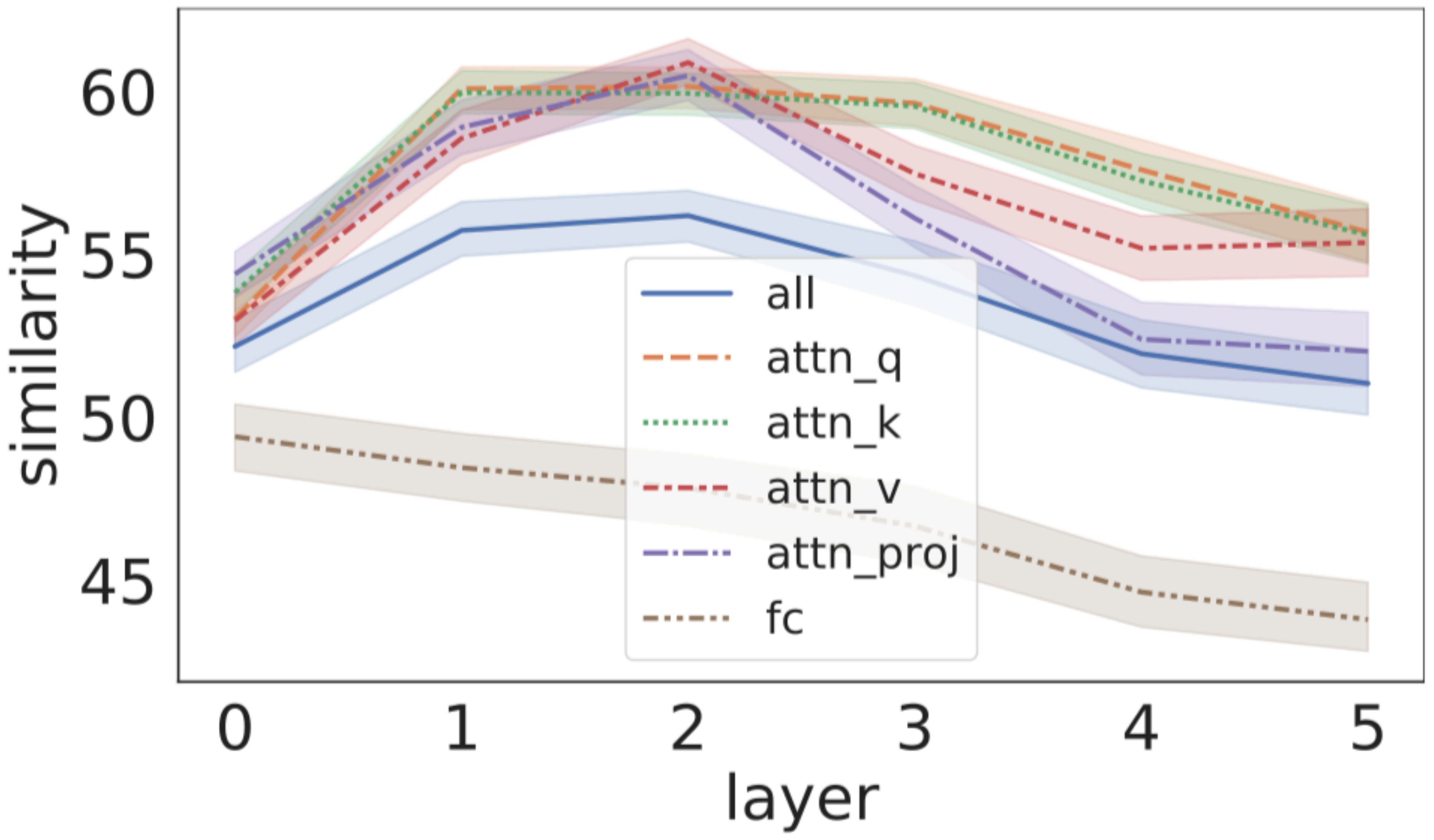
Only updates the corresponding parameters avoids catastrophic forgetting



Top/bottom layers prefer language specific capacity



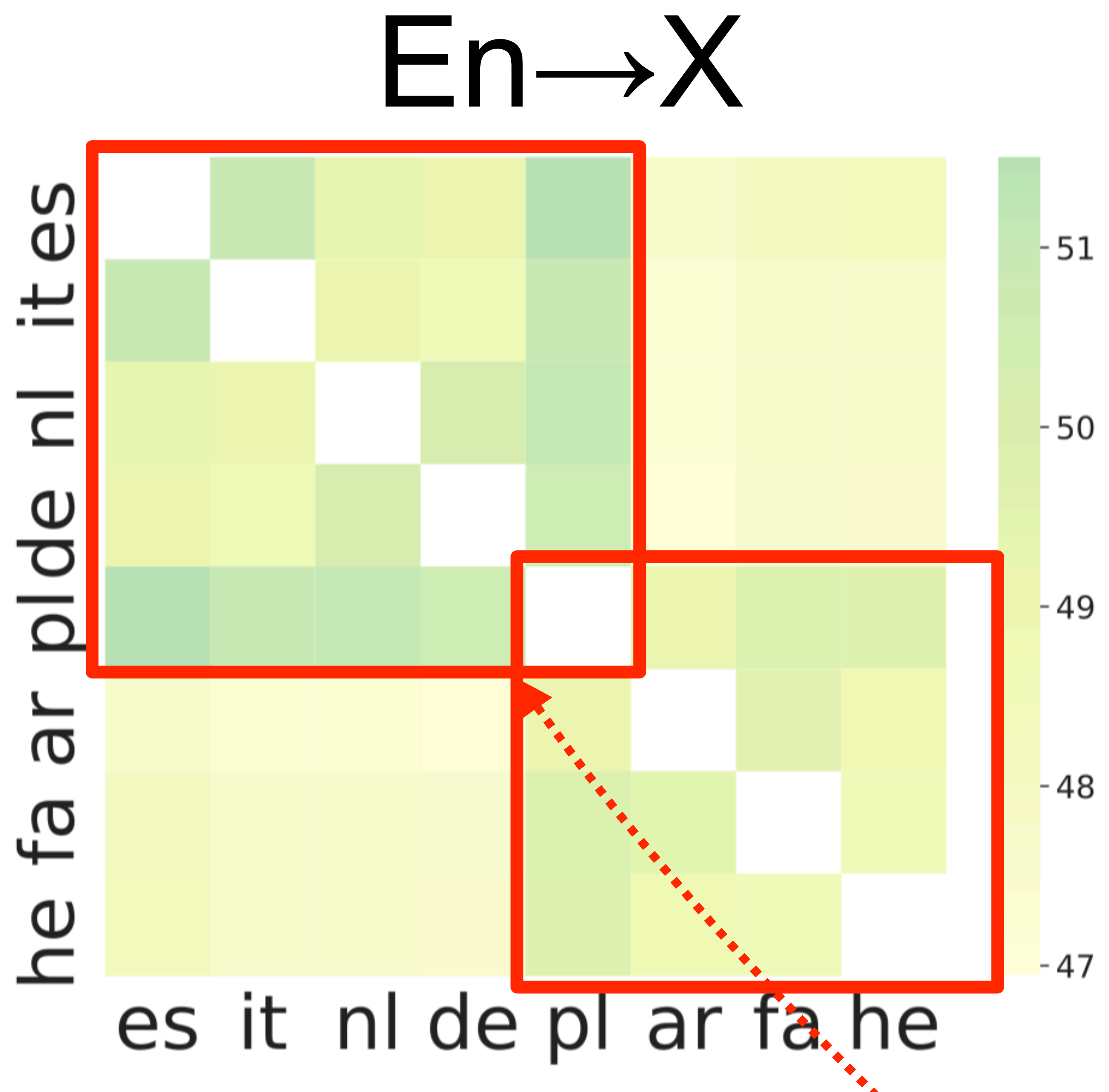
(a) Encoder

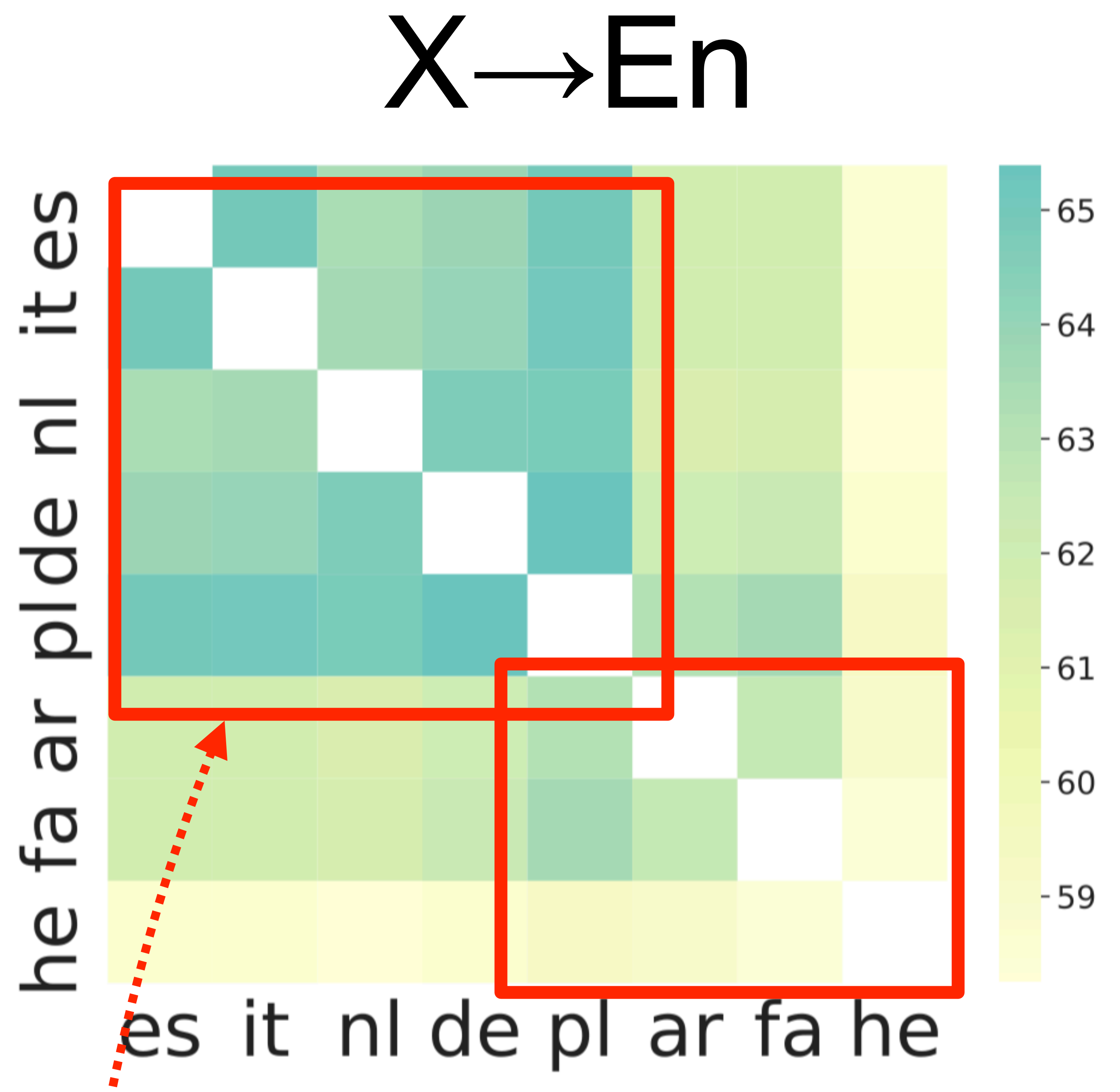


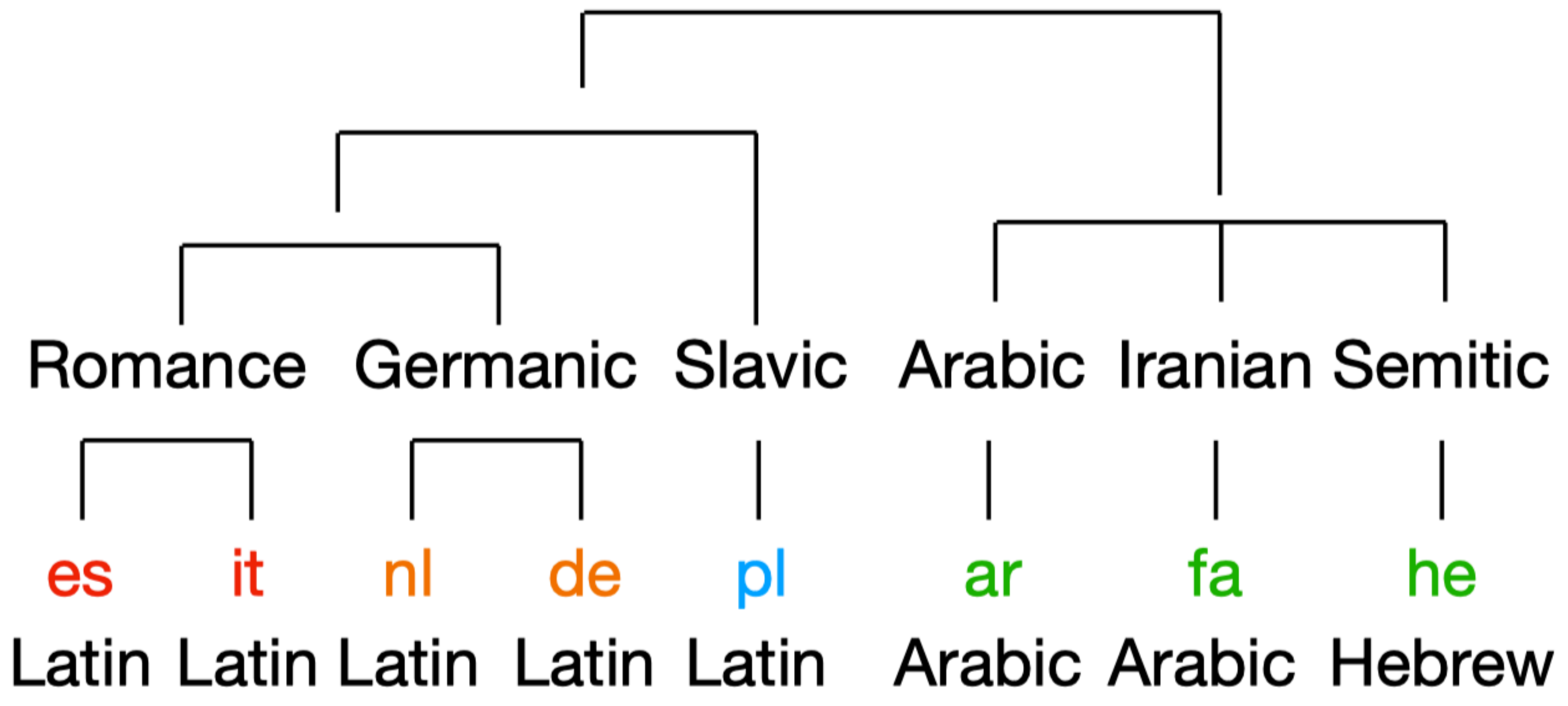
(b) Decoder

The top deals with output projection layer and the bottom is related to embedding layer, which are both language-specific.

Mask similarity is positively correlated to language family







Similar languages tends to group together for both En→X and X→En

Summary for Multilingual Pre-training

- Multilingual fused pre-training
  - Training encoder on masked sequences composed of multiple language, concatenated or mixed words.
- Multilingual sequence-to-sequence pre-training
  - mBart: Recover original sentence from noised ones in multiple languages.
  - mRASP & mRASP2: augmenting data with randomly substitute of words from bilingual lexicon + monolingual reconstruction + contrastive learning
  - LaSS: use pre-training and fine-tuning to discover language-common sub-nets and language-specific sub-nets for MT

Reading

- Song et al. MASS: Pre-train for Sequence to Sequence Generation, 2019.
- Lewis et al. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, 2020