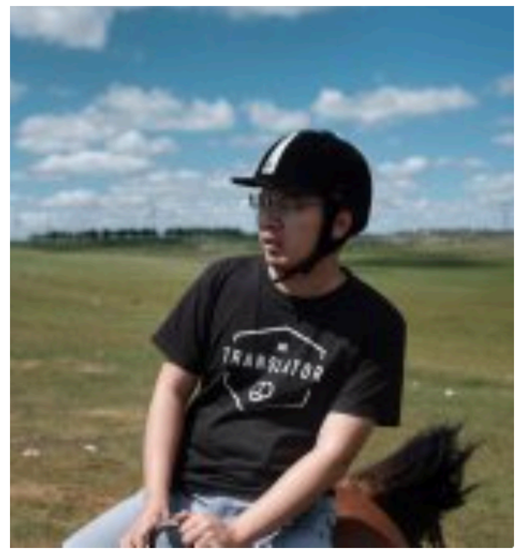
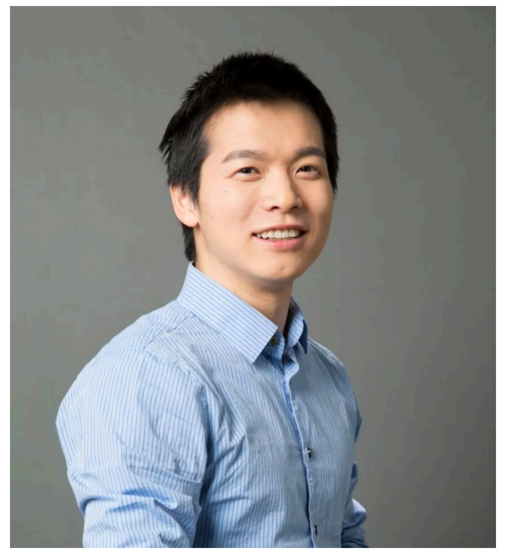
The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)

Pre-training Methods for Neural Machine Translation

Mingxuan Wang
ByteDance AI Lab

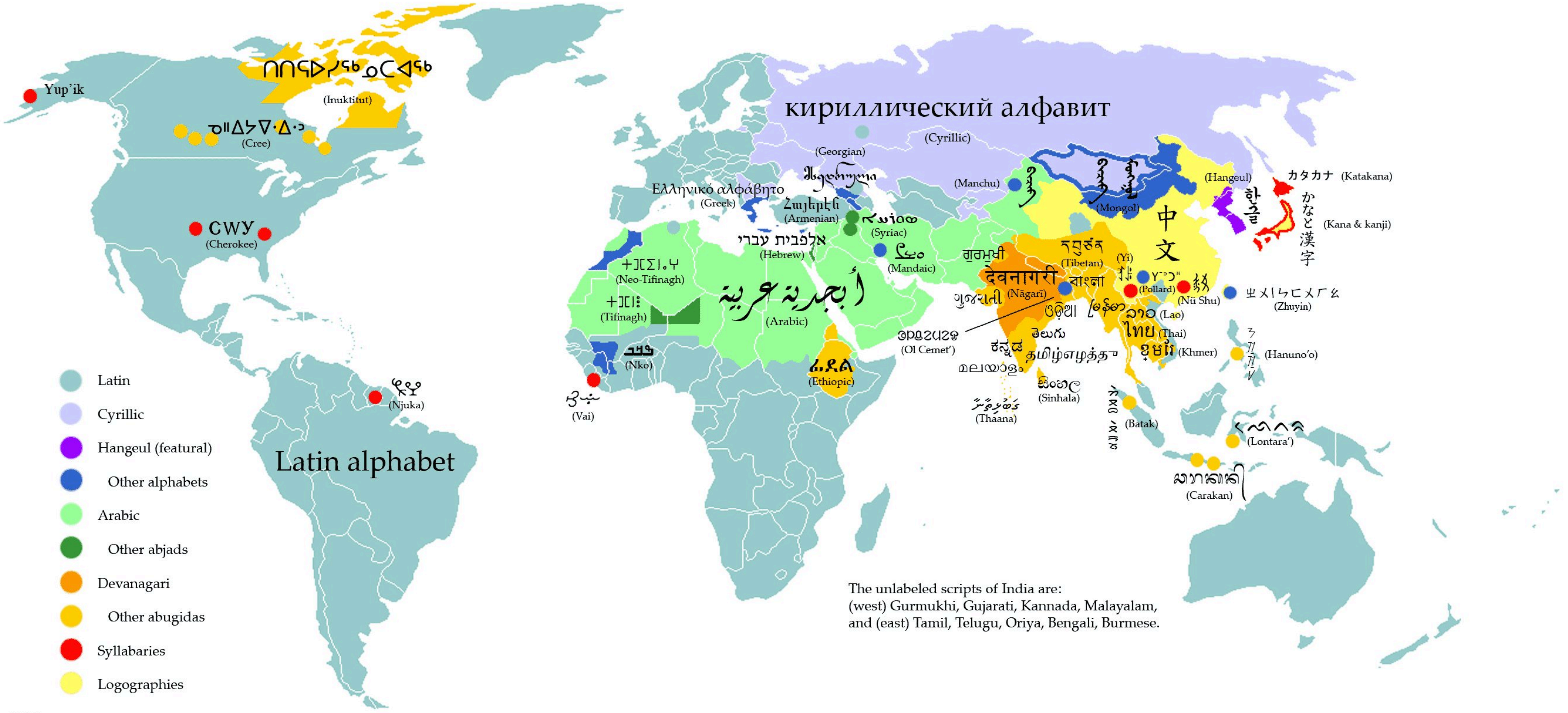


Lei Li
University of California, Santa Barbara

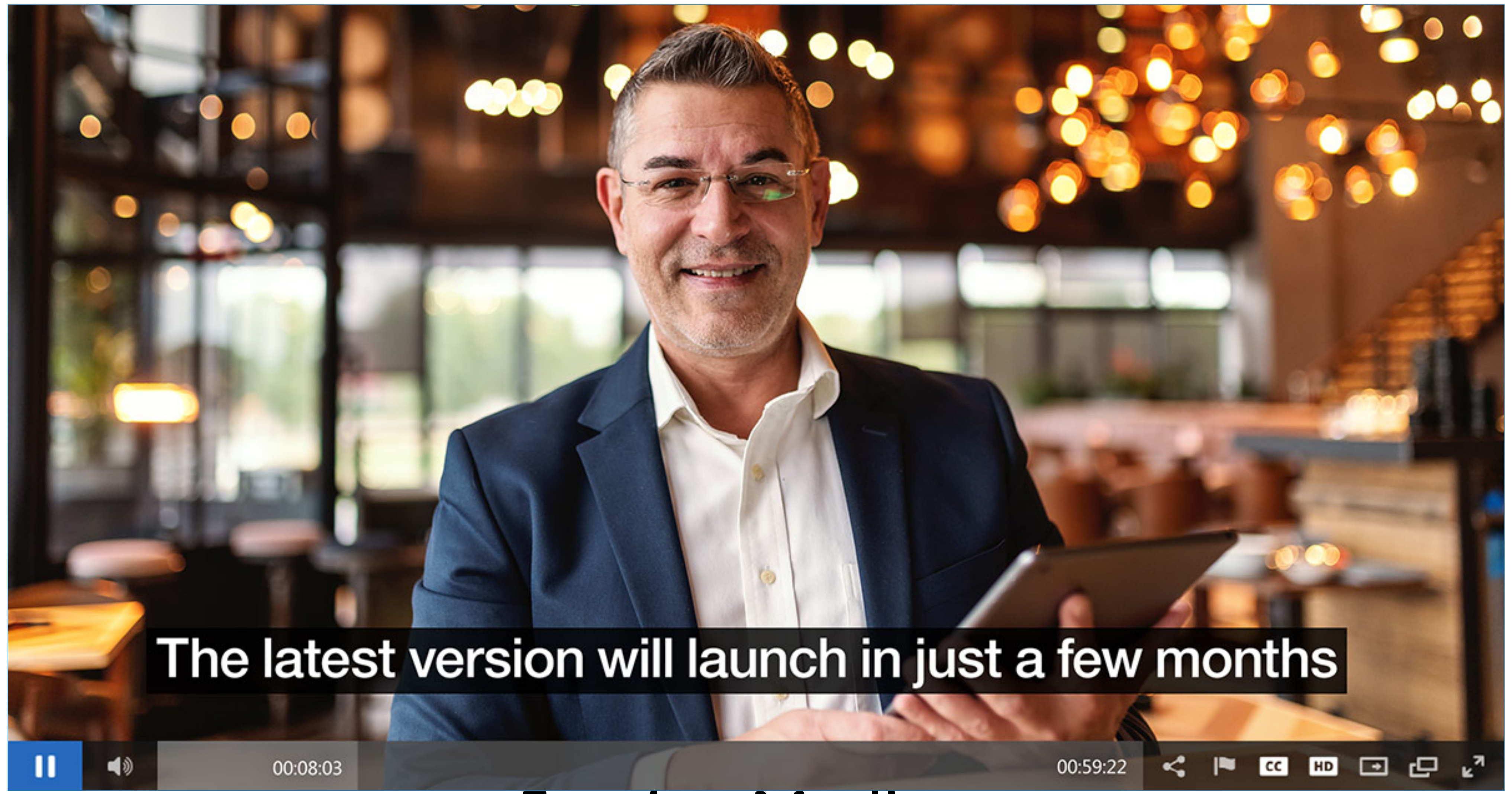


MT helps global information flow

7000 languages in the world



Cross Language Barrier with Machine Translation



Foreign Media

Global Conferences

Tourism

International Trade

Machine Translation has increased international trade by over 10%

informs
http://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Vol. 65, No. 12, December 2019, pp. 5449–5460
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Does Machine Translation Affect International Trade? Evidence from a Large Digital Platform

Erik Brynjolfsson,[a] Xiang Hui,[b] Meng Liu[b]

[a] Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142; [b] Marketing, Olin School of Business, Washington University in St. Louis, St. Louis, Missouri 63130

Contact: erikb@mit.edu, http://orcid.org/0000-0002-8031-6990 (EB); hui@wustl.edu, http://orcid.org/0000-0001-7595-3461 (XH); mengl@wustl.edu, http://orcid.org/0000-0002-5512-7952 (ML)

Received: April 18, 2019
Revised: April 18, 2019
Accepted: April 18, 2019
Published Online in Articles in Advance: September 3, 2019

https://doi.org/10.1287/mnsc.2019.3388

Copyright: © 2019 INFORMS

Abstract. Artificial intelligence (AI) is surpassing human performance in a growing number of domains. However, there is limited evidence of its economic effects. Using data from a digital platform, we study a key application of AI: machine translation. We find that the introduction of a new machine translation system has significantly increased international trade on this platform, increasing exports by 10.9%. Furthermore, heterogeneous treatment effects are consistent with a substantial reduction in translation costs. Our results provide causal evidence that language barriers significantly hinder trade and that AI has already begun to improve economic efficiency in at least one domain.

History: Accepted by Joshua Gans, business strategy.
Supplemental Material: The online appendix is available at https://doi.org/10.1287/mnsc.2019.3388.

Keywords: artificial intelligence • international trade • machine translation • machine learning • digital platforms

Equivalent to make the world smaller than 26%

Outline

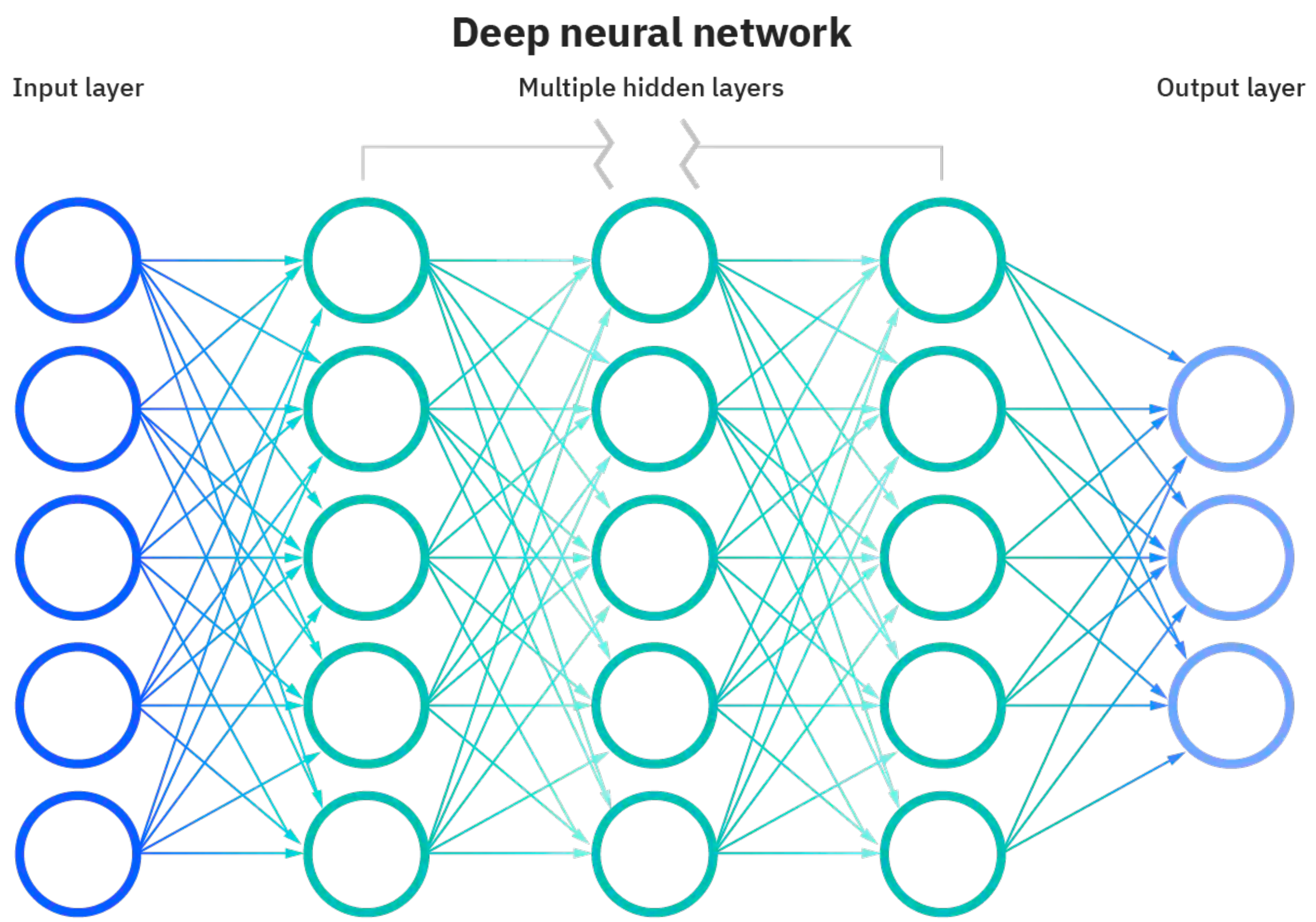
- Basics
  - NMT
  - Pre-training paradigm
- Monolingual Pre-training for NMT
  - Pre-training style
  - Contrast to other data augmentation methods
- Multilingual Pre-training for NMT
- Pre-training for Speech Translation

PART I: Basics

What is Neural Machine Translation

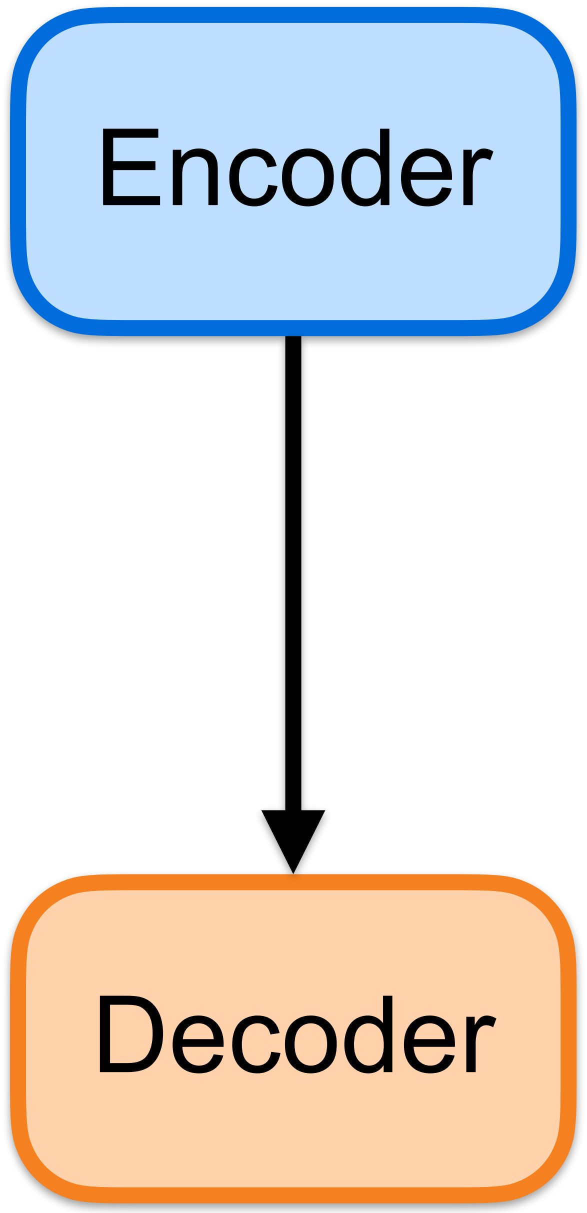
Automatic conversion of text/speech from one natural language to another with a single neural network

French: Quand tu souris, le monde entier s’arrête et se fige un instant.

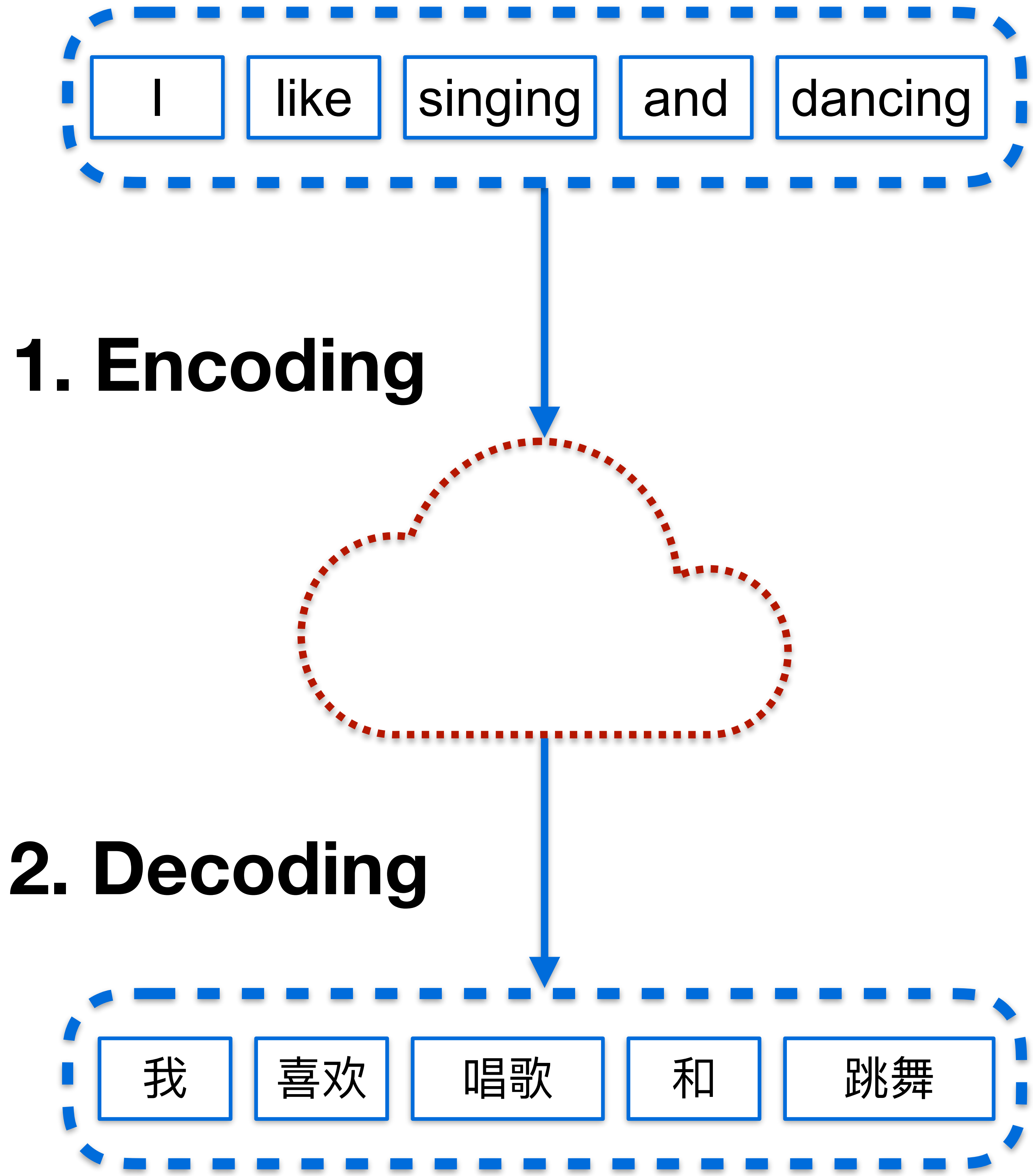


English: When you smile, the whole world stops and freezes for a moment.

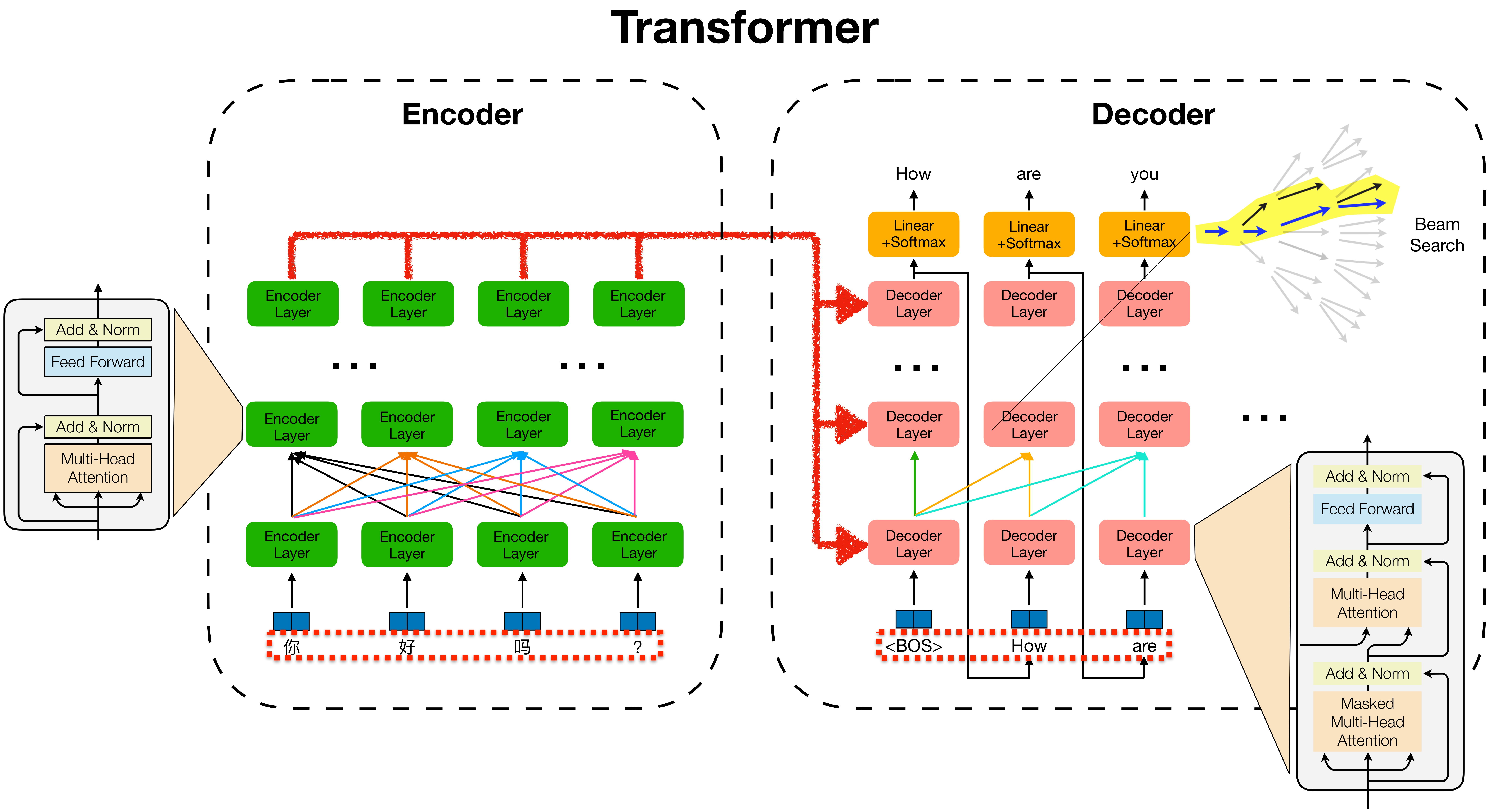
Encoder-Decoder Paradigm



Encoder-Decoder Paradigm

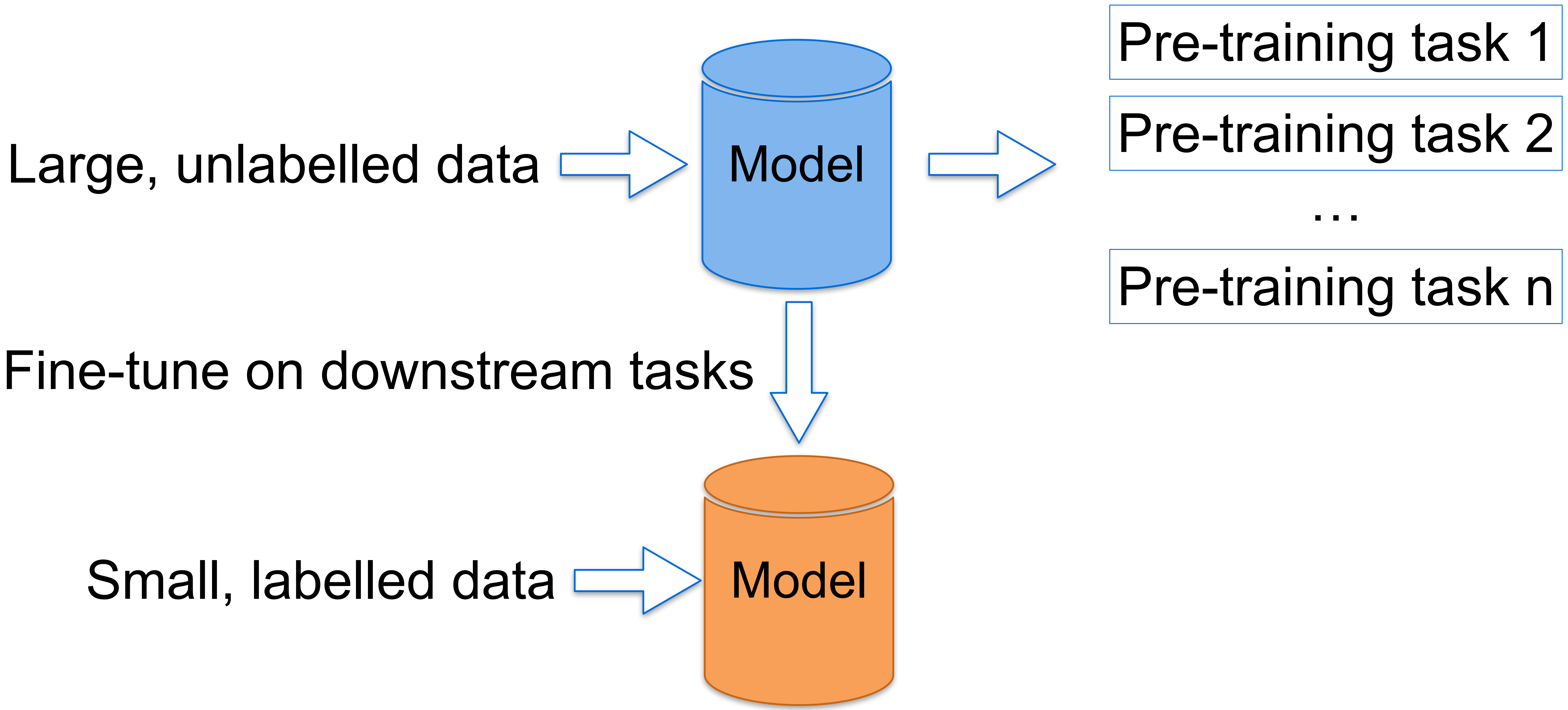


Transformer Architecture

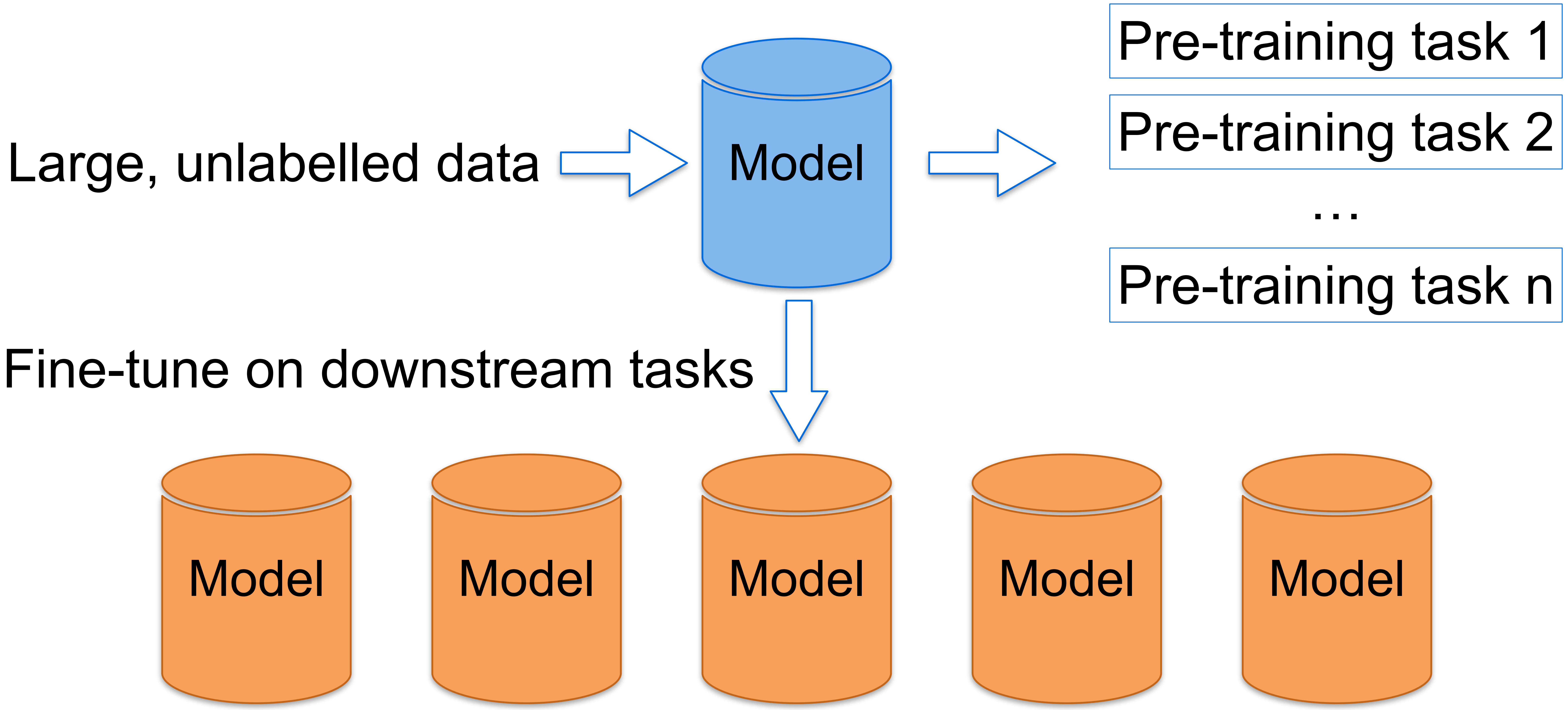


Pre-training & Fine-tuning

Self-supervised learning without labels

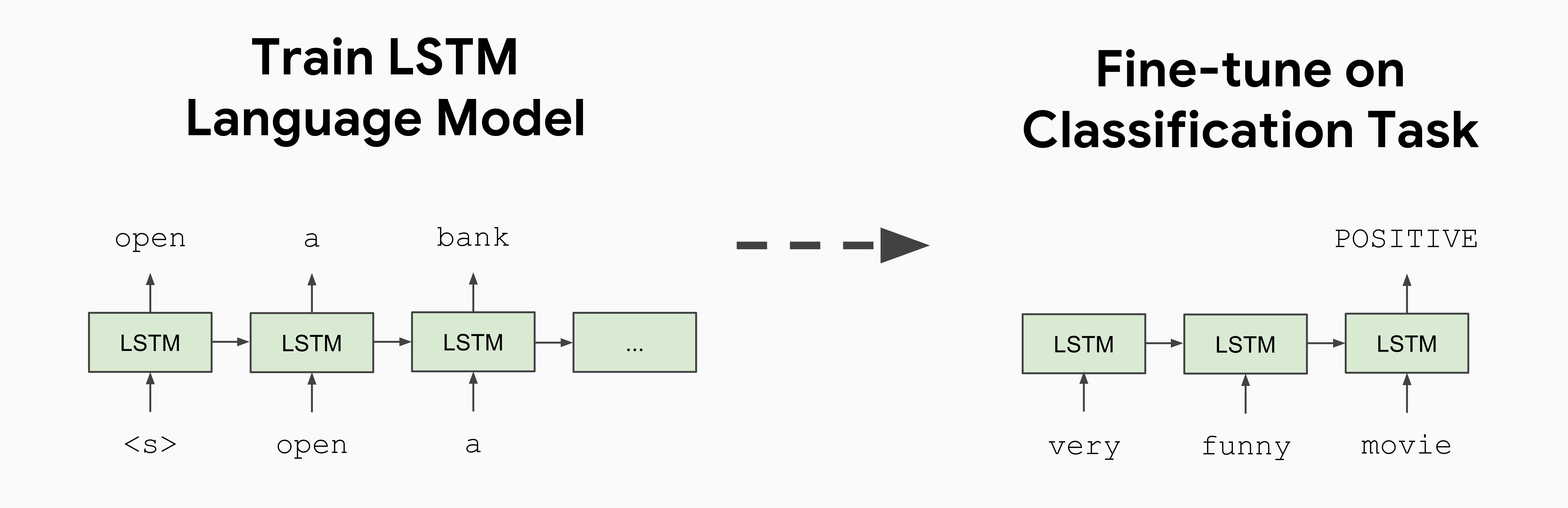


Pre-training & Fine-tuning



Context Representations

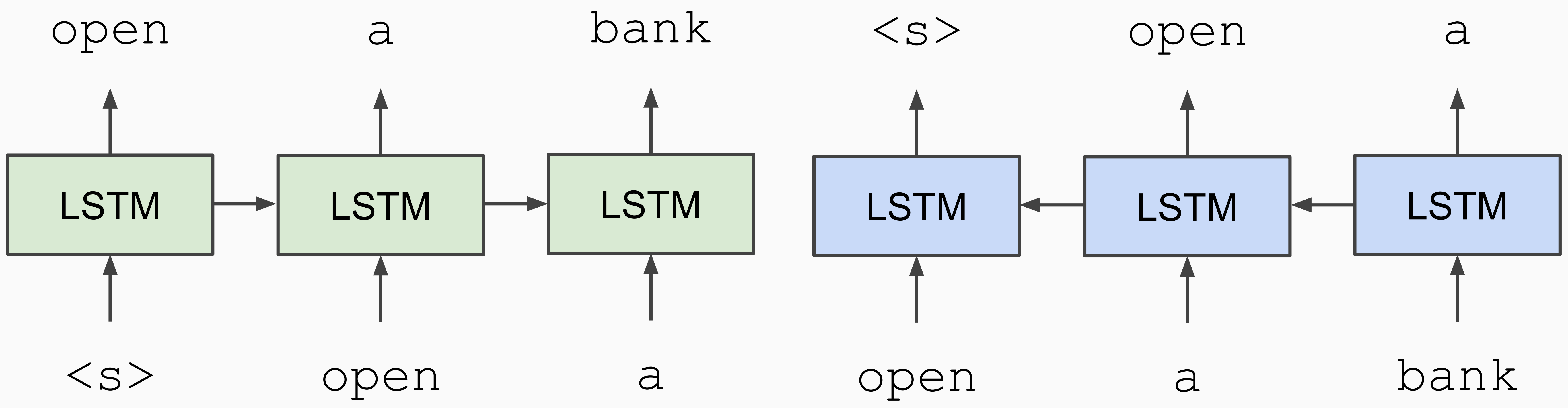
- Semi-supervised sequence learning, Google 2015



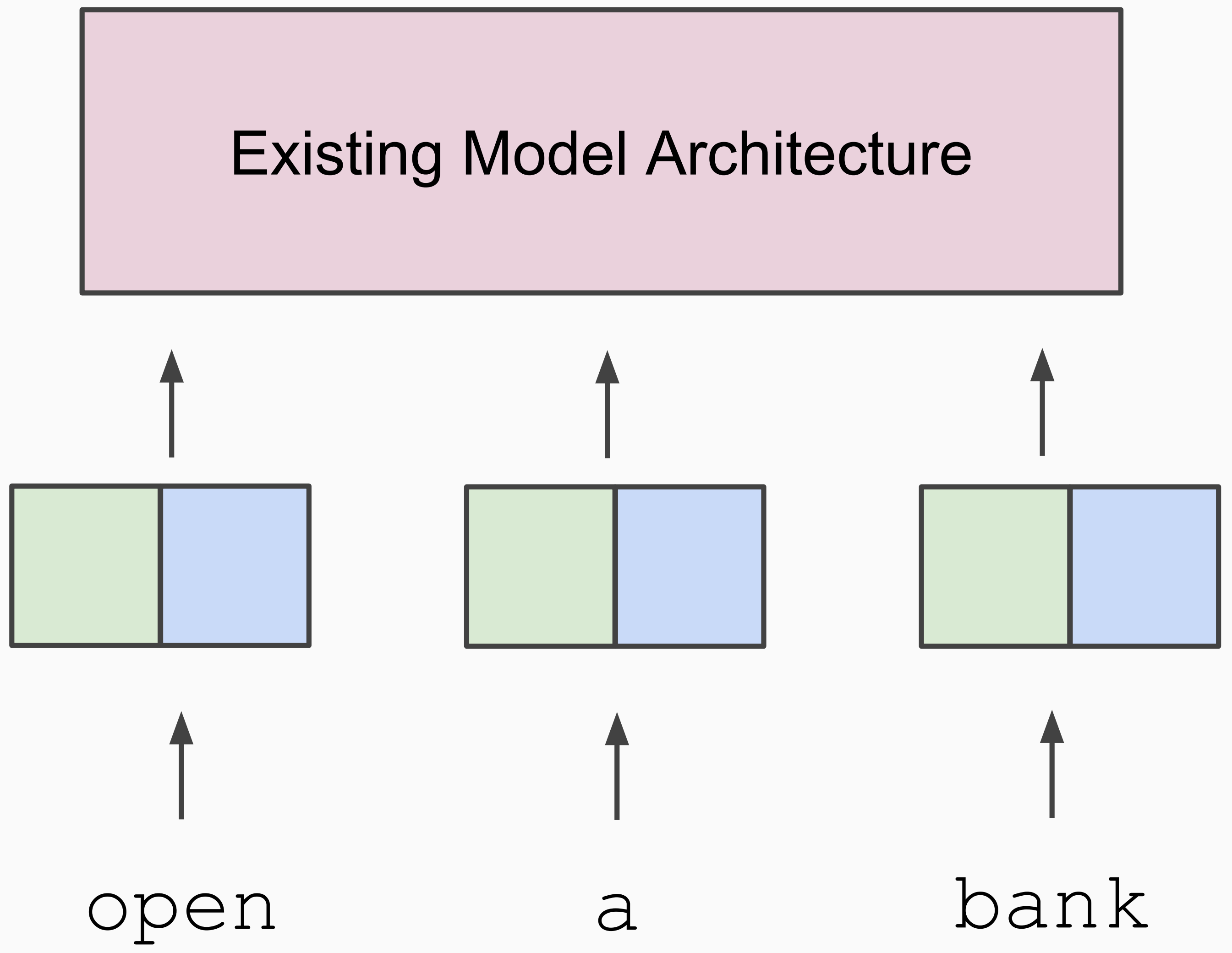
Context Representations

- Elmo: Deep contextual word embeddings

Train Separate Left-to-Right and Right-to-Left LMs

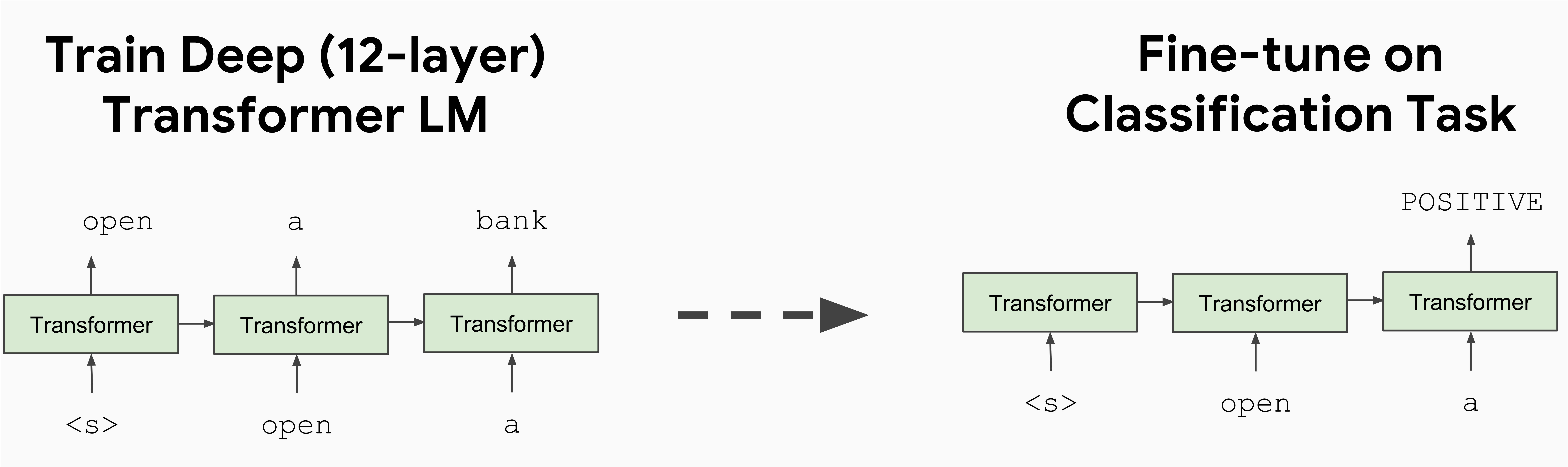


Apply as “Pre-trained Embeddings”



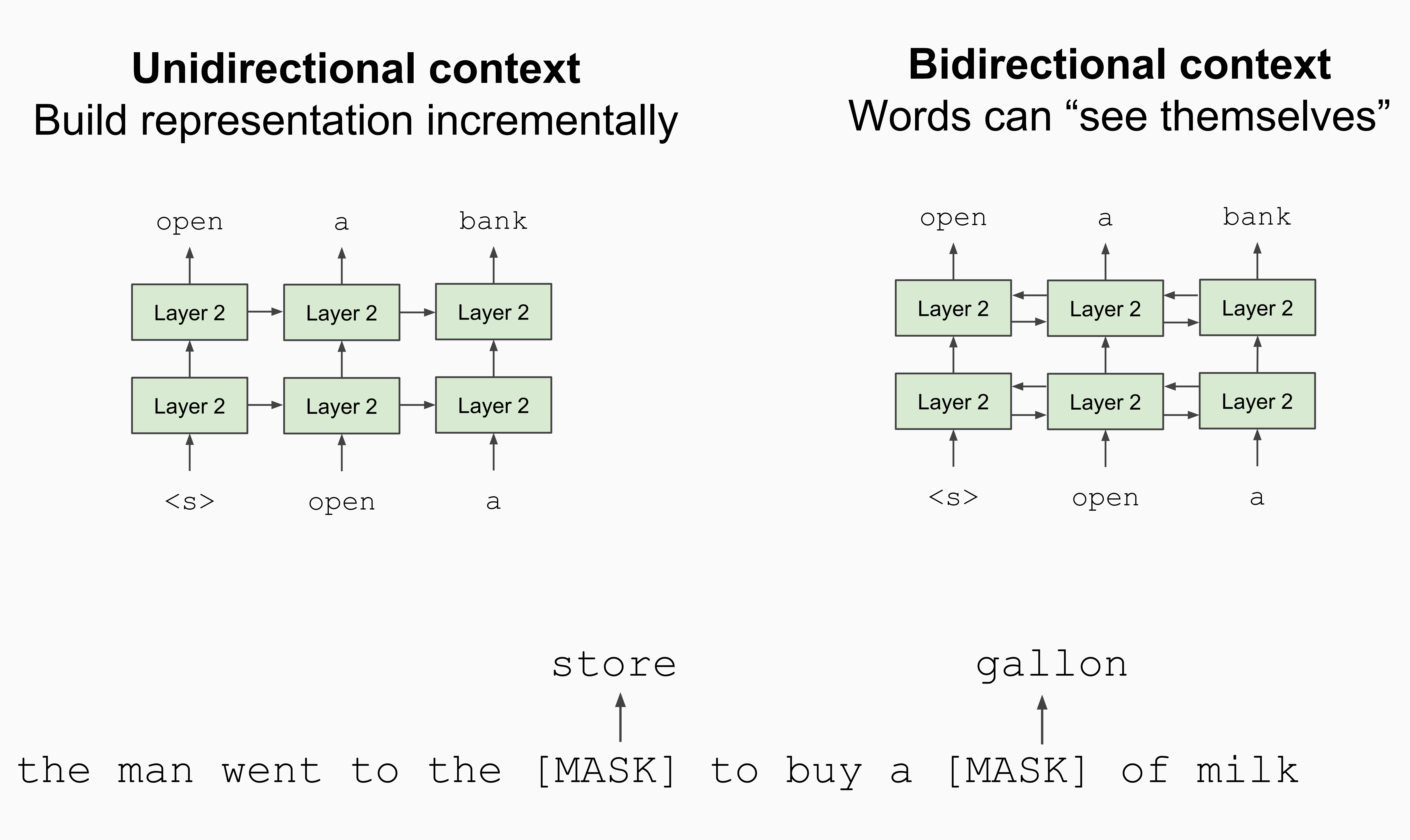
Context Representations

- GPT: improve language understanding by generative pre-training

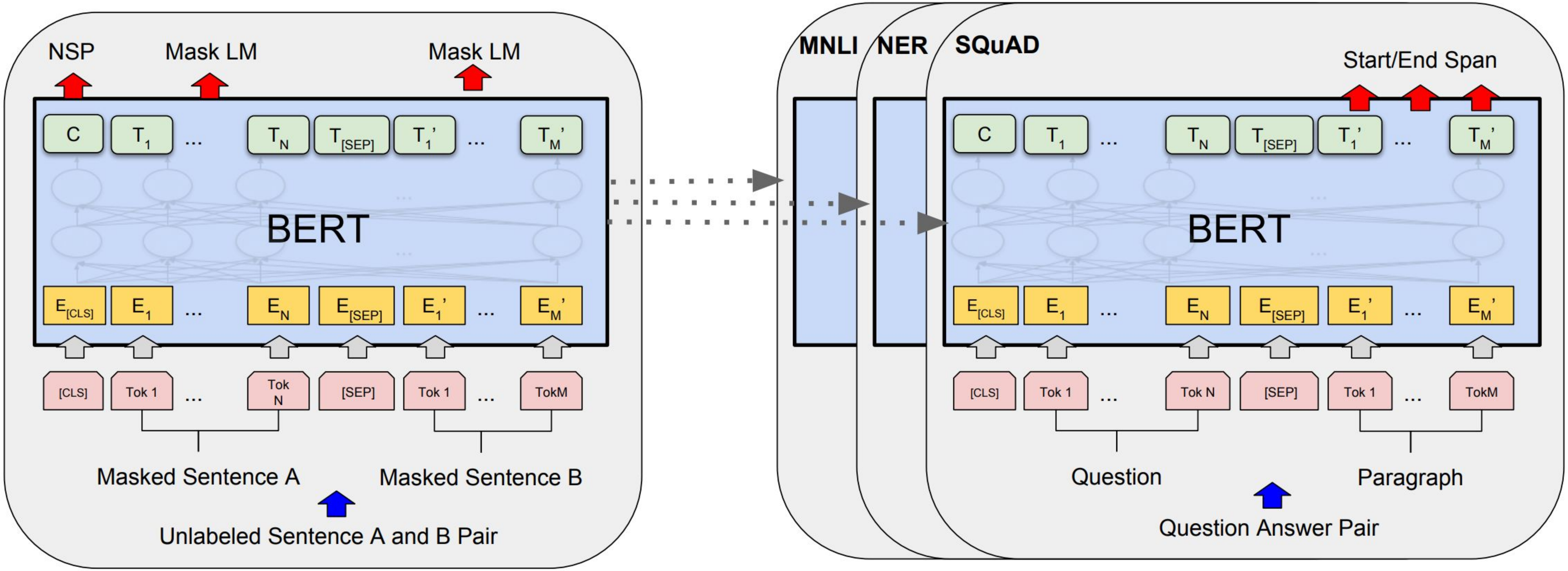


Context Representations

- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
  - Bidirectional
  - Random mask



BERT: Pre-training and Fine-tuning



Pre-training

Fine-Tuning

BERT: Pre-training and Fine-tuning

System	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	**Average**
	392k	363k	108k	67k	8.5k	5.7k	3.5k	2.5k	-
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.9	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	88.1	91.3	45.4	80.0	82.3	56.0	75.2
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.1	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**91.1**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**81.9**

BERT achieves SOTA results on a huge number of NLP benchmarks.

Pre-training & Fine-tuning

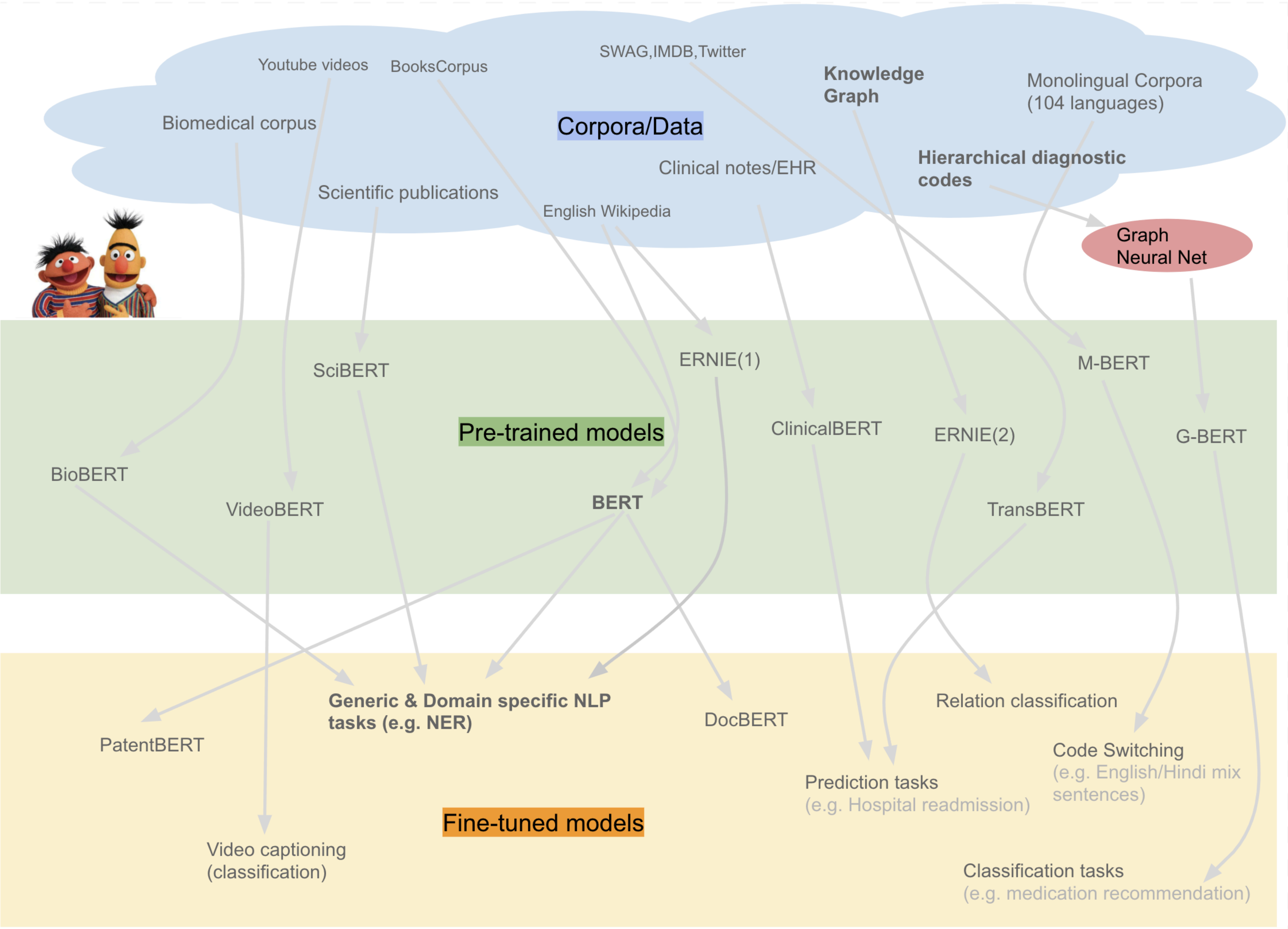


Image Source

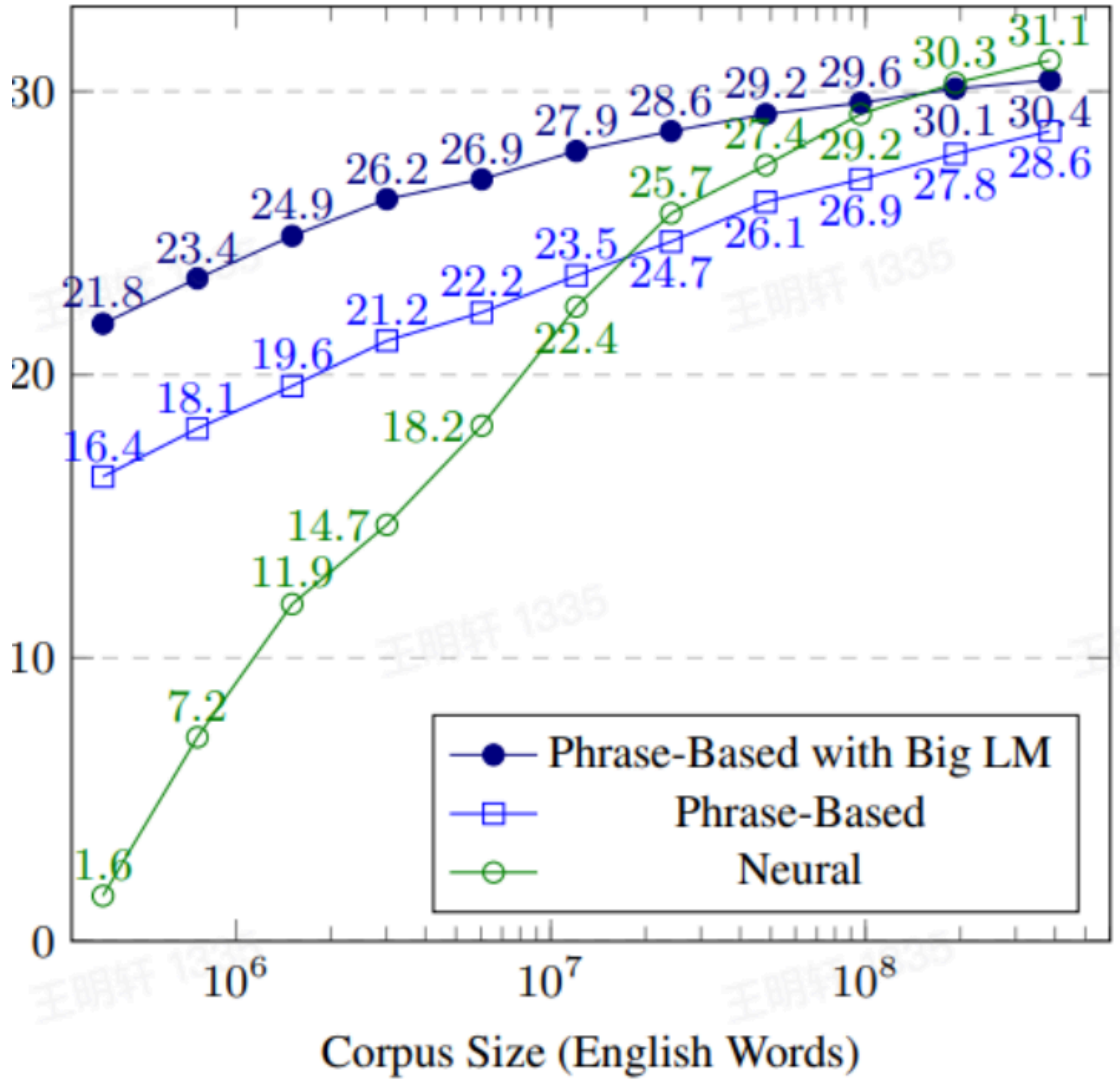
Pre-training & Fine-tuning

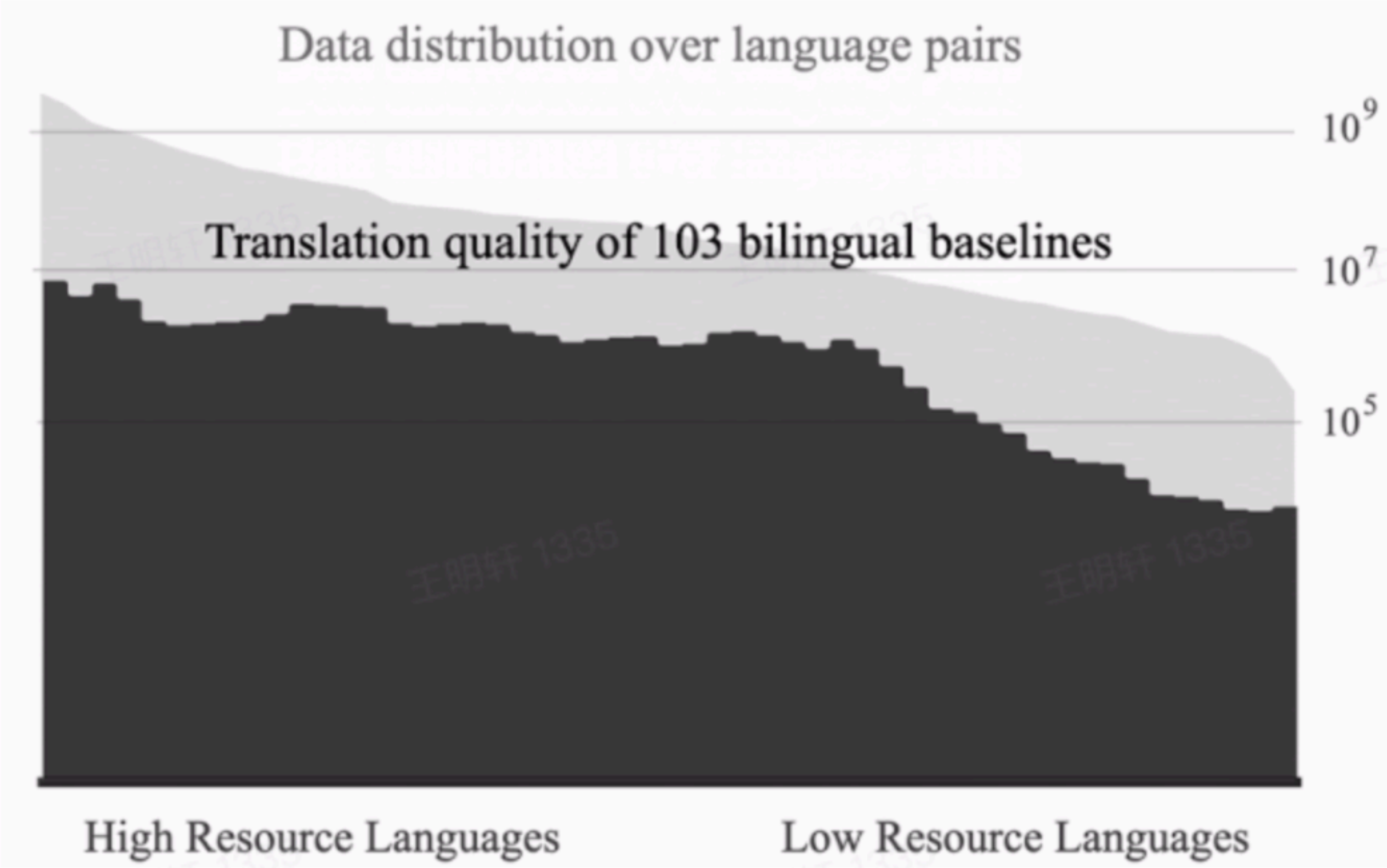
Does pre-training matter in NMT?

PART II: Monolingual Pre-training for NMT

Why Monolingual

MT: More data is better



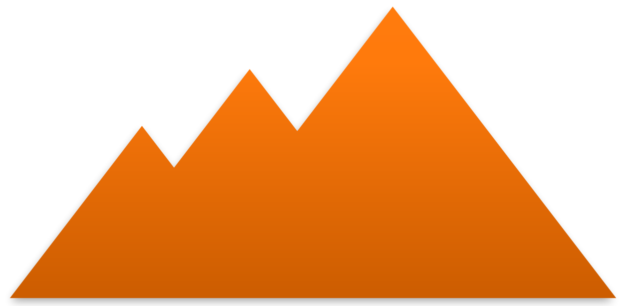


(from Google)

王明轩 1335

Why Monolingual

MT: Parallel data is limited



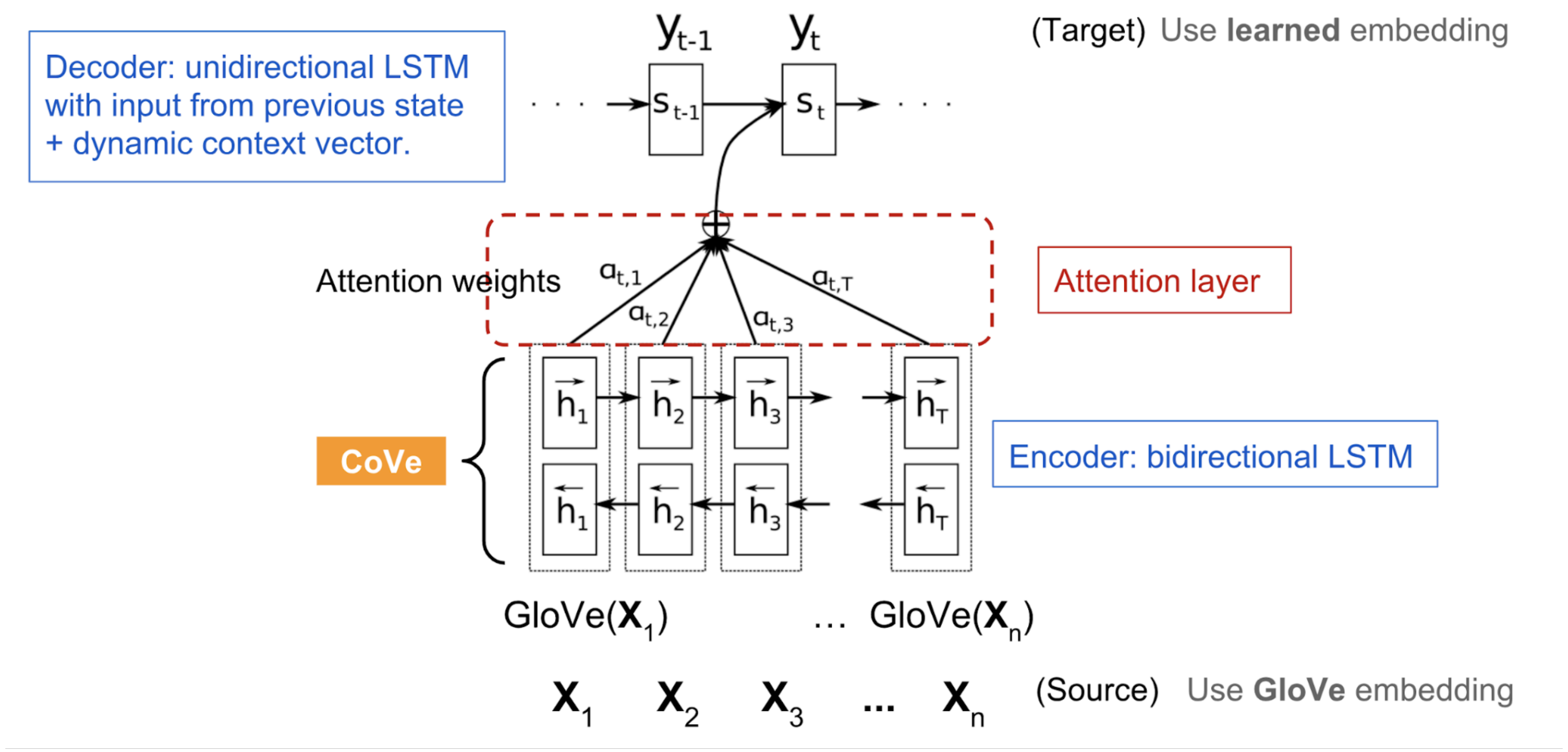
Parallel

Monolingual

PART2: Monolingual Pre-training for NMT

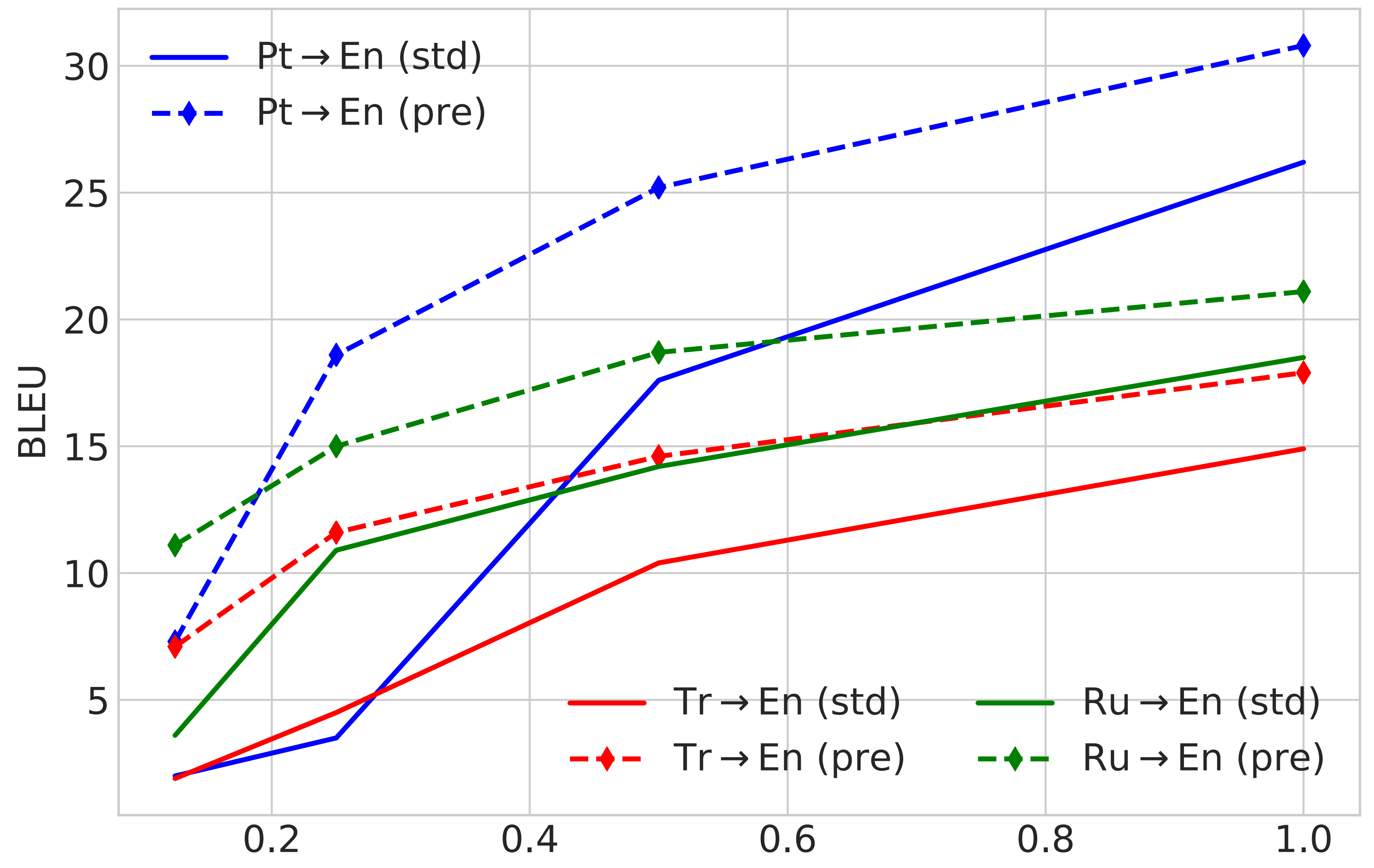
- The early stage
  - NMT initialized with word2vec [ACL 2017, NAACL 2018, AI 2020]
  - NMT initialized with language model [EMNLP 2017]
- BERT fusion
  - BERT Incorporating methods [ICLR 2020, AAAI 2020a]
  - BERT Tuning methods [AAAI 2020b]
- Unified sequence to sequence pre-training
  - MASS: Masked Sequence-to-Sequence Pre-training [ICML 2019]
  - BART: Denoising Sequence-to-Sequence Pre-training [ACL 2020]

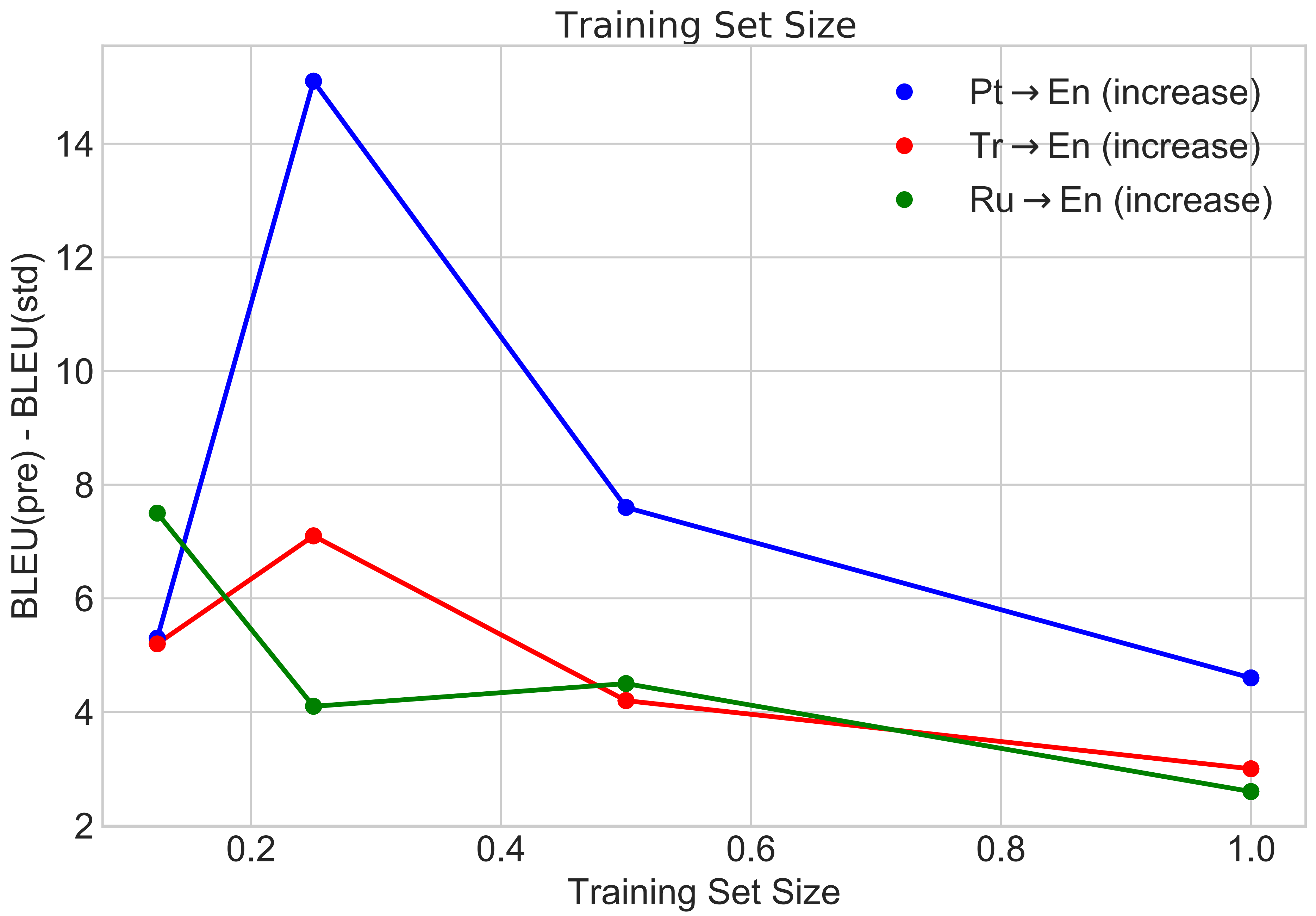
NMT initialized with word2vec



- When and Why are Pre-trained Word Embeddings Useful for Neural Machine Translation [NAACL 2018]
- Improve Neural Machine Translation by Building Word Vector [AI 2020]
- A bag of useful tricks for practical neural machine translation: Embedding layer initialization and large batch size [ACL 2017]

When and Why are Pre-trained Word Embeddings Useful for Neural Machine Translation



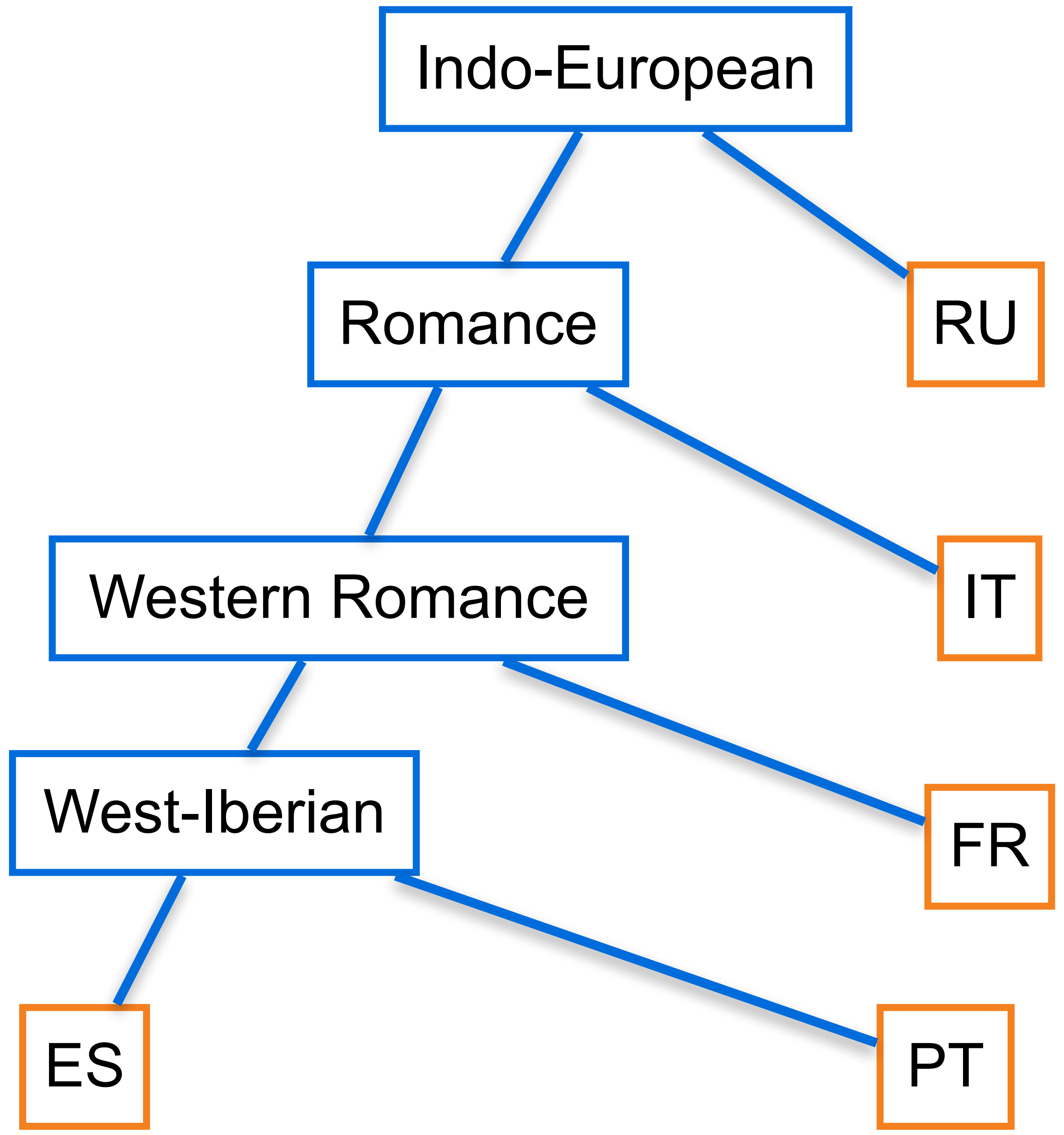


- The pre-trained embeddings help more when the size of the training data is small

When and Why are Pre-trained Word Embeddings Useful for Neural Machine Translation, [Qi et al NAACL 2018]

Effect of language similarity

Dataset	Lang. Family	std	pre
ES → PT	West-Iberian	17.8	24.8 (+**7.0**)
FR → PT	Western Romance	12.4	18.1 (+5.7)
IT → PT	Romance	14.5	19.2 (+4.7)
RU → PT	Indo-European	2.4	8.6 (+6.2)
HE → PT	*No Common*	3.0	11.9 (+**8.9**)



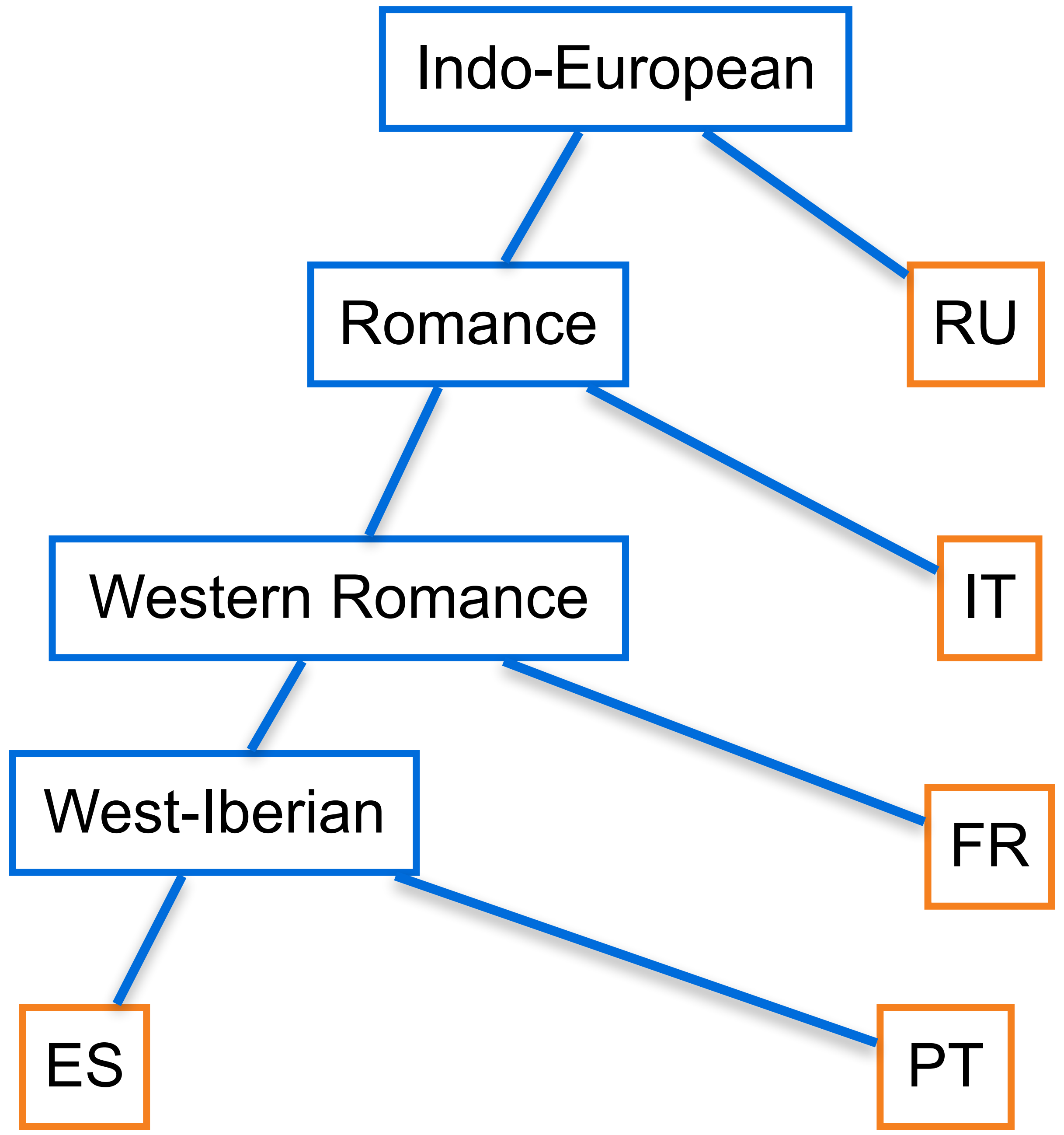
Language family tree

- All pairs are trained on 40,000 sentences
- Language similarity with PT: ES>FR>IT>RU
  - BLEU improves: ES>FR>IT
- RU and HE have very low baseline BLEU scores, so it makes sense that their increases would be larger

When and Why are Pre-trained Word Embeddings Useful for Neural Machine Translation, [Qi et al NAACL 2018]

Effect of language similarity

Dataset	Lang. Family	std	pre
Es → Pt	West-Iberian	17.8	24.8 (+**7.0**)
Fr → Pt	Western Romance	12.4	18.1 (+5.7)
It → Pt	Romance	14.5	19.2 (+4.7)
Ru → Pt	Indo-European	2.4	8.6 (+6.2)
He → Pt	*No Common*	3.0	11.9 (+**8.9**)



Language family tree

- All pairs are trained on 40,000 sentences
- Language similarity with PT: ES>FR>IT>RU
  - BLEU improves: ES>FR>IT
- RU and HE have very low baseline BLEU scores, so it makes sense that their increases would be larger

When and Why are Pre-trained Word Embeddings Useful for Neural Machine Translation, [Qi et al NAACL 2018]

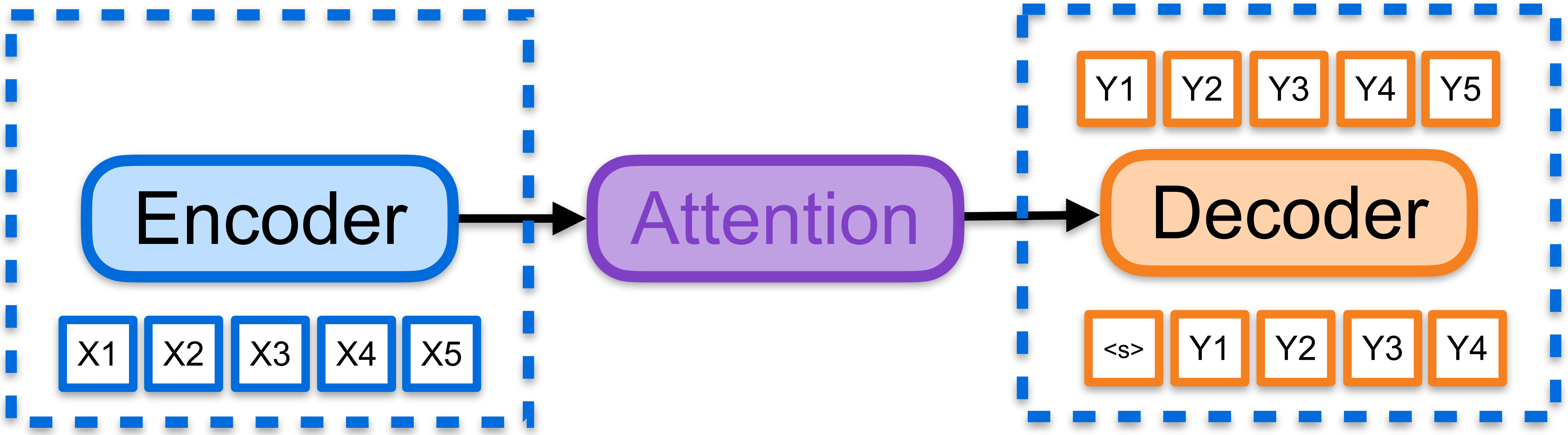
Effect of multilingual alignmet

Train	Eval	bi	std	pre	align
GL + PT	GL	2.2	17.5	20.8	**22.4**
AZ + TR	AZ	1.3	5.4	5.9	**7.5**
BE + RU	BE	1.6	**10.0**	7.9	9.6

- Training on both low-resource and higher-resource languages, and test on only the low-resource language
  - bi: the bilingual baseline
  - std: the multilingual baseline
  - pre: pre-training word embedding
  - align: convert the word embeddings of multiple languages to a single space [Smith et al., 2017]
- Alignment ensures that the word embeddings of the two source languages are put into similar vector spaces, and improves the performance

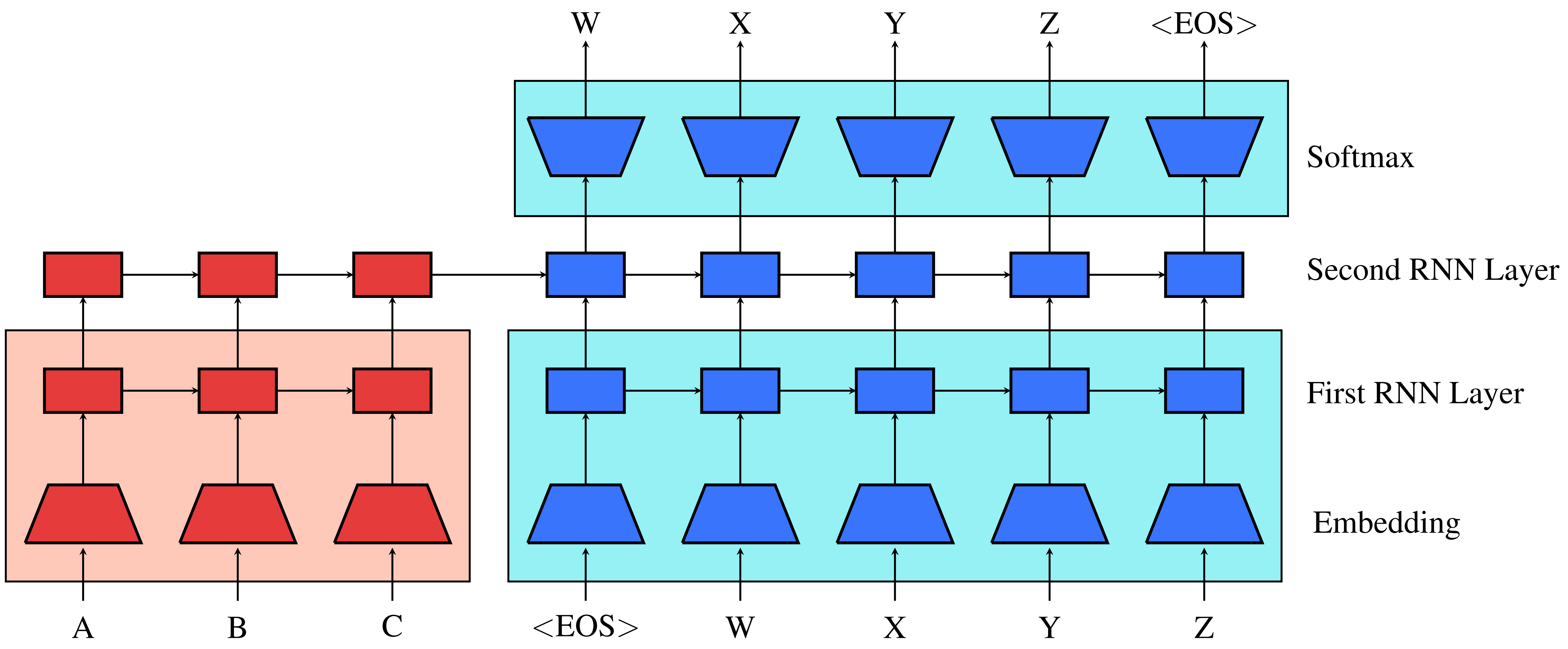
When and Why are Pre-trained Word Embeddings Useful for Neural Machine Translation, [Qi et al NAACL 2018]

NMT initialized with language model



- Unsupervised pretraining for sequence to sequence learning [EMNLP 2017]
- Exploiting Source-side Monolingual Data in Neural Machine Translation [EMNLP 2016]
- Semi-Supervised Learning for Neural Machine Translation [ACL 2016]

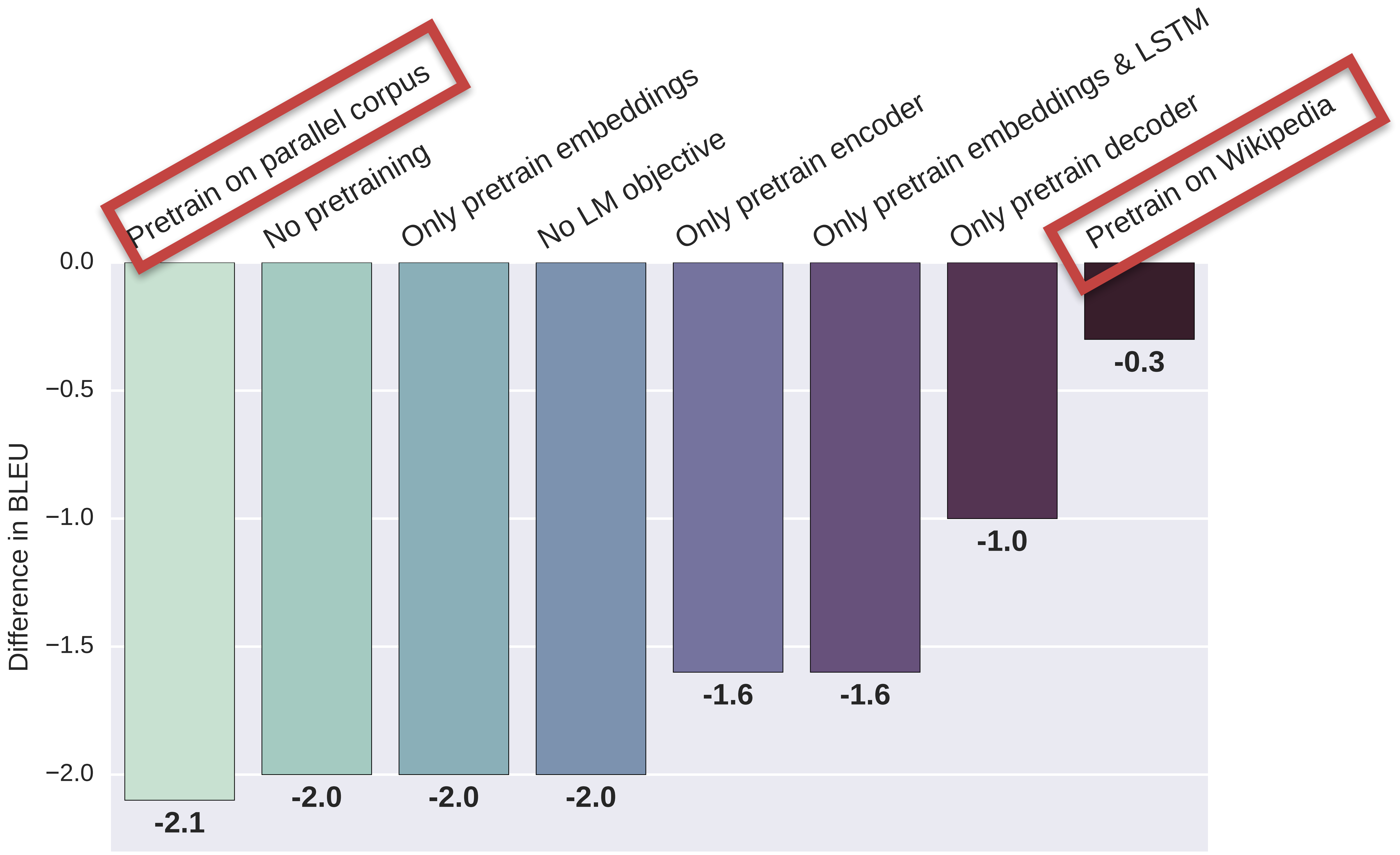
Unsupervised pretraining for sequence to sequence learning



- The red parameters are the encoder and the blue parameters are the decoder.
- All parameters in a shaded box are pre-trained with RNN language models
- Otherwise, randomly initialized.

Unsupervised pretraining for sequence to sequence learning, [Ramachandran et al EMNLP 2017]

Unsupervised pretraining for sequence to sequence learning

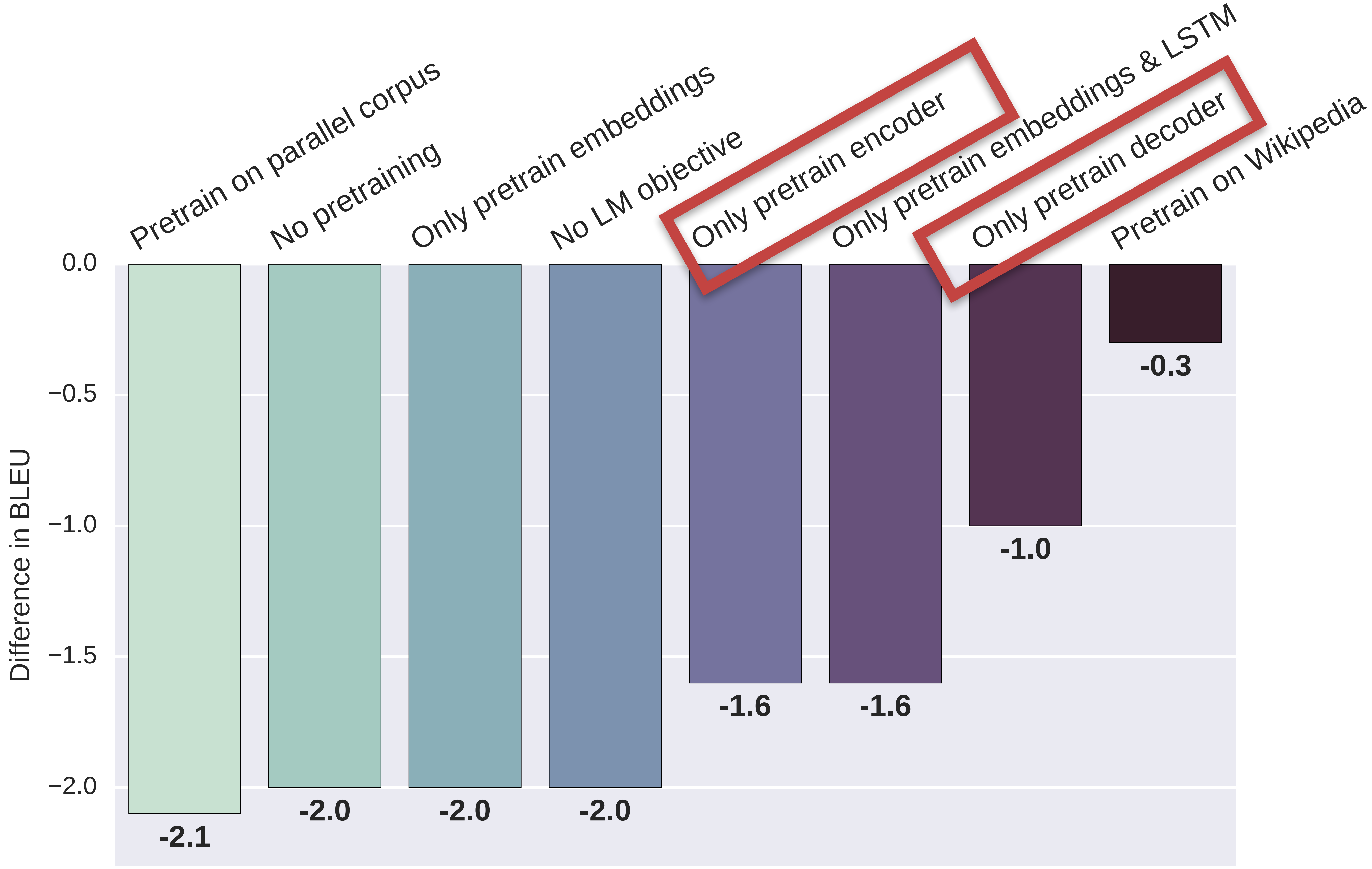


Pretraining on a lot of unlabeled data is essential.

If the model is initialized with LMs that are pretrained on the source part and target part of the *parallel* corpus

Unsupervised pretraining for sequence to sequence learning, [Ramachandran et al EMNLP 2017]

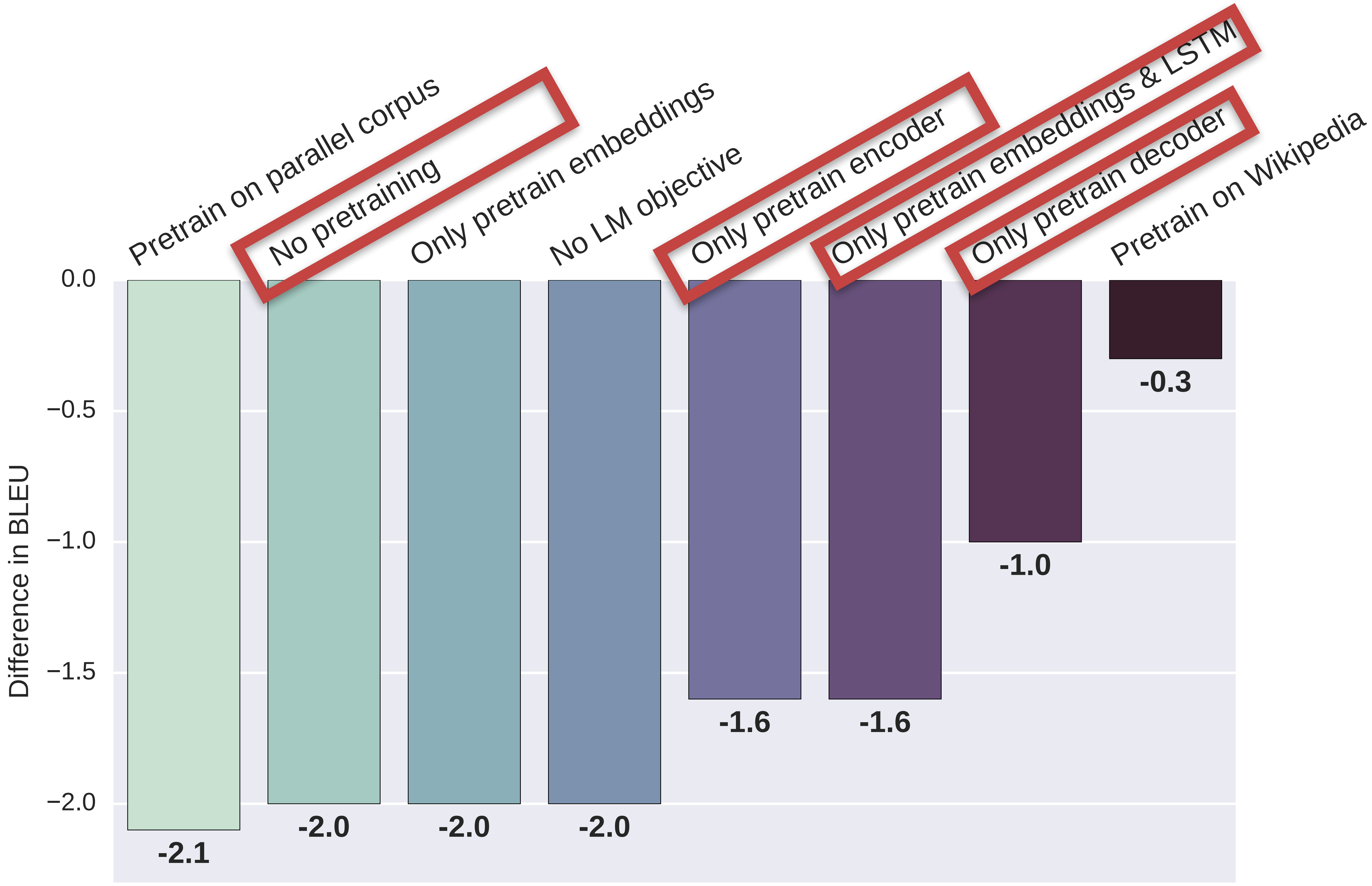
Unsupervised pretraining for sequence to sequence learning



Only pretraining the decoder is better than only pretraining the encoder

Unsupervised pretraining for sequence to sequence learning, [Ramachandran et al EMNLP 2017]

Unsupervised pretraining for sequence to sequence learning



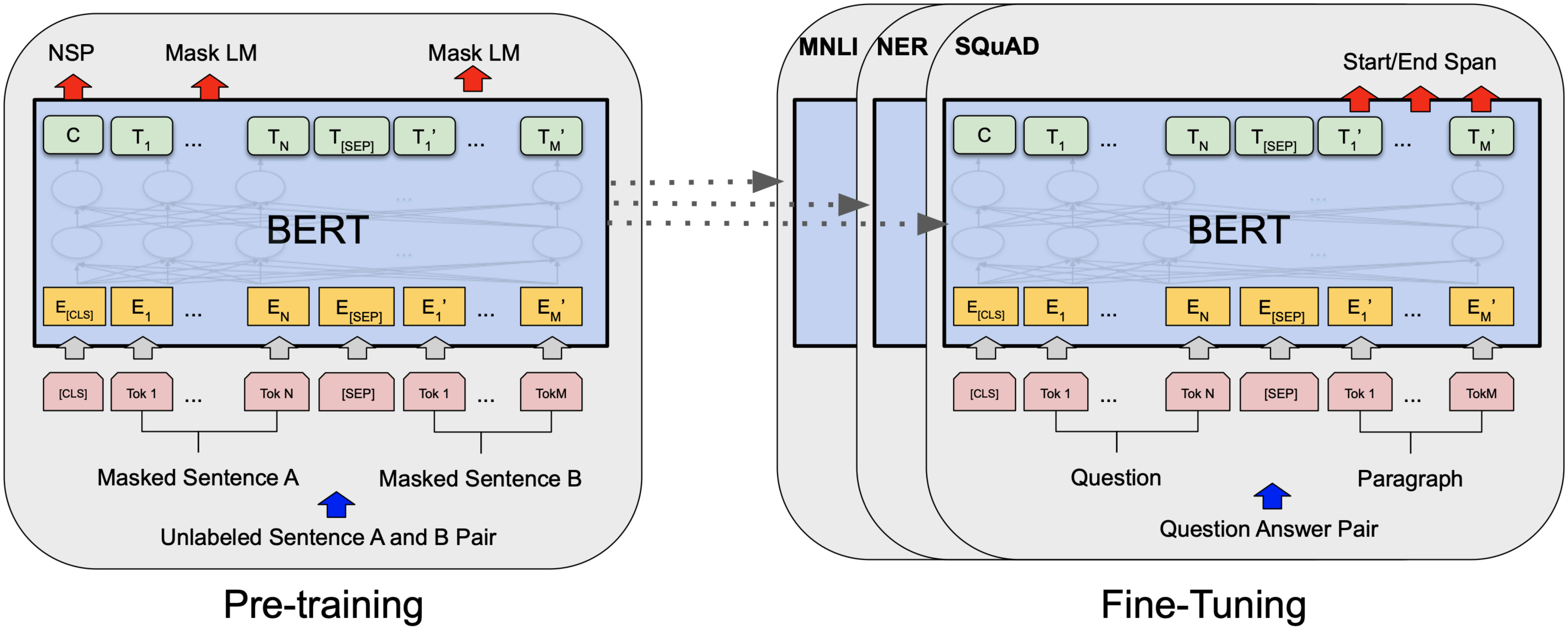
Pretrain as much as possible because the benefits compound.

Unsupervised pretraining for sequence to sequence learning, [Ramachandran et al EMNLP 2017]

Summary

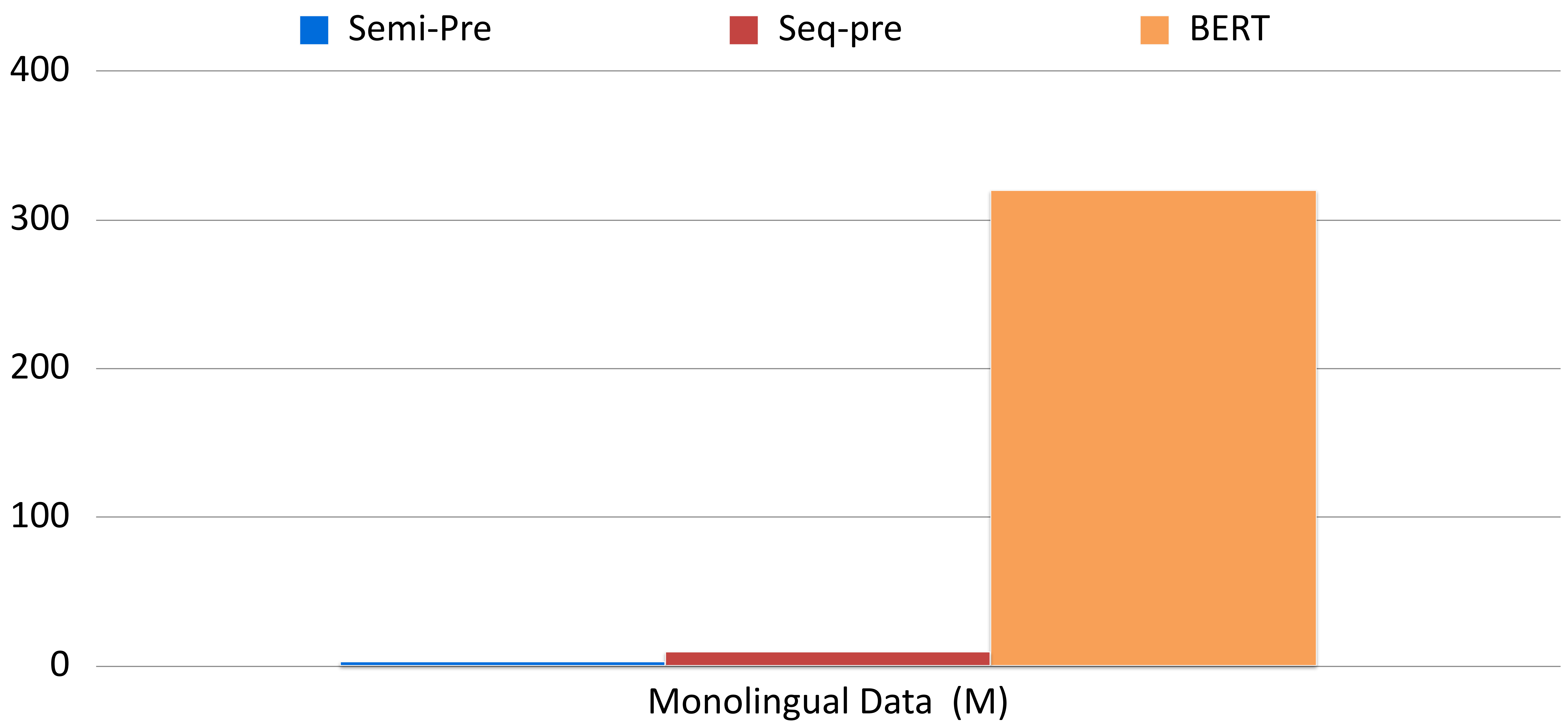
- Insight
  - Pre-training is effective on low-resource NMT
  - Pre-training as much as components
  - Pre-training as much as training data
  - Cross-lingual information helps
- Limitations:
  - The improvements on rich resource NMT is not large enough
  - The pre-training model is trained on limited training corpus, e.g. the monolingual part of the parallel data
  - Only a subset of parameters are pre-trained

Then, BERT comes…



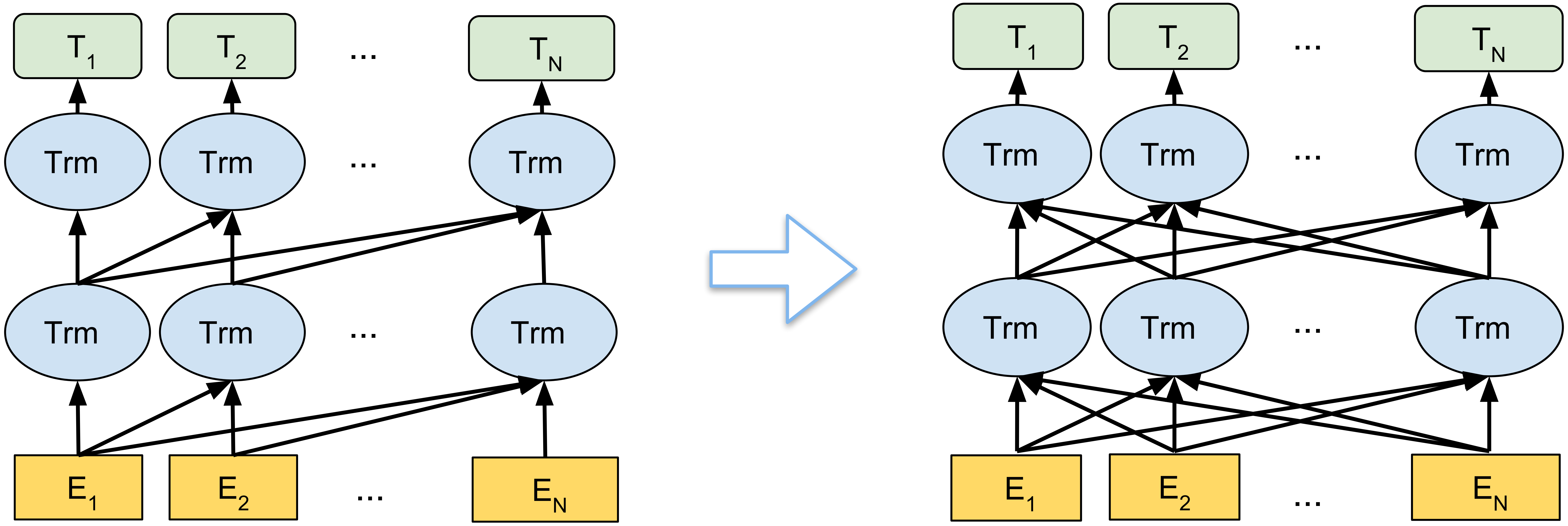
What happens?

Pre-training data scale increased



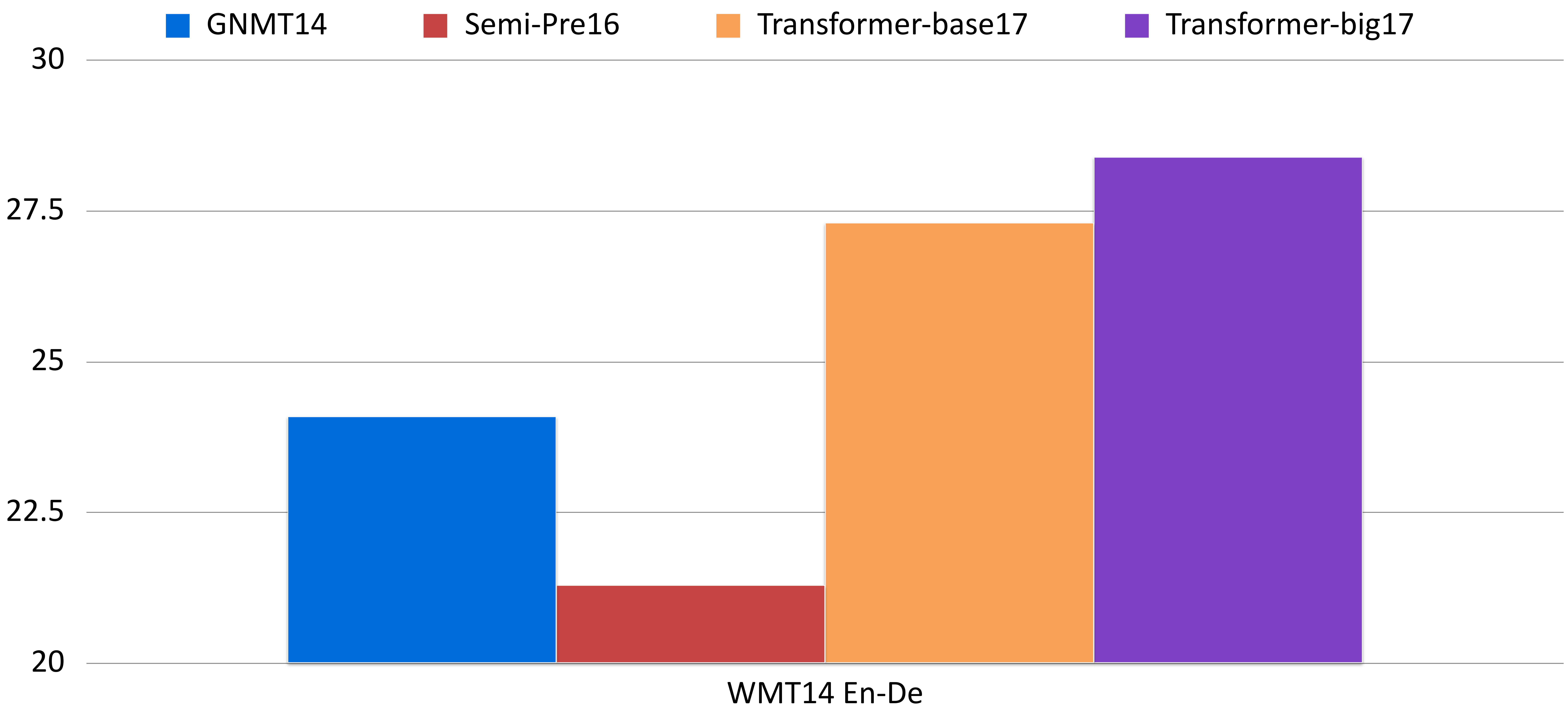
What happens?

Pre-training framework changed



What happens?

Baseline improved



What happens?

Does BERT matter in NMT?

PART2: Monolingual Pre-training for NMT

- The Bronze Age
  - NMT initialized with word2vec [ACL 2017, NAACL 2018, AI 2020]
  - NMT initialized with language model [EMNLP 2017]
- BERT fusion
  - BERT Incorporating methods [ICLR 2020, AAAI 2020a]
  - BERT Tuning methods [AAAI 2020b]
- Unified sequence to sequence pre-training
  - MASS: Masked Sequence-to-Sequence Pre-training [ICML 2019]
  - BART: Denoising Sequence-to-Sequence Pre-training [ACL 2020]

Incorporate BERT into Neural Machine Translation

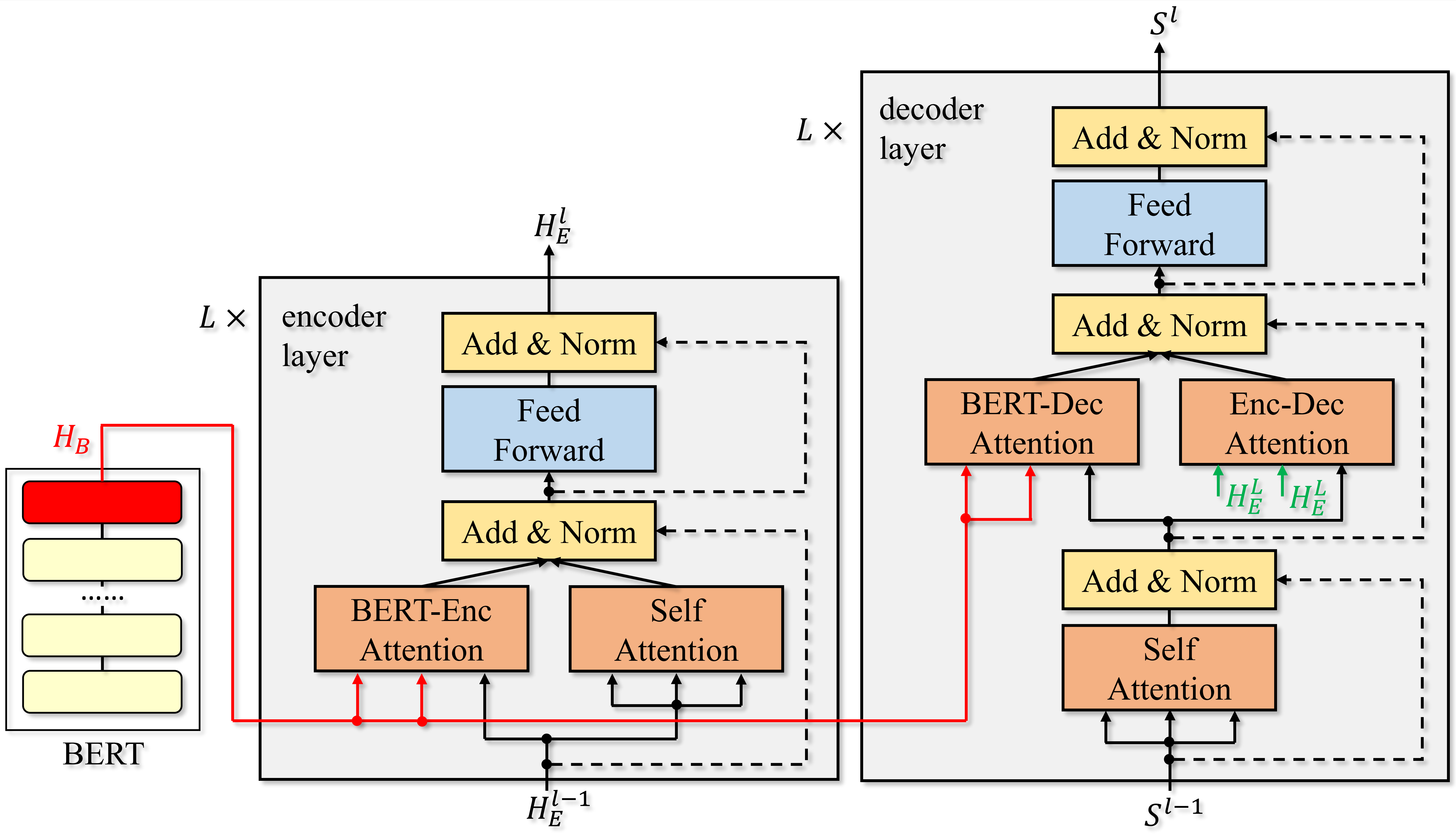
Table 1: Preliminary explorations on IWSLT'14 English→German translatio

Algorithm	BLEU score
Standard Transformer	28.57
Use BERT to initialize the encoder of NMT	27.14
Use XLM to initialize the encoder of NMT	28.22
Use XLM to initialize the decoder of NMT	26.13
Use XLM to initialize both the encoder and decoder of NMT	28.99
Leveraging the output of BERT as embeddings	29.67

- Fine-tuning BERT does NOT work !
  - BERT and XLM pre-training for the encoder decreased the performance
  - XLM pre-training for the decoder enlarged the performance gap
- BERT-Frozen achieved improvements

Incorporate BERT into Neural Machine Translation, [Zhu et al ICLR 2017]

Incorporate BERT into Neural Machine Translation



- BERT features are directly fed to both encoder and decoder layers
- Additional attention model to incorporate BERT features

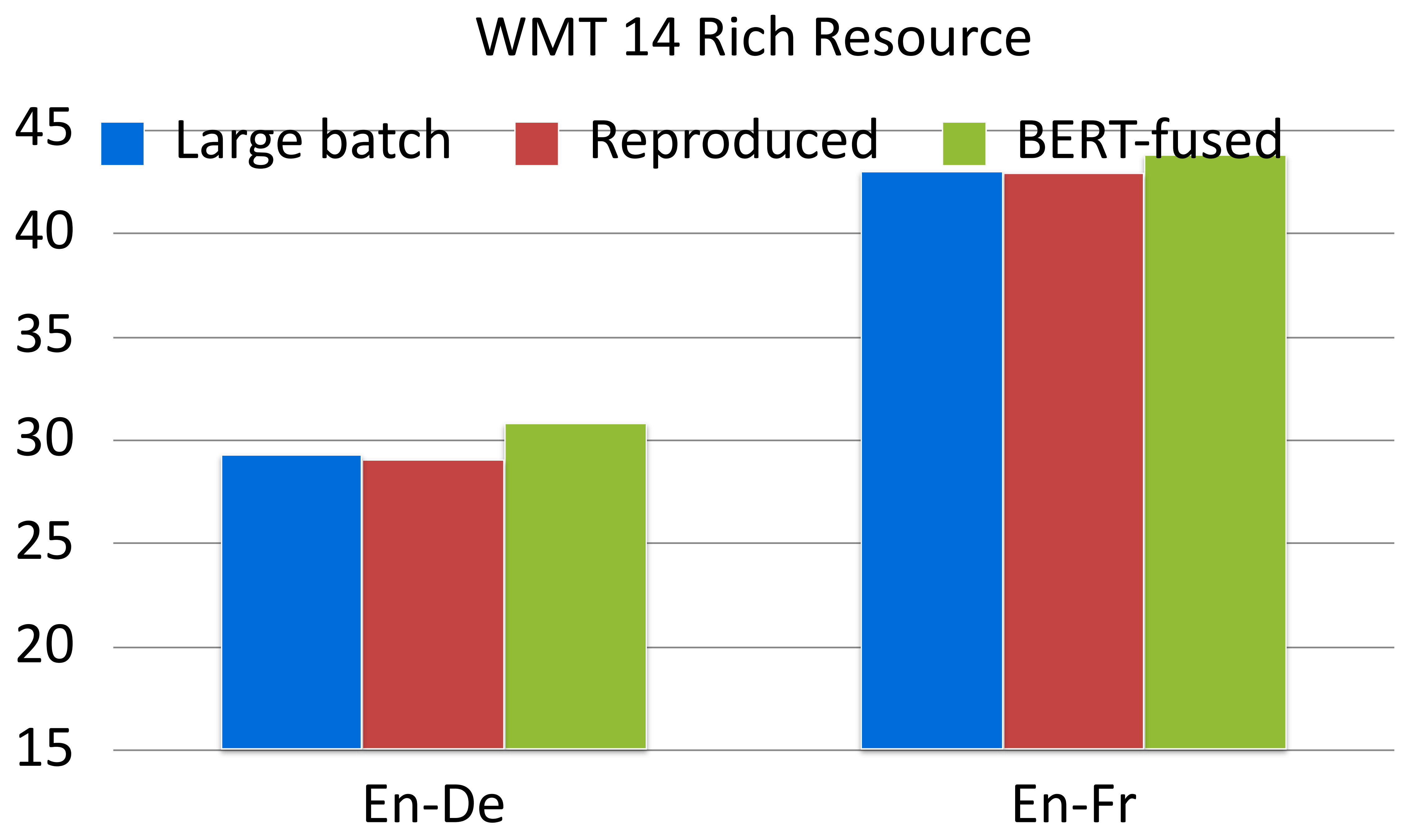
Incorporate BERT into Neural Machine Translation, [Zhu et al ICLR 2020]

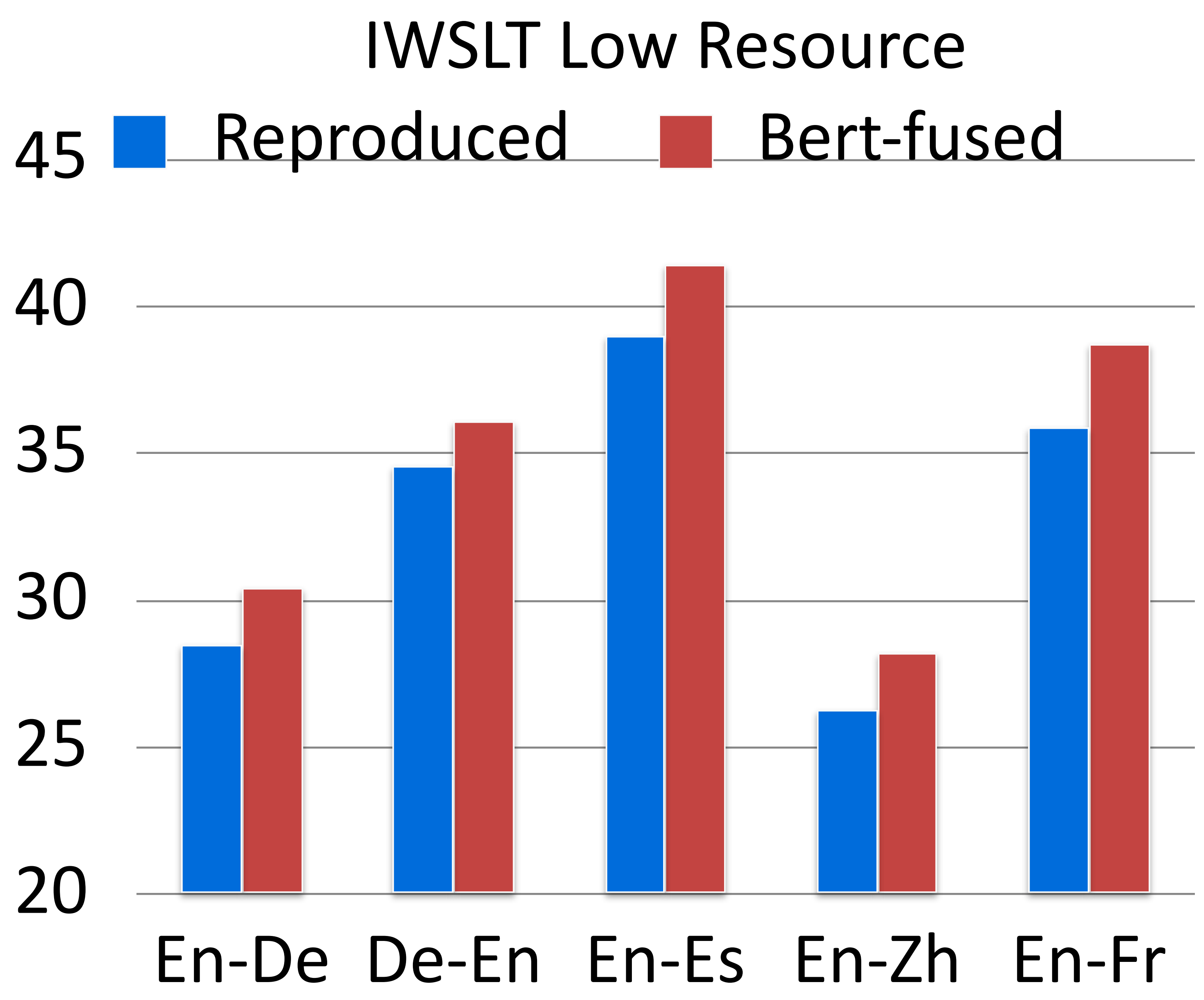
Datasets and settings

- Fine-tuning dataset
  - Low resource: IWSLT En-De, En-FR, En-Zh, En-Es (less than 250 k sentence pairs)
  - Rich resource: WMT14 En-De and En-Fr (4 M and 36 M sentence pairs)
- Settings
  - BERT base for IWSLT
  - BERT large for WMT
  - Both the BERT-encoder and BERTdecoder attention are randomly initialized

Incorporate BERT into Neural Machine Translation, [Zhu et al ICLR 2020]

Main results on supervised MT

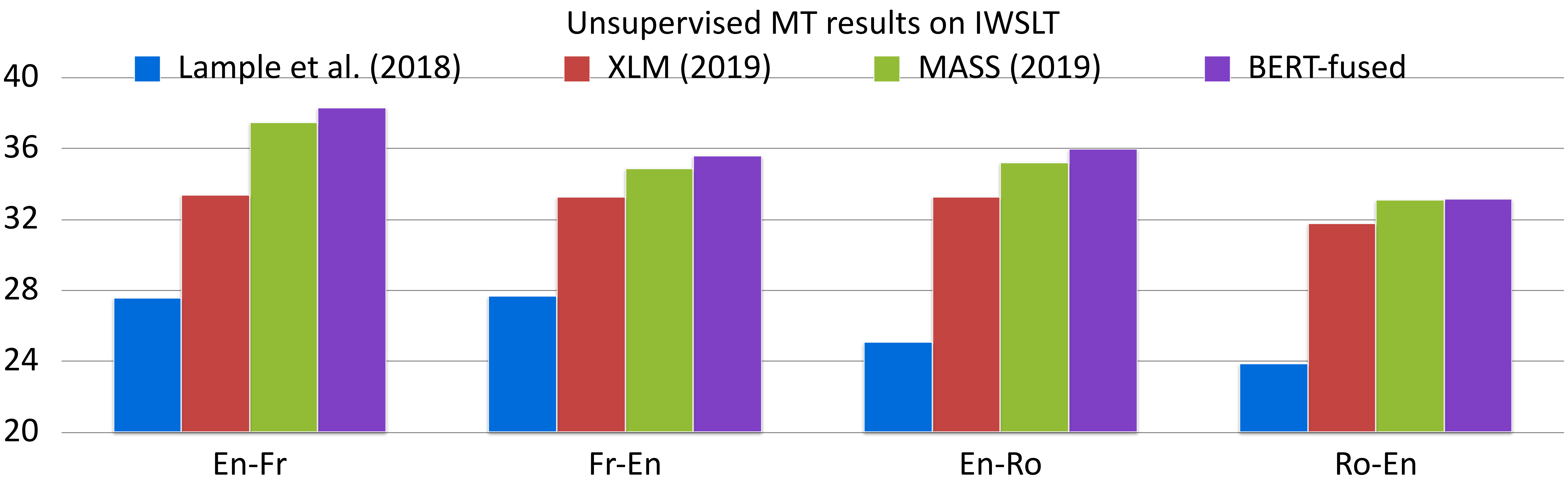




- Experiments on a strong baseline
- BERT-fused model outperforms transformer baseline in all settings

Incorporate BERT into Neural Machine Translation, [Zhu et al ICLR 2020]

Main results on unsupervised MT



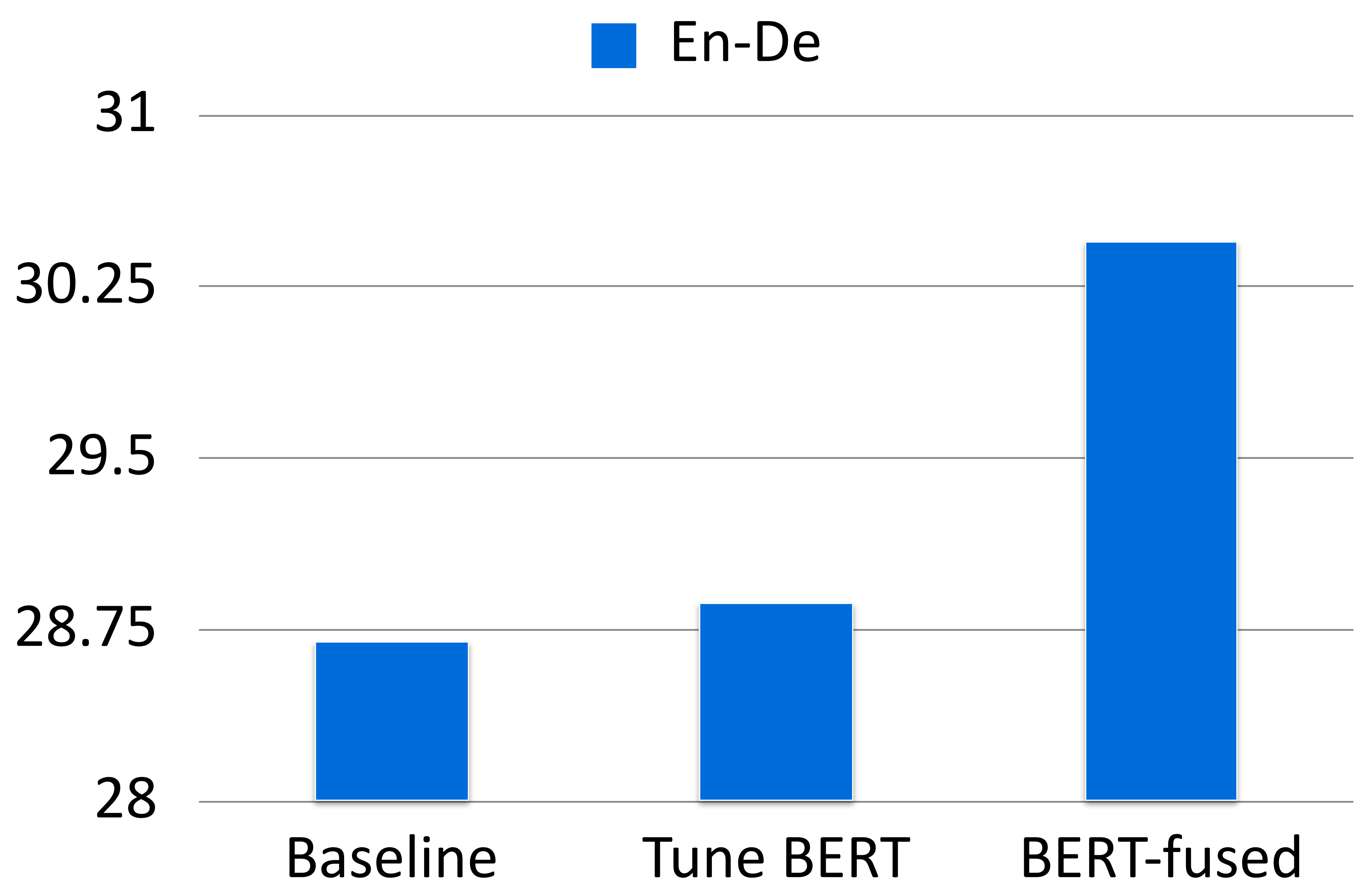
- Pre-training plays an crucial role in unsupervised NMT (Lample v.s. xml, mass and BERT-fused)
- BERT-fused outperforms XLM and MASS
- The comparison is slightly unfair, since BERT-fused introduced additional parameters

Incorporate BERT into Neural Machine Translation, [Zhu et al ICLR 2020]

NOT Tune BERT

Table 6: Ablation study on IWSLT'14 En→De.

Standard Transformer	28.57
BERT-fused model	30.45
Randomly initialize encoder/decoder of BERT-fused model	27.03
Jointly tune BERT and encoder/decoder of BERT-fused model	28.87
Feed BERT feature into all layers without attention	29.61
Replace BERT output with random vectors	28.91
Replace BERT with the encoder of another Transformer model	28.99
Remove BERT-encoder attention	29.87
Remove BERT-decoder attention	29.90



Jointly train BERT model with the NMT can also boost the baseline from 28.57 to 28.87.

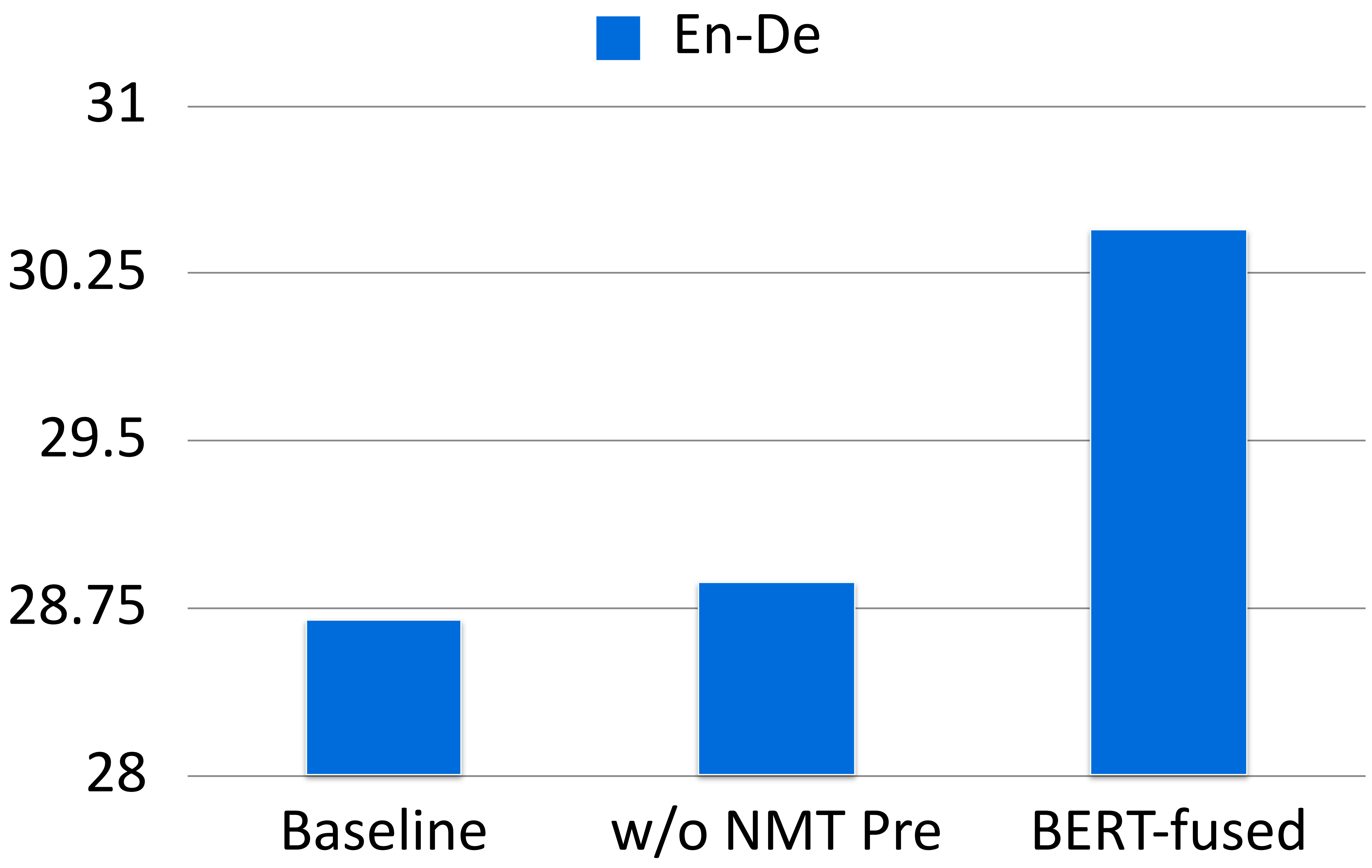
But it is not as good as fixing the BERT part, whose BLEU is 30.45

Incorporate BERT into Neural Machine Translation, [Zhu et al ICLR 2020]

NMT pre-training matters

Table 6: Ablation study on IWSLT'14 En→De.

Standard Transformer	28.57
BERT-fused model	30.45
Randomly initialize encoder/decoder of BERT-fused model	27.03
Jointly tune BERT and encoder/decoder of BERT-fused model	28.87
Feed BERT feature into all layers without attention	29.61
Replace BERT output with random vectors	28.91
Replace BERT with the encoder of another Transformer model	28.99
Remove BERT-encoder attention	29.87
Remove BERT-decoder attention	29.90



NMT Pre-training is also important to the success of BERT-fused model

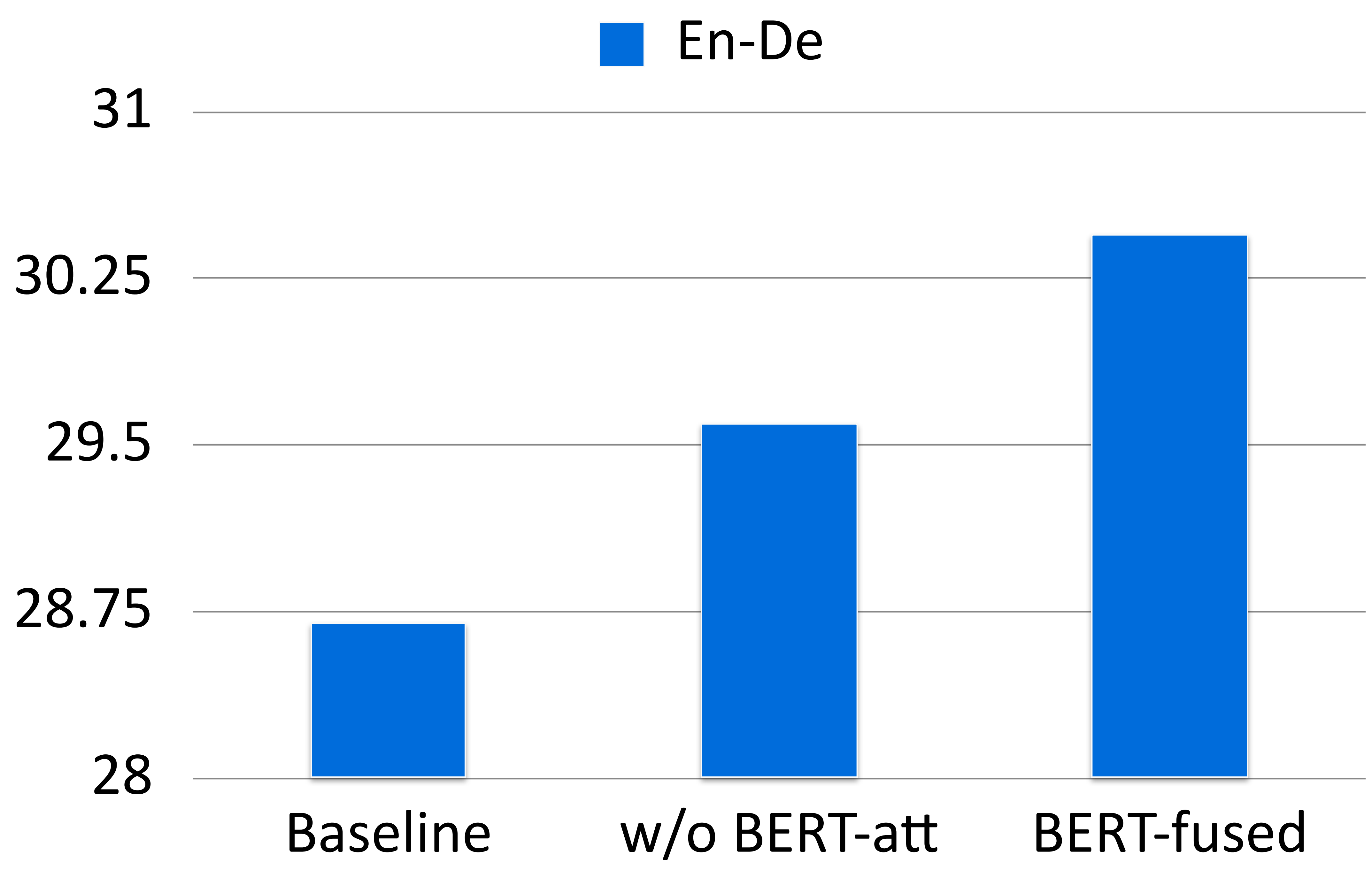
Without NMT pre-training, the performance lags behind the baseline model

Incorporate BERT into Neural Machine Translation, [Zhu et al ICLR 2020]

BERT attention module matters

Table 6: Ablation study on IWSLT'14 En→De.

Standard Transformer	28.57
BERT-fused model	30.45
Randomly initialize encoder/decoder of BERT-fused model	27.03
Jointly tune BERT and encoder/decoder of BERT-fused model	28.87
Feed BERT feature into all layers without attention	29.61
Replace BERT output with random vectors	28.91
Replace BERT with the encoder of another Transformer model	28.99
Remove BERT-encoder attention	29.87
Remove BERT-decoder attention	29.90



Remove attention module, the performance still outperforms baseline, but falls behind BERT-fused model

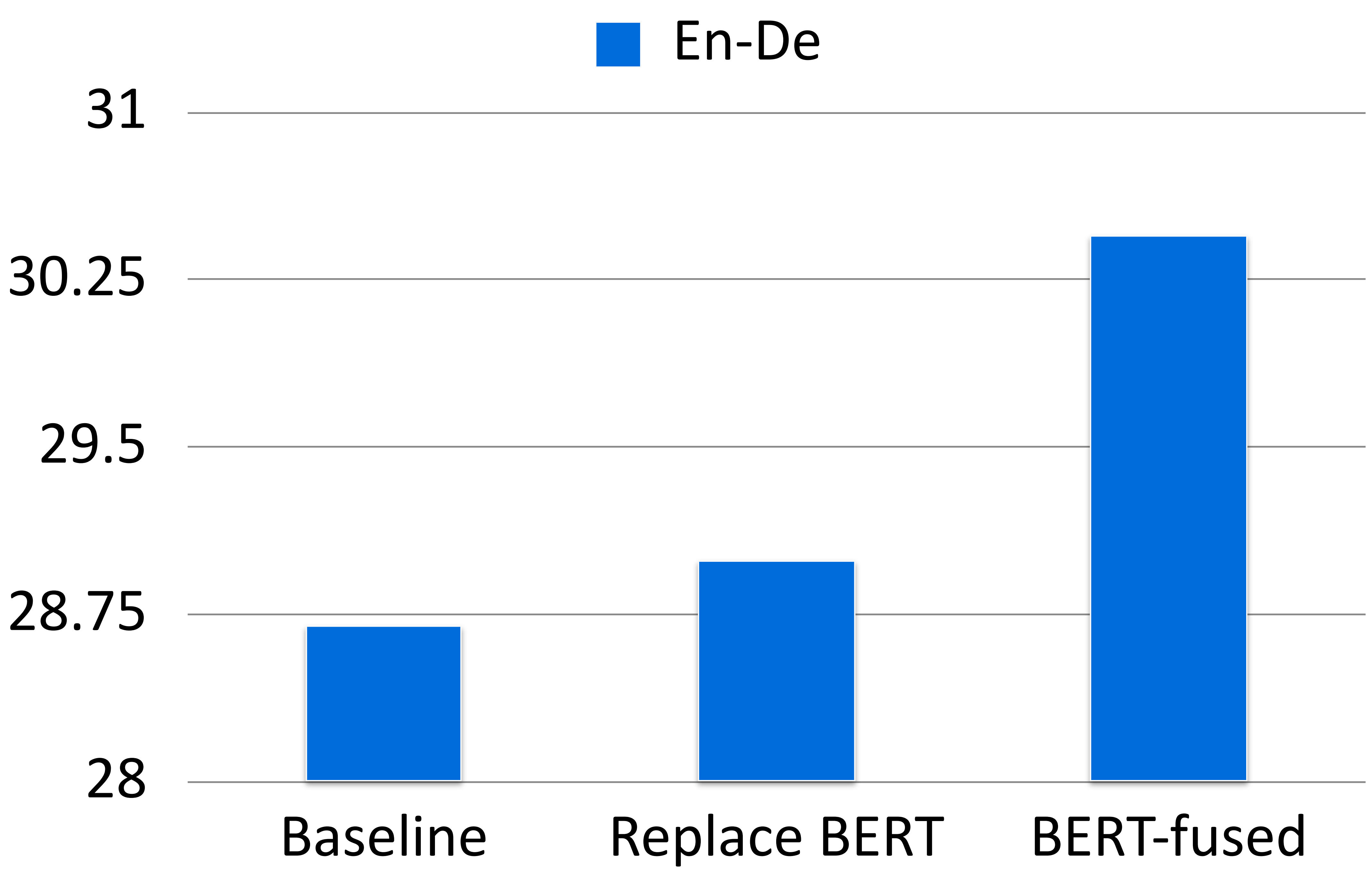
It suggest that separate BERT model provides additional gains

Incorporate BERT into Neural Machine Translation, [Zhu et al ICLR 2020]

Of course, BERT matters

Table 6: Ablation study on IWSLT'14 En→De.

Standard Transformer	28.57
BERT-fused model	30.45
Randomly initialize encoder/decoder of BERT-fused model	27.03
Jointly tune BERT and encoder/decoder of BERT-fused model	28.87
Feed BERT feature into all layers without attention	29.61
Replace BERT output with random vectors	28.91
Replace BERT with the encoder of another Transformer model	28.99
Remove BERT-encoder attention	29.87
Remove BERT-decoder attention	29.90

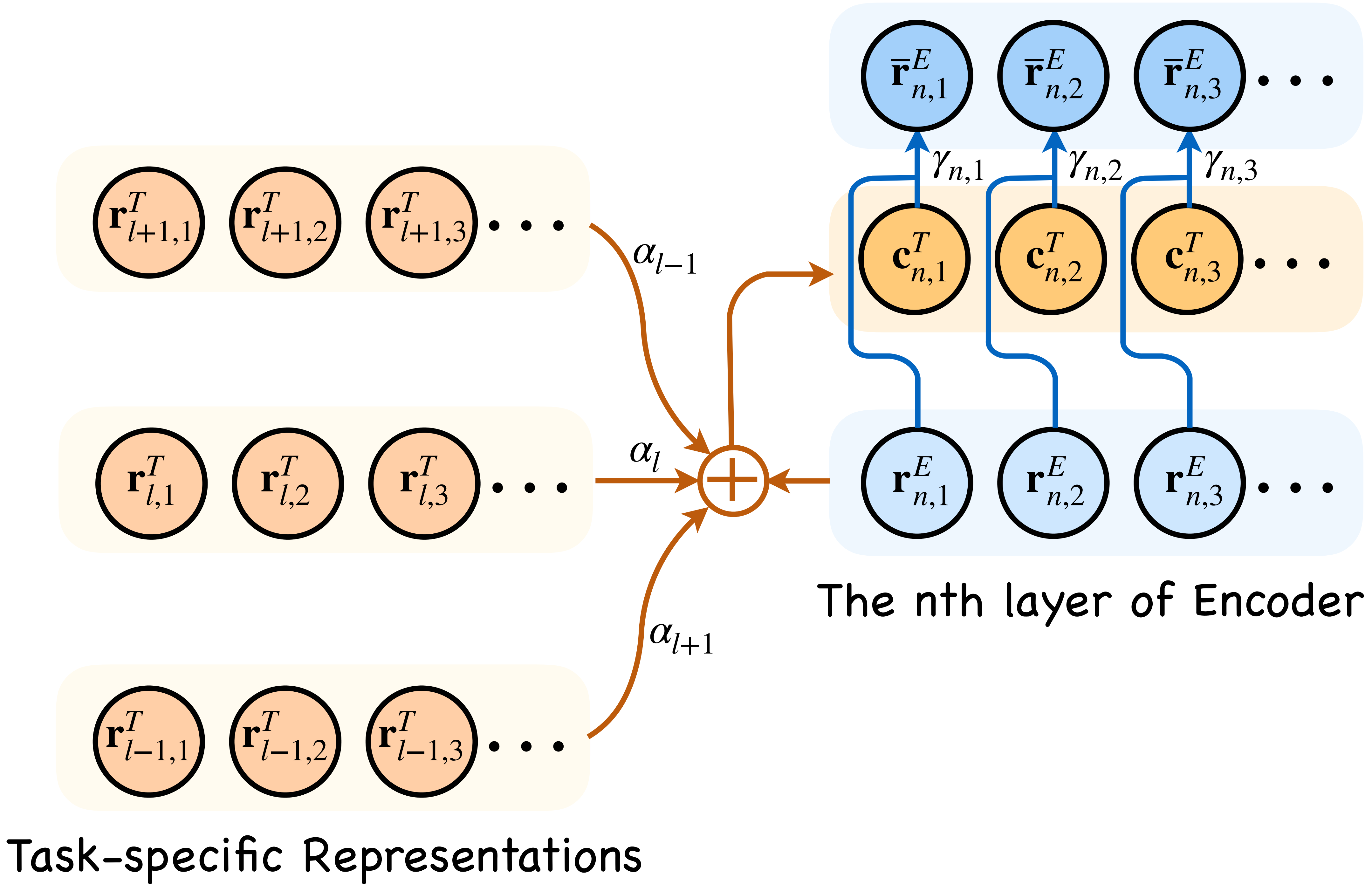


Replace BERT representation with another transformer model, the performance drops significantly

It indicates BERT provides meaningful information and the improvements is not from the additional parameters.

Incorporate BERT into Neural Machine Translation, [Zhu et al ICLR 2020]

Acquiring Knowledge from Pre-trained Model to Neural Machine Translation



- Key idea
  - Dynamic fusion of different BERT layers, while BERT-fused model only uses the last layer of BERT
  - Incorporate BERT into all encoder layers and decoder layers with adaptive weight
  - Experiments including both BERT & GPT

Acquiring Knowledge from Pre-trained Model to Neural Machine Translation, [Weng et al AAAI 2020]

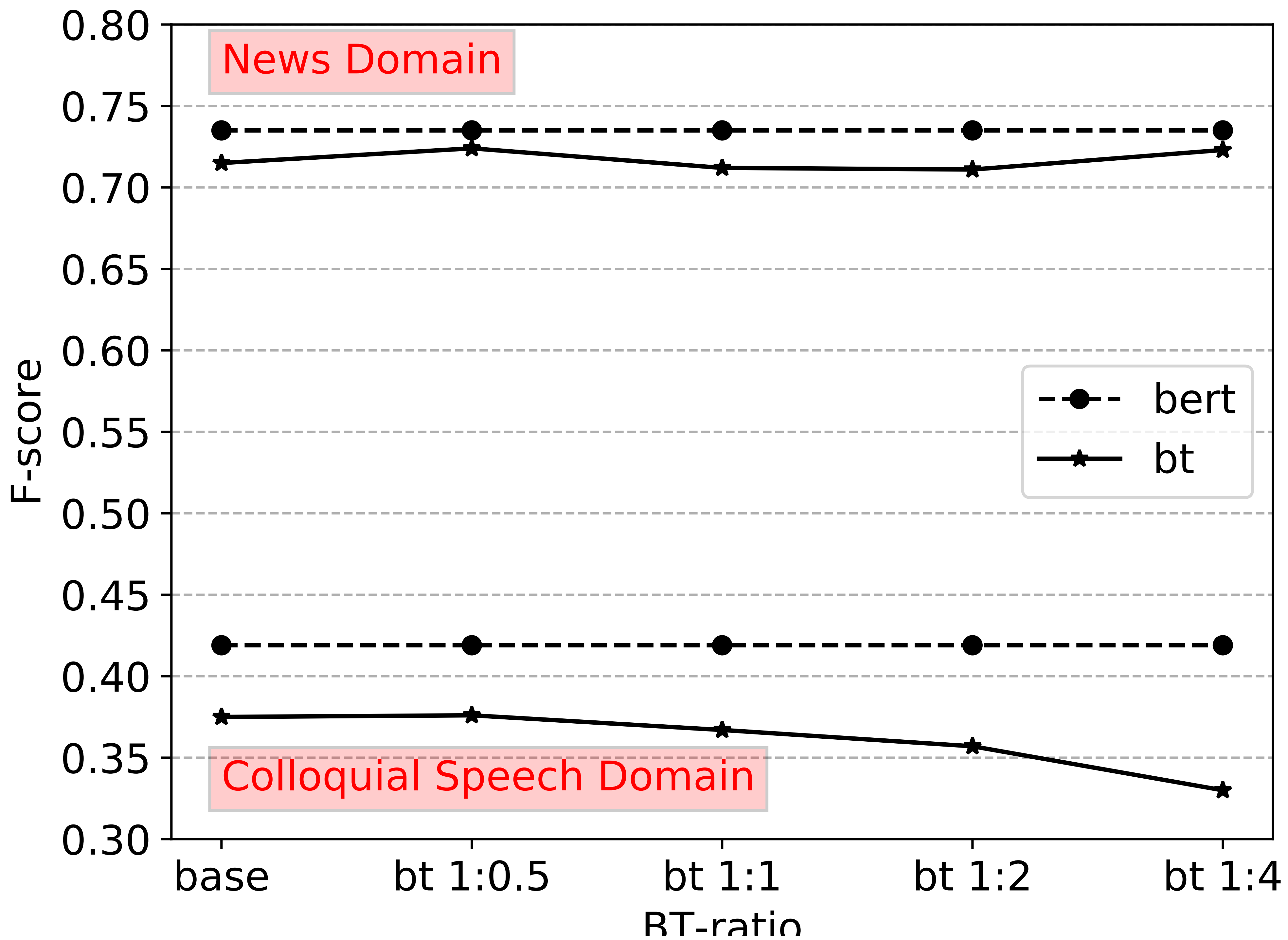
GPT v.s. BERT

Model	Pre-trained Model Encoder	Pre-trained Model Decoder	EN→DE BLEU	EN→DE Δ	DE→EN BLEU	DE→EN Δ	ZH→EN BLEU	ZH→EN Δ
Transformer (Vaswani et al. 2017)	N/A	N/A	27.3	–	N/A	–	N/A	–
Transformer (Zheng et al. 2019)	N/A	N/A	27.14	–	N/A	–	N/A	–
Transformer (Dou et al. 2018)	N/A	N/A	27.31	–	N/A	–	24.13	–
Transformer	N/A	N/A	27.31	–	32.51	–	24.47	–
w/ Fine-tuning	GPT	N/A	27.82	+0.51	33.17	+0.66	25.11	+0.64
	N/A	GPT	27.45	+0.14	32.87	+0.36	24.59	+0.12
	GPT	GPT	27.85	+0.54	32.79	+0.28	25.21	+0.74
	BERT	N/A	28.22	+0.91	33.64	+1.13	25.33	+0.86
	N/A	BERT	27.42	+0.11	33.13	+0.62	24.78	+0.31
	BERT	BERT	28.32	+1.01	33.57	+1.06	25.45	+0.98
	GPT	BERT	28.29	+0.98	33.33	+0.82	25.42	+0.95
	BERT	GPT	28.32	+1.01	33.57	+1.05	25.46	+0.99
	MASS		28.07	+0.76	33.29	+0.78	25.11	+0.64
	DAE		27.63	+0.33	33.03	+0.52	24.67	+0.20
w/ APT Framework	GPT	BERT	28.89	+1.58	34.32	+1.81	25.98	+1.51
	BERT	GPT	**29.23**	**+1.92**	**34.84**	**+2.33**	26.21	+1.74
	GPT	GPT	28.97	+1.66	34.26	+1.75	26.01	+1.54
	BERT	BERT	29.02	+1.71	34.67	+2.16	**26.46**	**+1.99**

Acquiring Knowledge from Pre-trained Model to Neural Machine Translation, [Weng et al AAAI 2020]

Pre-training has better generalization ability

System	En→De	Zh→En
Standard Transformer	29.20	45.15
+ back translation (1:0.5)	**30.41**	46.70
+ back translation (1:1)	30.25	**47.23**
+ back translation (1:2)	30.18	47.04
+ back translation (1:4)	30.25	46.39
BERT-fused model	30.03	46.55



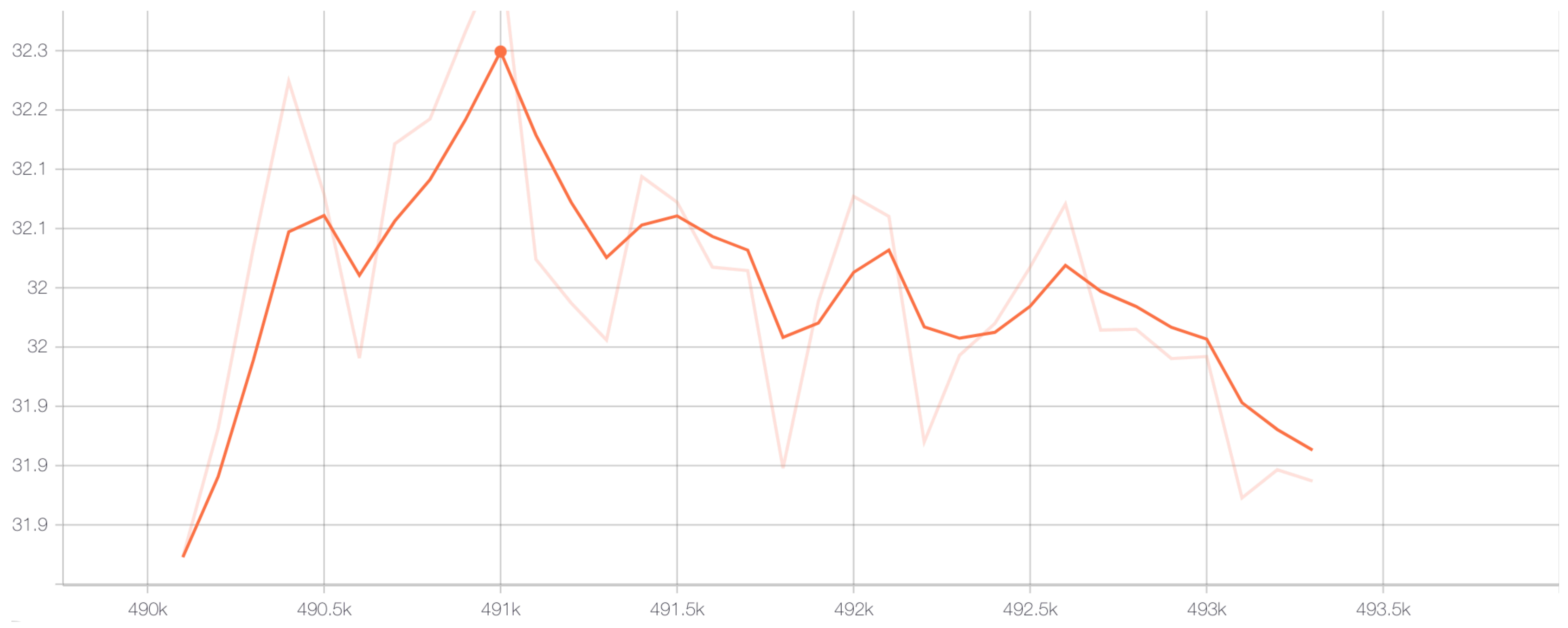
- Pre-training is much more promising
  - – better generalization ability
  - – Back translation is limited with data scale

Comparison between Pre-training and Large-scale Back-translation, [Huang et al ACL 2021]

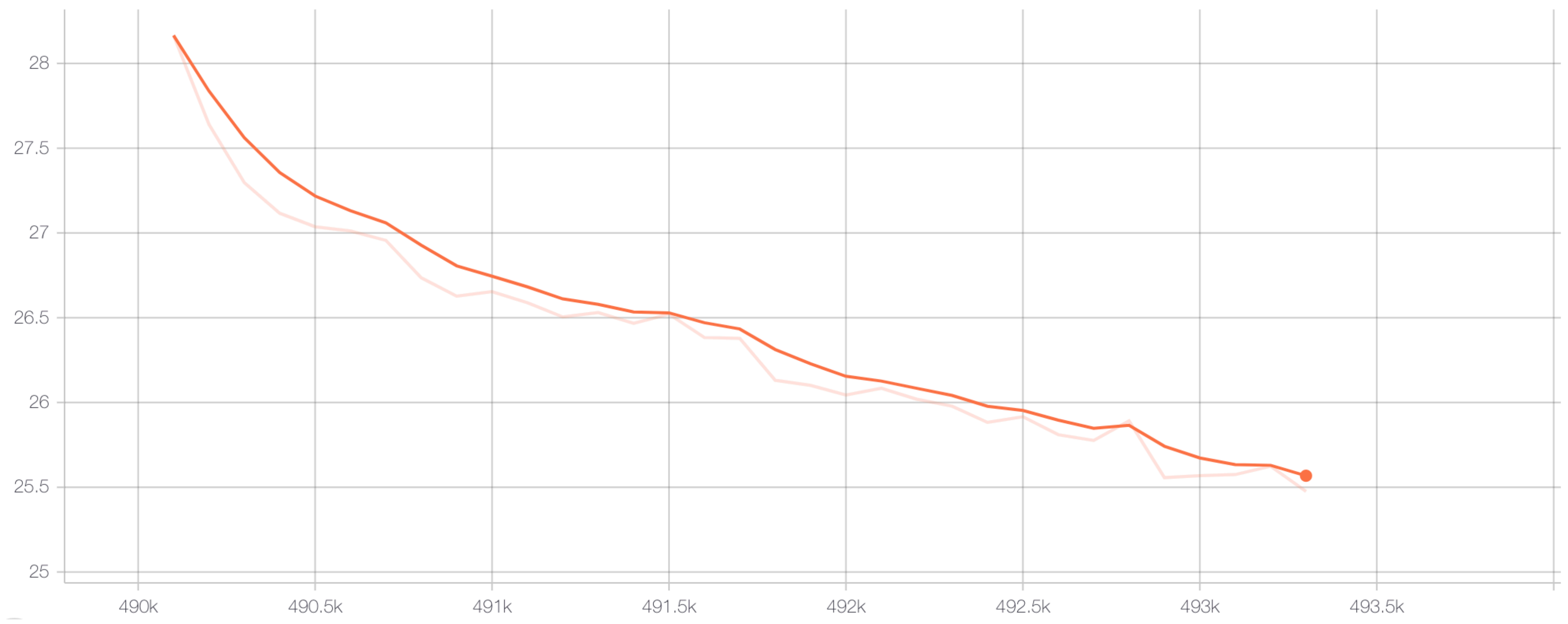
Summary

- Advantages
  - BERT features are fused in all layers
  - Additional attention model adaptively determine how to leverage BERT feature
- Limitions
  - Additional cost including training storage and inference time
  - Why not tune BERT?

Towards Making Most of BERT for NMT



Performance on fine-tuning NMT



Performance on other BERT tasks

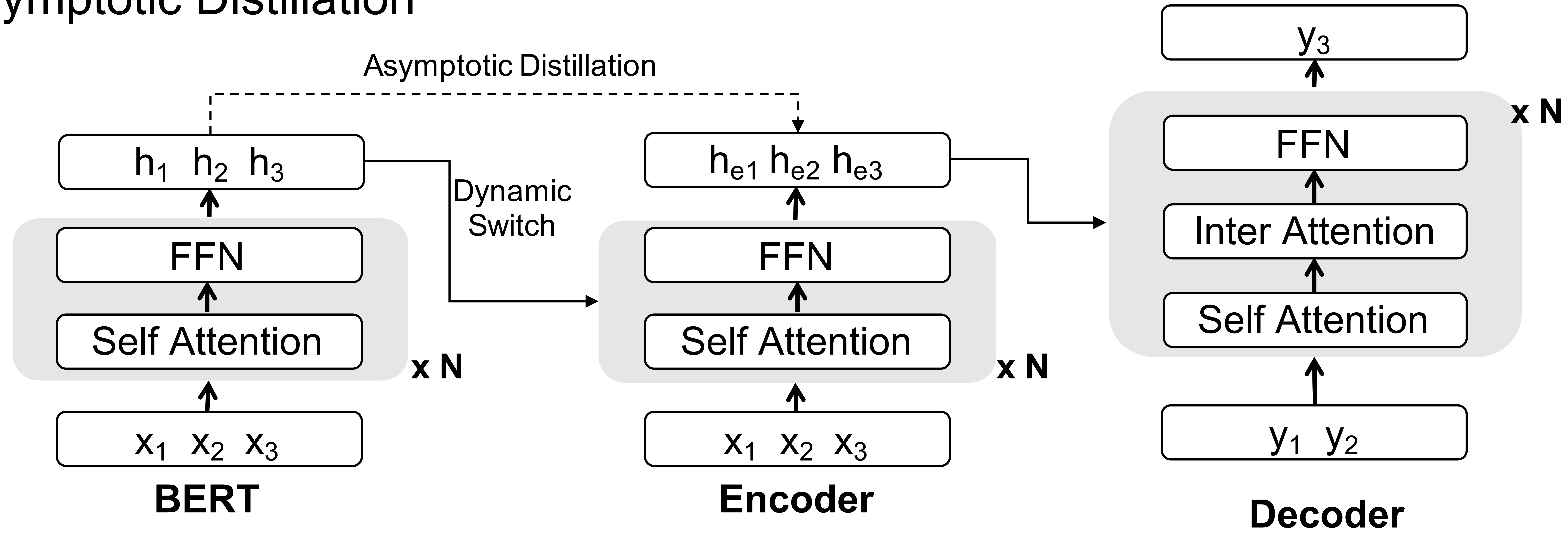
Why simply incorporating BERT does not work as expectation

- Fine-tuning leads to performance degradation on the original task
- The situation is more severe on NMT fine-tuning
  - High capacity of baseline needs much updating
  - Updating to much makes the model forgets its universal knowledge from pre-training

Towards Making Most of BERT for NMT, [Yang et al AAAI 2020]

Not tuning too much

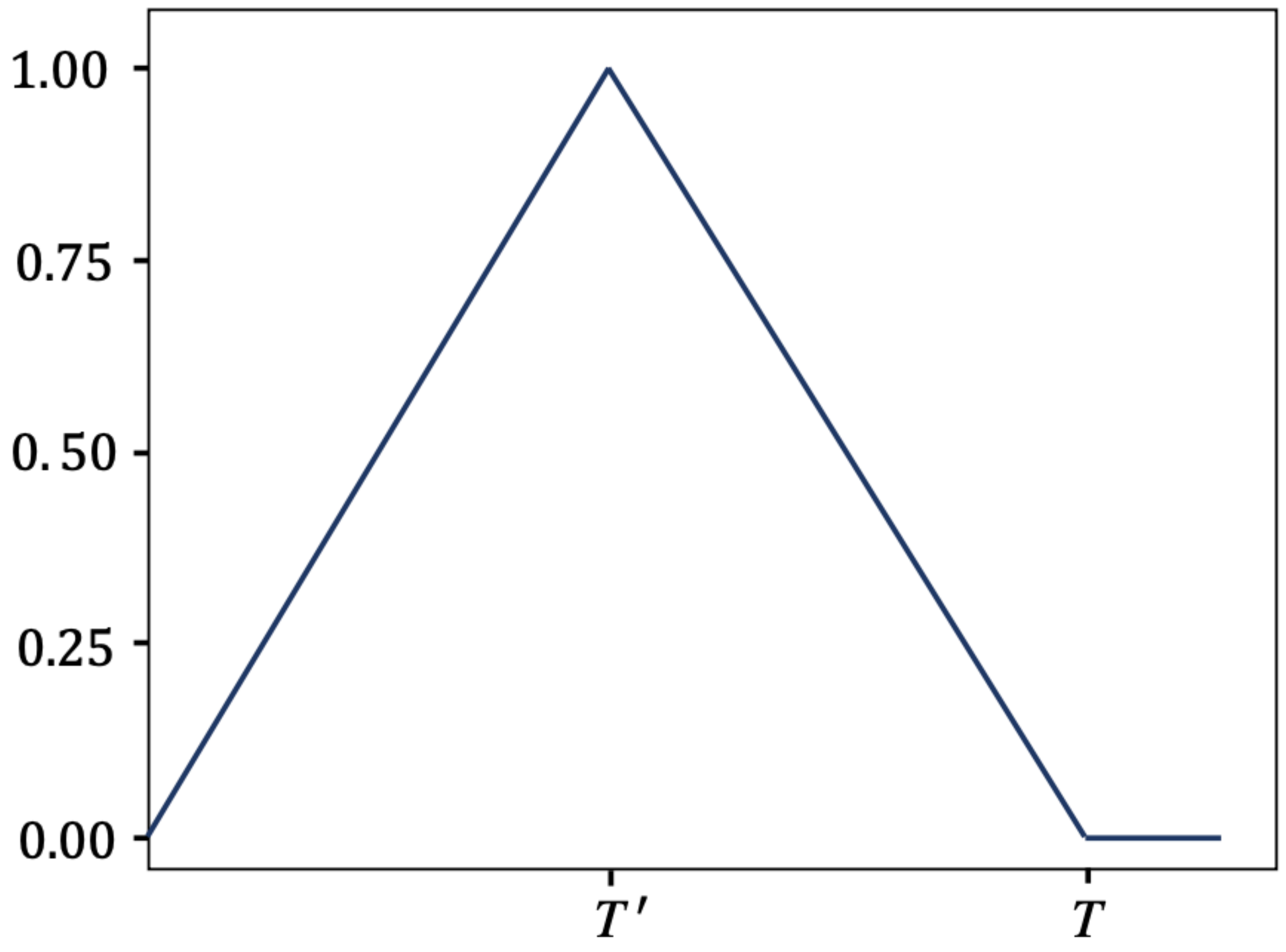
- Concerted training framework
    - Rate-scheduled Learning
    - Dynamic Switch
    - Asymptotic Distillation



Towards Making Most of BERT for NMT, [Yang et al AAAI 2020]

Not tuning too much

- Rate-scheduled Learning rate
  - Gradually increase the learning rate of BERT parameters from 0 to 1
  - Then, decrease the learning rate of BERT parameters from 1 to 0
  - Keep the BERT parameters frozen



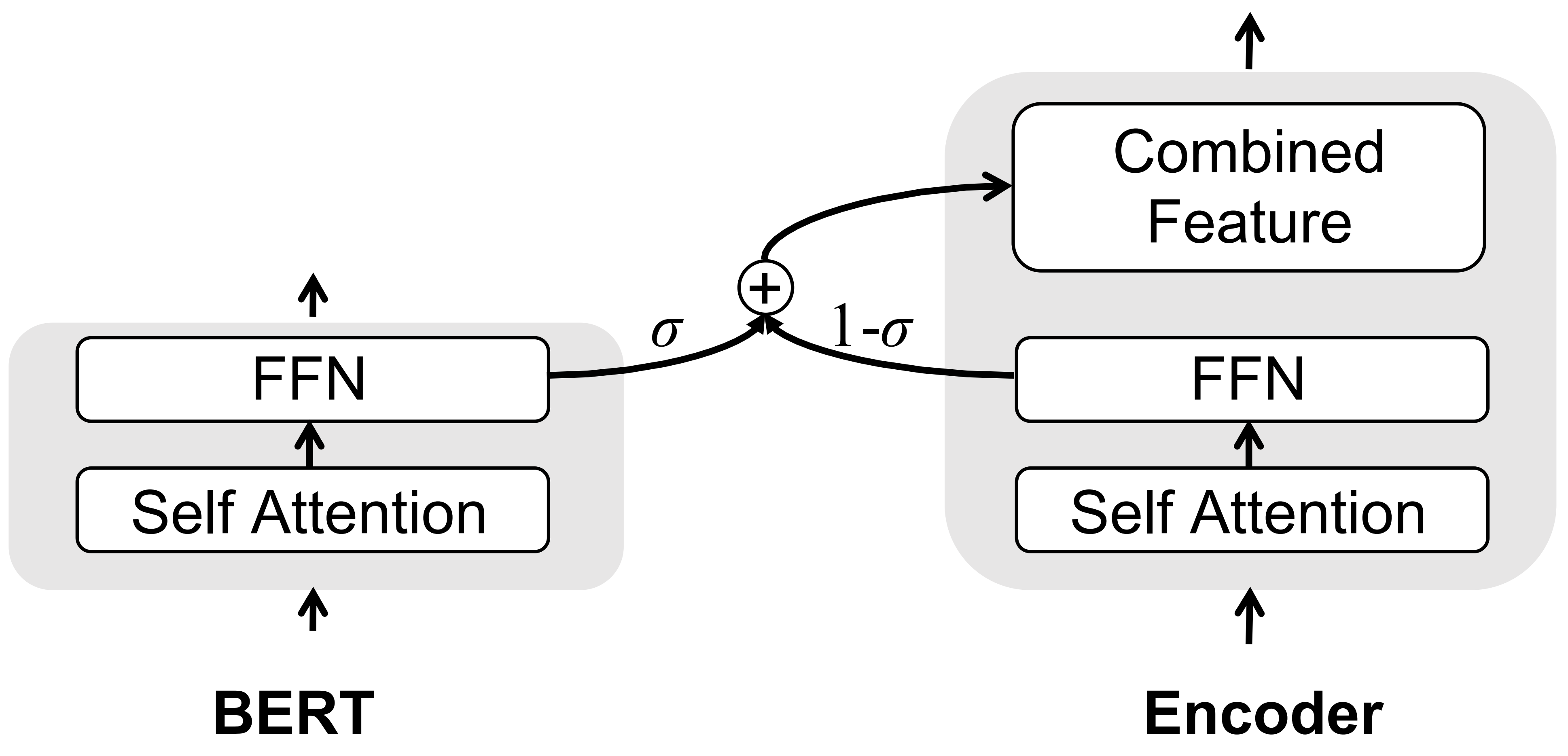
Learning rate scalar for BERT parameter

Rate-scheduled learning rate is actually a trade off between fine-tuning and BERT frozen

Towards Making Most of BERT for NMT, [Yang et al AAAI 2020]

Not tuning too much

- Dynamic Switch
  - Use a gate to dynamically decide which part is more important
  - If σ is learned to 0, it degrade to the NMT model
  - If σ is learned to 1, it simply act as Bert fine-tune approach

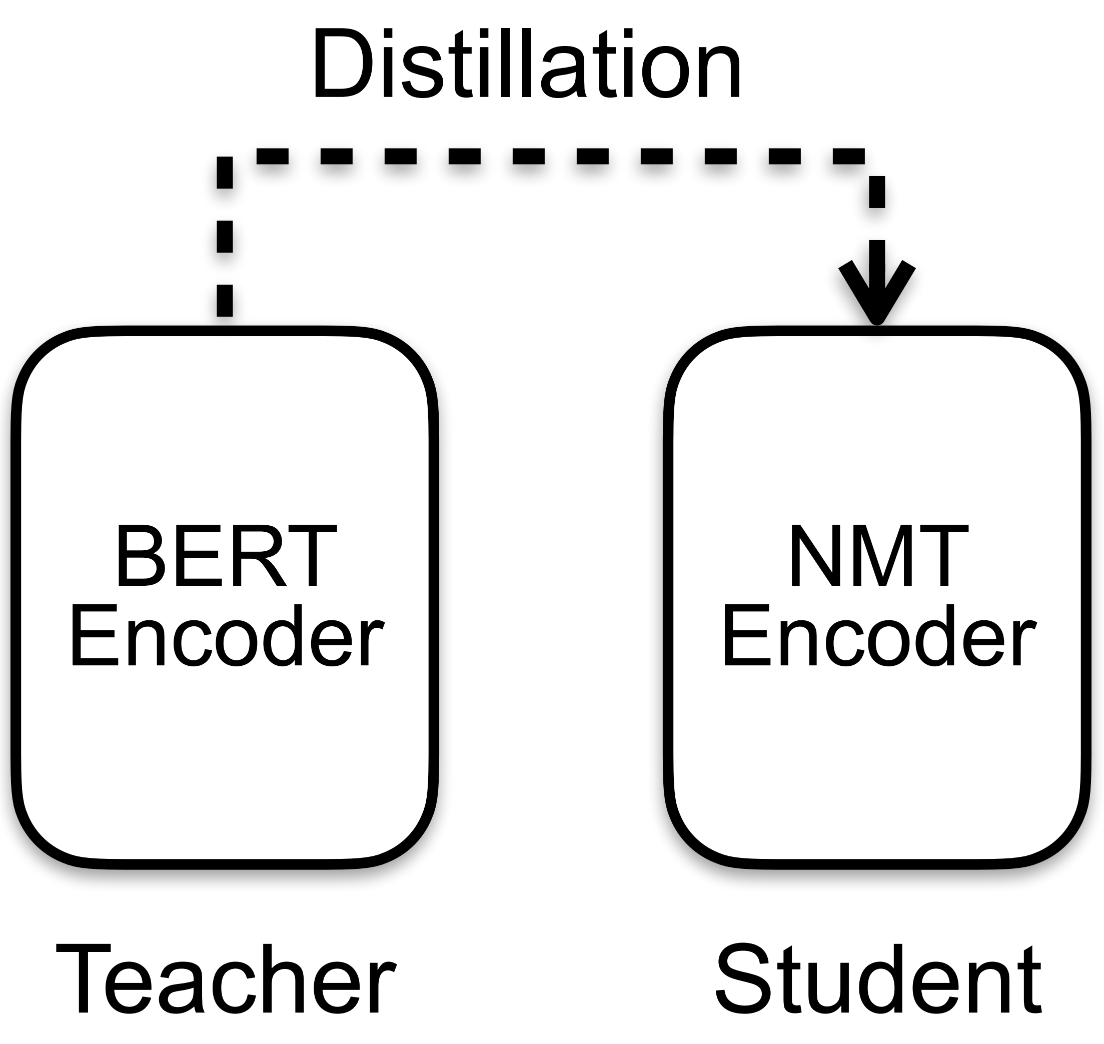


Dynamic Switch is more flexible than rate-scheduled learning rate

Towards Making Most of BERT for NMT, [Yang et al AAAI 2020]

Not tuning too much

- Asymptotic Distillation
  - The pre-trained BERT serves as a teacher network while the encoder of the NMT model serves as a student
  - Minimize MSE loss of hidden states between NMT encoder and BERT to retain the pre-trained information
  - Use a hyper-parameter to balances the preference between pre-training distillation and NMT objective



$$\mathcal{L}_{KD} = \| h_{\text{bert}} - h_{\text{nmt}} \|^2$$

Distillation Without introducing of additional parameters!

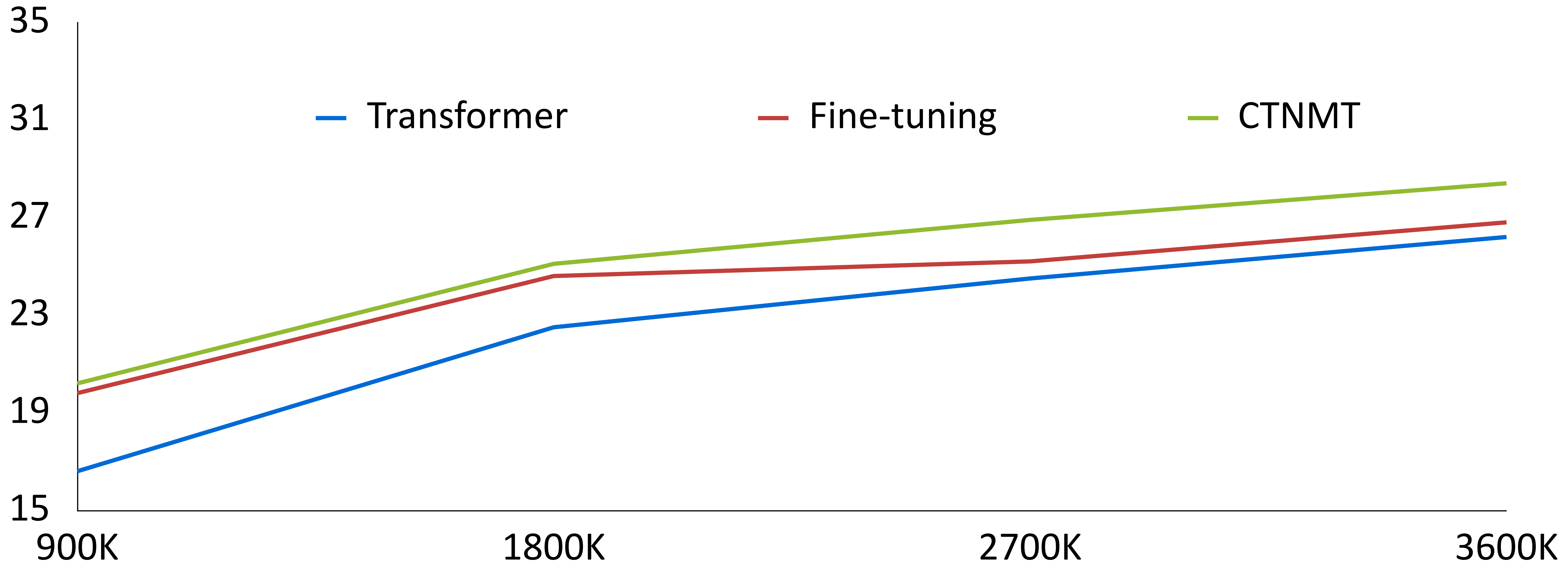
Not tuning too much

System	Architecture	En-De	En-Fr	En-Zh
	Existing systems			
Vaswani et al. (2017)	Transformer base	27.3	38.1	-
Vaswani et al. (2017)	Transformer big	28.4	41.0	-
Lample and Conneau (2019)	Transformer big + Fine-tuning	27.7	-	-
Lample and Conneau (2019)	Transformer big + Frozen Feature	28.7	-	-
Chen et al. (2018)	RNMT+ + MultiCol	28.7	41.7 -	
	Our NMT systems			
CTNMT	Transformer (base)	27.2	41.0	37.3
CTNMT	Rate-scheduling	29.7	41.6	38.4
CTNMT	Dynamic Switch	29.4	41.4	38.6
CTNMT	Asymptotic Distillation	29.2	41.6	38.3
CTNMT	+ ALL	**30.1**	**42.3**	**38.9**

- Three strategies can independently work well on WMT14 En-De, En-Fr and WMT18 En-Zh
- CTNMT base model achieves even better results than Transformer big model

Towards Making Most of BERT for NMT, [Yang et al AAAI 2020]

Not tuning too much



- CTNMT outperforms fine-tuning on all training steps
- The performance gaps is enlarged as the fine-tuning steps increasing

Towards Making Most of BERT for NMT, [Yang et al AAAI 2020]

Summary

- Advantage
  - Simple and effective, obtains +3 BLEU on WMT14 en-de benchmark
  - Three methods can be used separately or jointly
- Limitation
  - Introducing pre-training method for decoder is promising but still difficult
  - Cross attention is import but not pre-trained

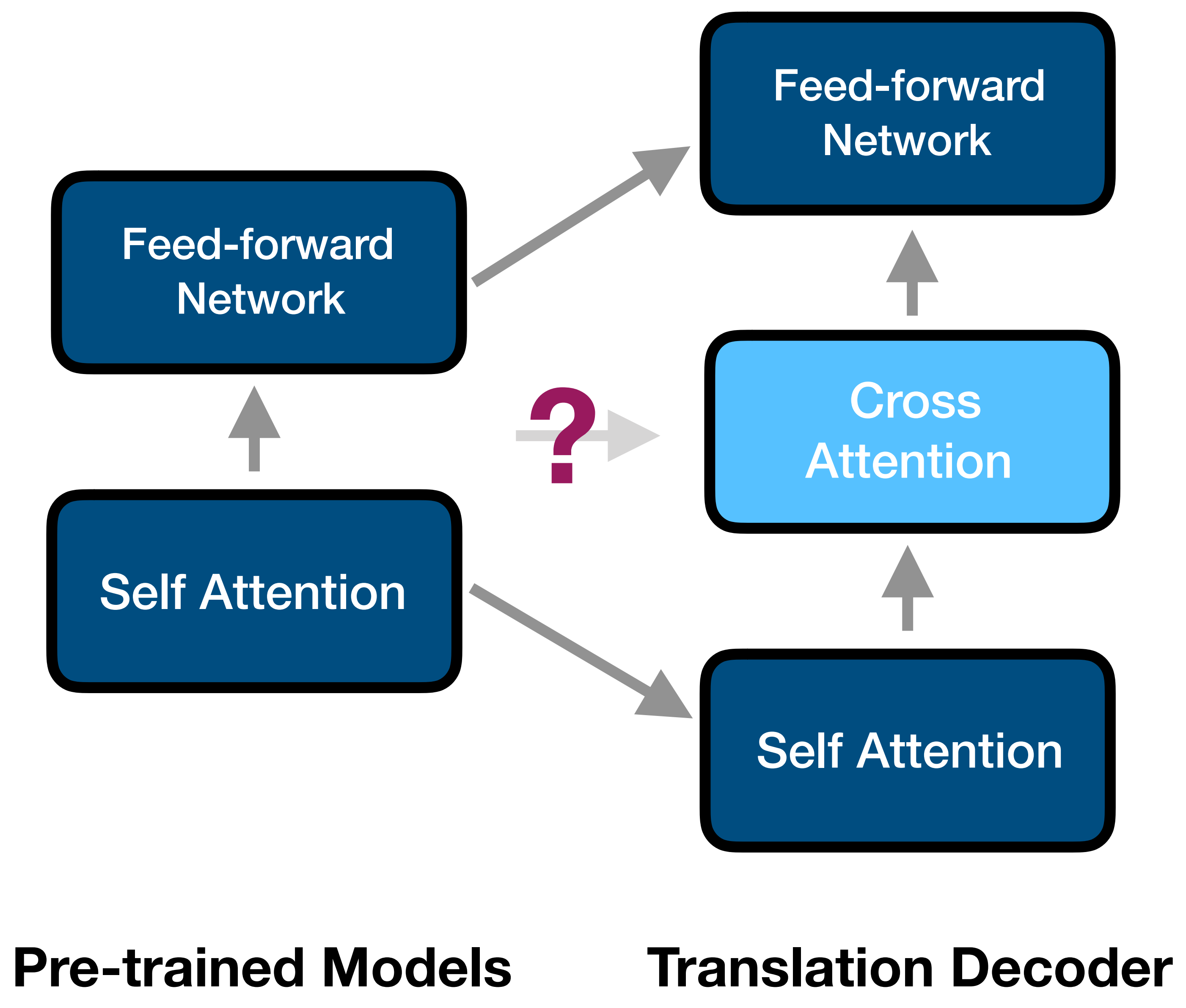
Models	En→De BLEU
BERT Enc	29.2
BERT Dec	26.1
GPT-2 Enc	27.7
GPT-2 Dec	27.4

	Encoder	Decoder
GPT	✗	✗
BERT	✓	✗

Towards Making Most of BERT for NMT, [Yang et al AAAI 2020]

Decode has cross attention but not GPT

- Cross attention plays a crucial role in NMT
- Pre-trained language models, such as BERT and GPT, have none
- This mismatch between the generation models and conditional generation models makes the pre-trained model usage for translation decoder pretty tricky

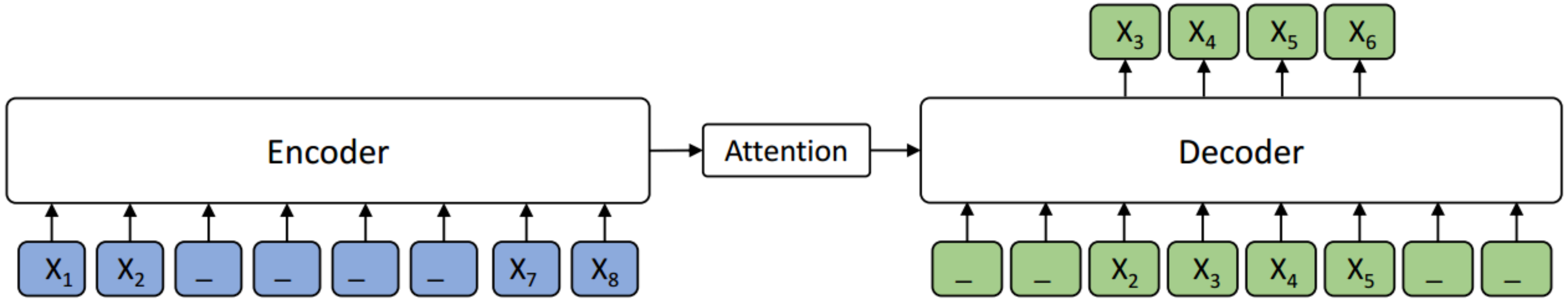


PART2: Monolingual Pre-training for NMT

- The Bronze Age
  - NMT initialized with word2vec [ACL 2017, NAACL 2018, AI 2020]
  - NMT initialized with language model [EMNLP 2017]
- BERT fusion
  - BERT Incorporating methods [ICLR 2020, AAAI 2020a]
  - BERT Tuning methods [AAAI 2020b]
- Unified sequence to sequence pre-training
  - MASS: Masked Sequence-to-Sequence Pre-training [ICML 2019]
  - BART: Denoising Sequence-to-Sequence Pre-training [ACL 2020]

MASS: Pre-train for Sequence to Sequence Generation

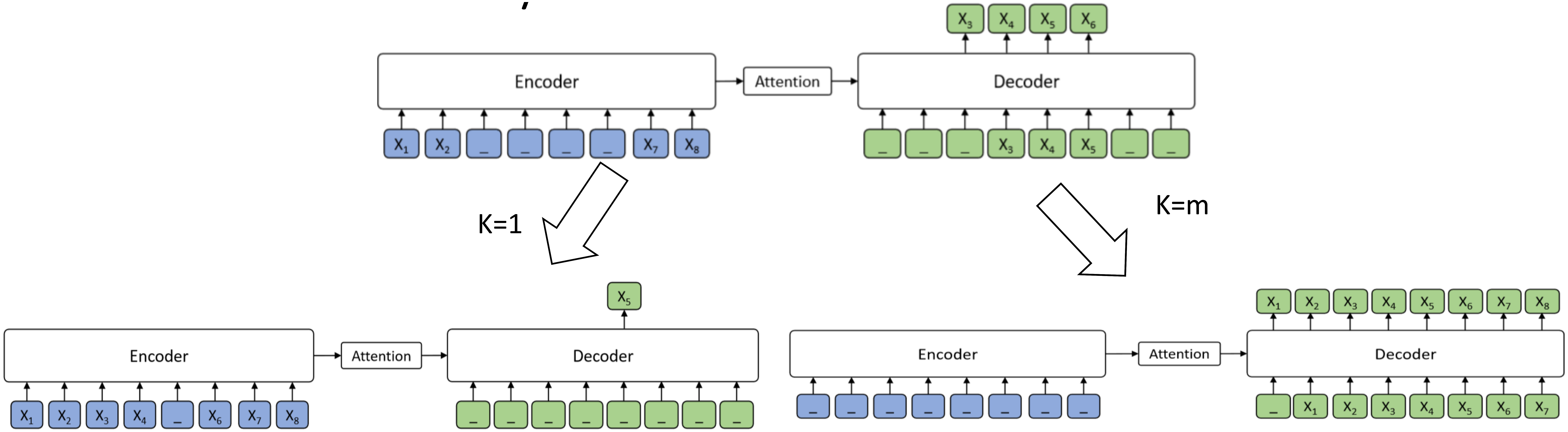
- MASS is carefully designed to jointly pre-train the encoder and decoder



- 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

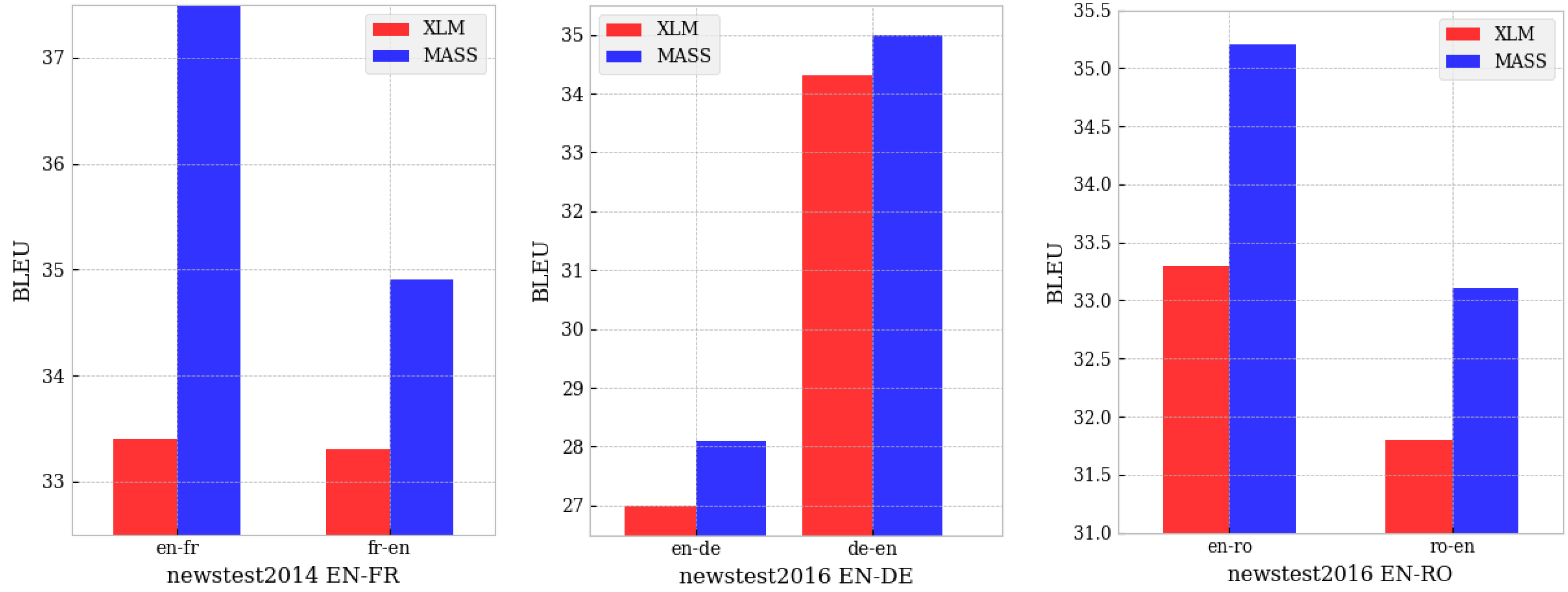


Length	Probability	Model
$k = 1$	$P(x^u \mid x^{\setminus u}; \theta)$	masked LM in BERT
$k \in [1, m]$	$P(x^{u:v} \mid x^{\setminus u:v}; \theta)$	MASS

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

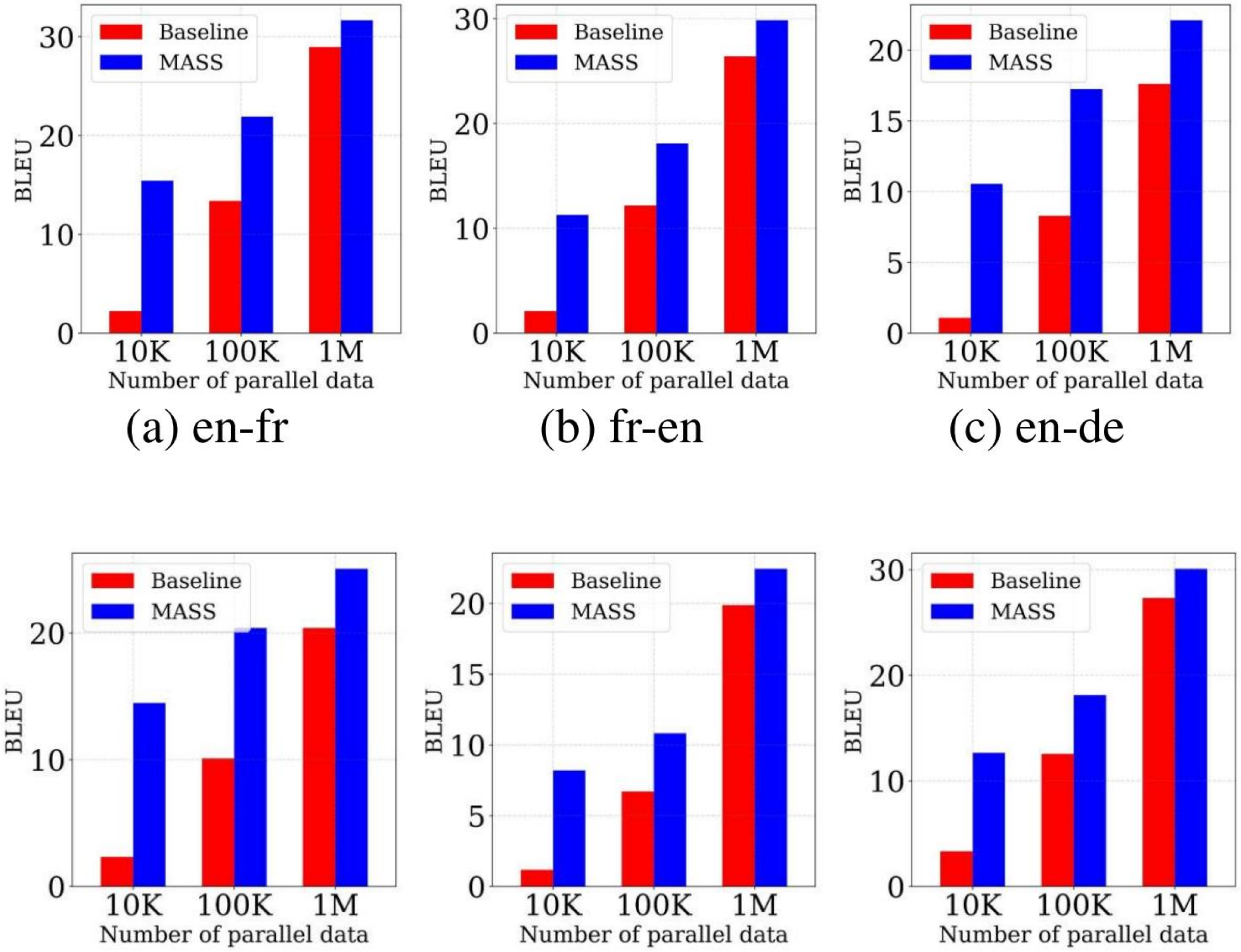
MASS: Pre-train for Sequence to Sequence Generation, [Song et al ICML 2019]

Unsupervised NMT



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

Low-resource NMT



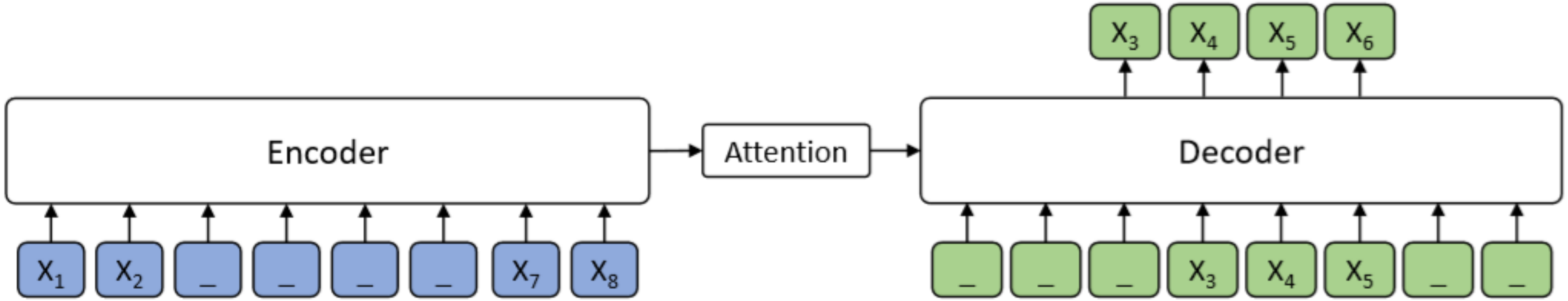
(a) en-fr (b) fr-en (c) en-de

(d) de-en (e) en-ro (f) ro-en

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

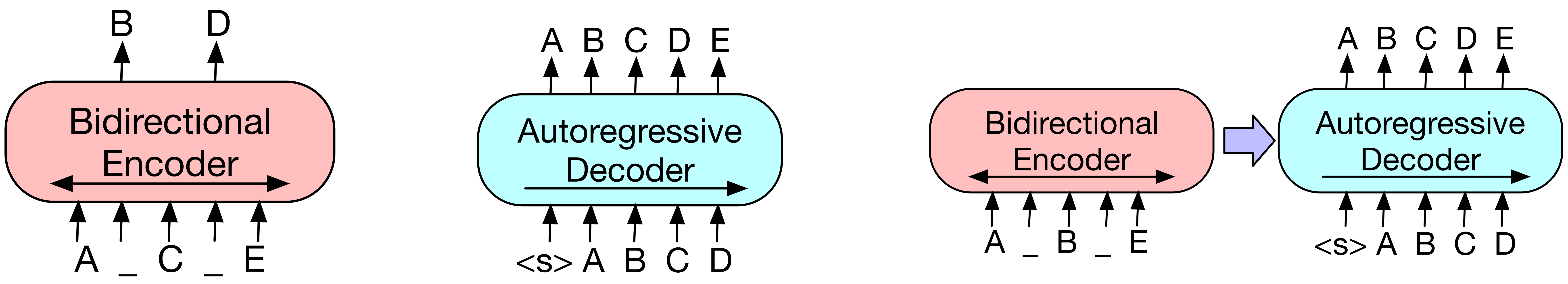
Summary

- Advantages
  - Unified sequence-to-sequence pretraining which jointly pretrains encoder, decoder and cross attention
  - Achieves improvements on zero-shot / unsupervised NMT
- Limitions
  - No experiments on rich resource NMT
  - Pretraing 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

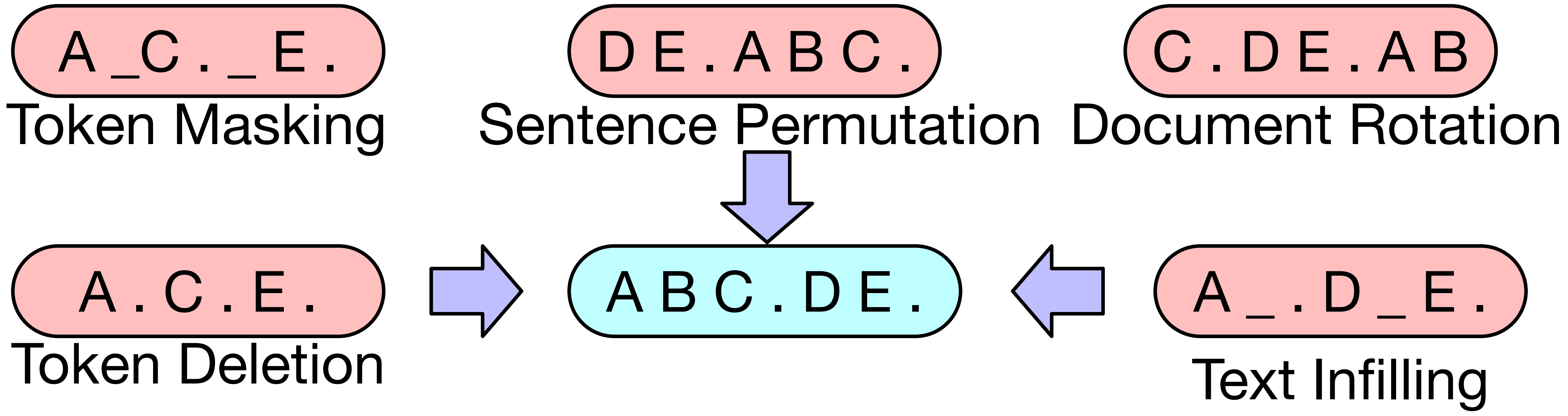


A schema comparison with BERT, GPT and BART.

- 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]

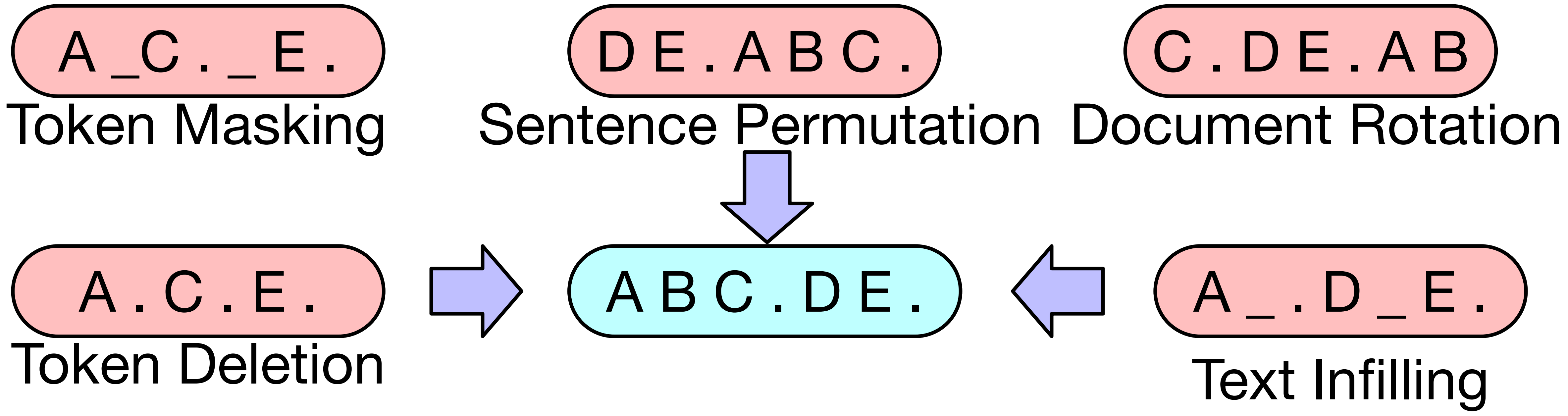
Noising the input



- 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]

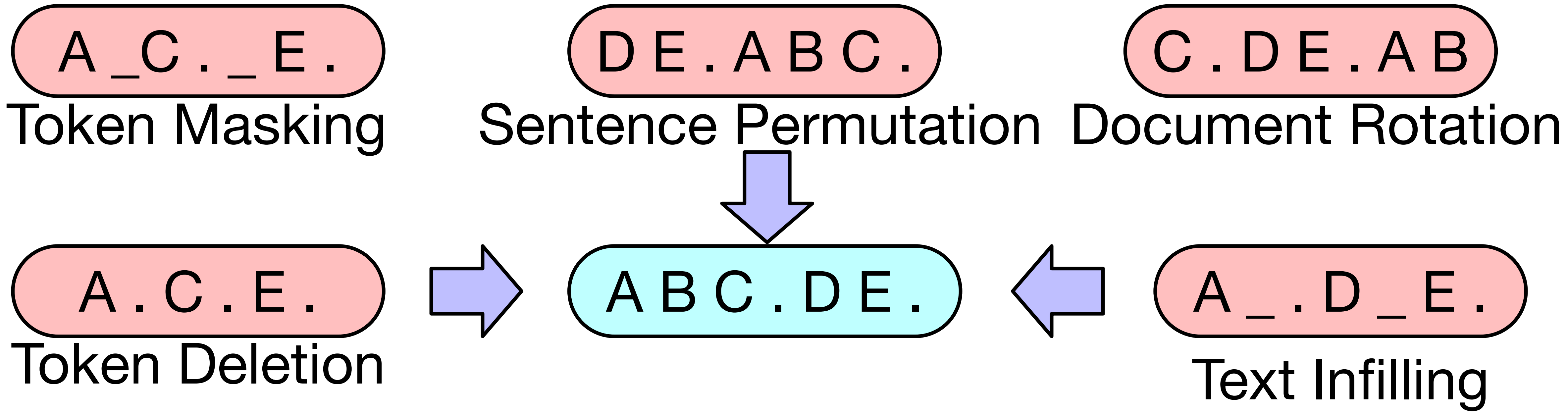
Noising the input



- 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]

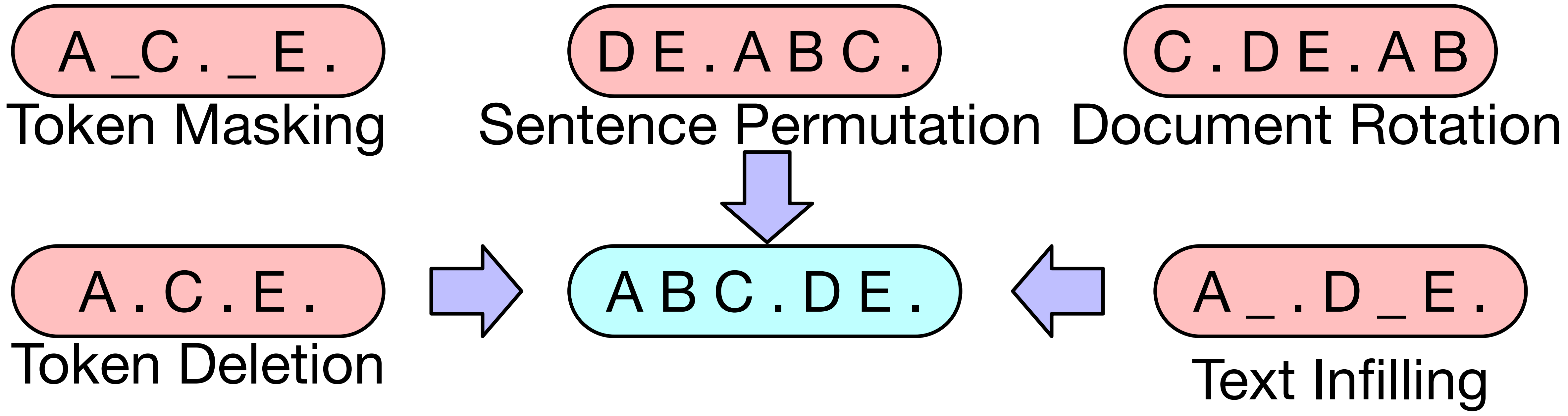
Noising the input



- 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]

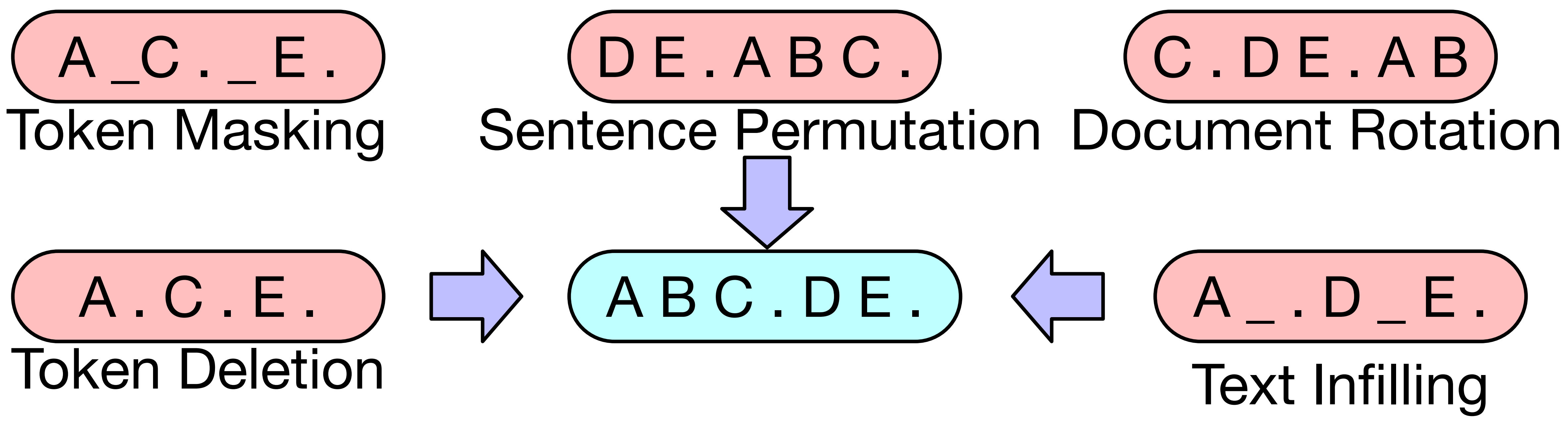
Noising the input



- 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]

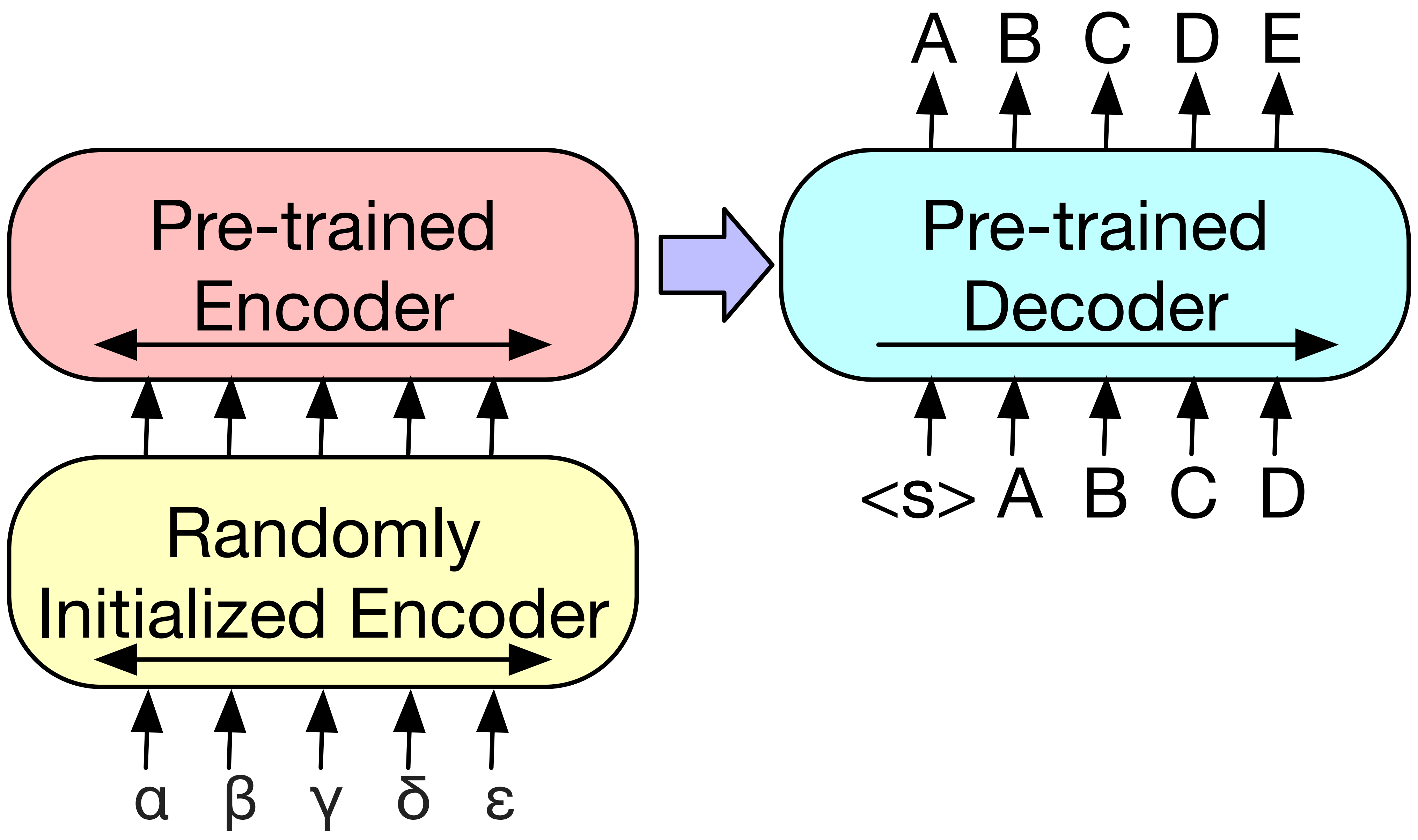
Noising the input



- 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]

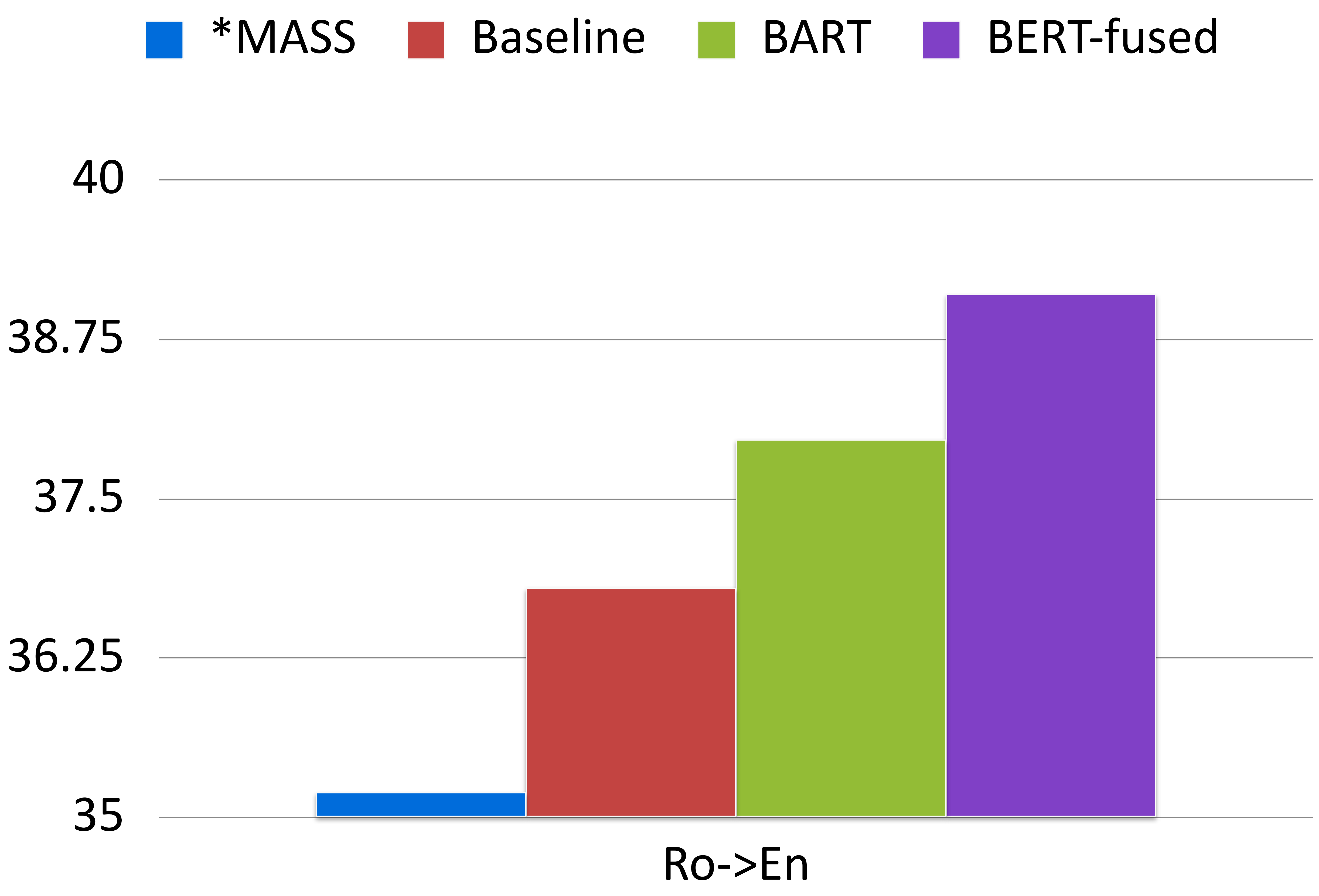
Fine-Tune on Neural Machine Translation



- 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]

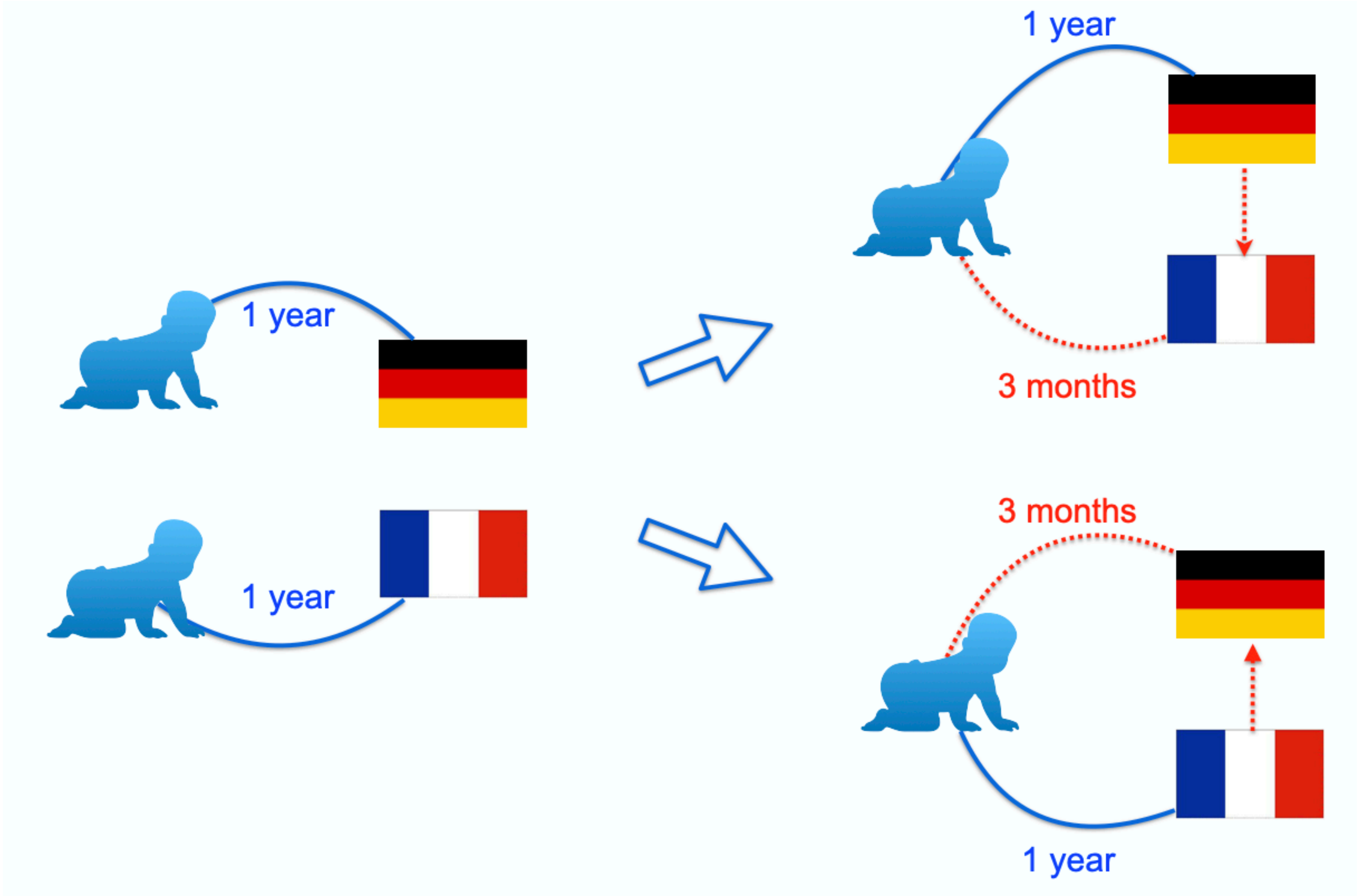
PART III: Multilingual Pre-training for NMT

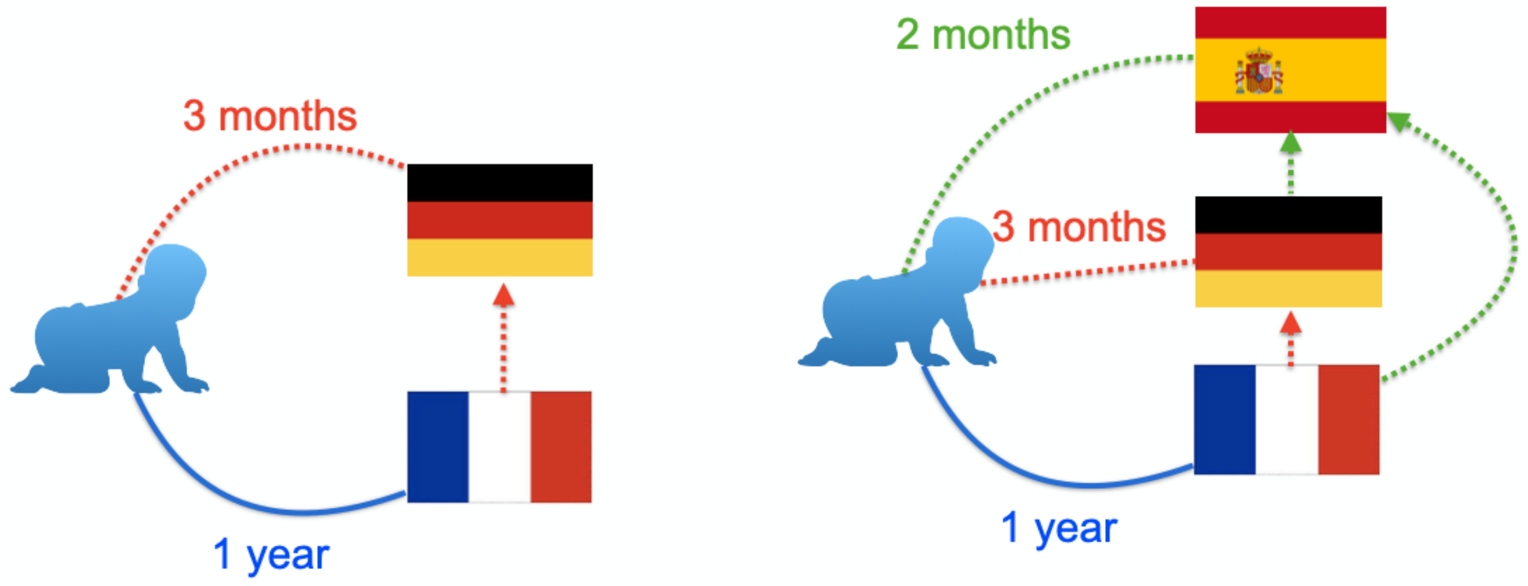
PART 3: Multilingual Pre-training for NMT

- 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]

Multi-lingual Pre-training for NMT

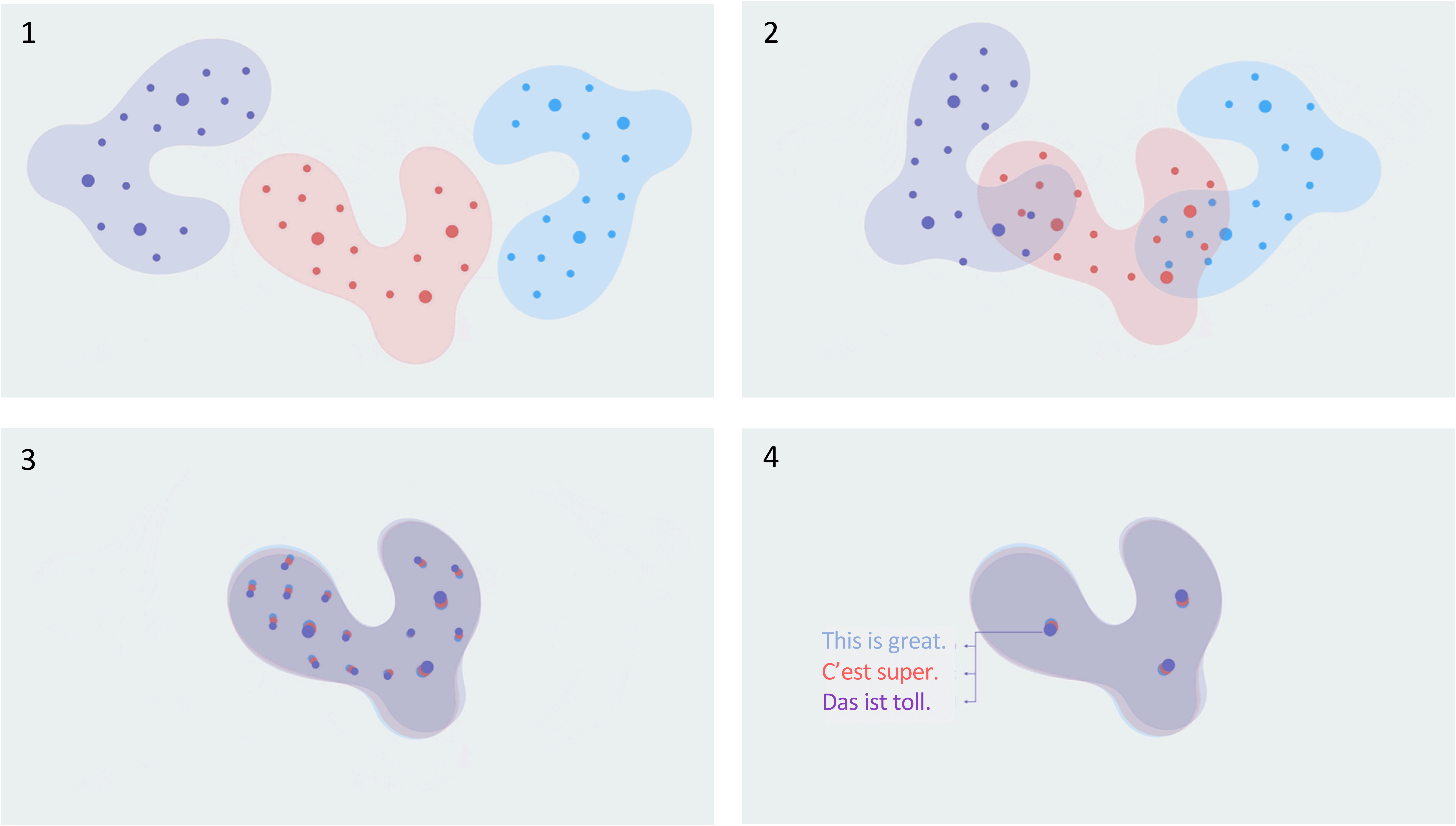
- Data scarcity for low/zero resource languages.
- Transfer knowledge between languages.





Cross-lingual Language Model Pretraining

Learning cross-lingual representation

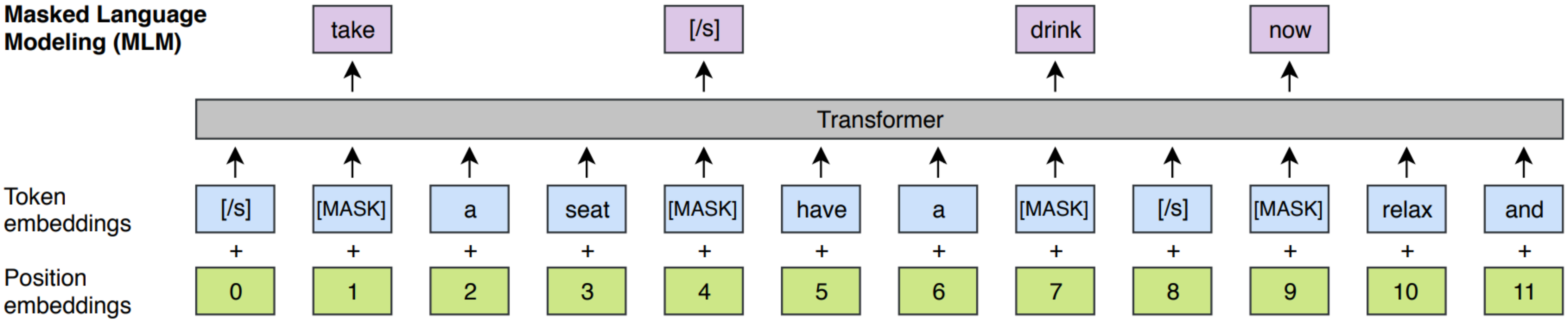


Cross-lingual Language Model Pre-training, [Conneau et al NeurIPS 2019]

Multiple masked language model (MLM)

Similar to BERT, but in many languages…

Multilingual representations emerge from a single model trained on many languages

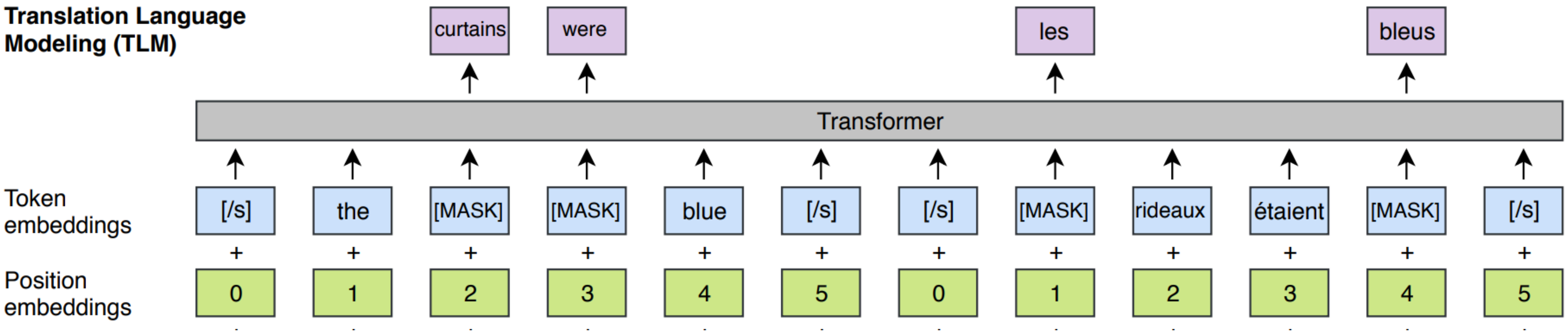


Multilingual Masked language modeling pretraining

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



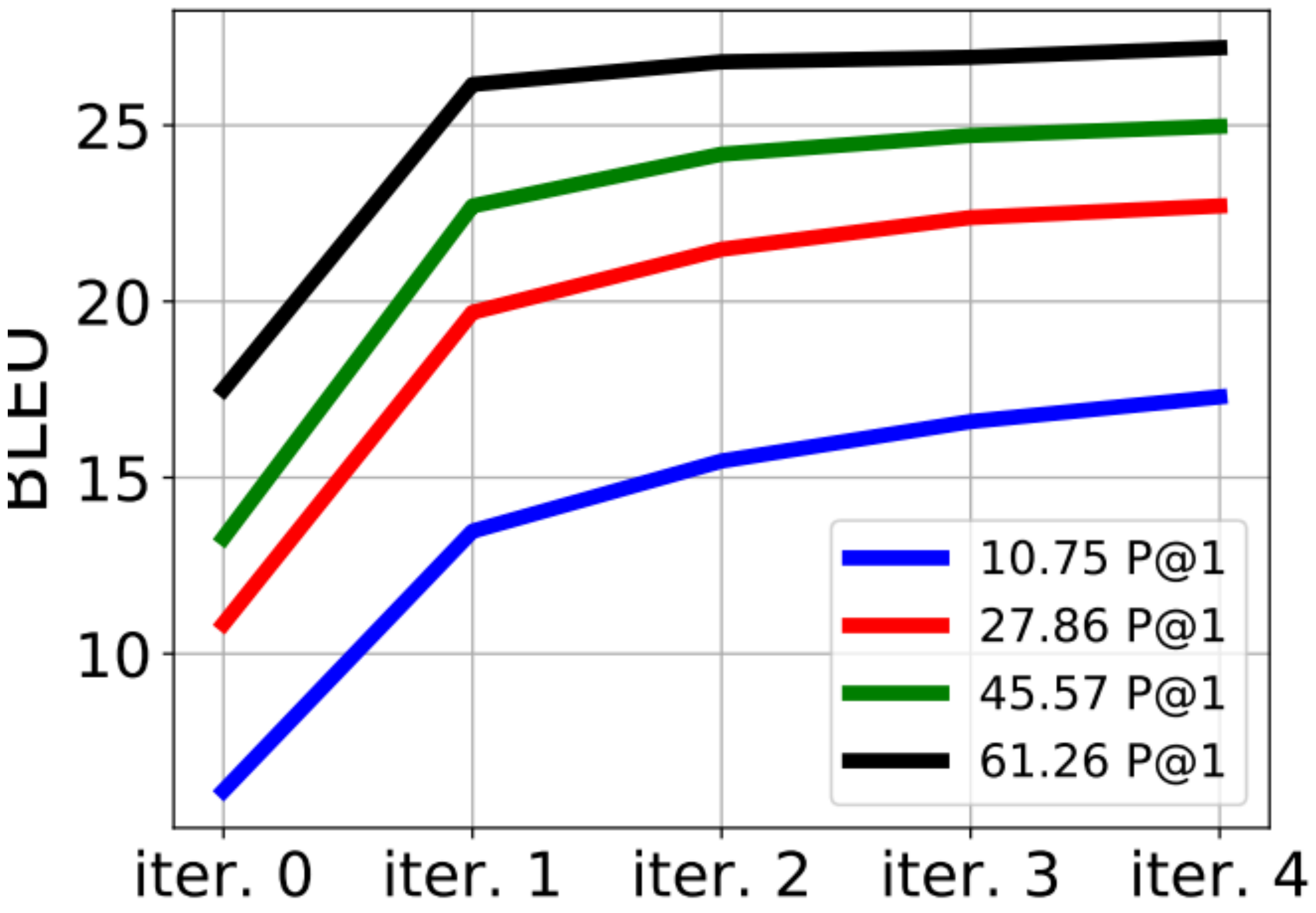
Translation language modeling (TLM) pretraining

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

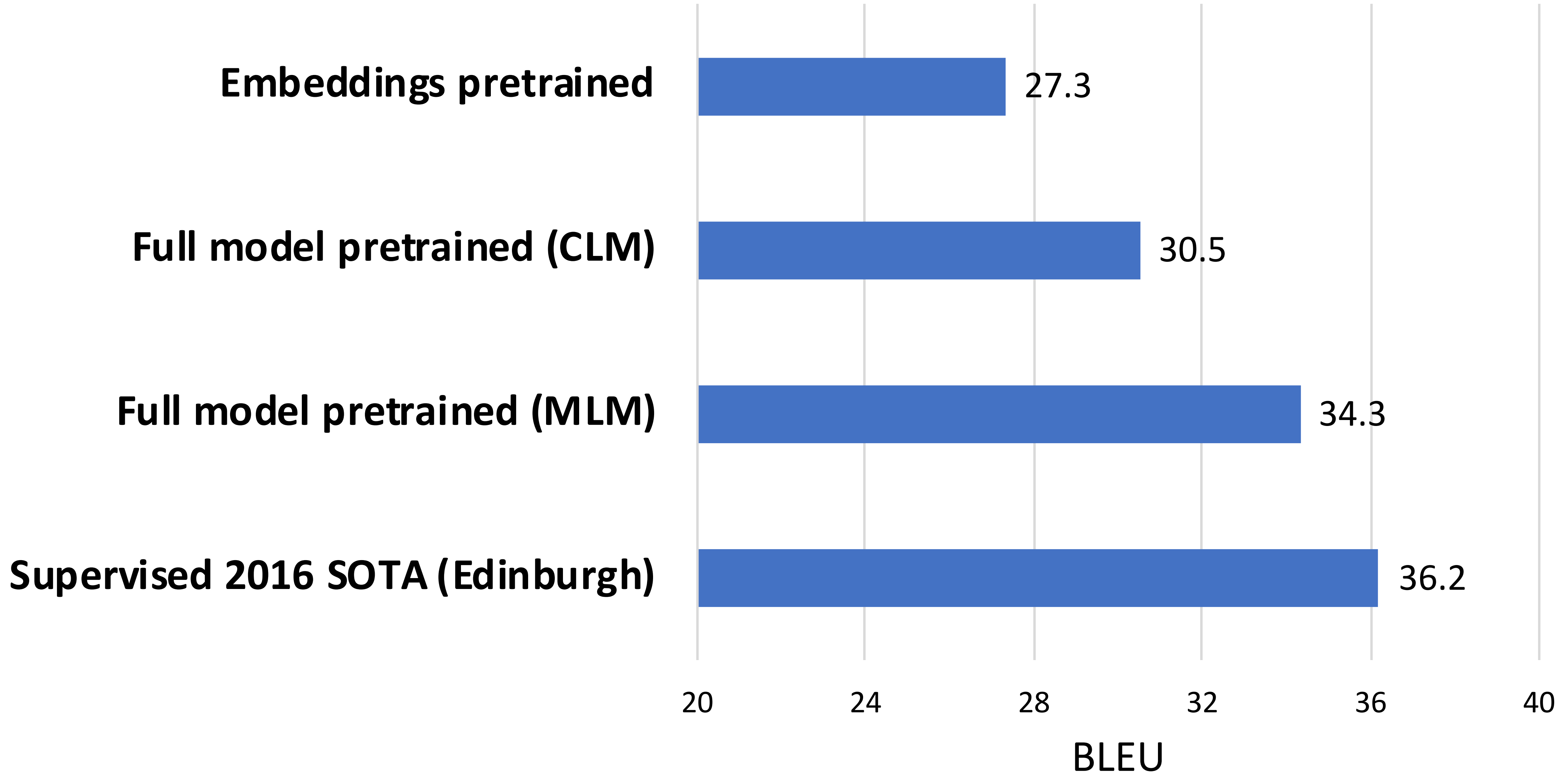
Results on Unsupervised Machine Translation

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

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



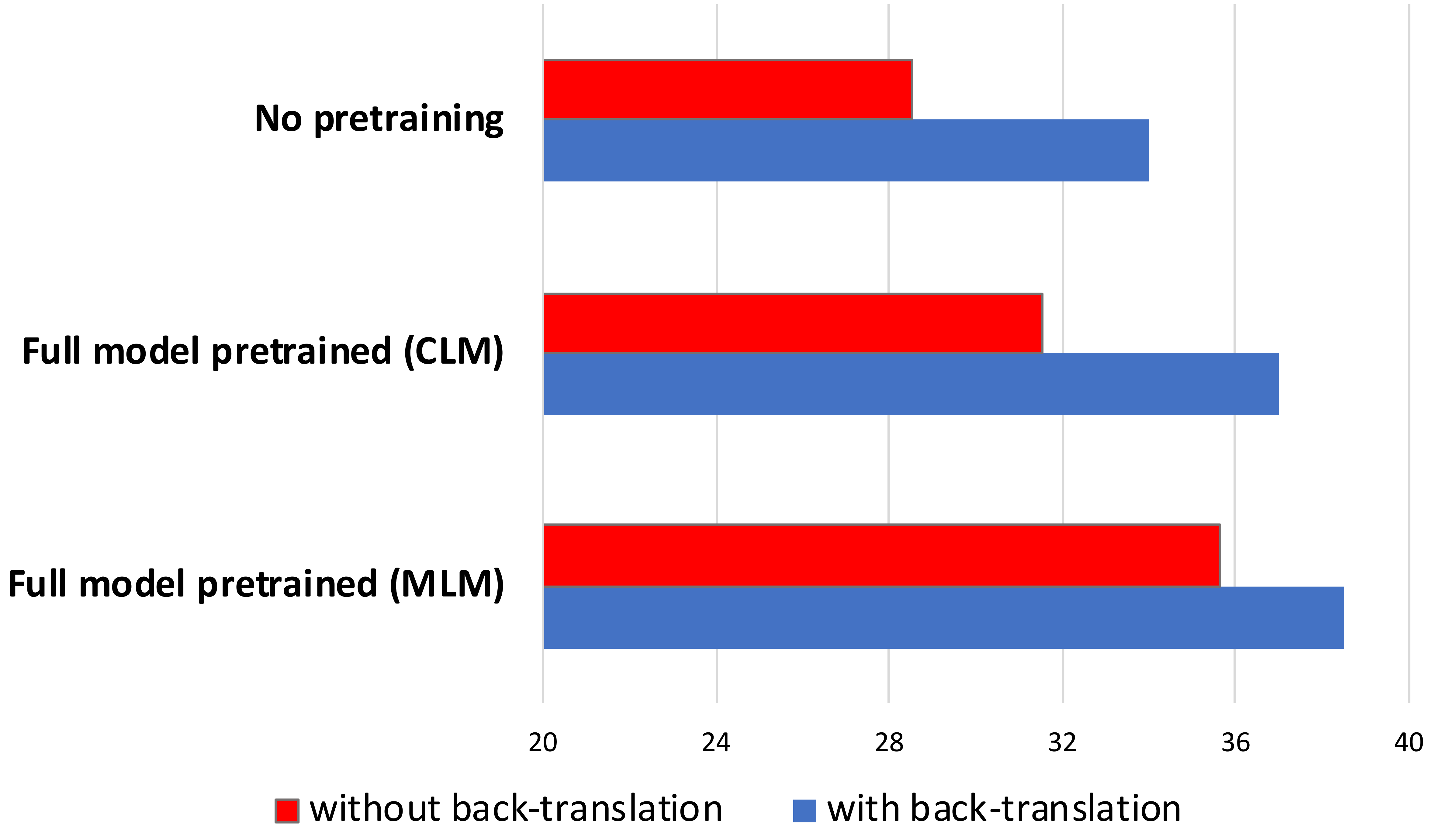
Full Transformer model initialization
significantly improves performance (+7 BLEU)



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

Results on supervised machine translation

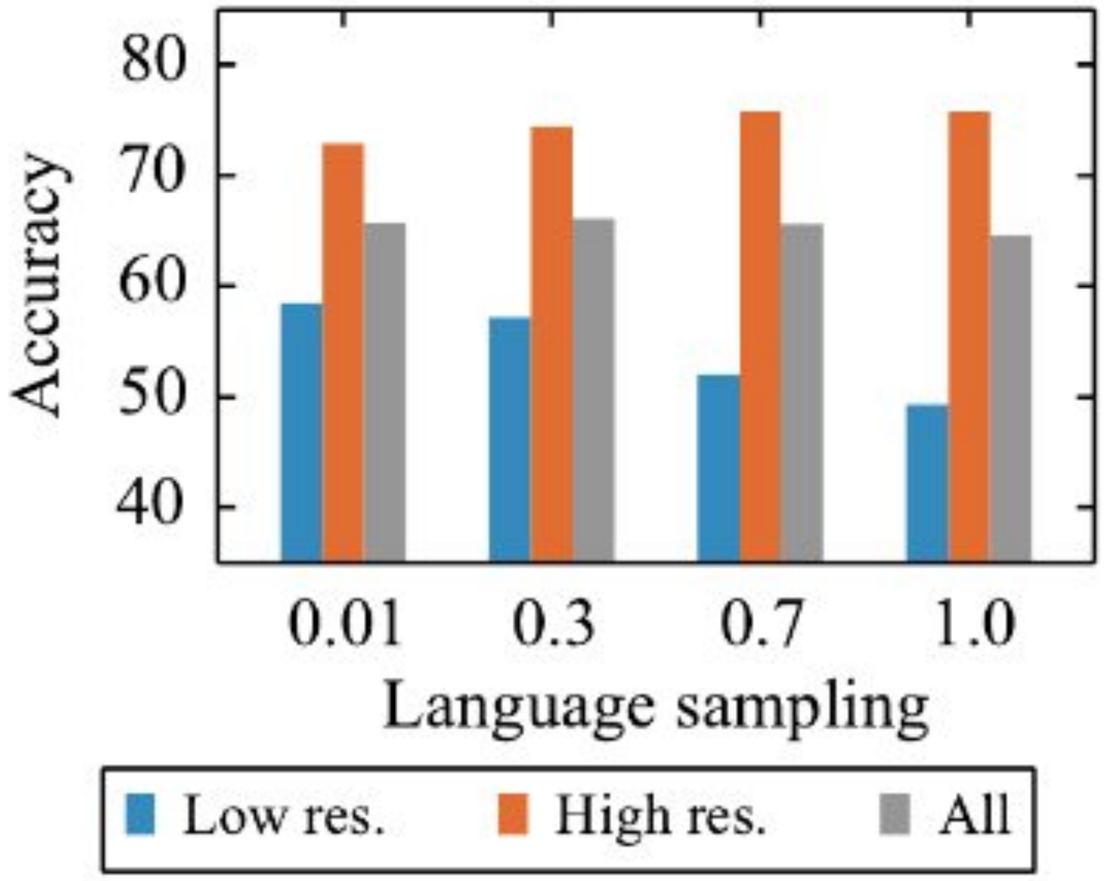
- Pre-training is important for translation
  - Pre-training both encoder and decoder improves
  - MLM is better than CLM
  - Back translation + Pre-training achieve the best



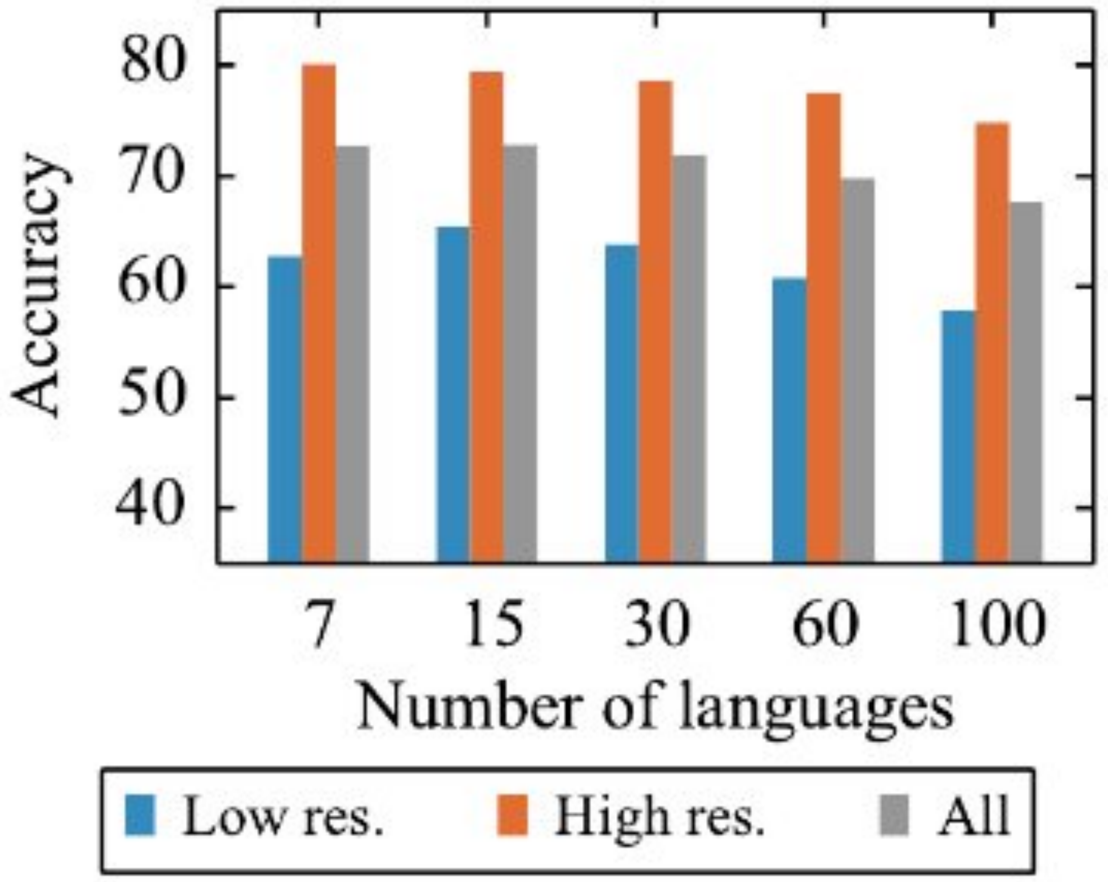
Cross-lingual Language Model Pre-training, [Conneau et al NeurIPS 2019]

Ablation study

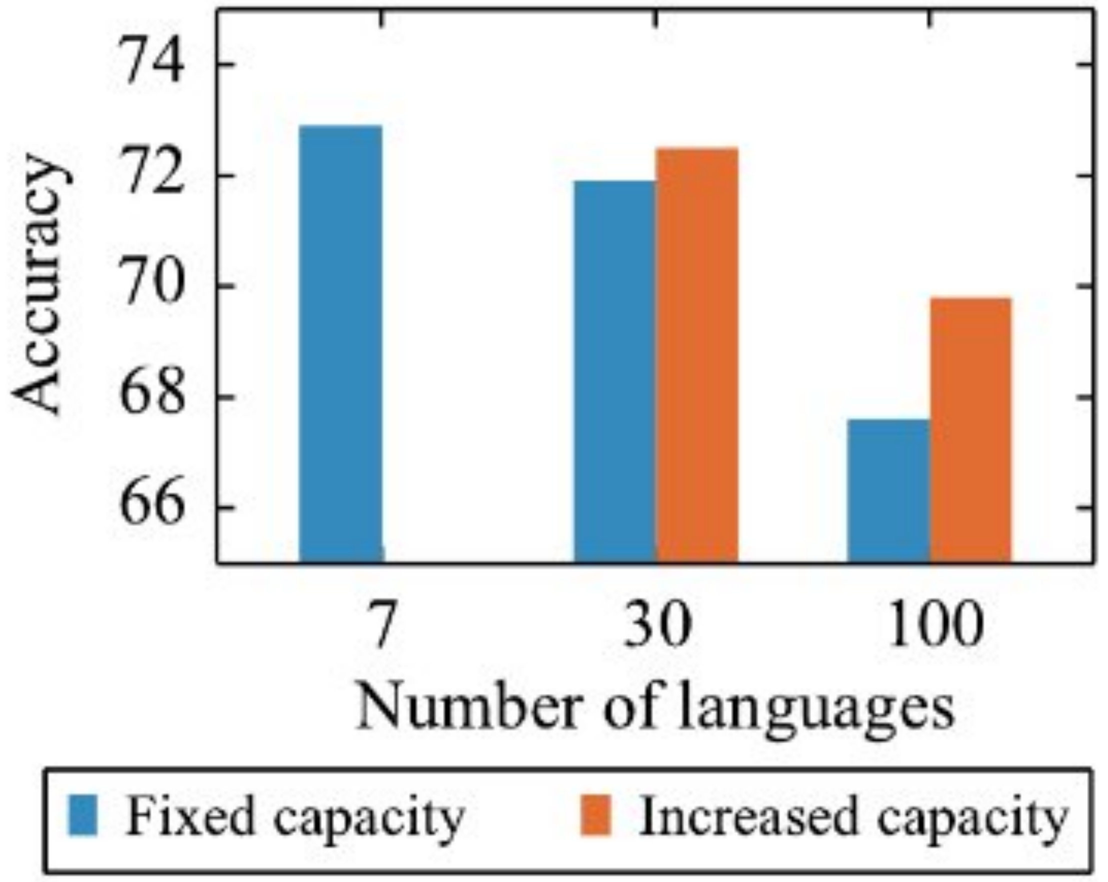
- Adding more languages improves performance on low-resource languages due to positive knowledge transfer
- Sampling batches more often in some languages improves performance in these languages but decrease performance in other languages (capacity allocation problem)



High-res/low-res trade-off



The transfer-interference trade-off

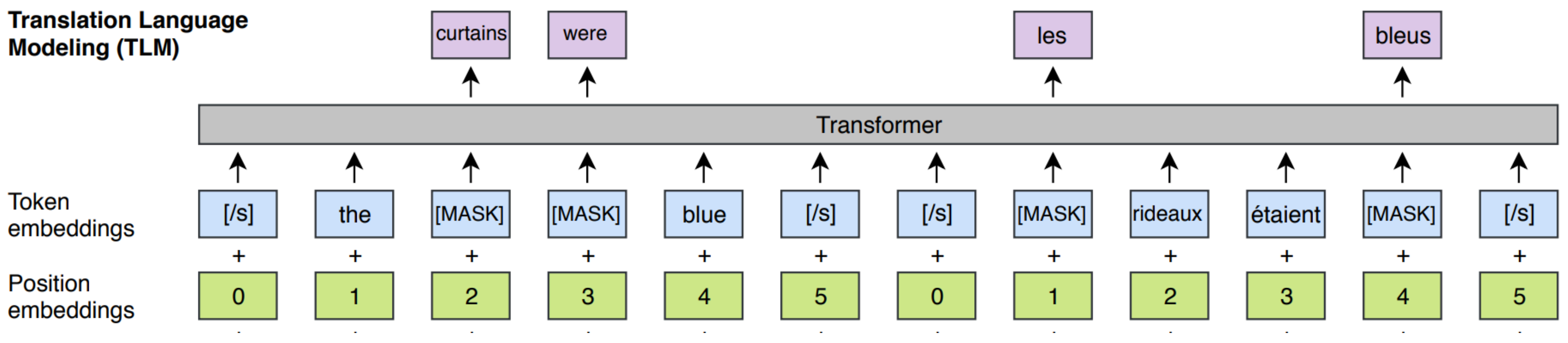


The curse of multilinguality

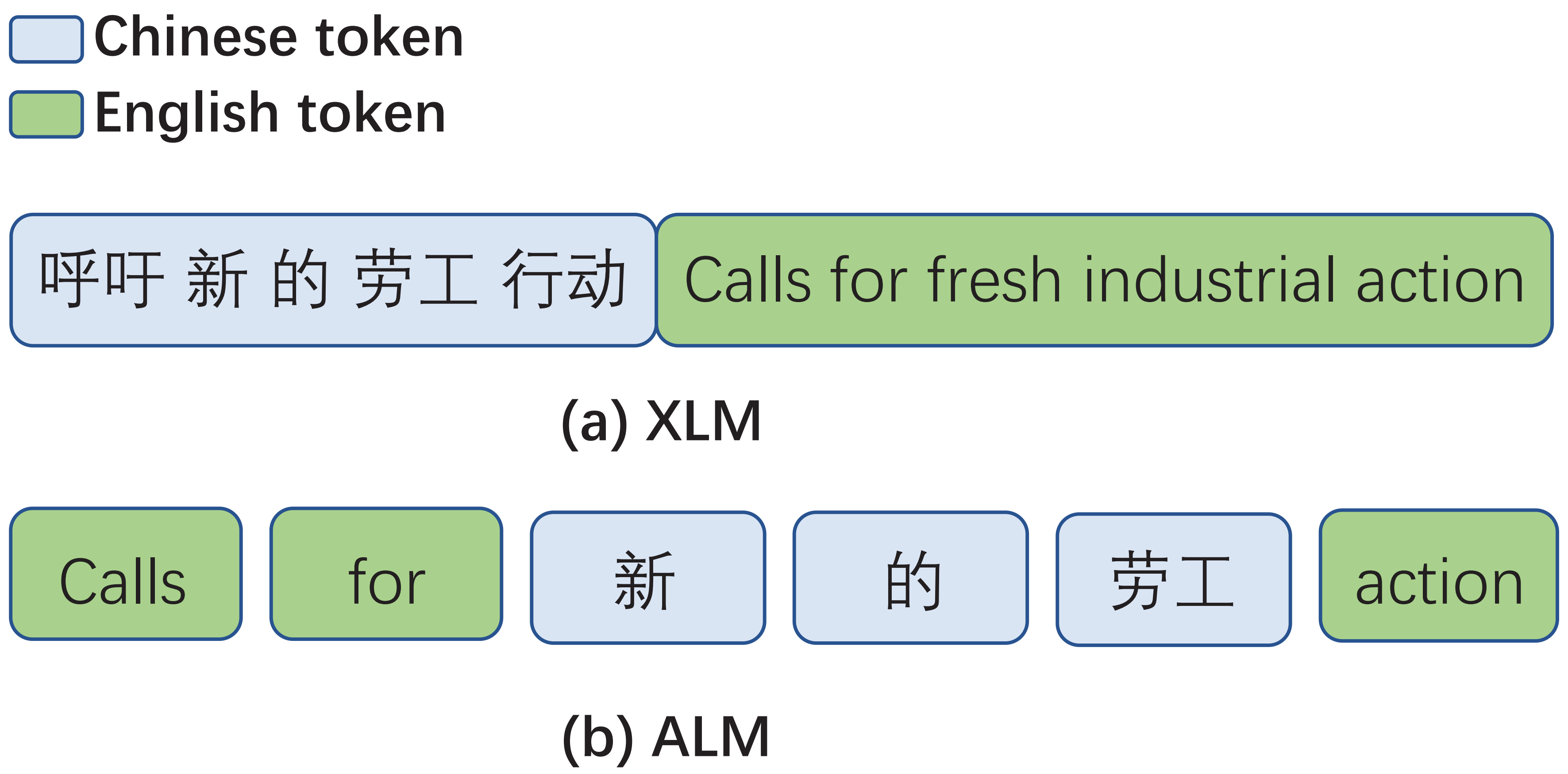
Summary

- 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

Alternating Language Modeling for Cross-Lingual Pre-Training



Translation language modeling (TLM) pretraining

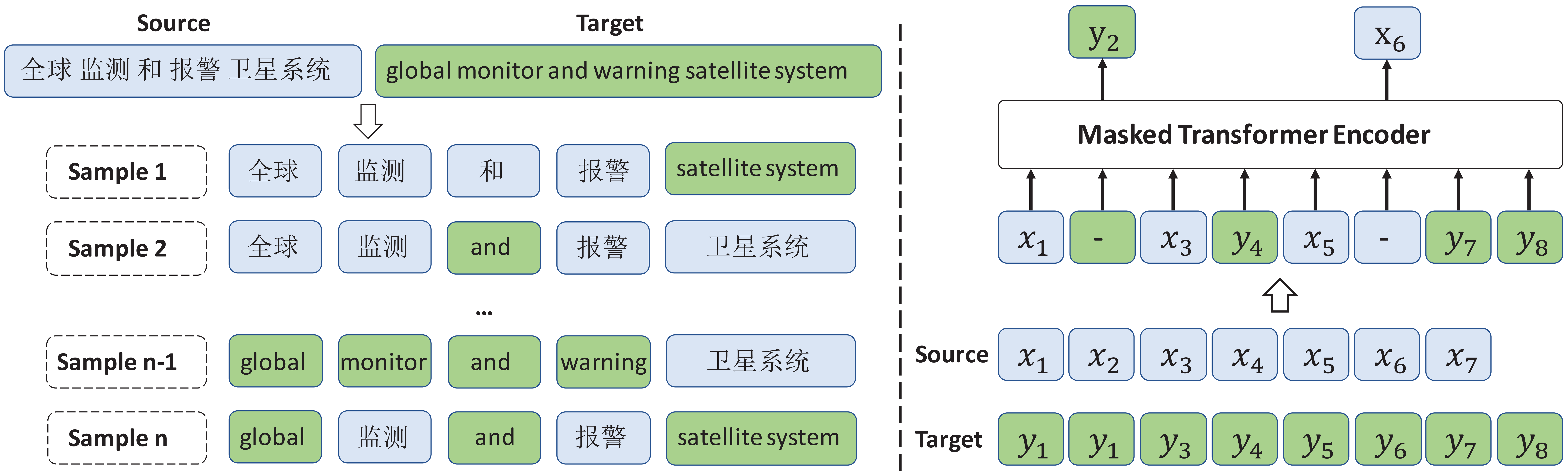


Sentence level mixing

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



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

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.

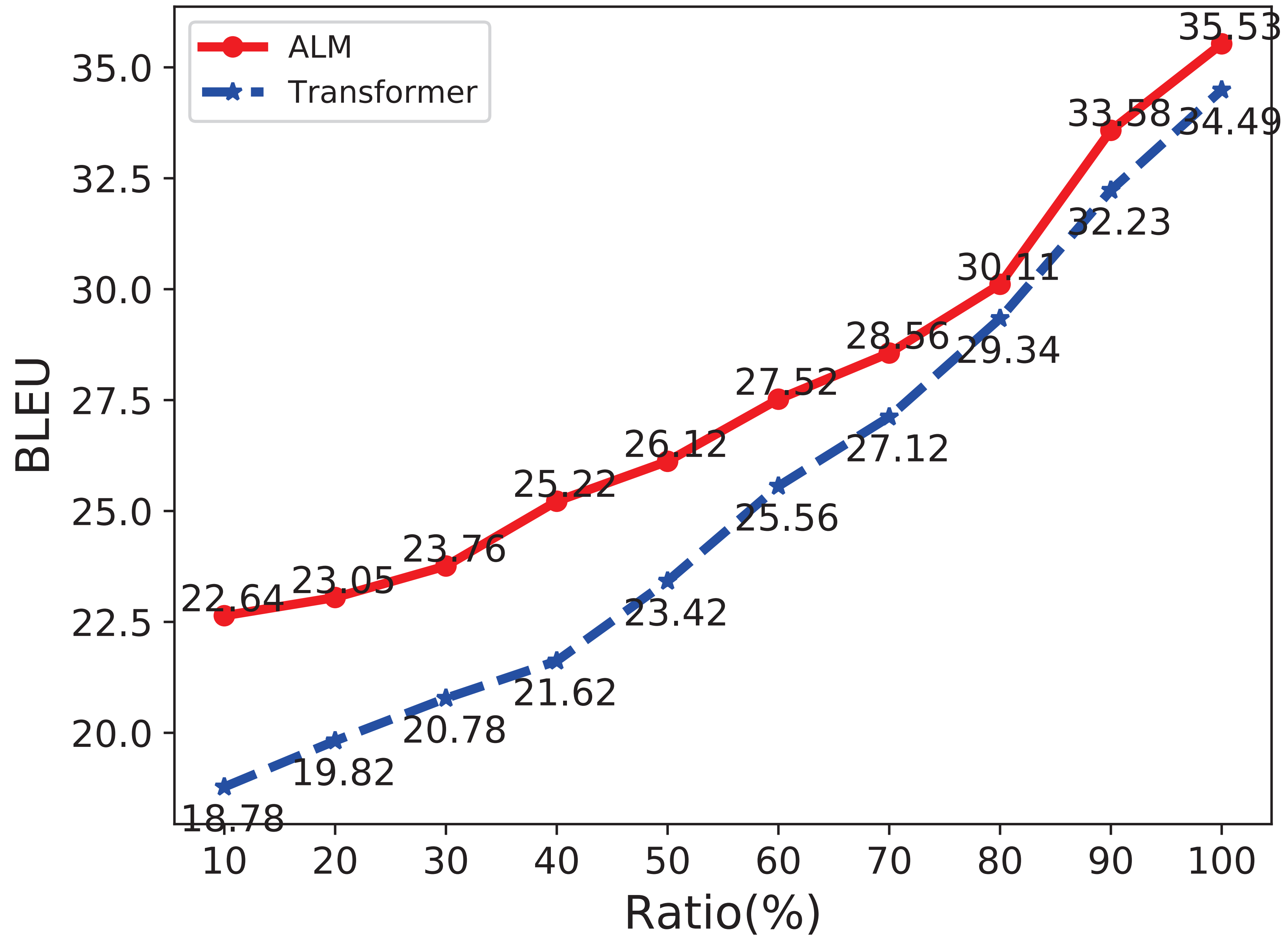
Results

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.

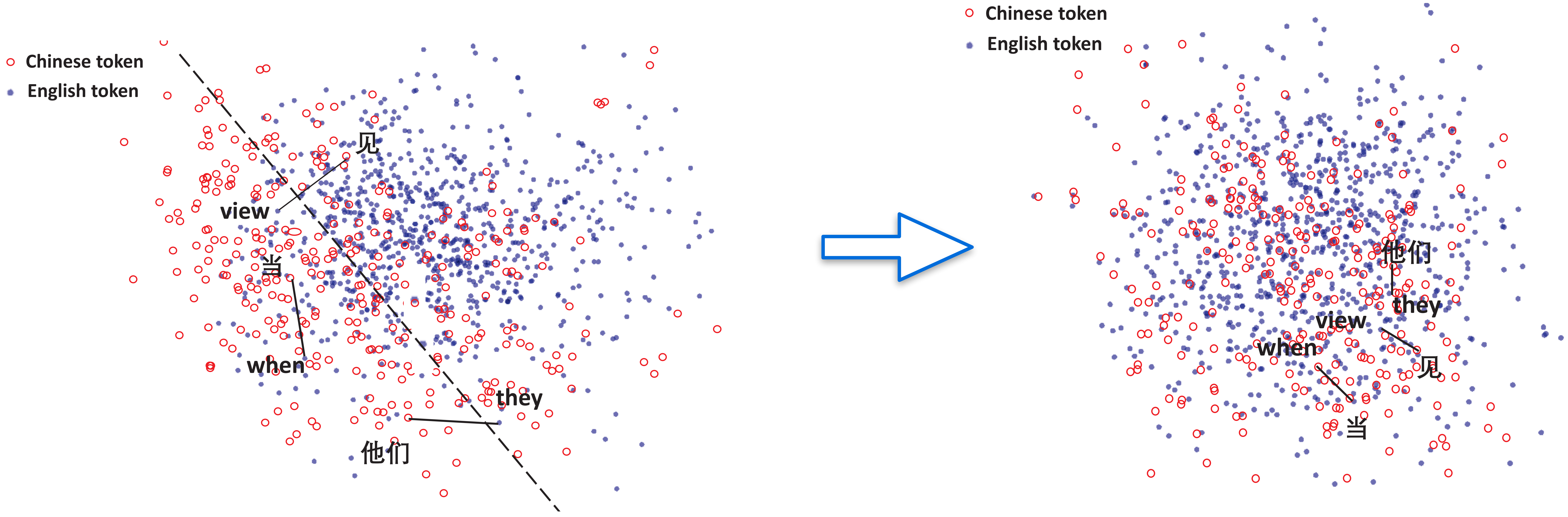
Results



- 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

Visualization of word embedding

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

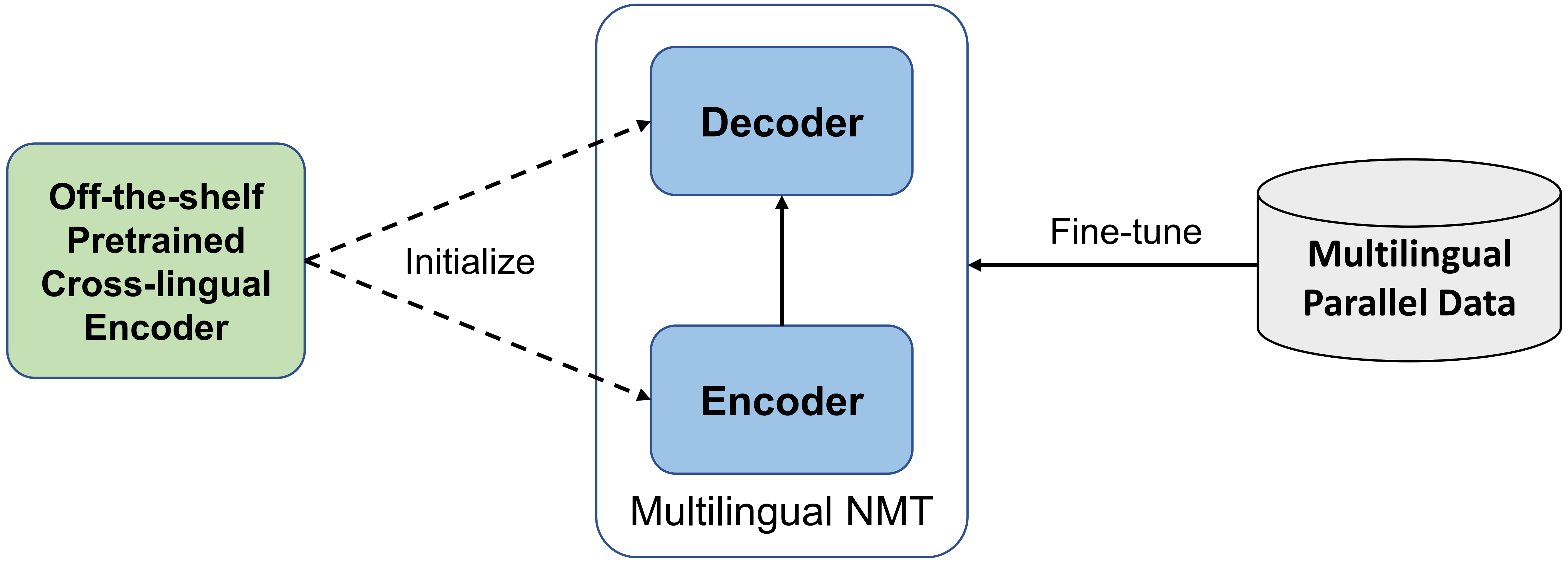


(a) Transformer

(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 [Ma et al]

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**

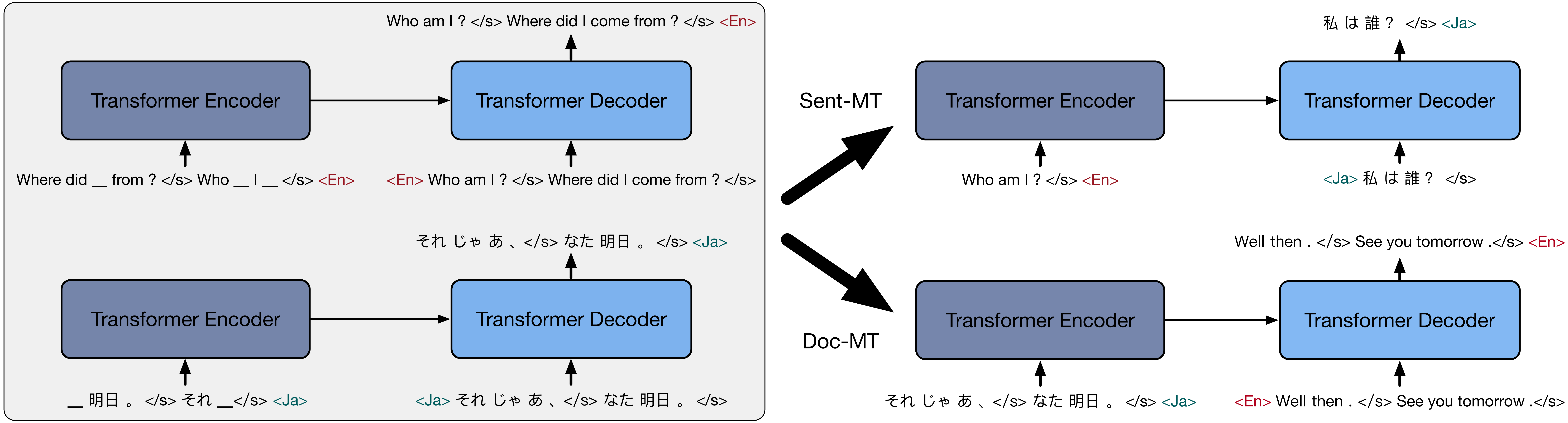
- Generally, the improvements are smaller than X → En
- The multilingual part of En → X is at the decoder side, which XLM-R is not an expert in.

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

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]

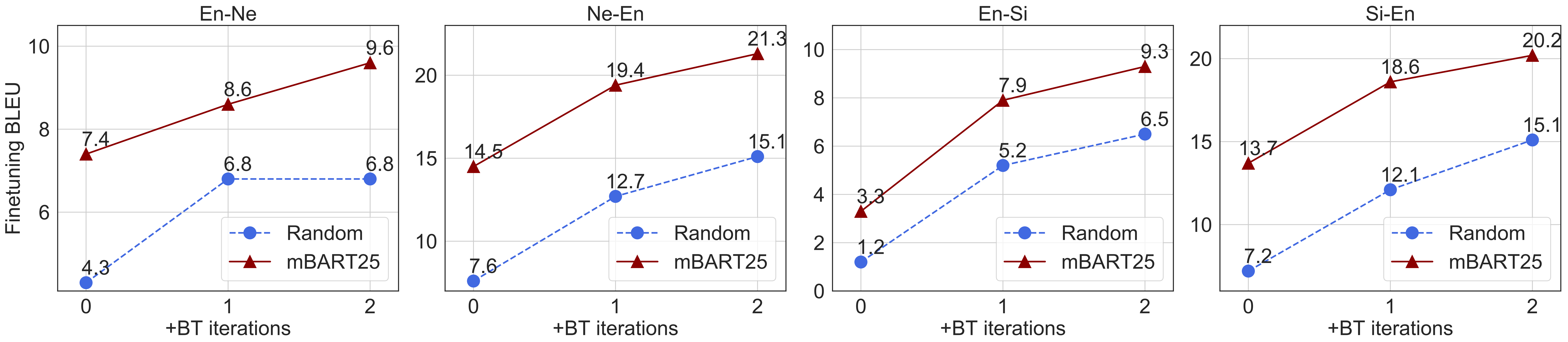
mBART: Rich-resource translation

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?

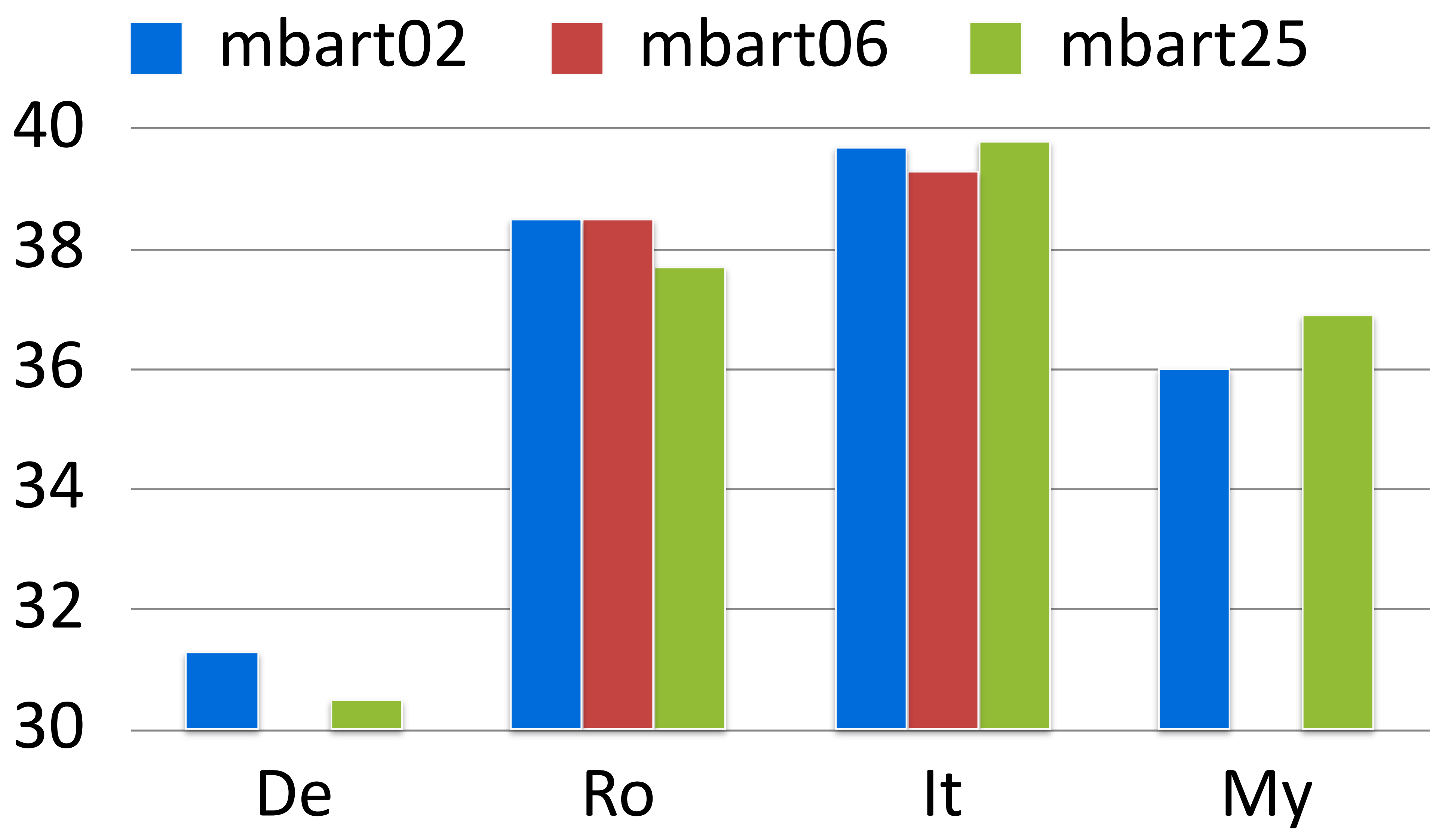
Pre-training		Fine-tuning		
Model	**Data**	**En→Ro**	**Ro→En**	**+BT**
Random	None	34.3	34.0	36.8
XLM (2019)	En Ro	-	35.6	38.5
MASS (2019)	En Ro	-	-	39.1
BART (2019)	En	-	-	38.0
XLM-R (2019)	CC100	35.6	35.8	-
BART-En	En	36.0	35.8	37.4
BART-Ro	Ro	37.6	36.8	38.1
mBART02	En Ro	**38.5**	**38.5**	**39.9**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

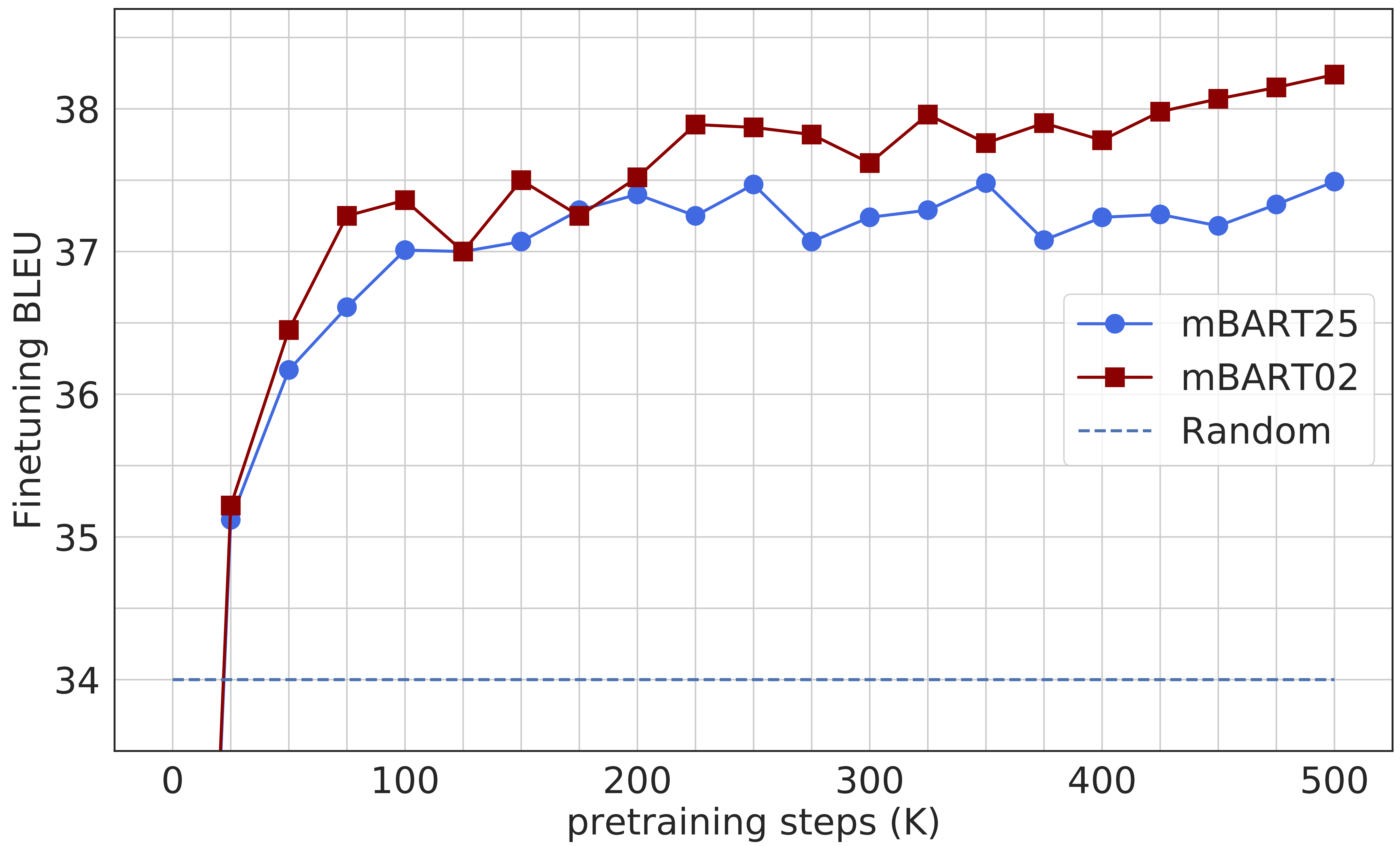
Languages	De	Ro	It	My	En
Size/GB	66.6	61.4	30.2	1.6	300.8
mBART02	**31.3**	**38.5**	39.7	36.5	
mBART06	-	**38.5**	39.3	-	
mBART25	30.5	37.7	**39.8**	**36.9**	



- 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]

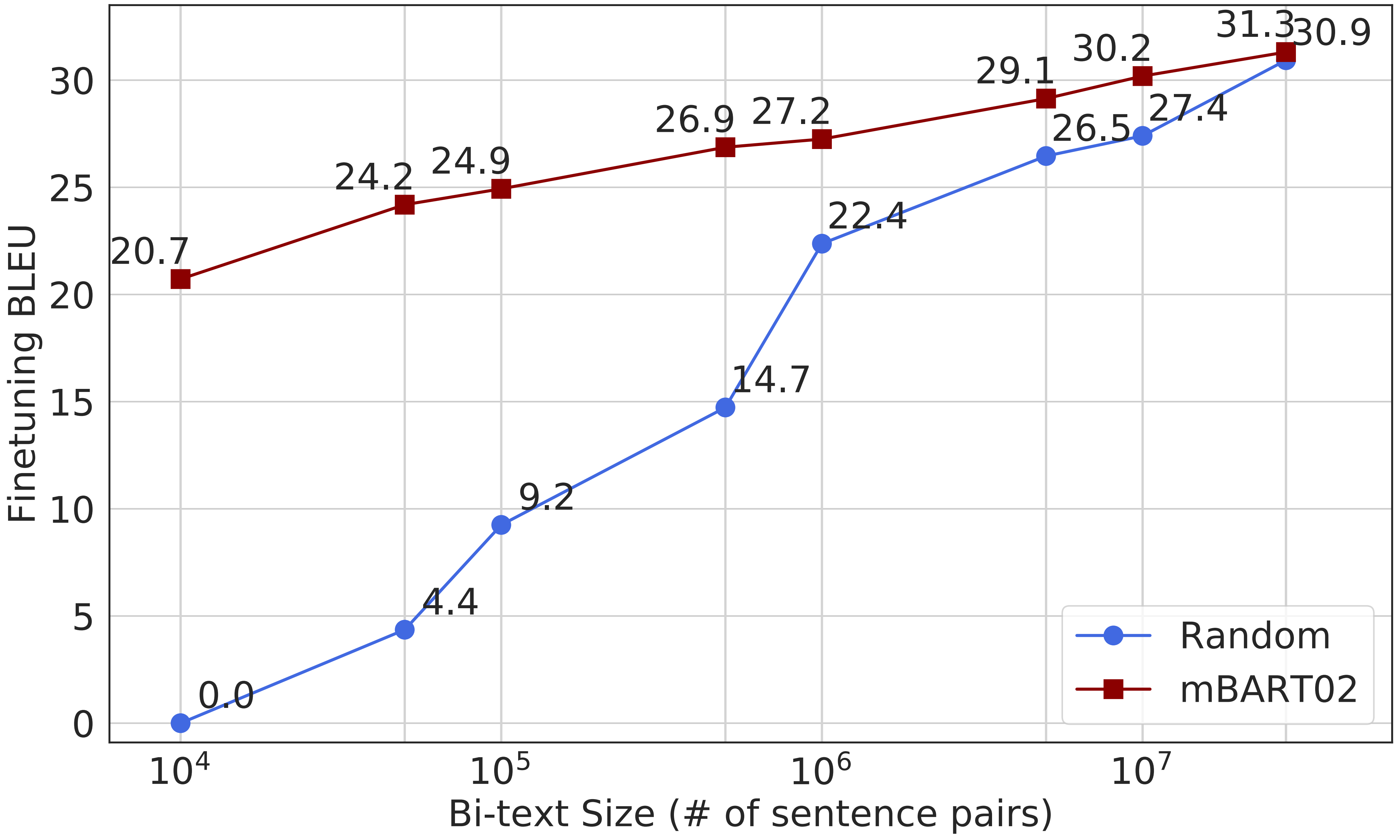
Analysis: Pre-training steps matters



- 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

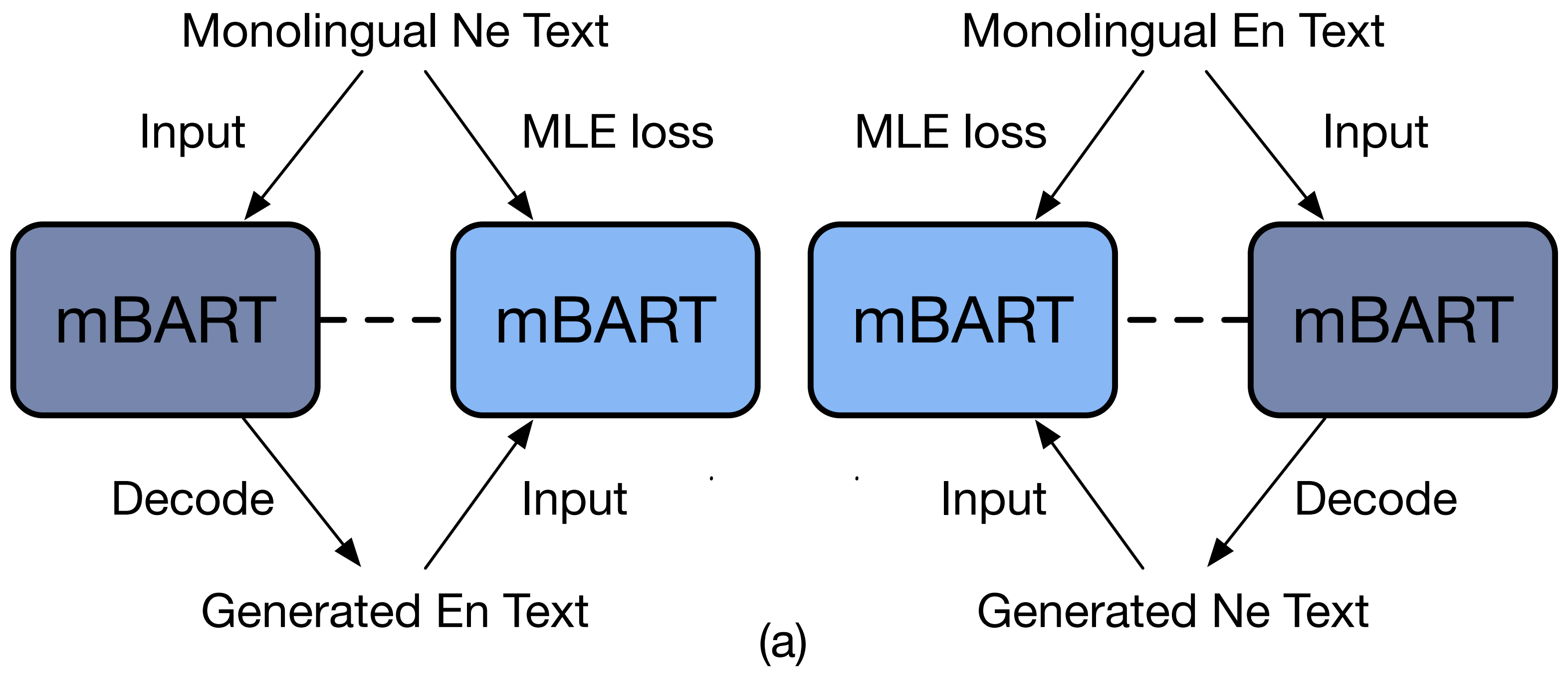
	Monolingual	Nl-En	En-Nl	Ar-En	En-Ar	Nl-De	De-Nl
Random	None	34.6 (-8.7)	29.3 (-5.5)	27.5 (-10.1)	16.9 (-4.7)	21.3 (-6.4)	20.9 (-5.2)
mBART02	En Ro	41.4 (-2.9)	34.5 (-0.3)	34.9 (-2.7)	21.2 (-0.4)	26.1 (-1.6)	25.4 (-0.7)
mBART06	En Ro Cs It Fr Es	43.1 (-0.2)	34.6 (-0.2)	37.3 (-0.3)	21.1 (-0.5)	26.4 (-1.3)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

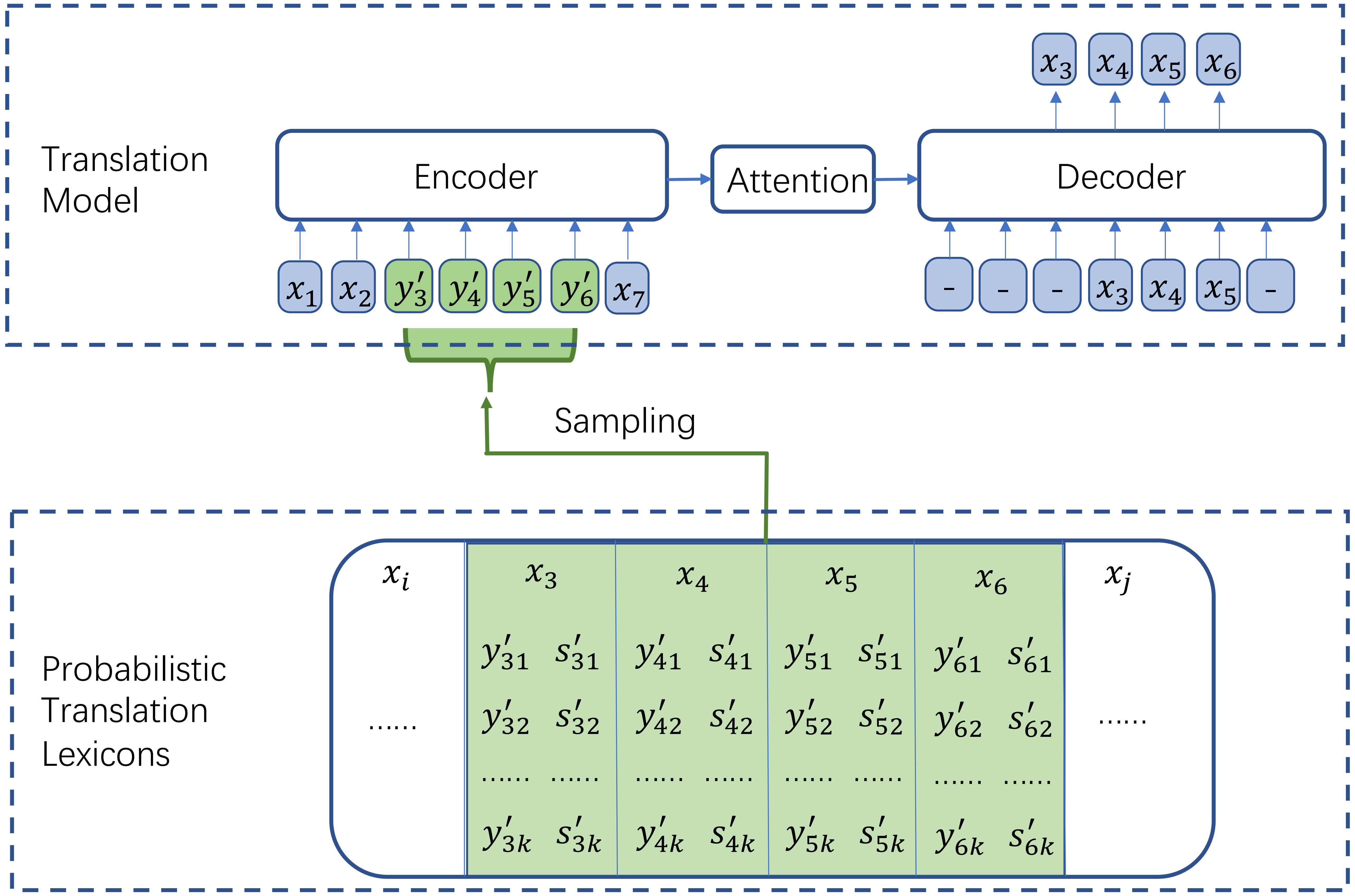
Model	Similar Pairs				Dissimilar Pairs			
	En-De		En-Ro		En-Ne		En-Si	
	←	→	←	→	←	→	←	→
Random	21.0	17.2	19.4	21.2	0.0	0.0	0.0	0.0
XLM (2019)	34.3	26.4	31.8	33.3	0.5	0.1	0.1	0.1
MASS (2019)	**35.2**	28.3	**33.1**	**35.2**	-	-	-	-
mBART	34.0	**29.8**	30.5	35.0	**10.0**	**4.4**	**8.2**	**3.9**

- 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]

CSP: Code-Switching Pre-training for Neural Machine Translation

- Sequence-level pre-training with only monolingual data
- Sub-span of the source sentence is replaced with their lexical translation



The training paradigm follows MASS

Lexical translation is build with only monolingual data. [Learning bilingual word embeddings with (almost) no bilingual data.]

CSP: Code-Switching Pre-training for Neural Machine Translation [Yang et al., EMNLP 2020]

CSP: Code-Switching Pre-training for Neural Machine Translation

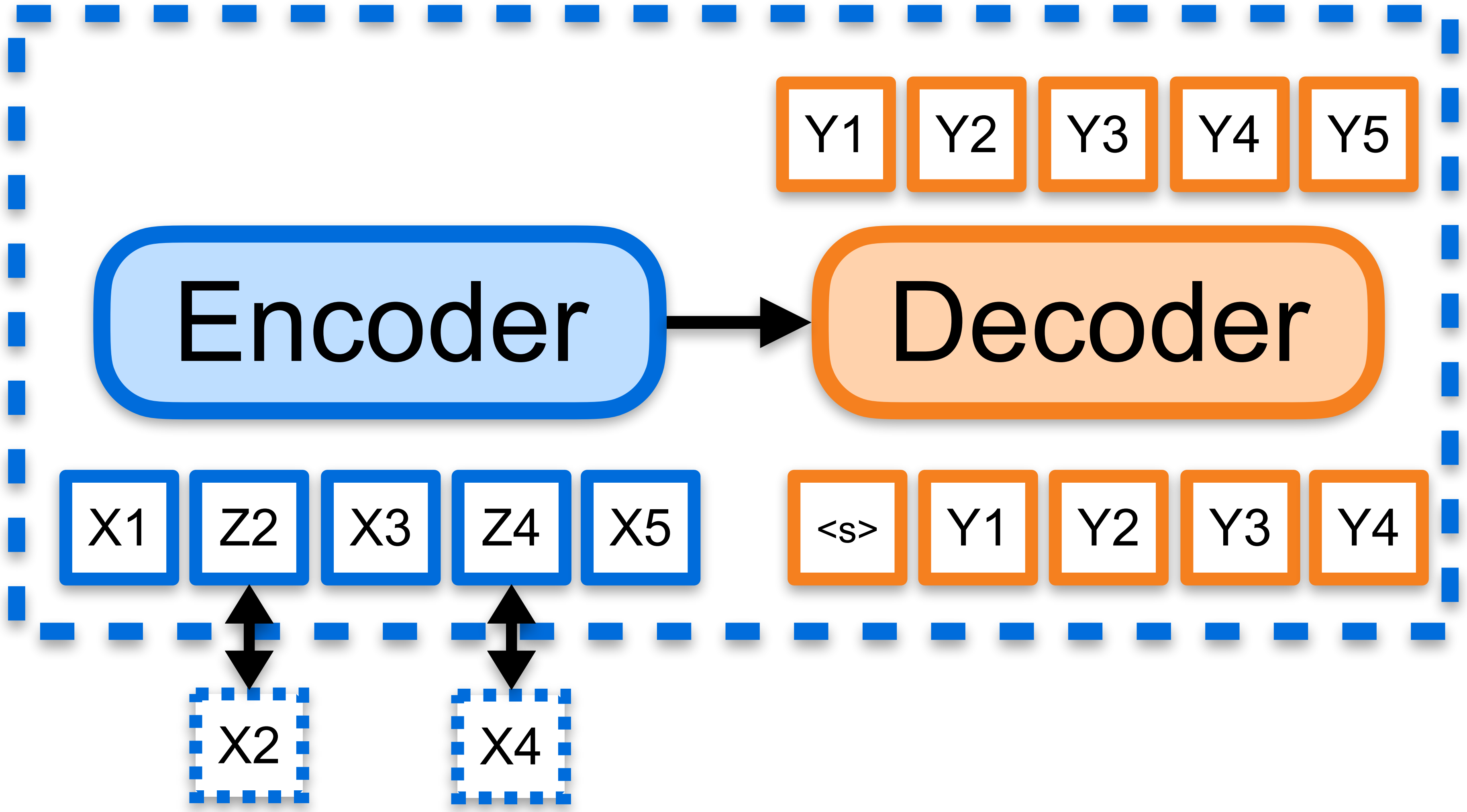
System	en-de	de-en	en-fr	fr-en	zh-en
Yang et al. (2018)	10.86	14.62	16.97	15.58	14.52
Lample et al. (2018b)	17.16	21.0	25.14	24.18	-
Lample and Conneau (2019)	27.0	34.3	33.4	33.3	-
Song et al. (2019b)	28.1	35.0	37.5	**34.6**	-
Lample and Conneau (2019) (our reproduction)	27.3	33.8	32.9	33.5	22.1
Song et al. (2019b) (our reproduction)	27.9	34.7	37.3	34.1	22.8
CSP and fine-tuning (ours)	**28.7**	**35.7**	**37.9**	34.5	**23.9**

System	en-de	en-fr	zh-en
Vaswani et al. (2017)	27.3	38.1	-
Vaswani et al. (2017) (our reproduction) / + BT	27.0 / 28.6	37.9 / 39.3	42.1 / 43.7
Lample and Conneau (2019) (our reproduction) / + BT	28.1 / 29.4	38.3 / 39.6	42.0 / 43.7
Song et al. (2019b) (our reproduction) / + BT	28.4 / 29.6	38.4 / 39.6	42.5 / 44.1
CSP and fine-tuning (ours) / + BT	**28.9 / 30.0**	**38.8 / 39.9**	**43.2 / 44.6**

CSP: Code-Switching Pre-training for Neural Machine Translation [Yang et al., EMNLP 2020]

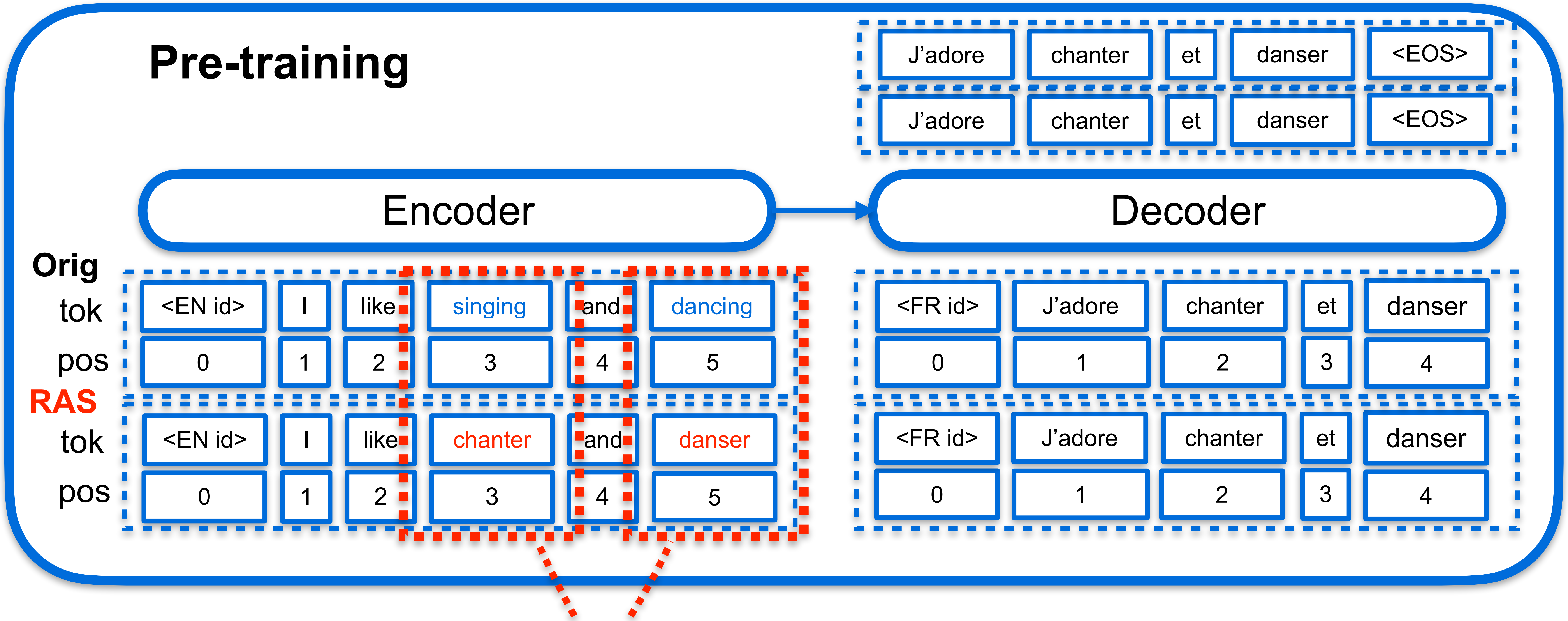
mRASP: multilingual Random Aligned Substitution Pre-training

- **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

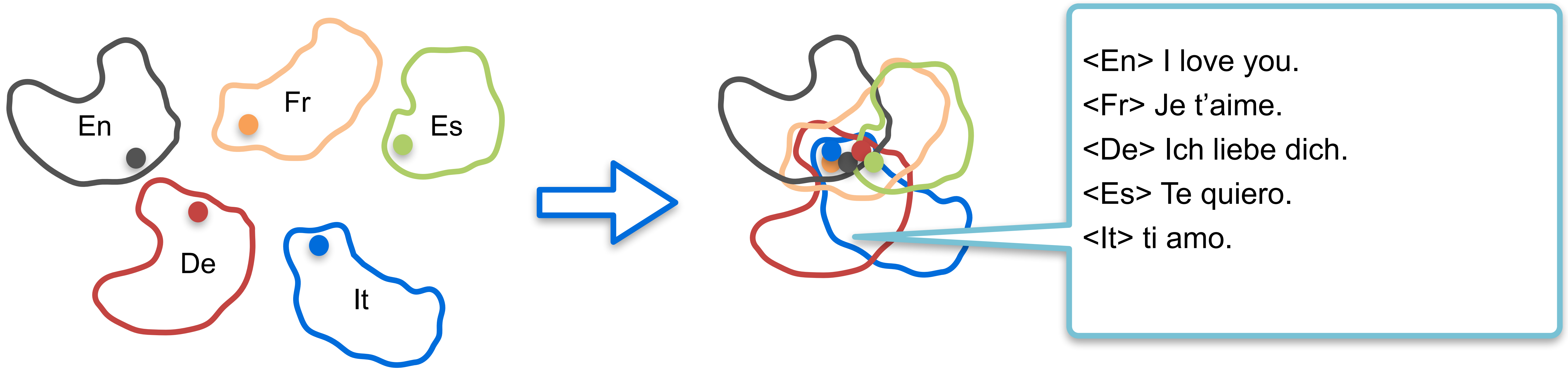


Pre-training Multilingual Neural Machine Translation by Leveraging Alignment Information [Lin et al., EMNLP 2020]

mRASP: Overview

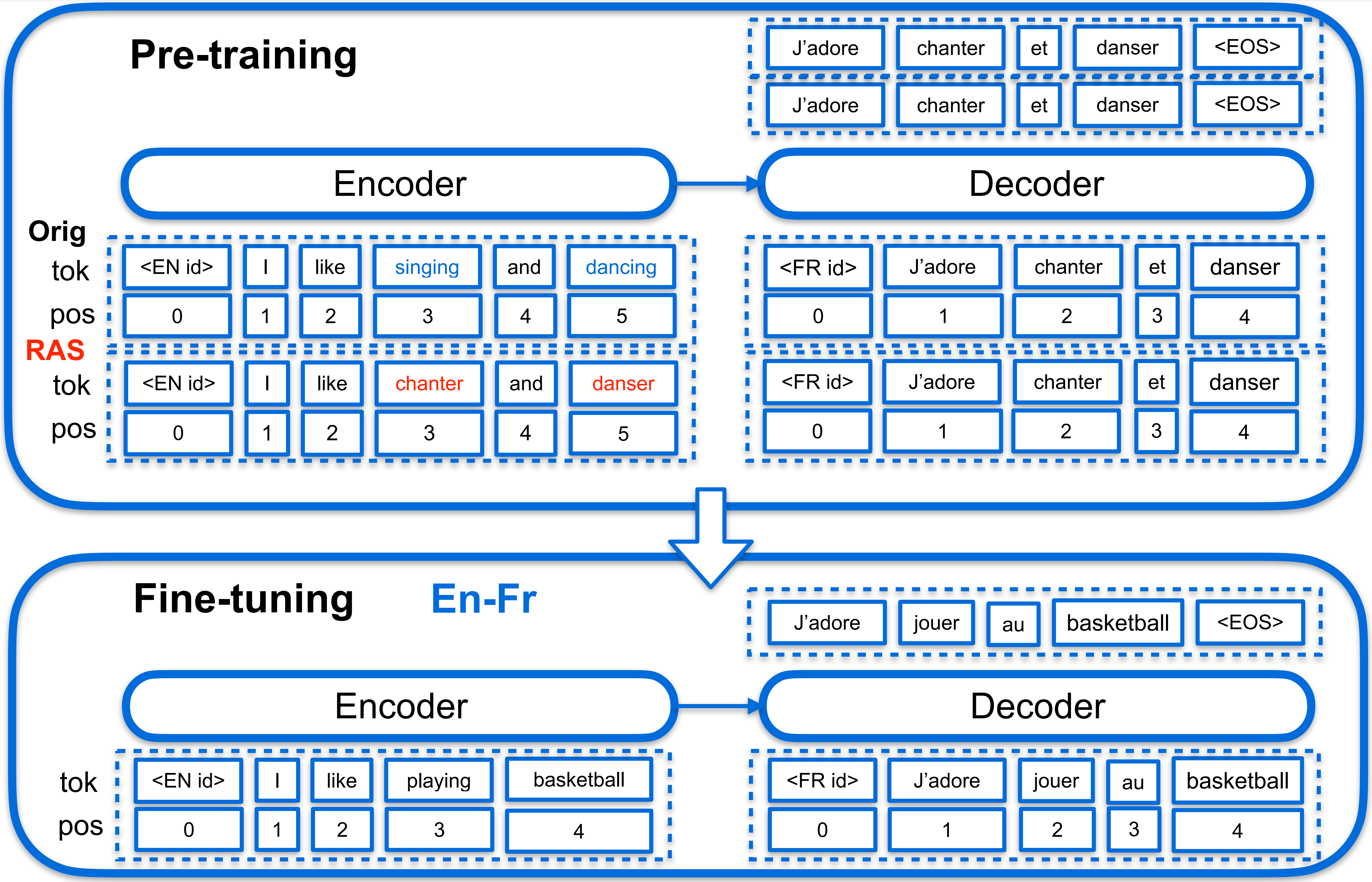


Random Aligned Substitution



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

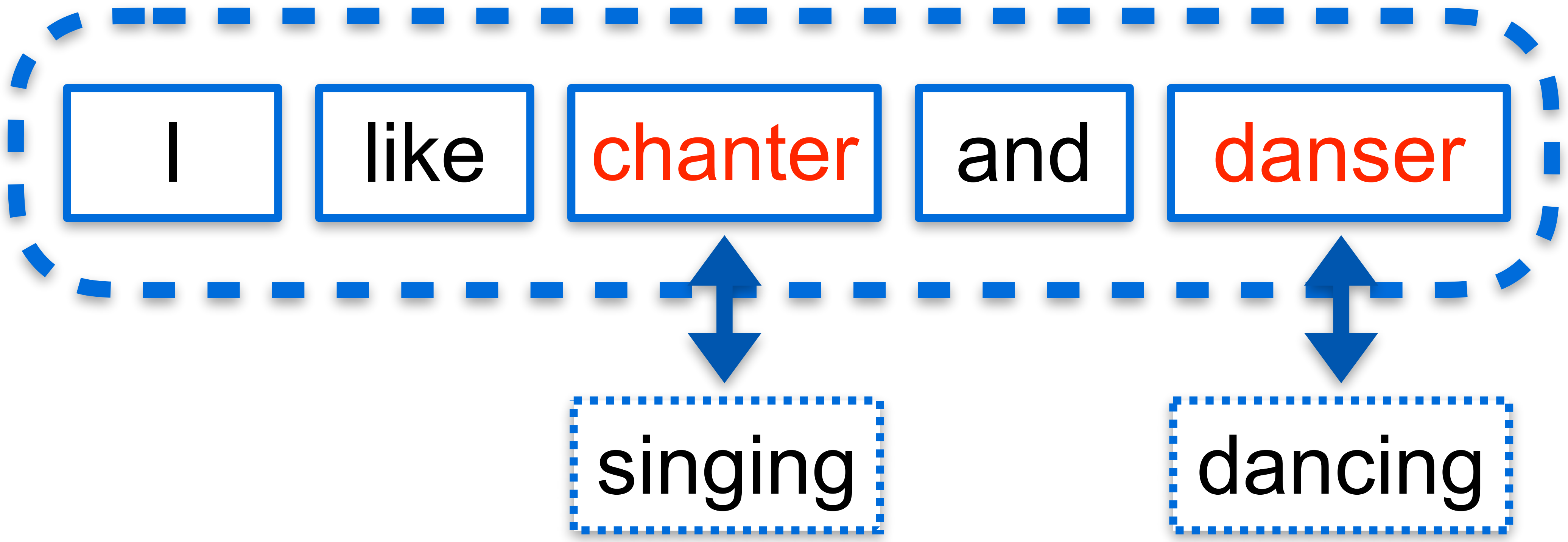
mRASP: Overview



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

mRASP: RAS method

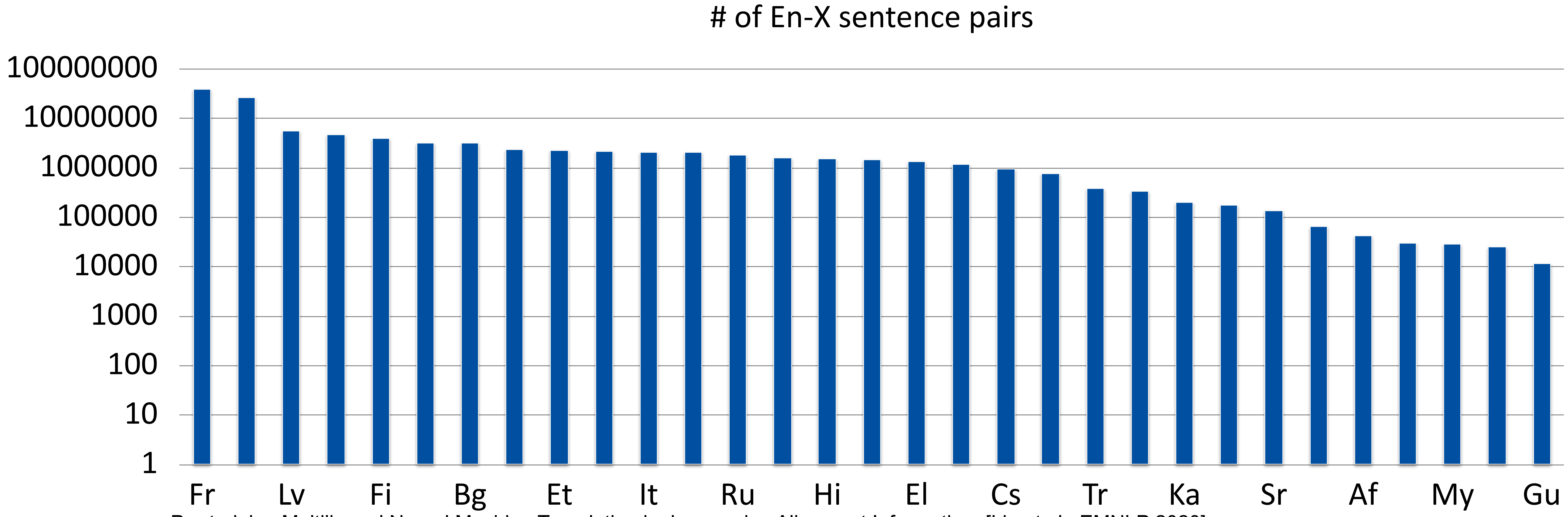
- **Random Aligned Substitution (RAS)**
  - Randomly replace a source word to its synonym in different language.
  - Draw the embedding space closer.



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

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]

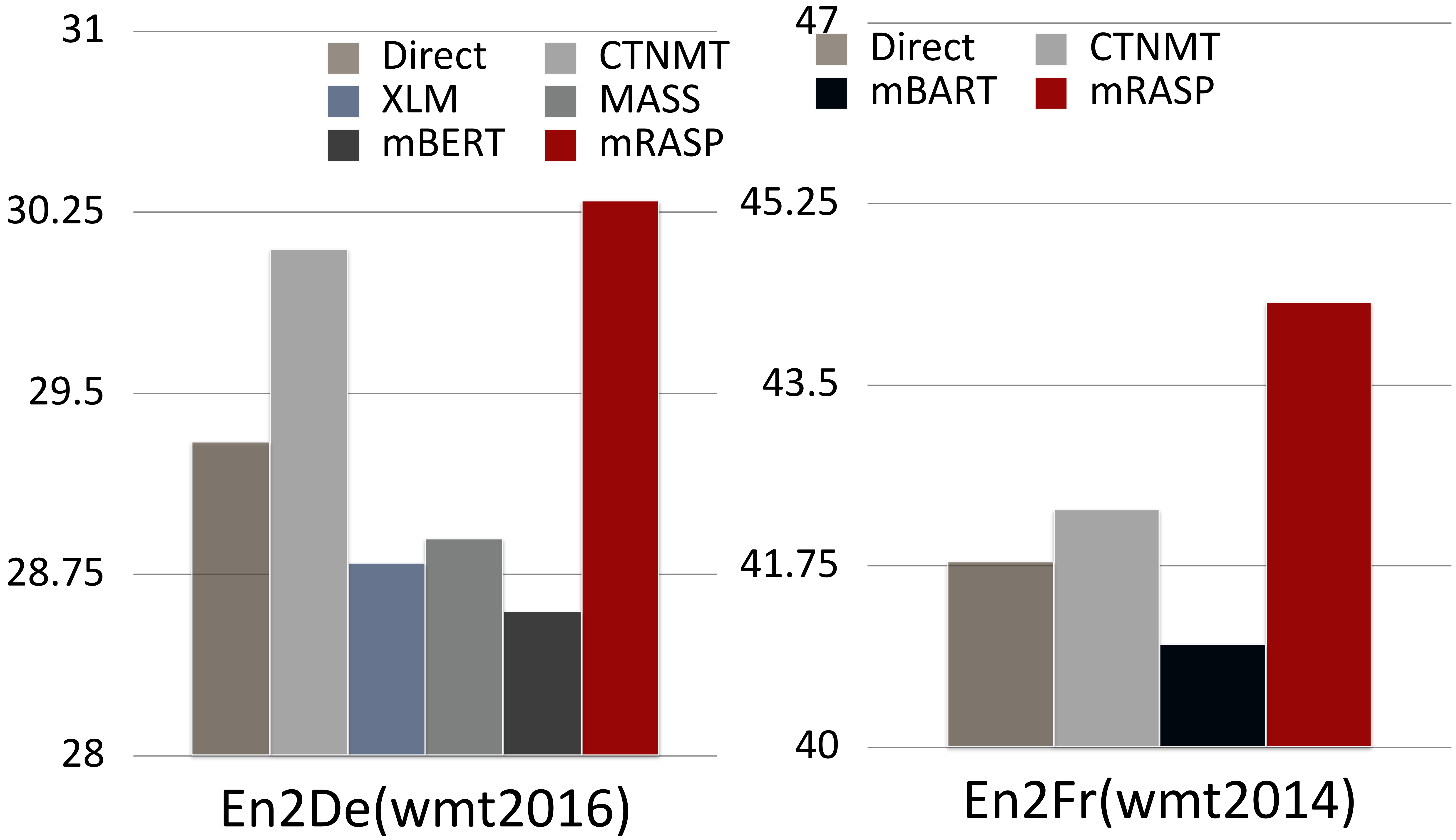
mRASP: Fine-tuning Dataset

- 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.)

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

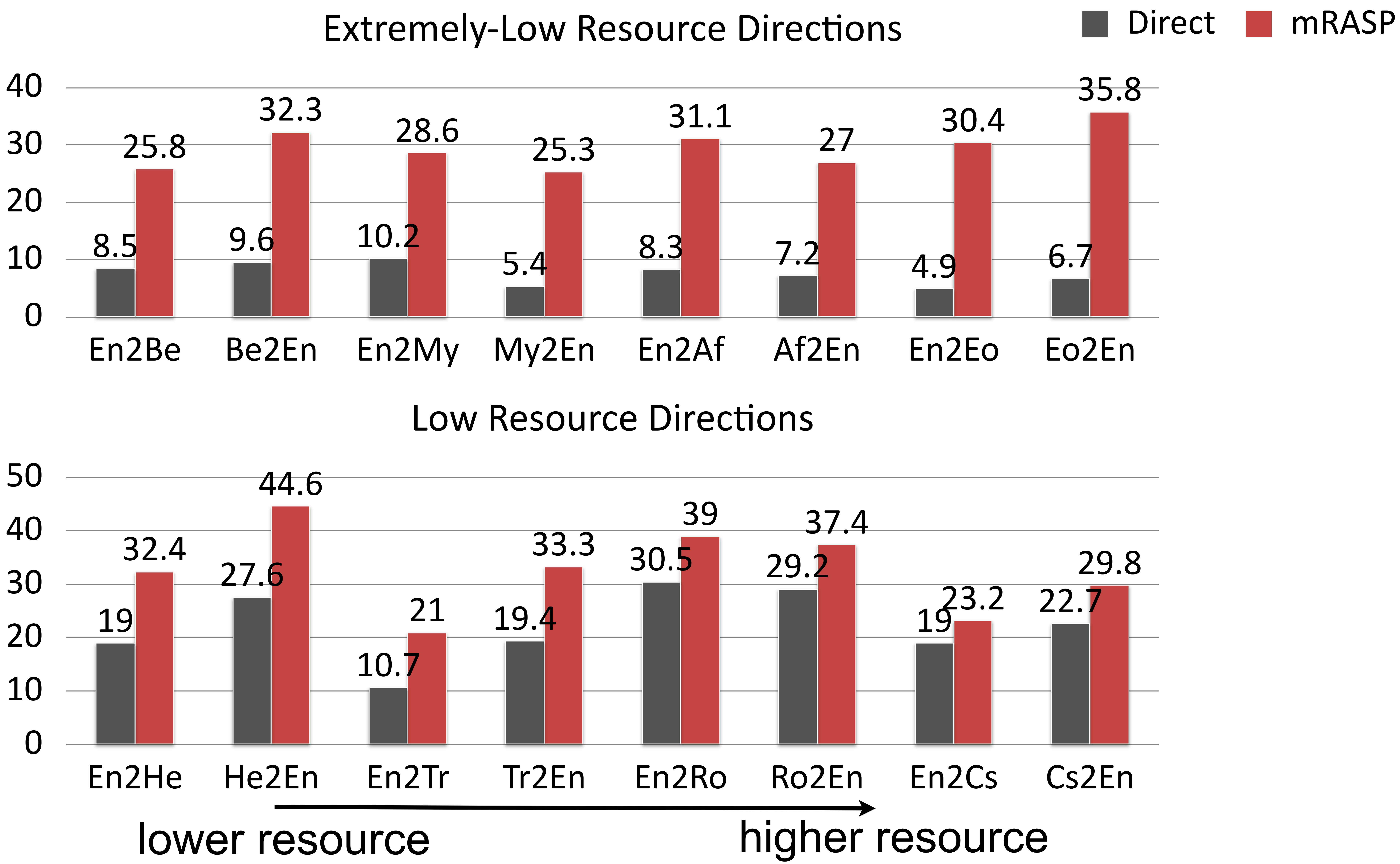
mRASP: Rich resource works

- 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.

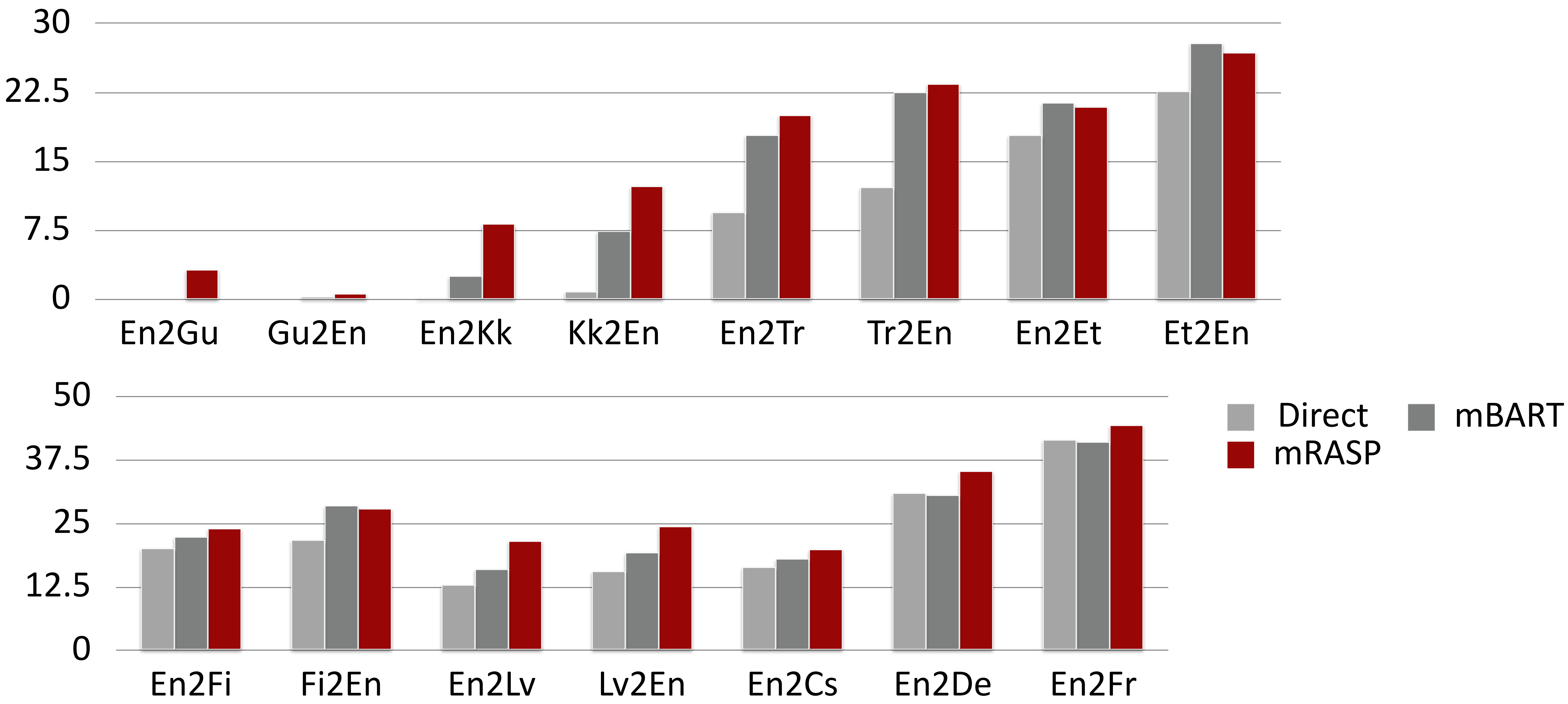
		Fr-Zh(20K)		De-Fr(9M)	
		—>	<—	—>	<—
Exotic Pair	Direct	0.7	3	23.5	21.2
	mRASP	25.8	26.7	29.9	23.4
		Nl-Pt(12K)		Da-El(1.2M)	
		—>	<—	—>	<—
Exotic Full	Direct	0.0	0.0	14.1	16.9
	mRASP	14.1	13.2	17.6	19.9
		En-Mr(11K)		En-Gl(1.2M)	
		—>	<—	—>	<—
Exotic Source/ Target	Direct	6.4	6.8	8.9	12.8
	mRASP	22.7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

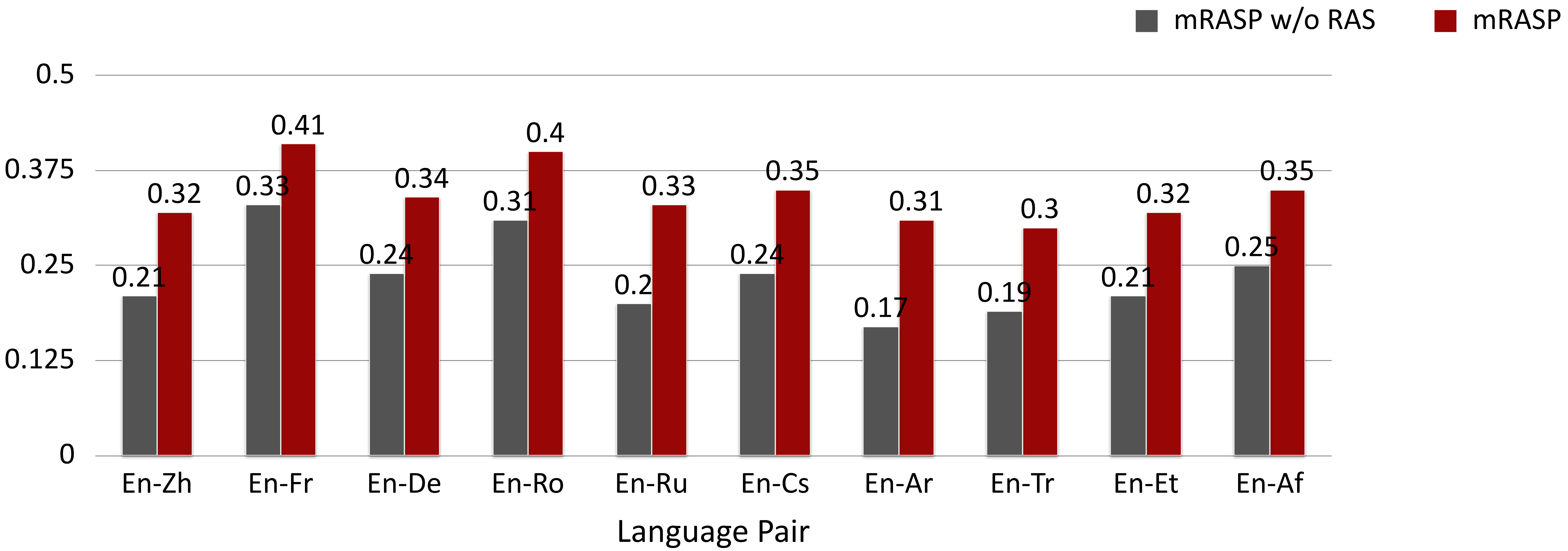
mRASP: Compare with other methods

- 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

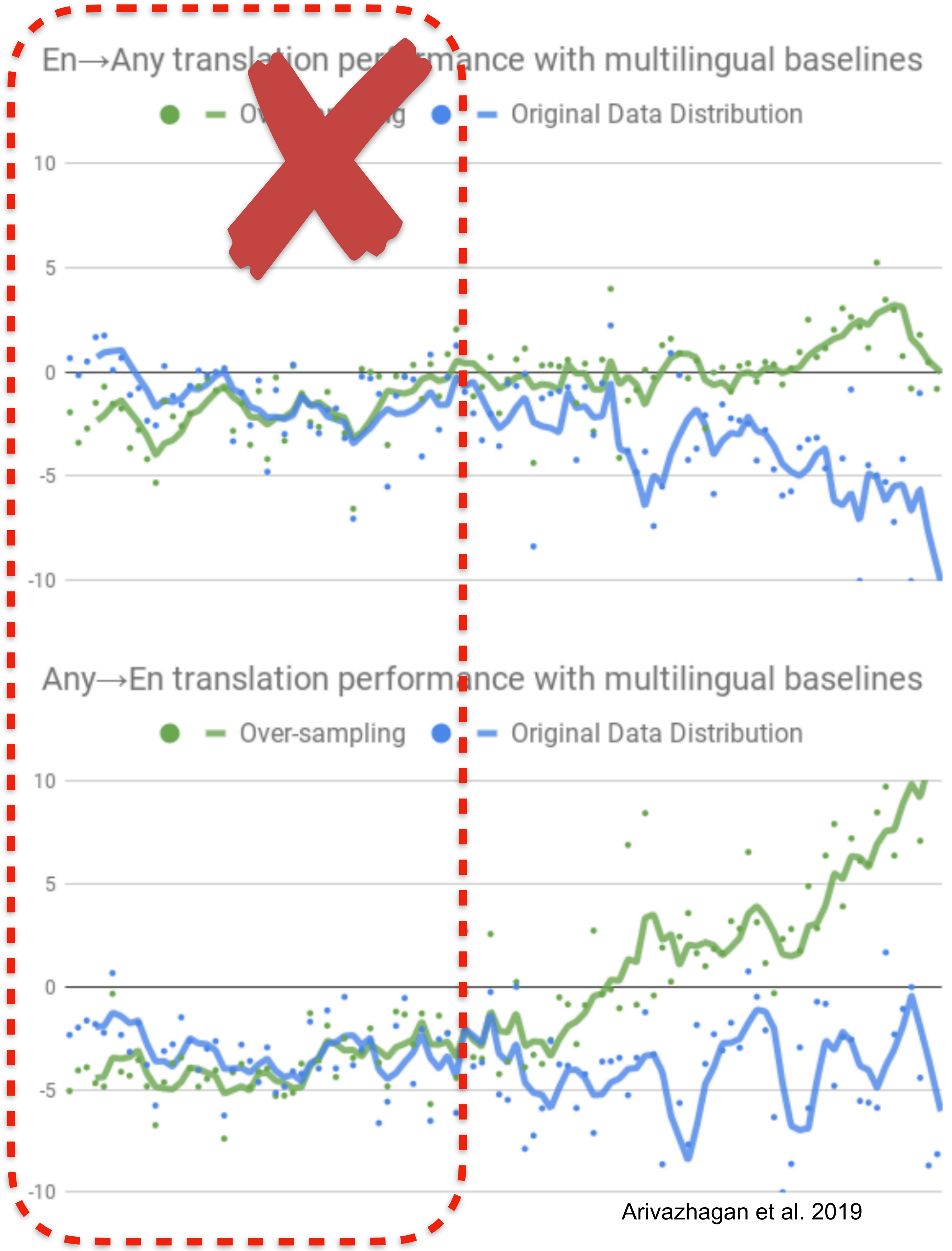
Zero-shot

Enabling unsupervised / zero-shot translation

Parallel

Monolingual

Leveraging both parallel & monolingual data

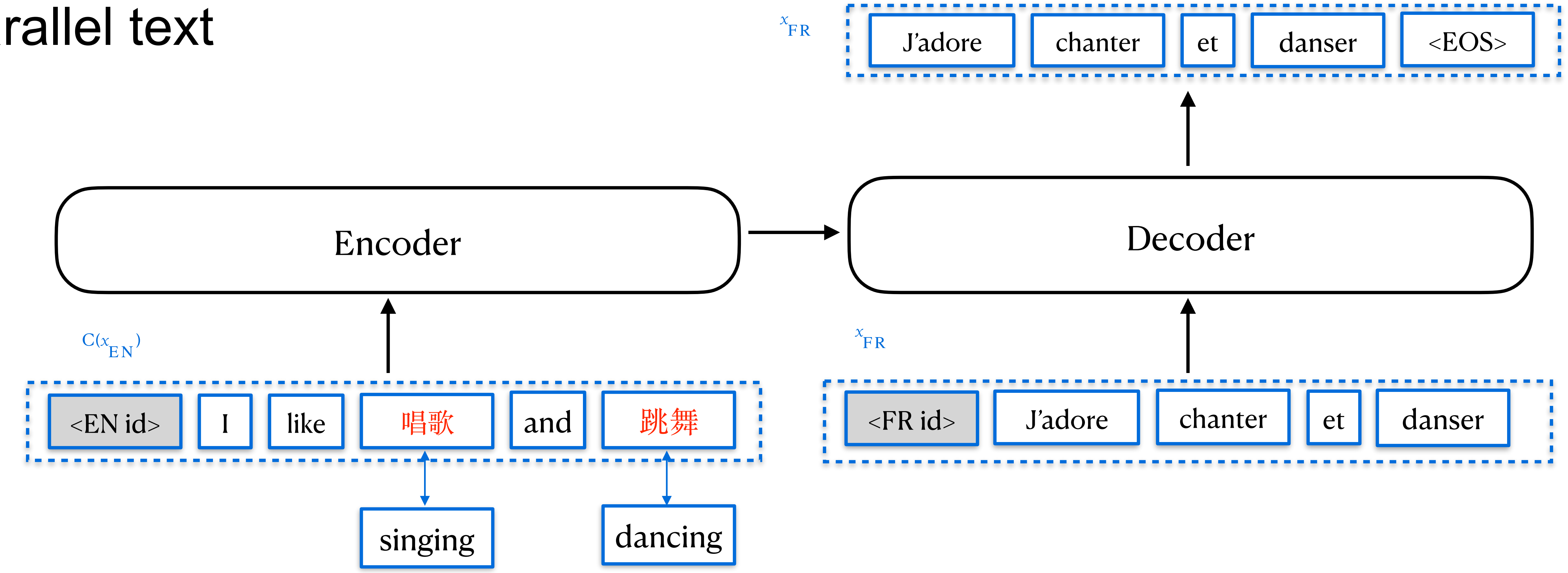


Comparable / better performance on high-resource directions

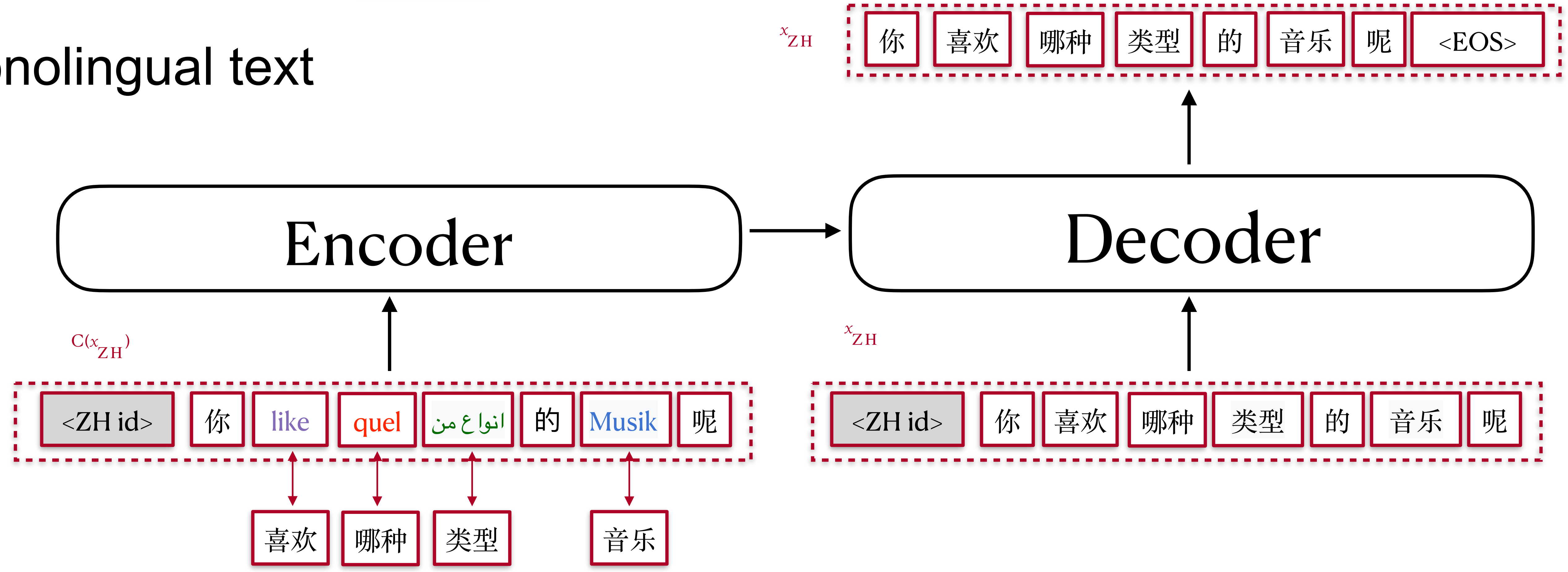
Contrastive Learning for Many-to-many Multilingual Neural Machine Translation [Pan et al., ACL 2021]

mRASP2 introduces monolingual data

- Parallel text

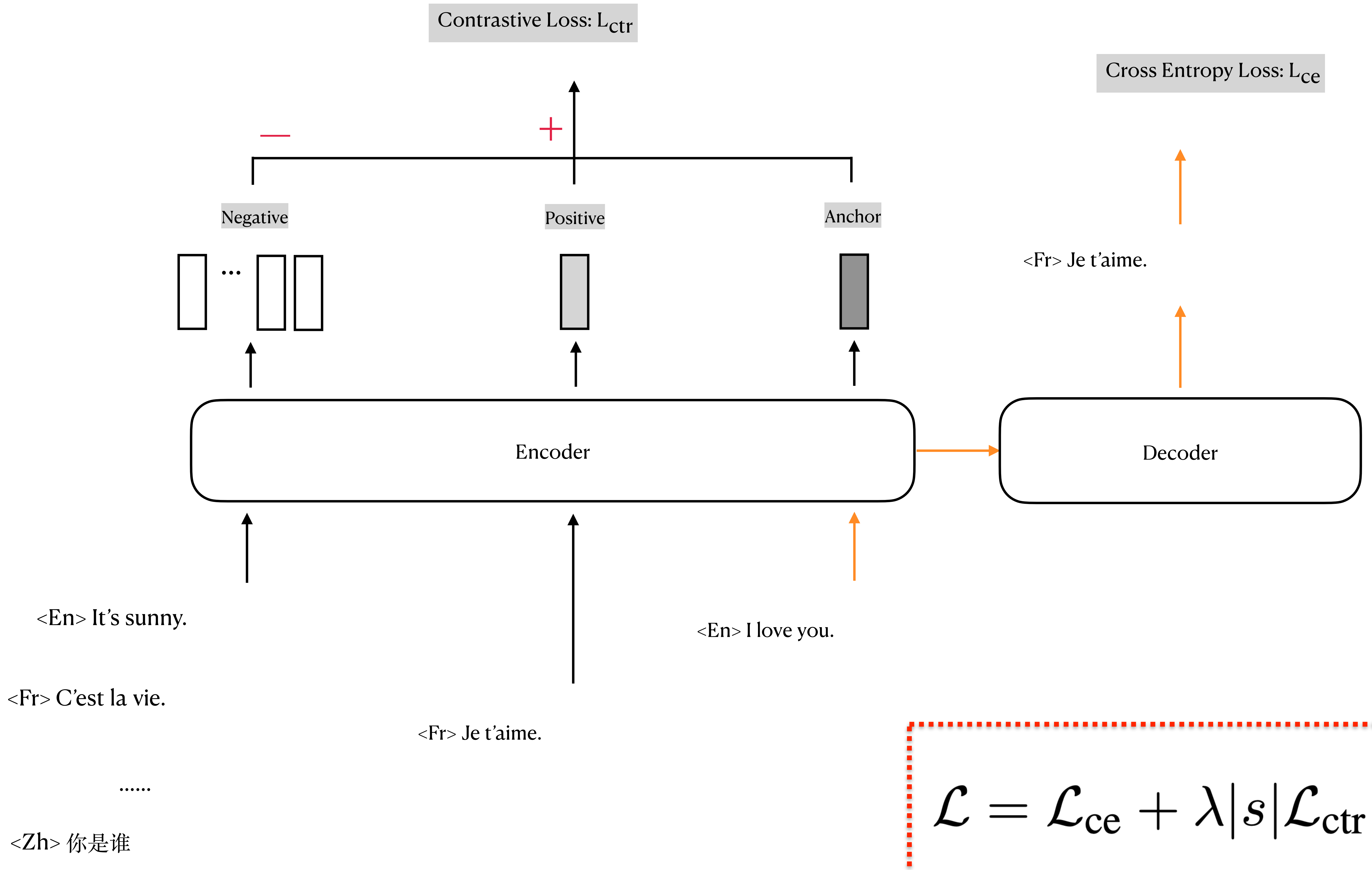


- Monolingual text



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

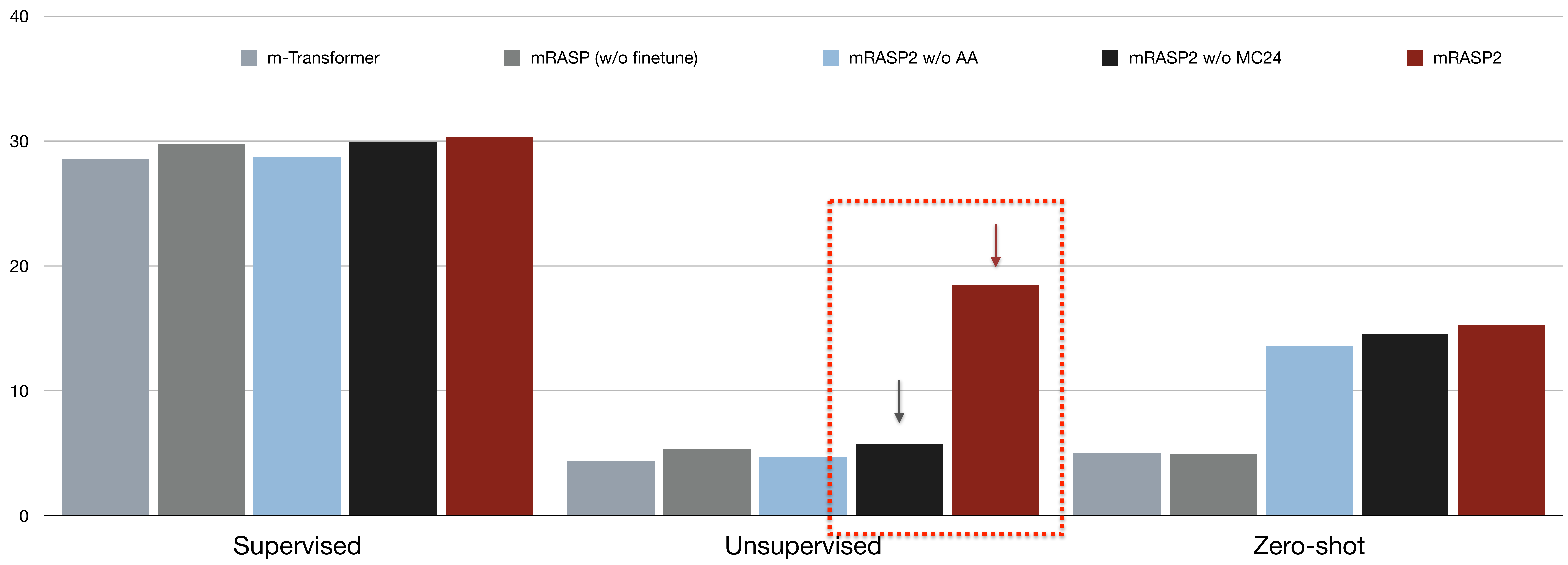
mRASP2 maps different languages in a same space



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

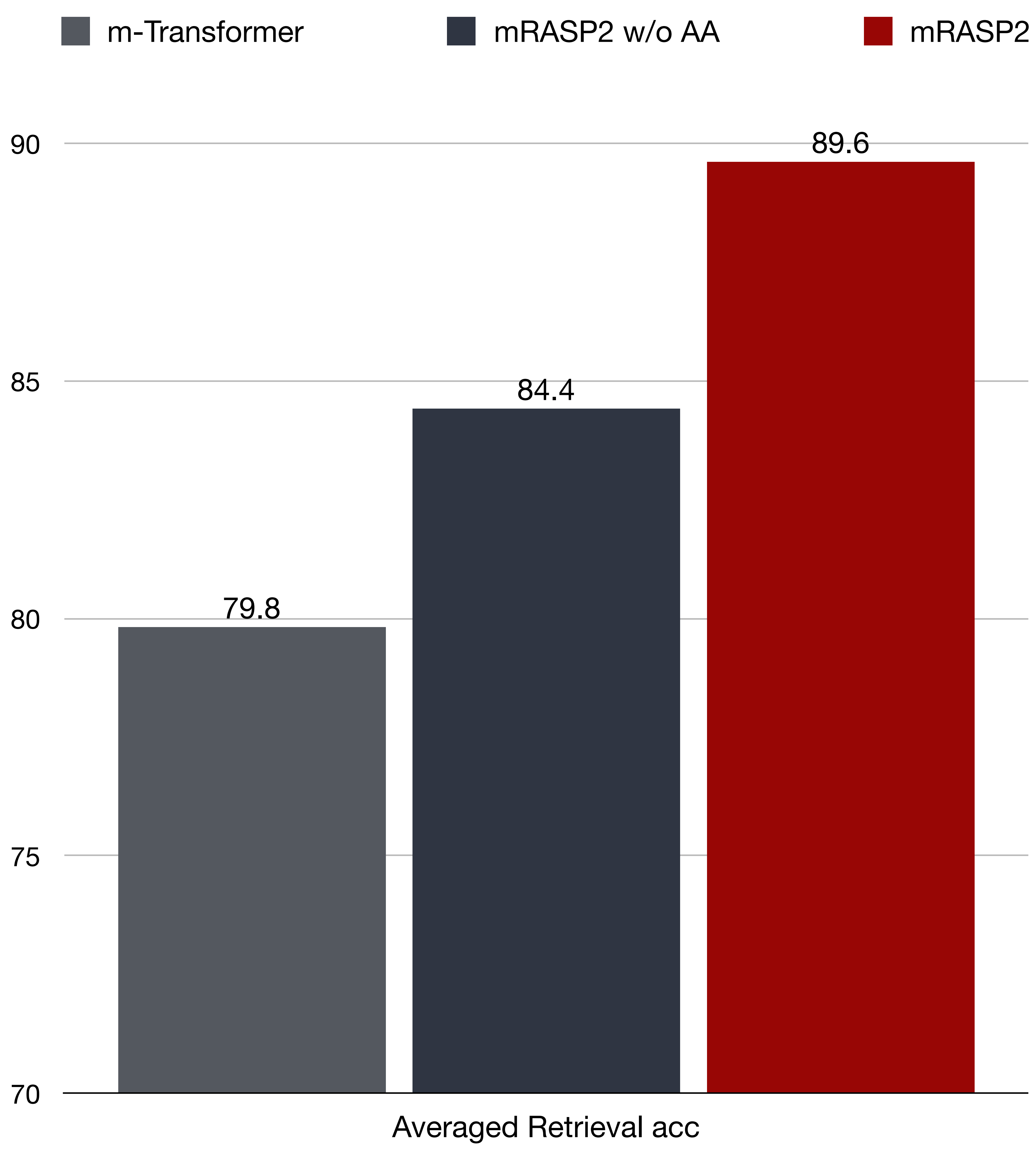
Contrastive Learning for Many-to-many Multilingual Neural Machine Translation [Pan et al., ACL 2021]

Experiments



Monolingual Corpus mainly contributes to unsupervised translation

Better Semantic Alignment: Sentence Retrieval

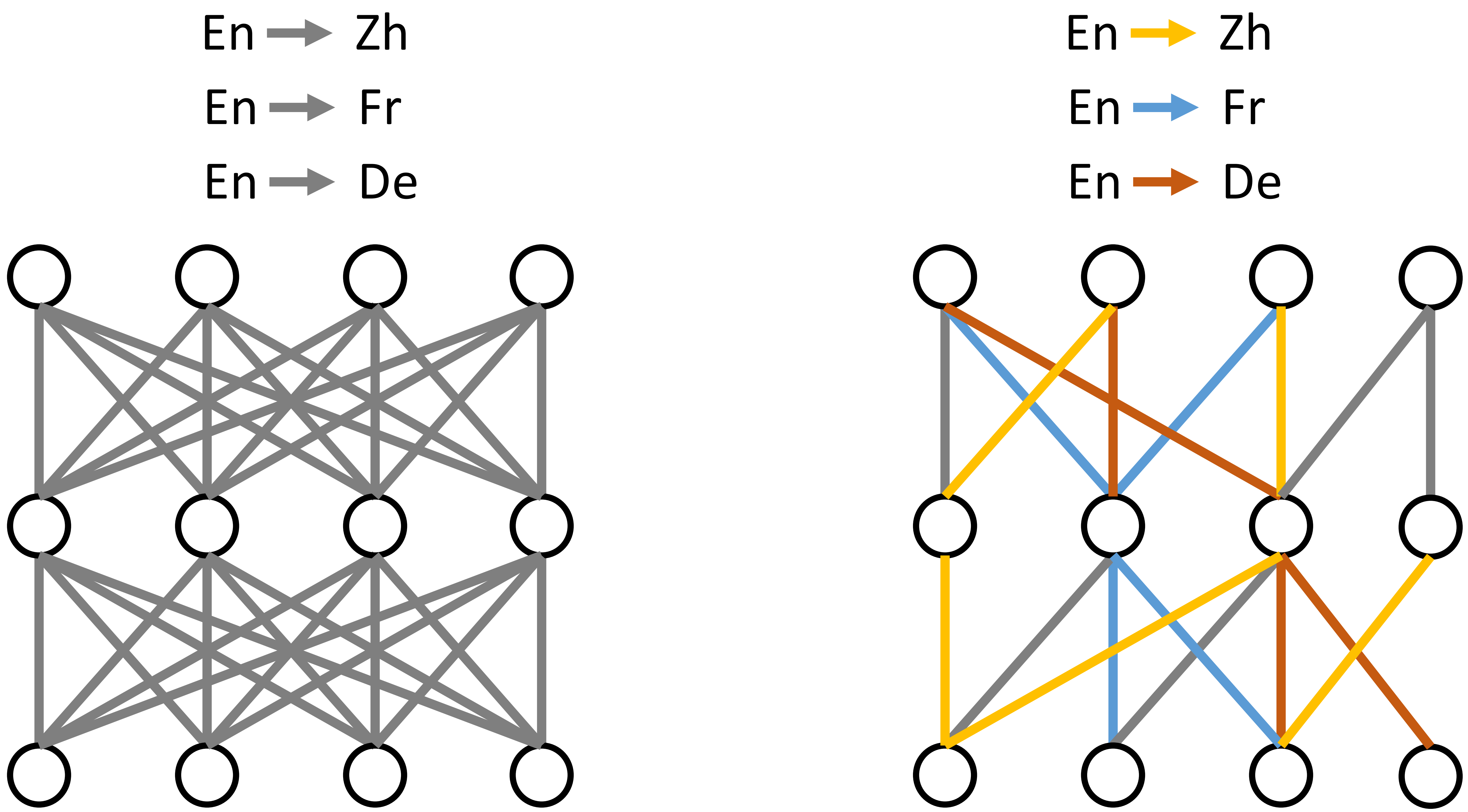


15-way parallel test set(Ted-M): 2284 samples

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

Learning Language Specific Sub-network for Multilingual Machine Translation

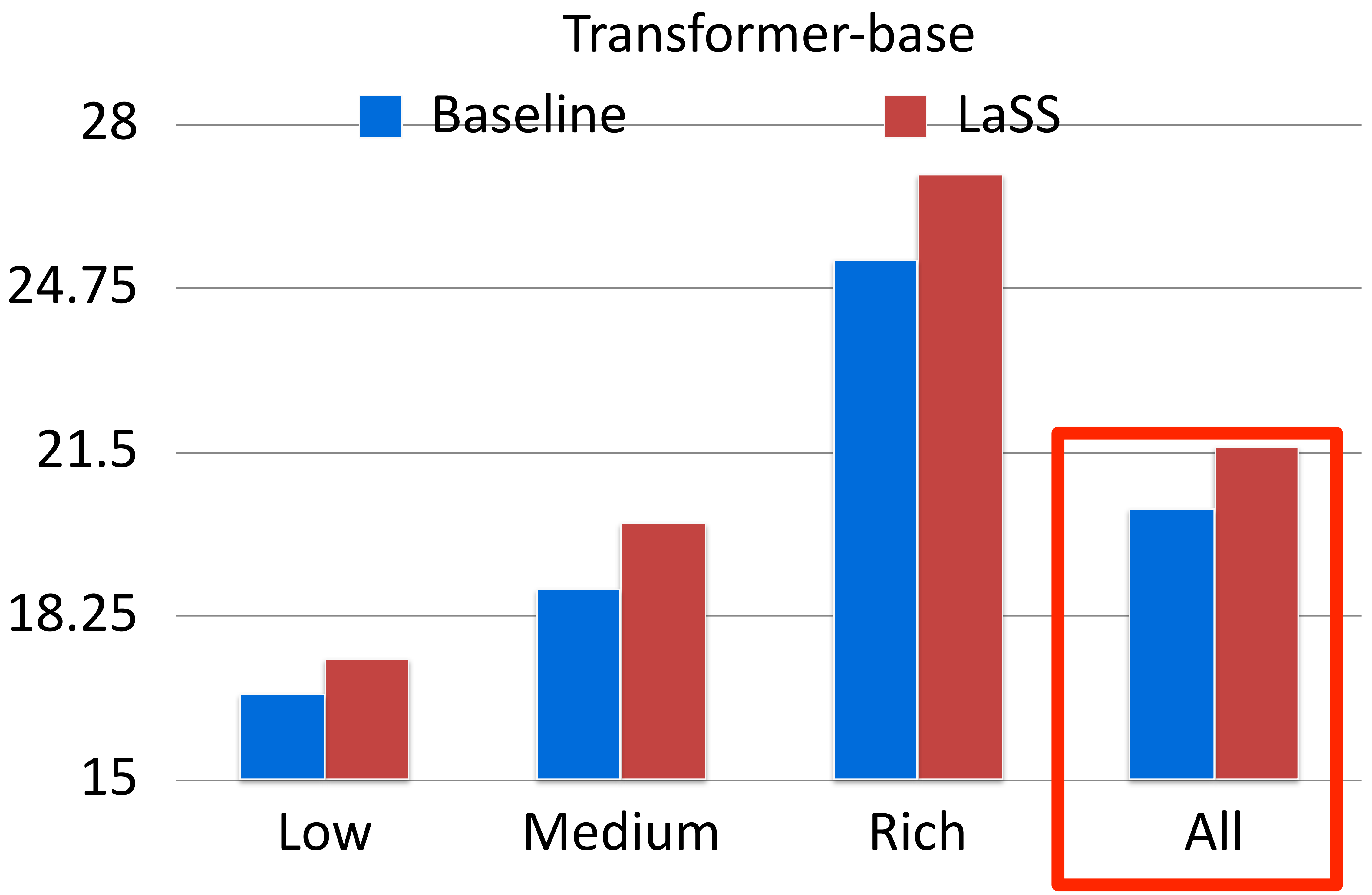
- LaSS accommodates one sub-network for each language pair.
  - Each language pair has shared parameters with some other language pairs and preserves its language-specific parameters
  - For fine-tuning, only updates the corresponding parameters

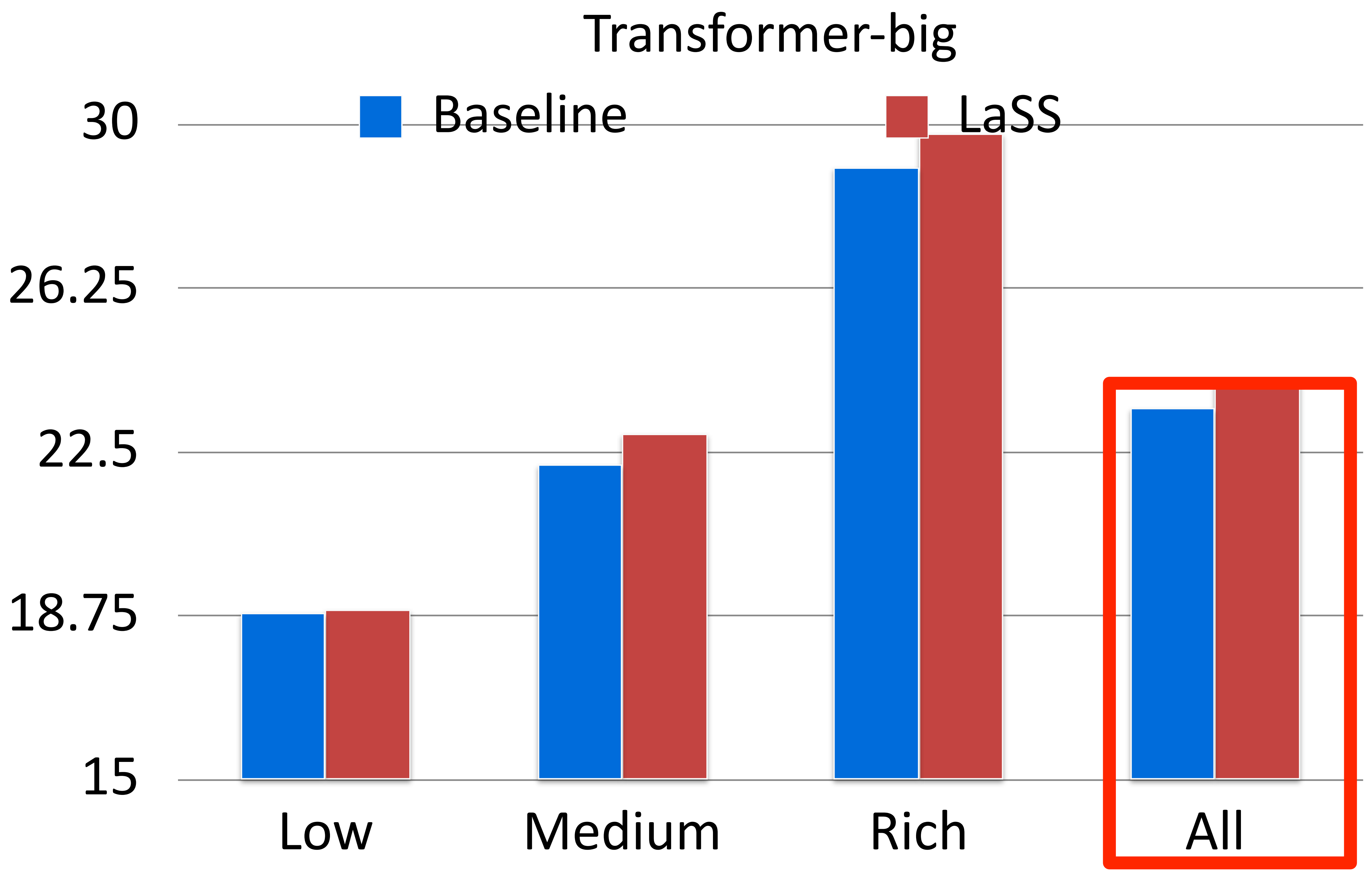


Learning Language Specific Sub-network for Multilingual Machine Translation [Lin et al., ACL 2021]

Efficacy in alleviating Parameter Interference

– WMT



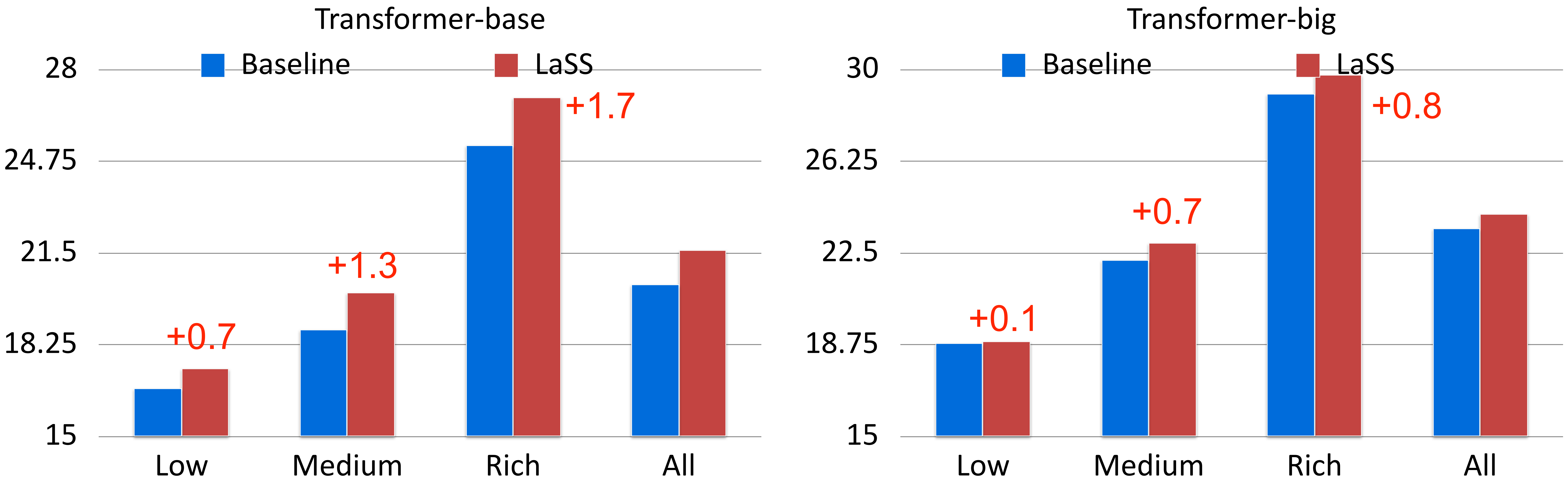


LaSS obtains consistent gains for both Transformer-base and Transformer-big

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

LaSS obtains more gains for rich resource

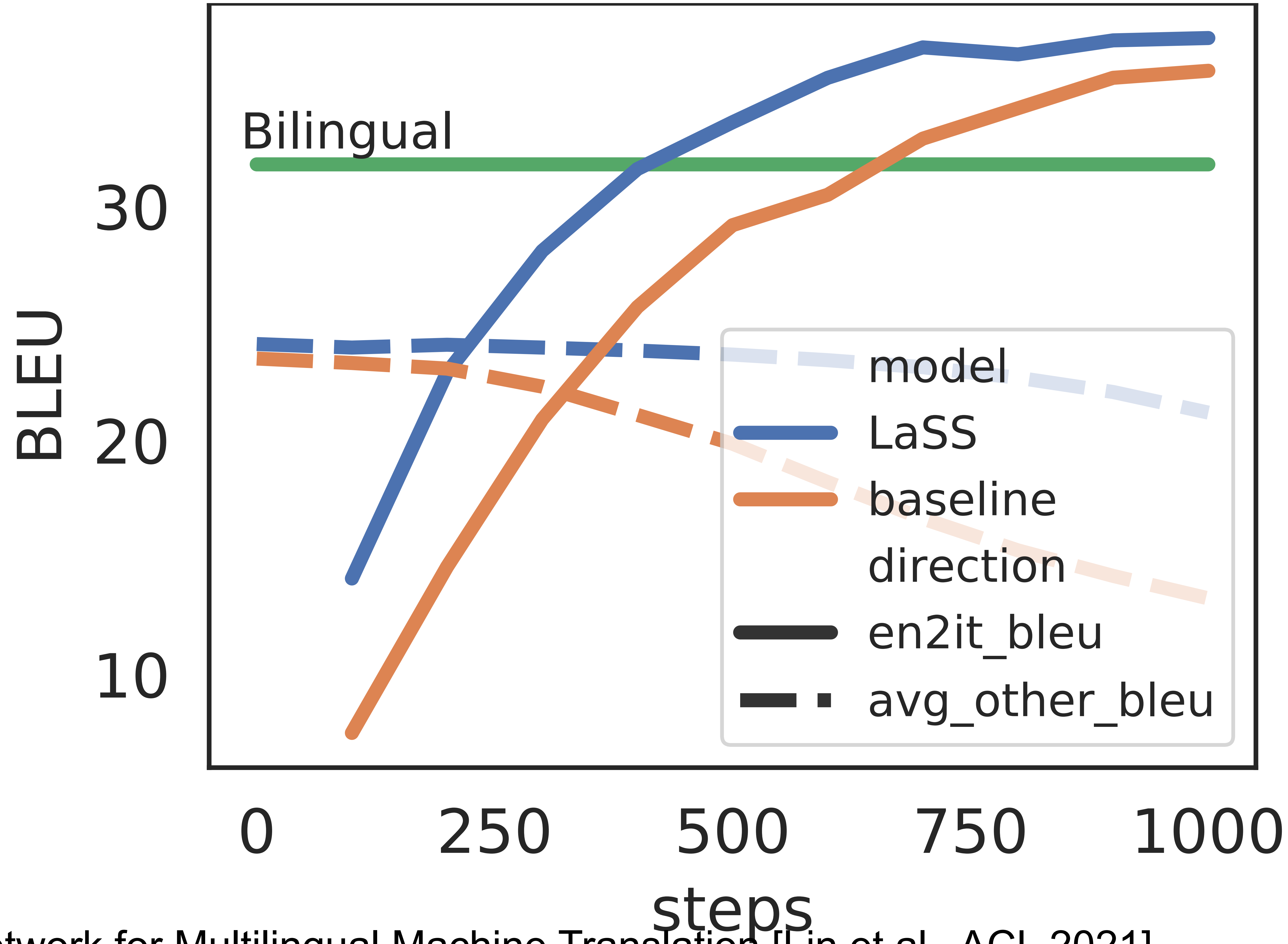
– WMT



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

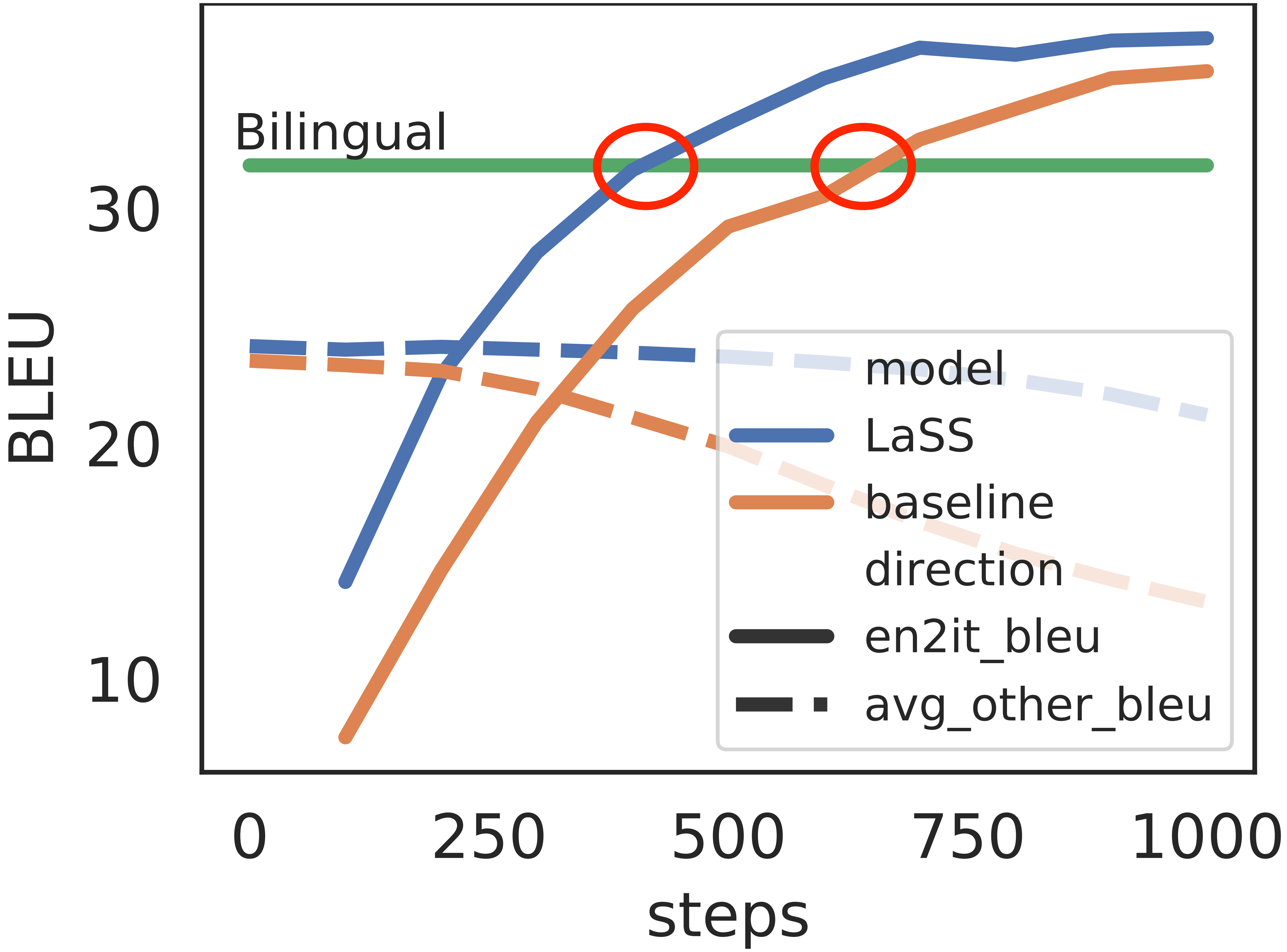


Learning Language Specific Sub-network for Multilingual Machine Translation [Lin et al., ACL 2021]

Adaptation to New Language Pairs

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

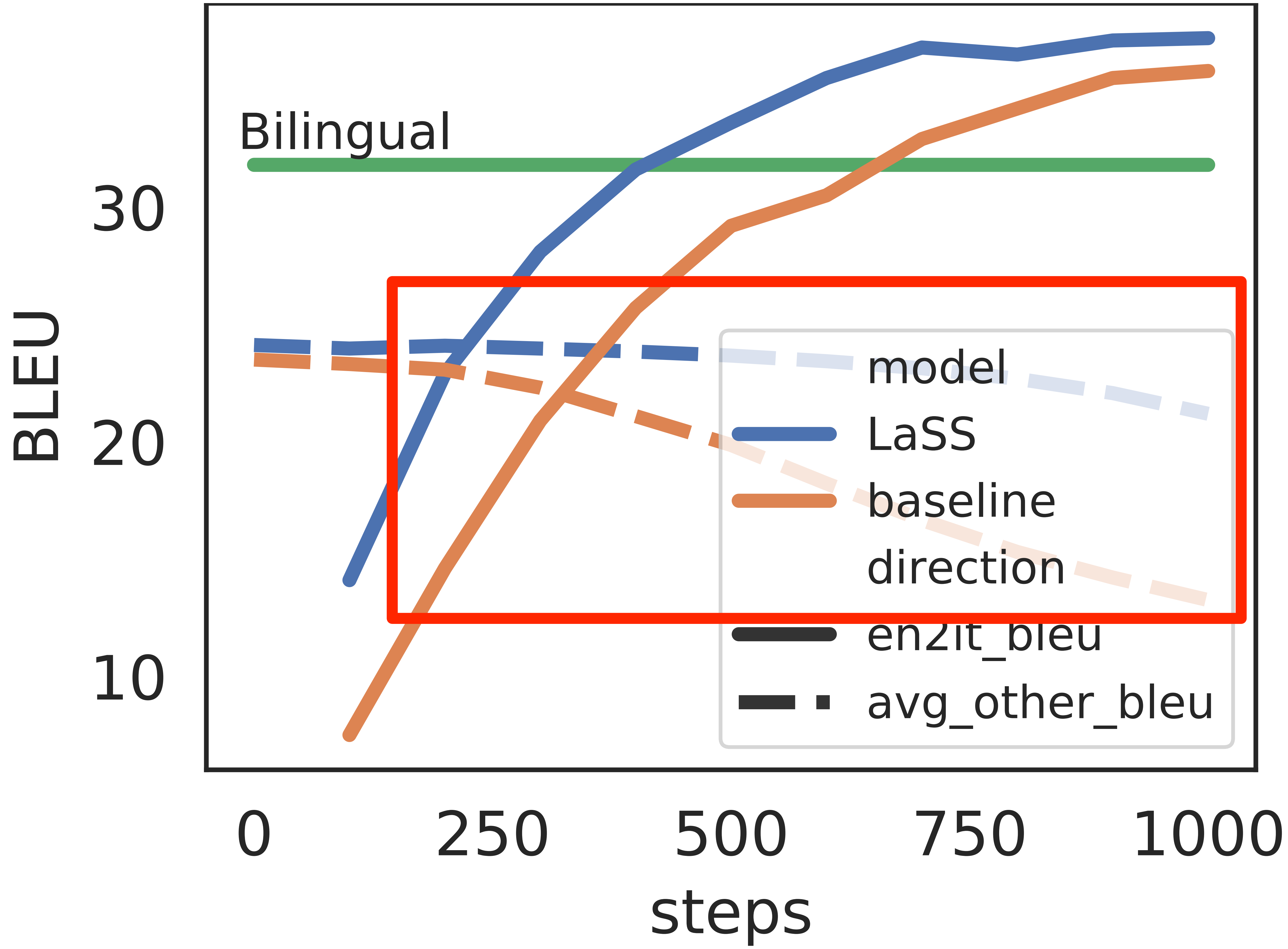
LaSS reaches the bilingual model performance with fewer steps.



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

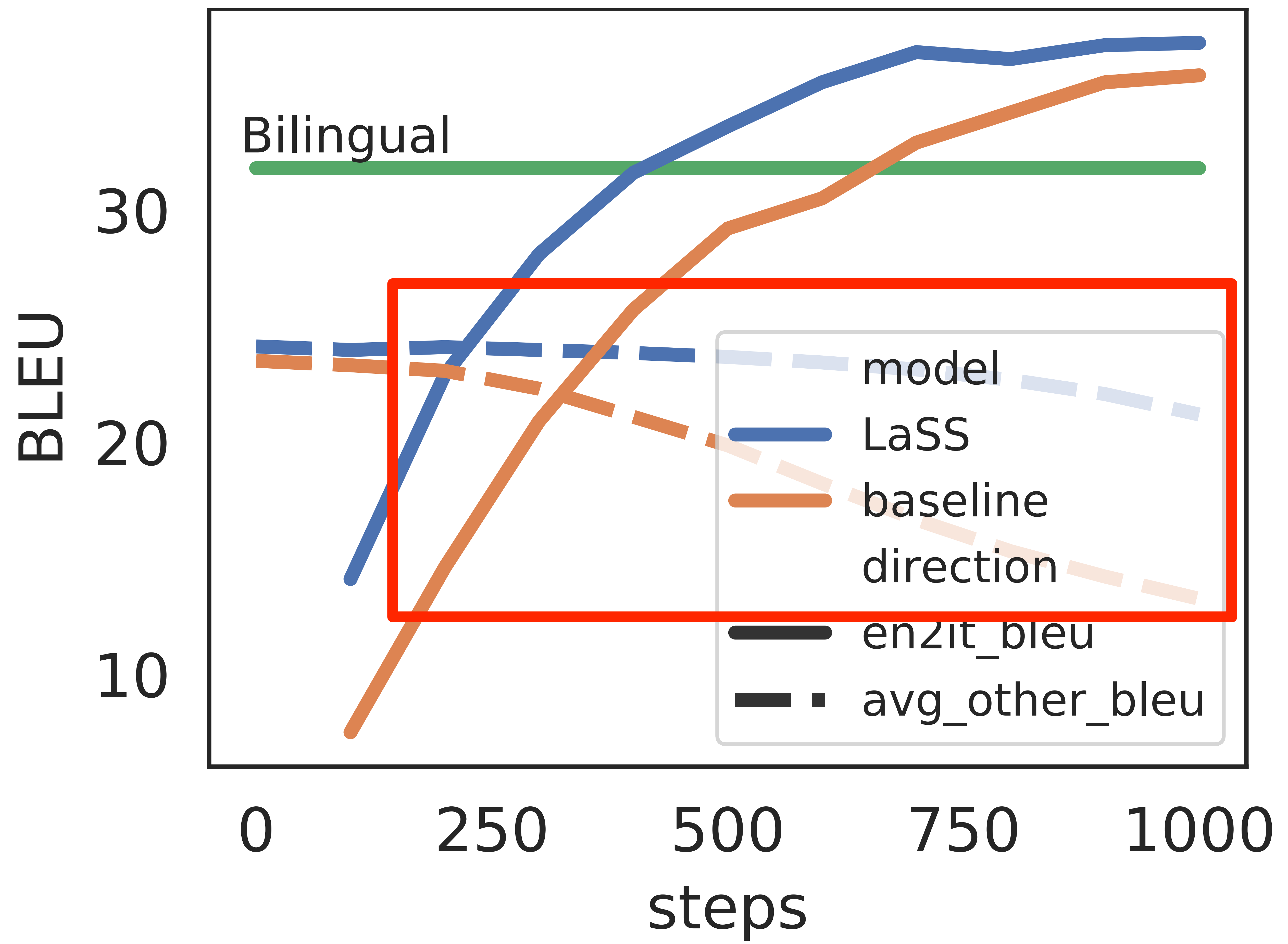


Adaptation to New Language Pairs

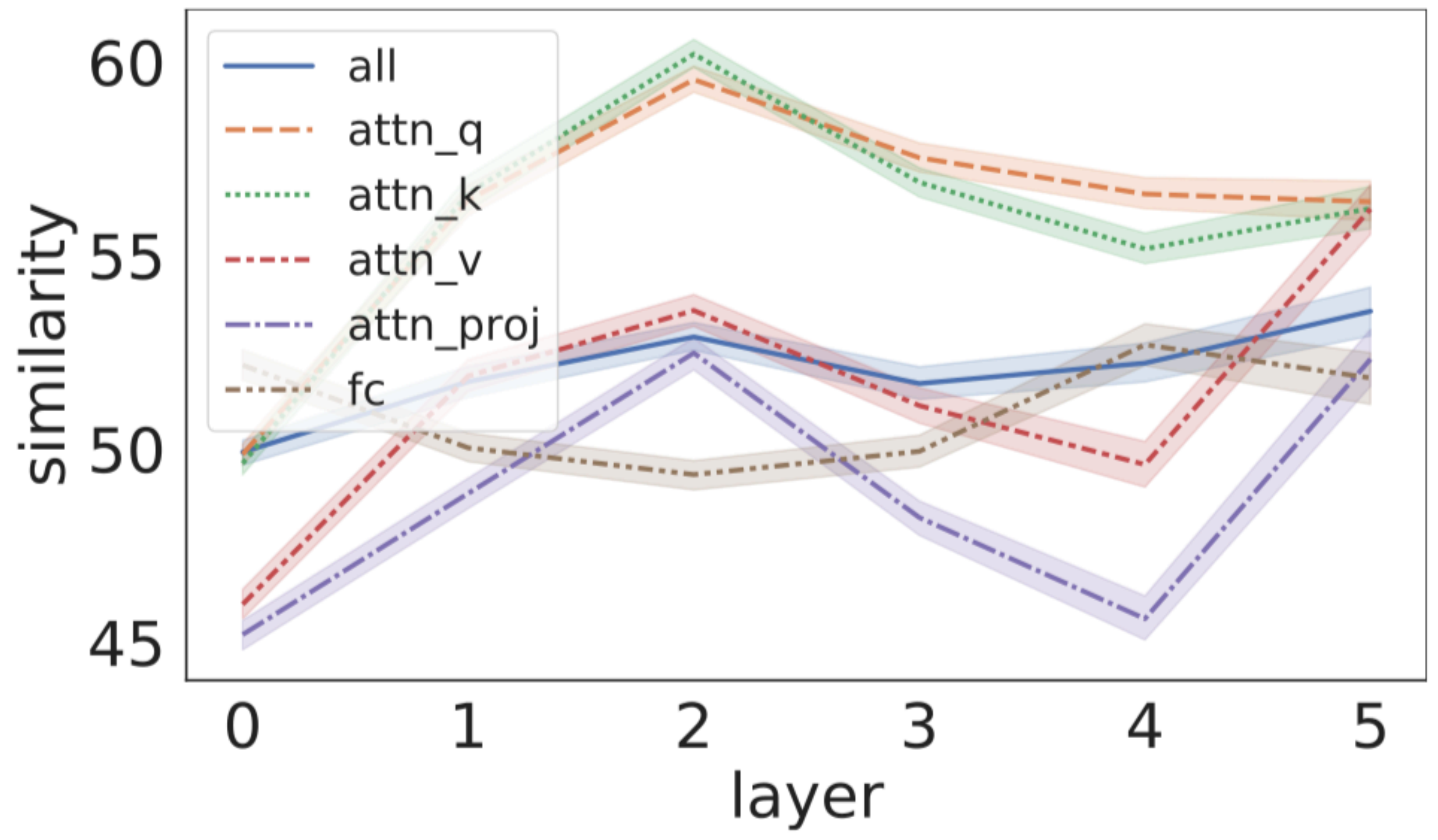
- Distribute a new sub-network for new language pair and train the sub-network for fixed steps

easy adaptation is attributed to the language specific sub-network

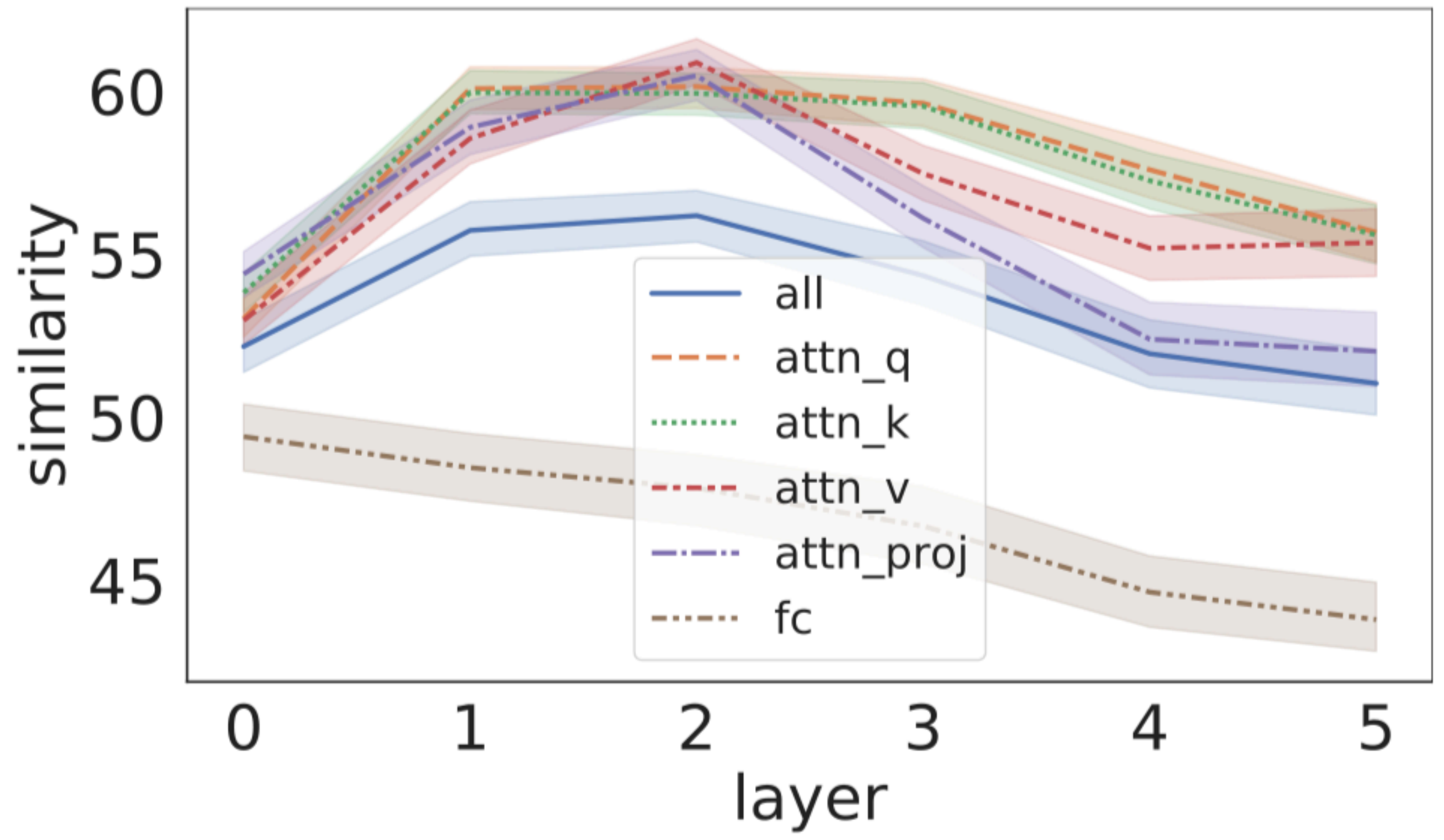
Only updates the corresponding parameters avoids catastrophic forgetting



Top/bottom layers prefer language specific capacity



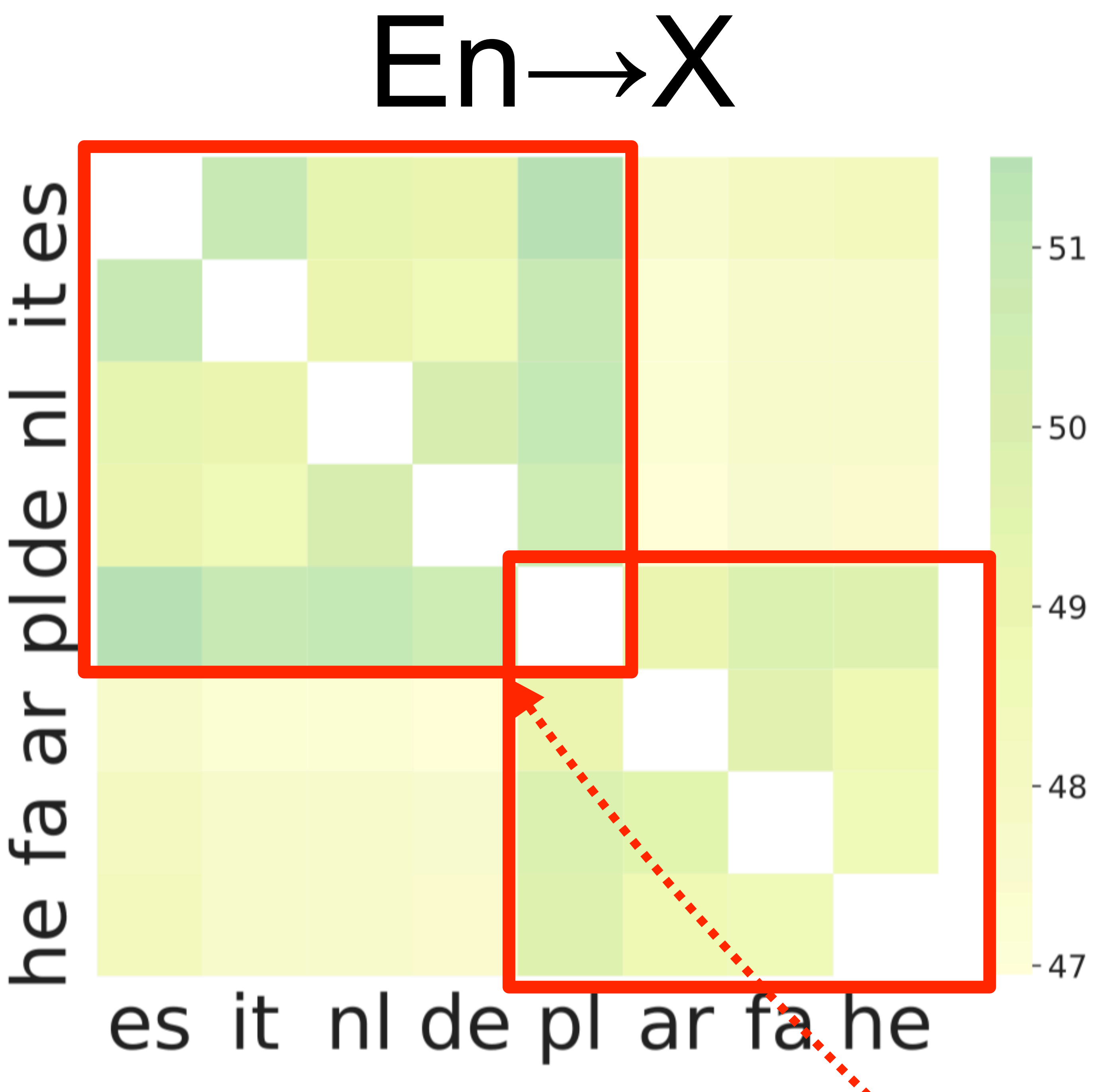
(a) Encoder

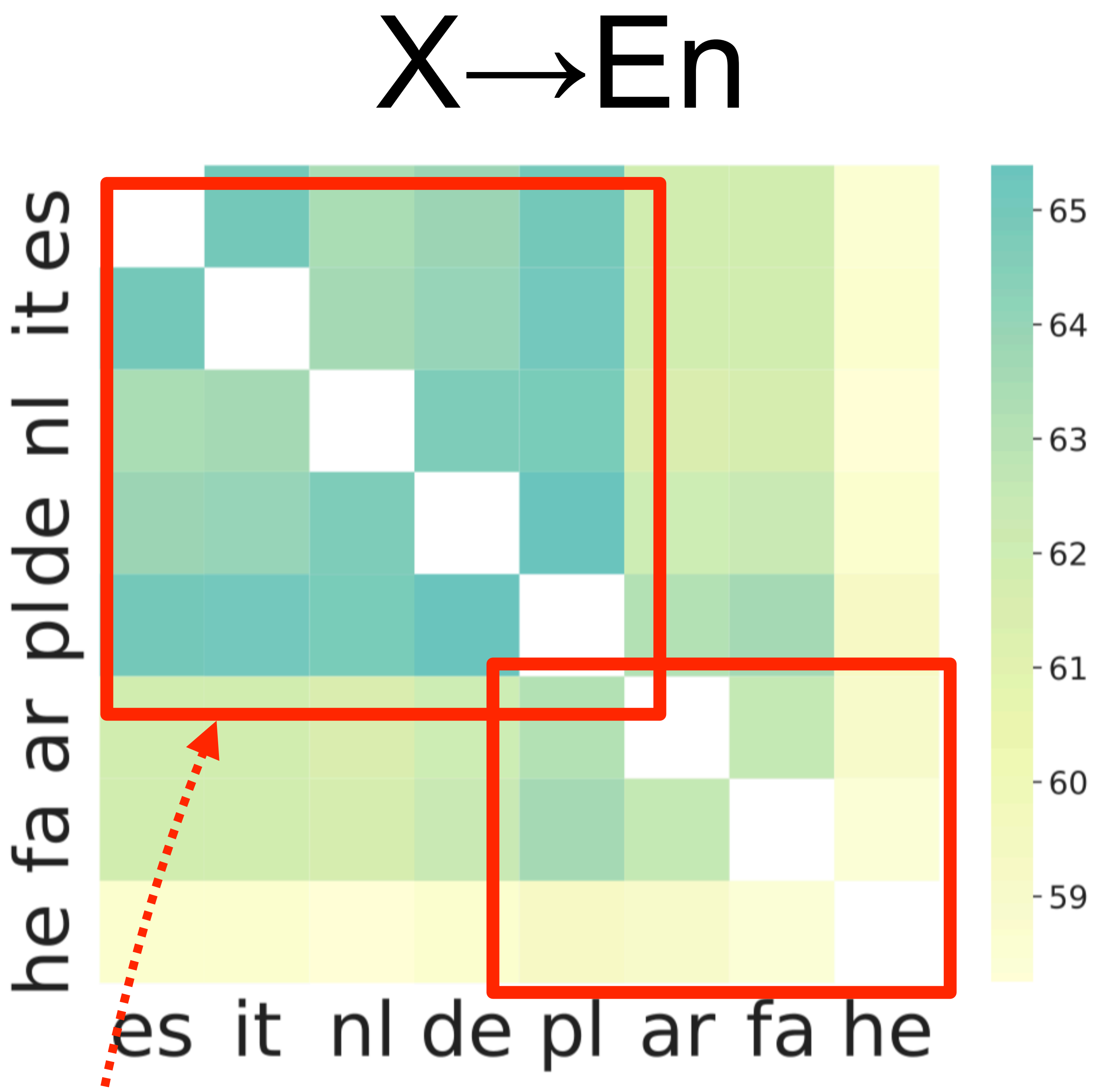


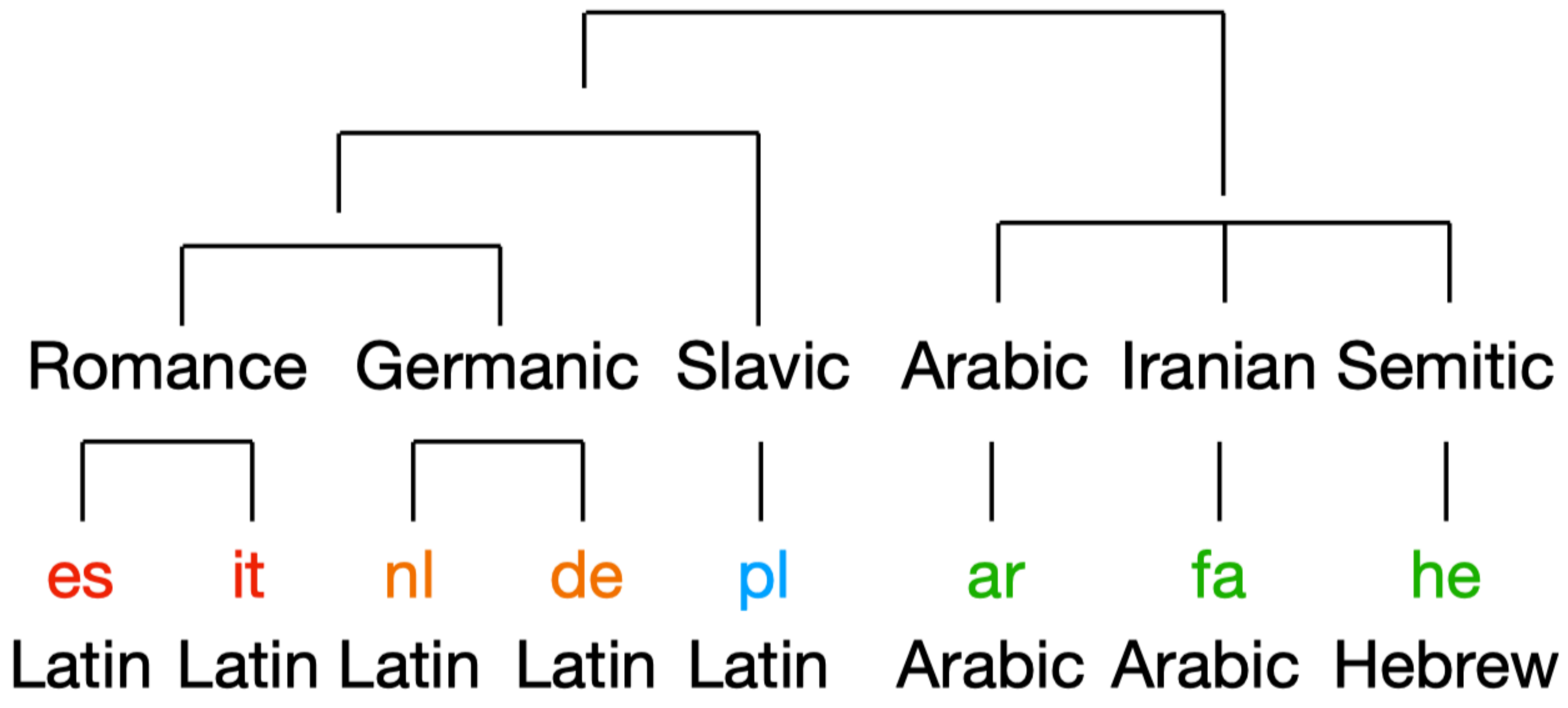
(b) Decoder

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

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

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

PART IV: Pre-training for Speech Translation

Speech-to-Text Translation(ST)

- source language ***speech(audio)*** ⟶ target lang ***text***

Application Type

- (Non-streaming) ST e.g. video translation
- Streaming ST e.g. realtime conference translation

System

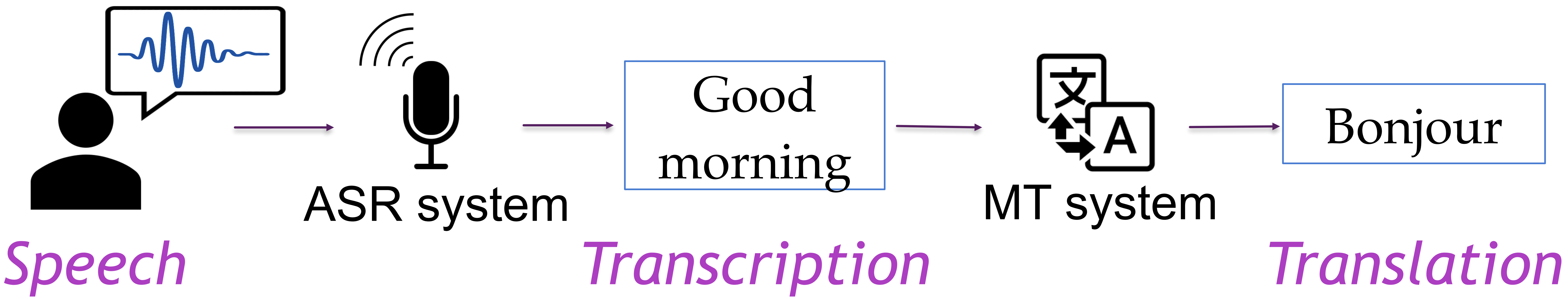
- Cascaded ST
- End-to-end ST

Cascaded ST System

- Challenges:

1.Computationally inefficient

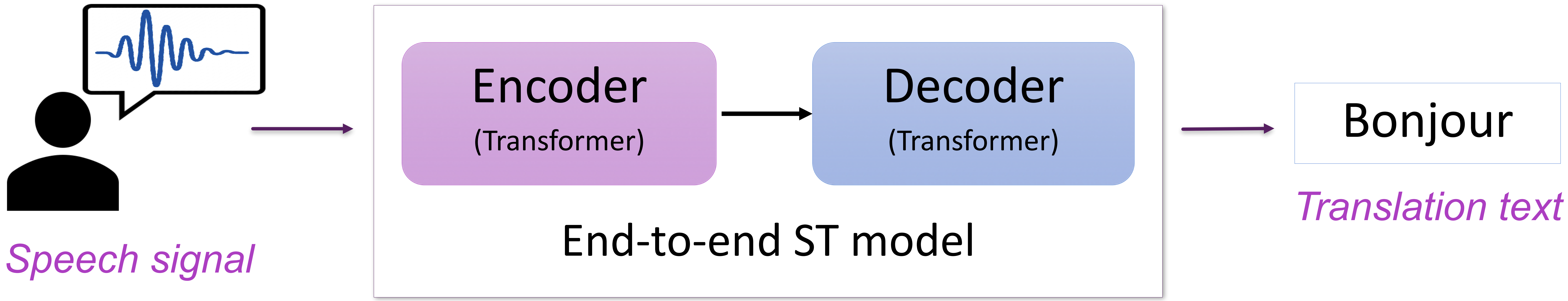
2.Error propagation: Wrong transcription ➡️ Wrong translation



do at this and see if it works for you ➡️ 这样做，看看它是否对你有用

duet this and see if it works for you ➡️ 二重奏一下，看看它是否对你有用

End-to-end ST Model



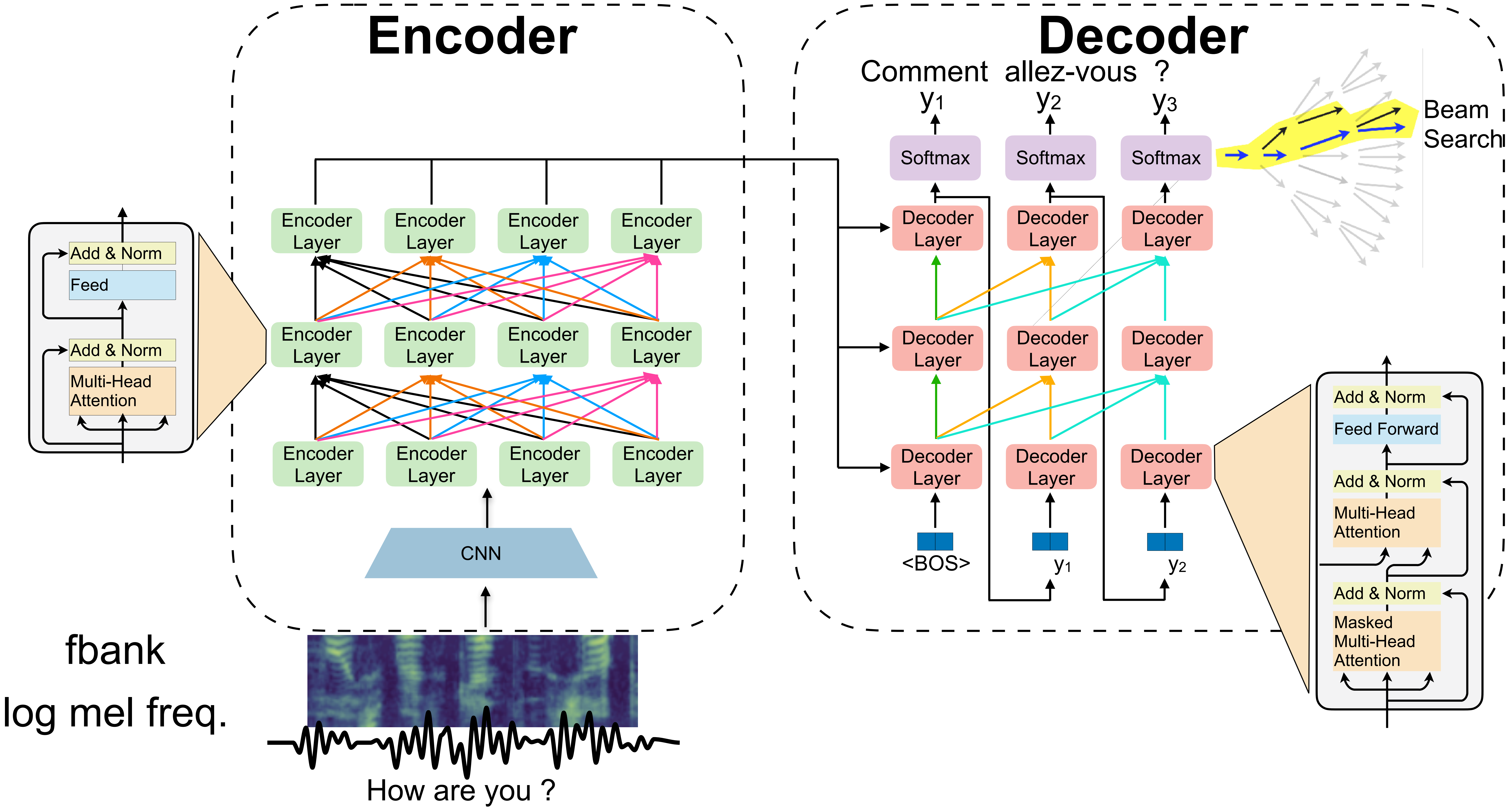
- Single model to produce text translation from speech
- Basic model: Encoder-Decoder architecture (e.g. Transformer)
- Advantage:
  - Reduced latency, simpler deployment
  - Avoid error propagation

[1] Bérard et al., Listen and translate: A proof of concept for end-to-end speech-to-text translation. 2016

Basic Speech Translation Model (Same as MT)

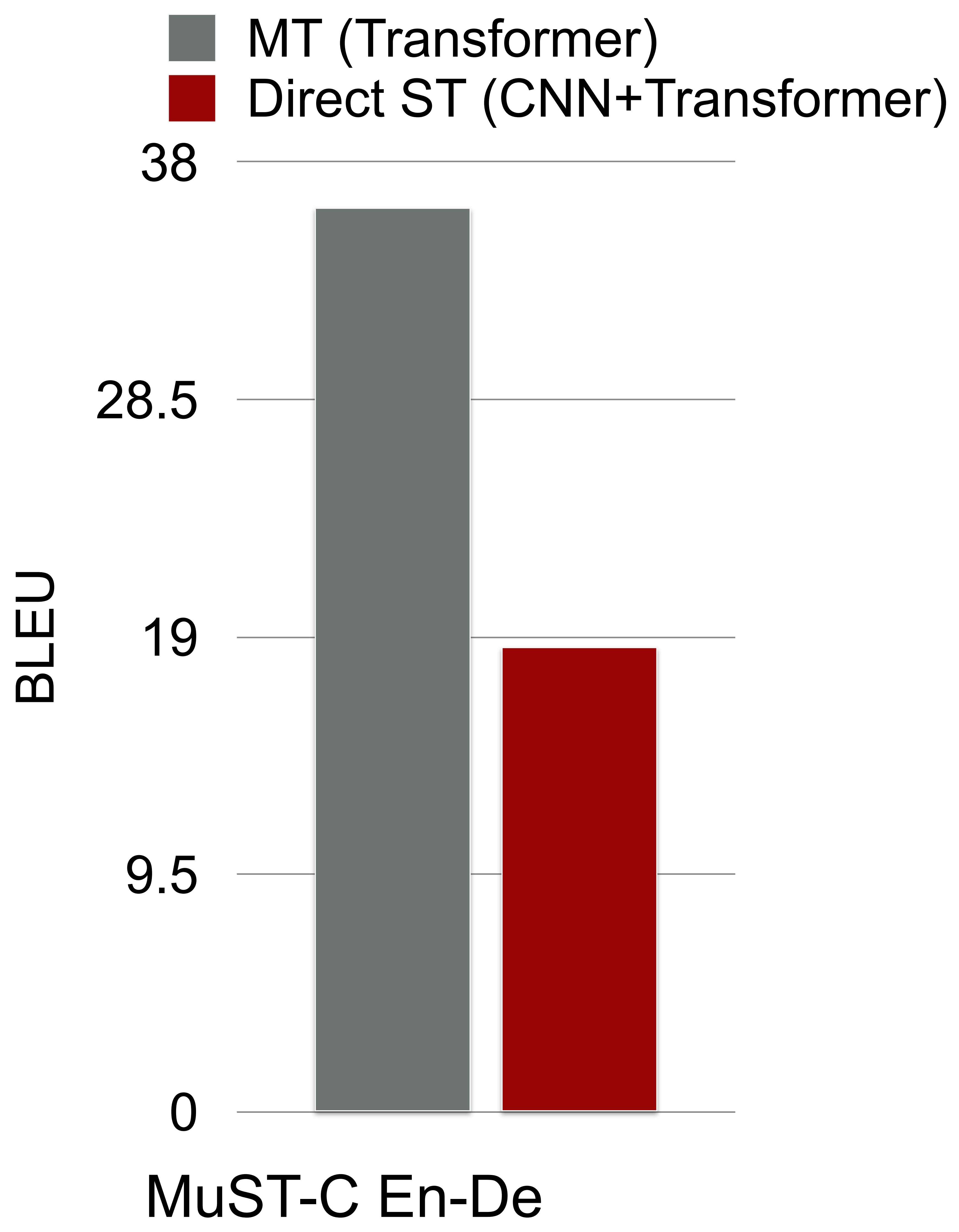
Transformer-based: N-layer convolution + attention encoder, M-layer decoder

Training data: <audio seq., translation text>



Challenge

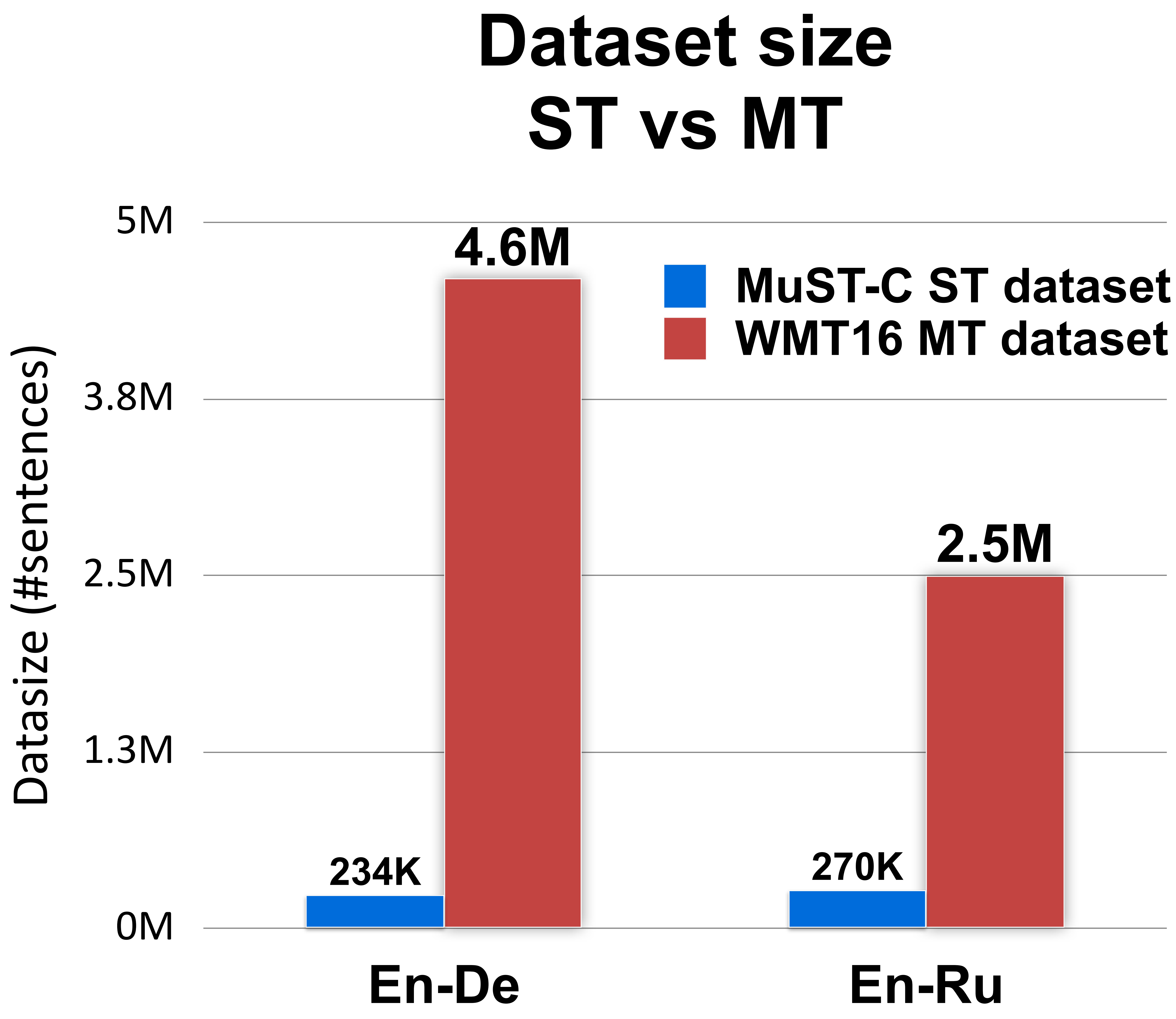
- Data scarcity - lack of large parallel audio-translation corpus
- Modality disparity between audio and text
- Performance gap of direct ST:
  - BLEU: ST 18.6 vs. MT 36.2 (on MuST-C En-De)



Pre-training for Speech Translation

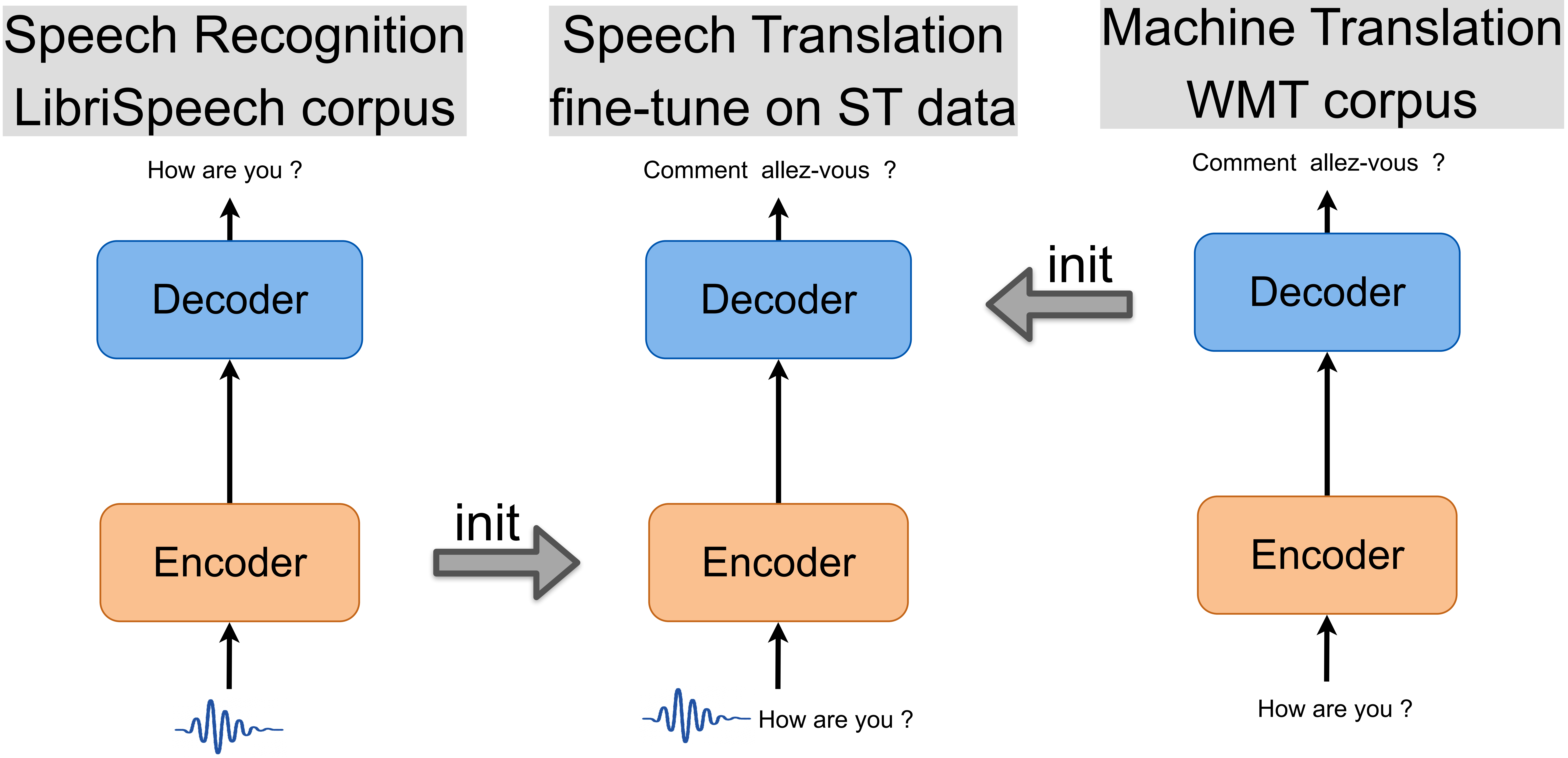
- MT Pre-training
  - Decoder initialization from separately trained MT model
  - Single-modal(audio) Encoder-Decoder: COSTT[Dong et al, AAAI 2021b]
- ASR Pre-training
  - Curriculum Pre-training [Wang et al, ACL 2020]
  - LUT [Dong et al, AAAI 2021a]
- Audio Pre-training
  - Wav2vec & Wav2Vec2.0 [Schneider et al. Interspeech 2019, Baevski et al NeurIPS2020]
  - Apply to ST [Wang et al, 2021, Zhao et al, ACL 2021, Wang et al, Interspeech 2021]
- Raw Text Pre-training
  - LUT [Dong et al, AAAI 2021a]
- Bi-modal Pre-training
  - TCEN-LSTM [Wang et al, AAAI 2020]
  - Chimera [Han et al, ACL 2021a]
  - XSTNet [Ye et al, Interspeech 2021]
  - Wav2vec2.0 + mBart + Self-training [Li et al, ACL 2021b]
  - FAT-ST [Zheng et al, ICML 2021]

Using external Parallel Text

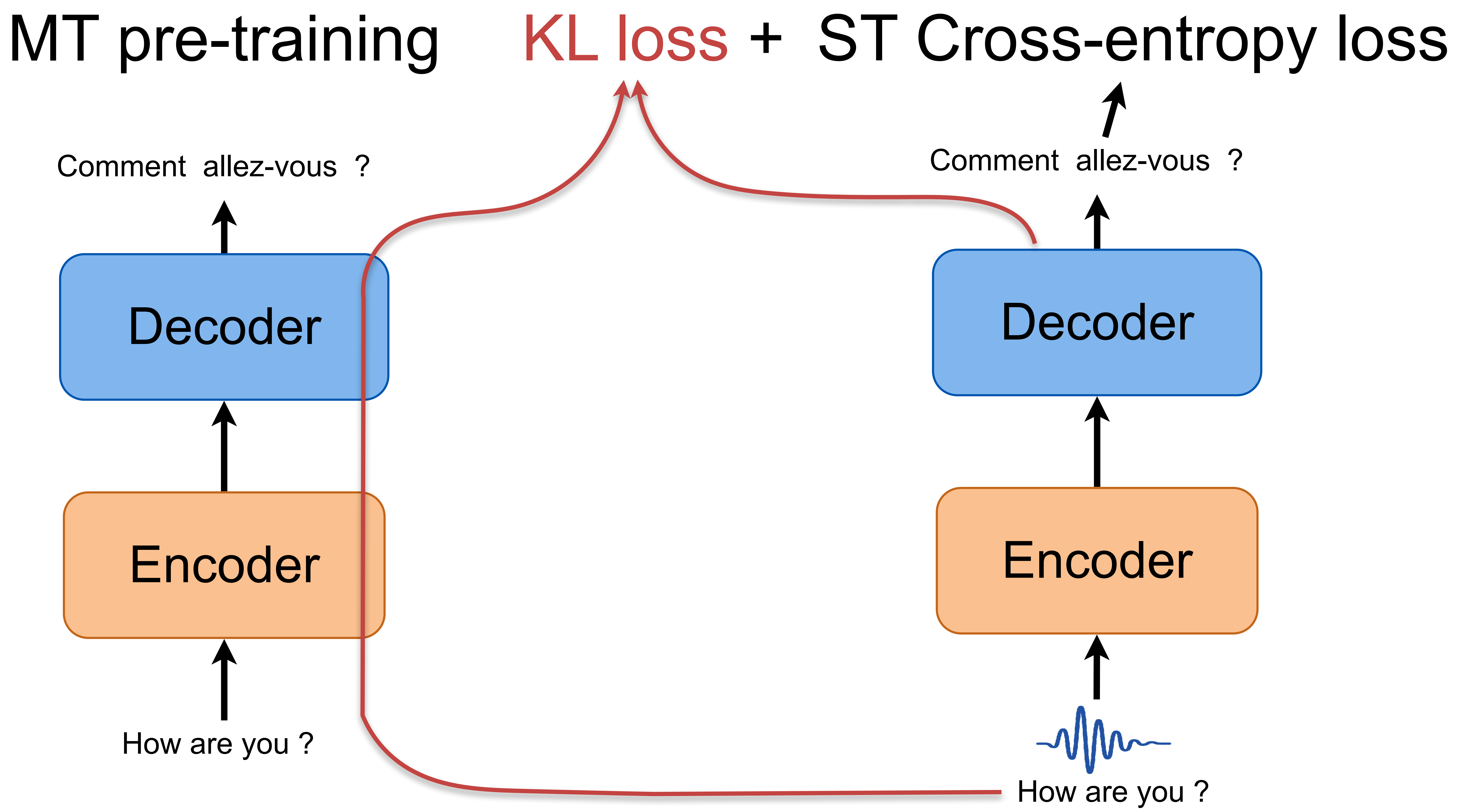


How to use MT data *with much larger scale* to improve ST performance?

Separate Encoder-Decoder Pre-train



Knowledge Distillation from MT model



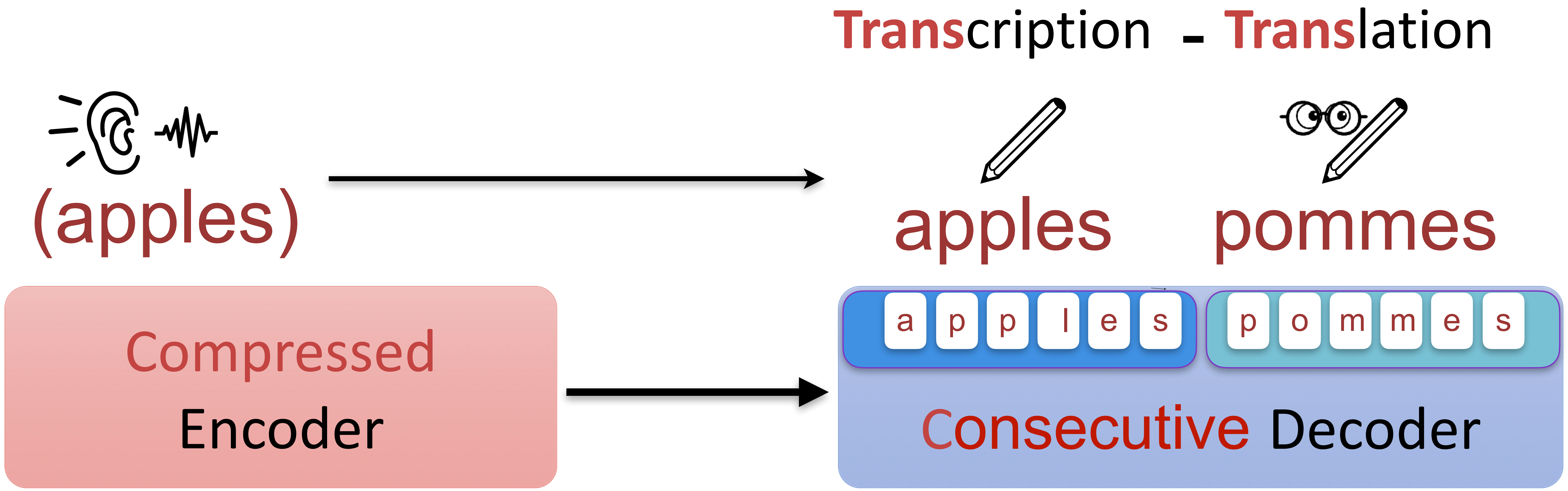
End-to-End Speech Translation with Knowledge Distillation [Liu et al, Interspeech 2019]

Pre-train ST's decoder with full MT

How to make a single model's decoder to perform text translation?

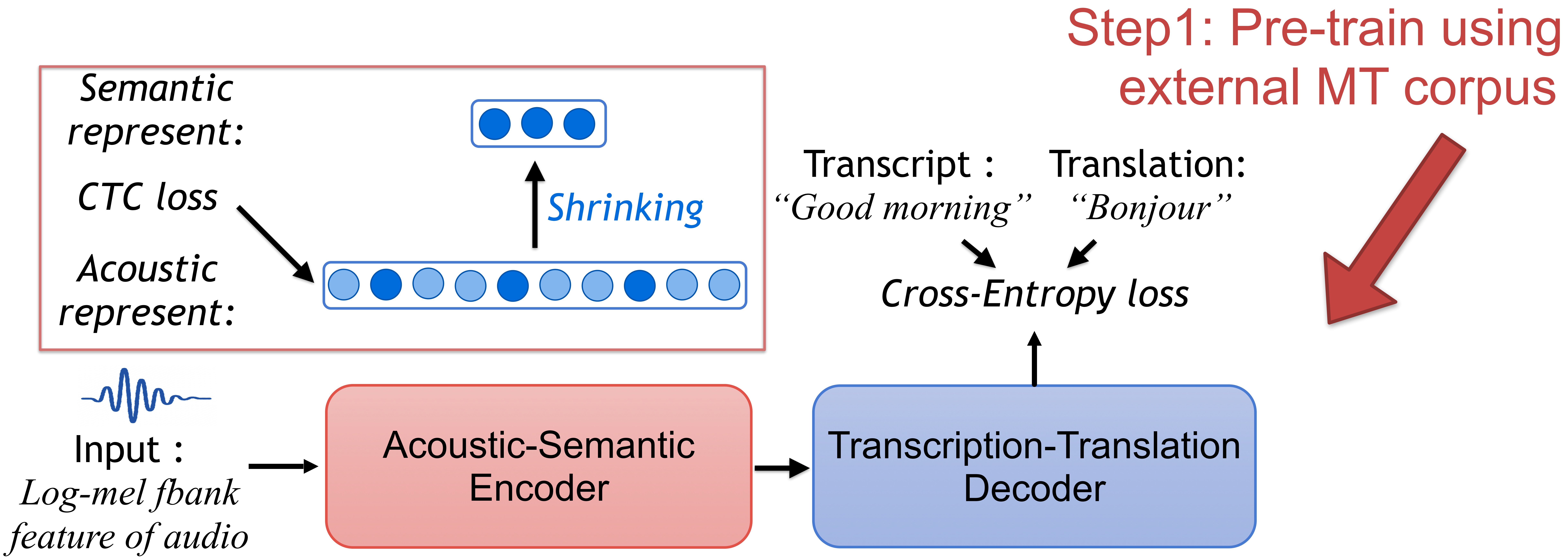
Decoder ==> translation

Encoder -> Decoder ==> transcribe and translation



Consecutive Decoding for Speech-to-text Translation [Q. Dong, M. Wang, H. Zhou, S. Xu, B. Xu, Lei Li, AAAI 2021]

COSTT for ST



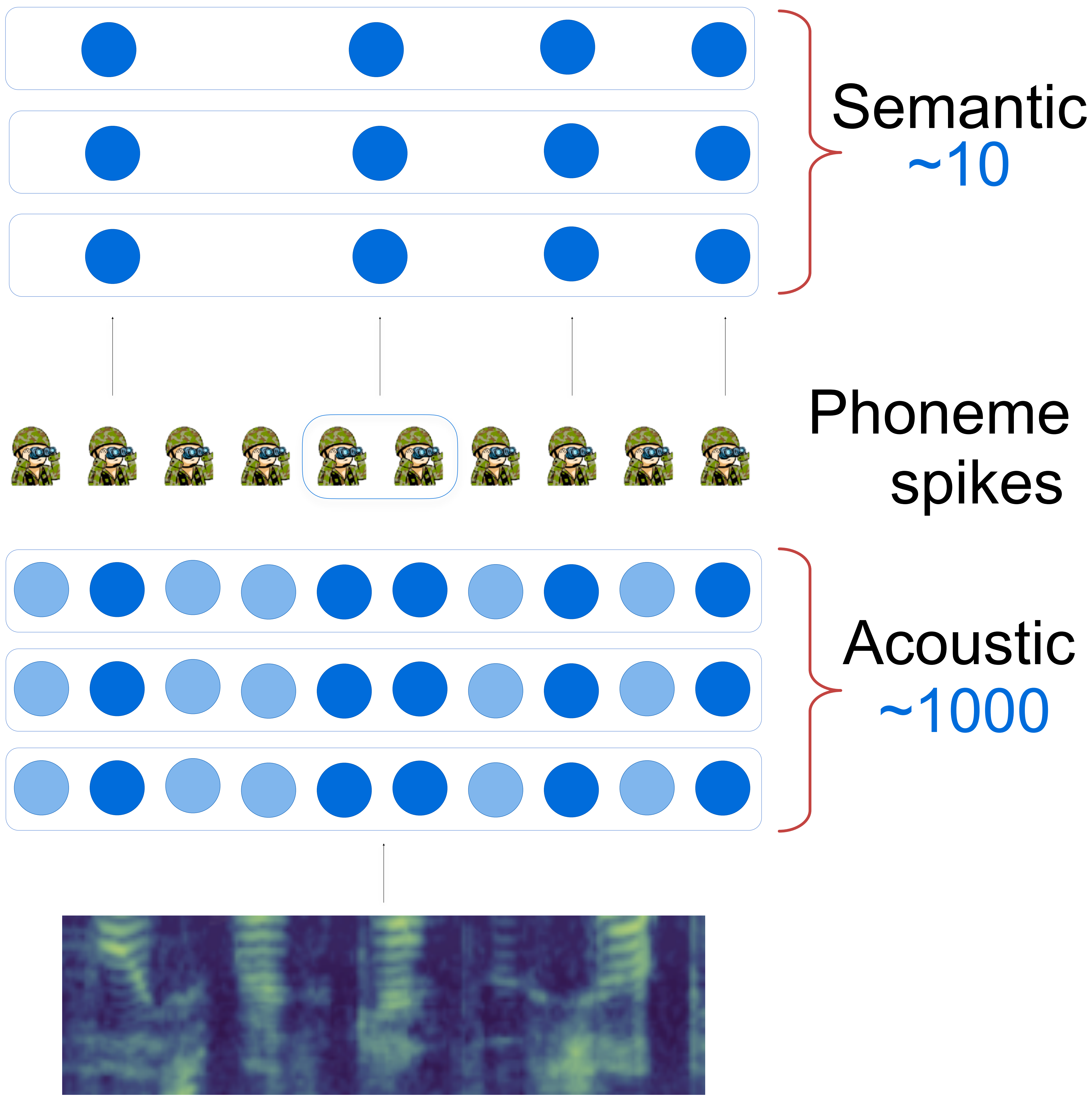
Step 2: Train encoder w/ shrinking module using CTC

Step 3: Train full model on ST data <audio, transcript, translation>

Consecutive Decoding for Speech-to-text Translation [Q. Dong, M. Wang, H. Zhou, S. Xu, B. Xu, Lei Li, AAAI 2021]

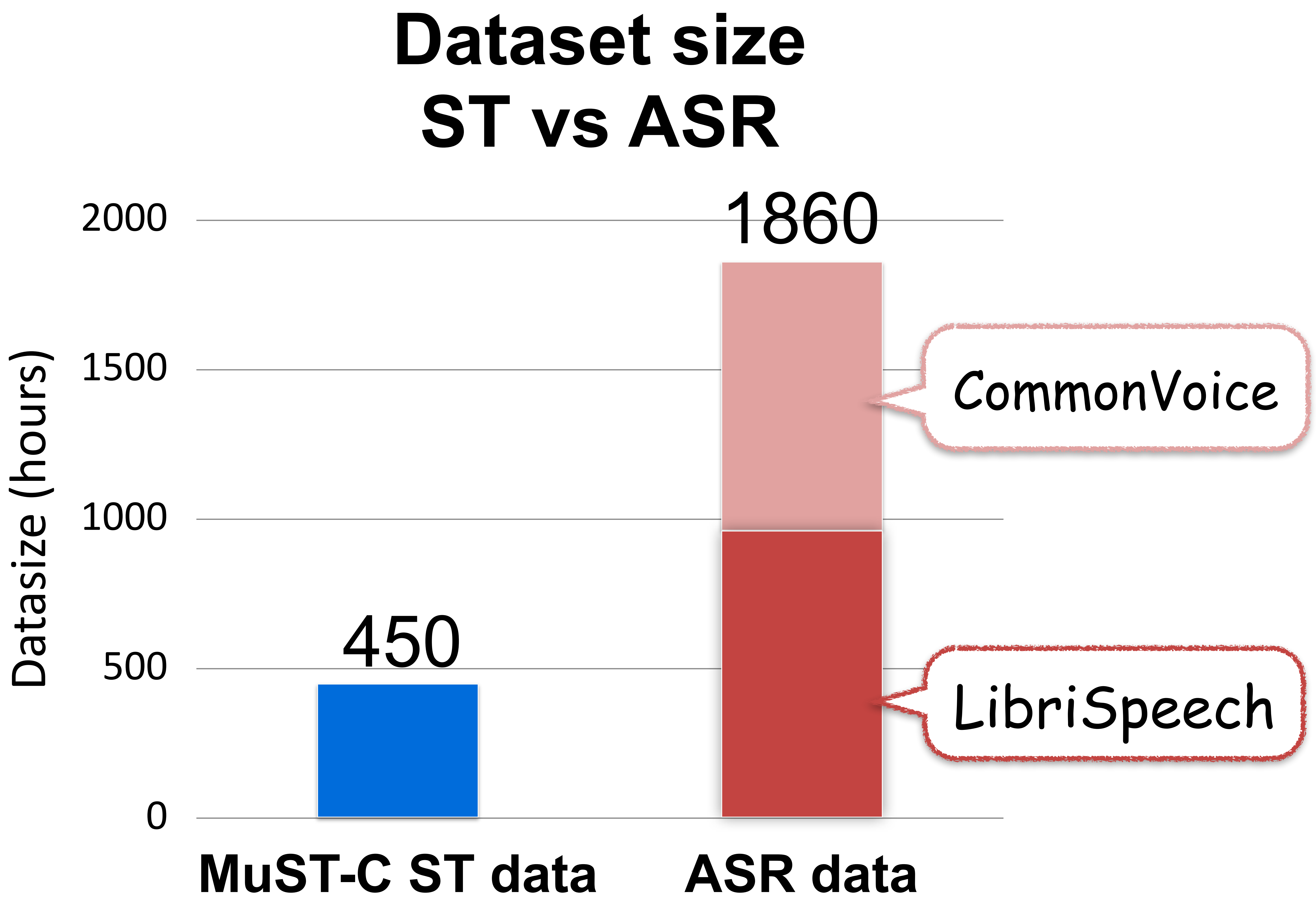
Advantages of COSTT

- Unified training with both transcript and translation text
- Reduced encoder output size with CTC-guided shrinking
- Able to pre-train the decoder with external MT parallel data



Consecutive Decoding for Speech-to-text Translation [Q. Dong, M. Wang, H. Zhou, S. Xu, B. Xu, Lei Li, AAAI 2021]

Using external ASR data

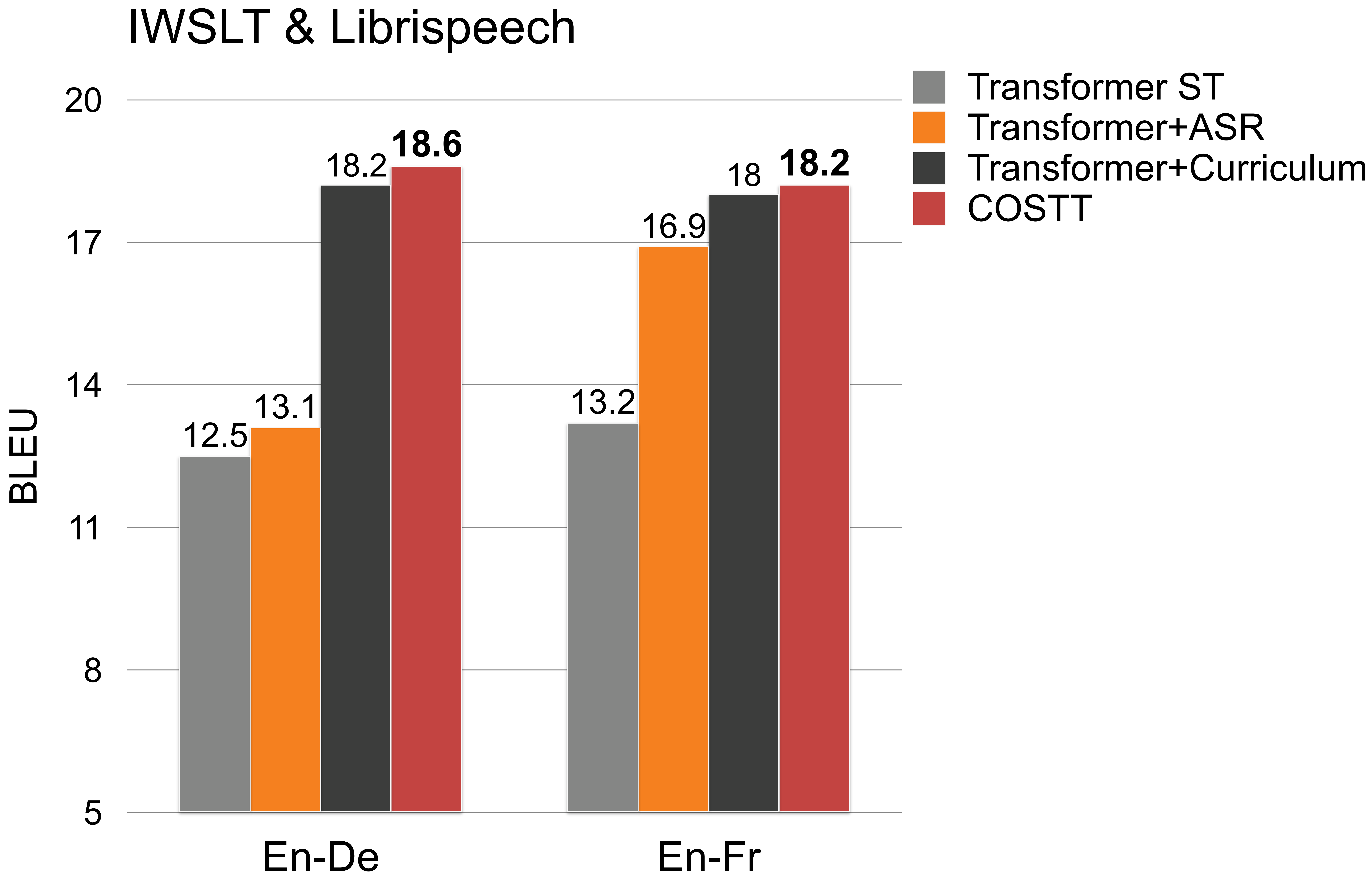


How to use larger external <u>ASR data</u> to improve ST performance?

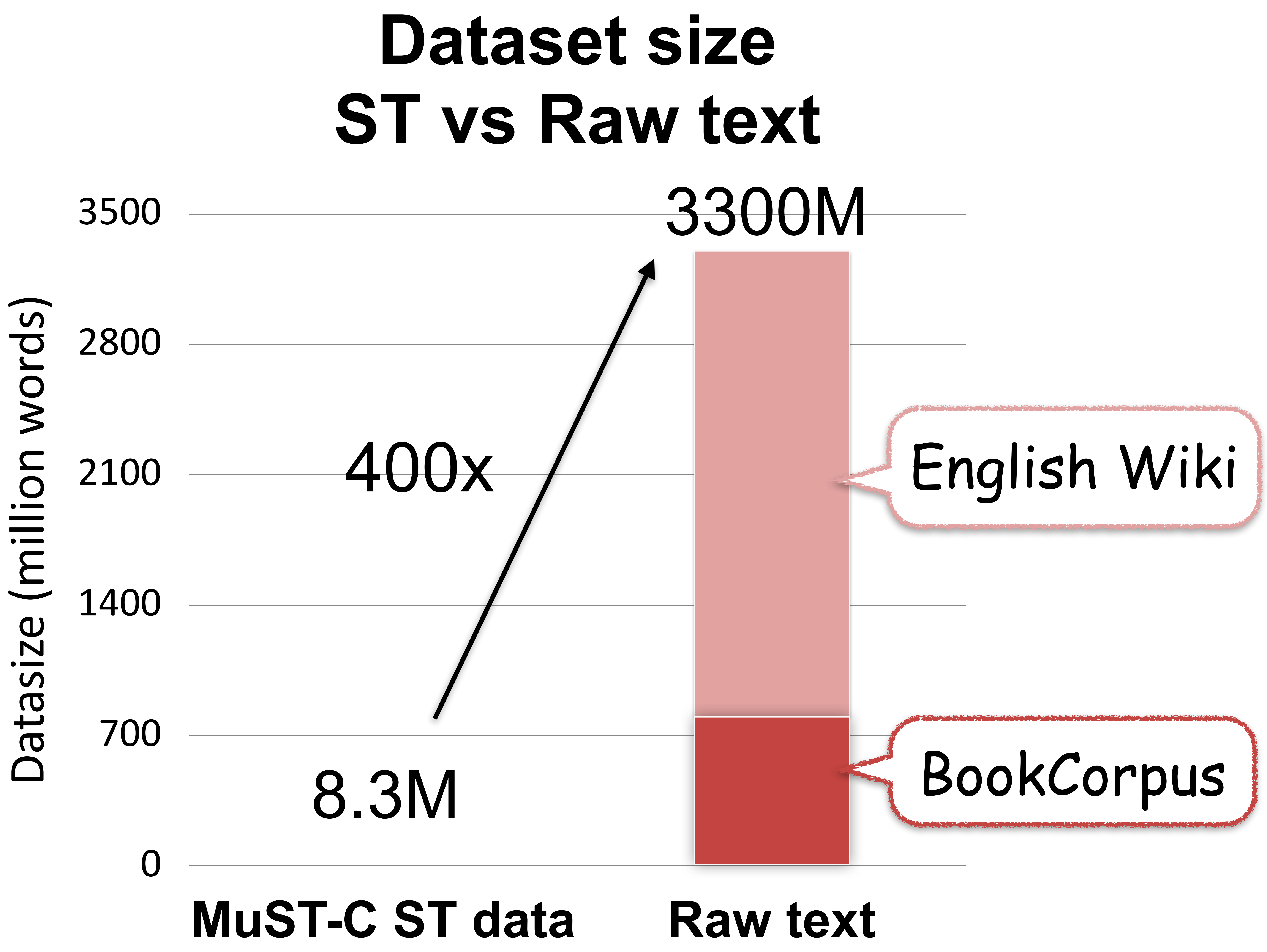
Curriculum Pre-training with ASR data

Curriculum Pre-training for End-to-end Speech Translation [Wang et al, ACL 2020]

ASR Pre-training helps ST



Raw Text Pre-training

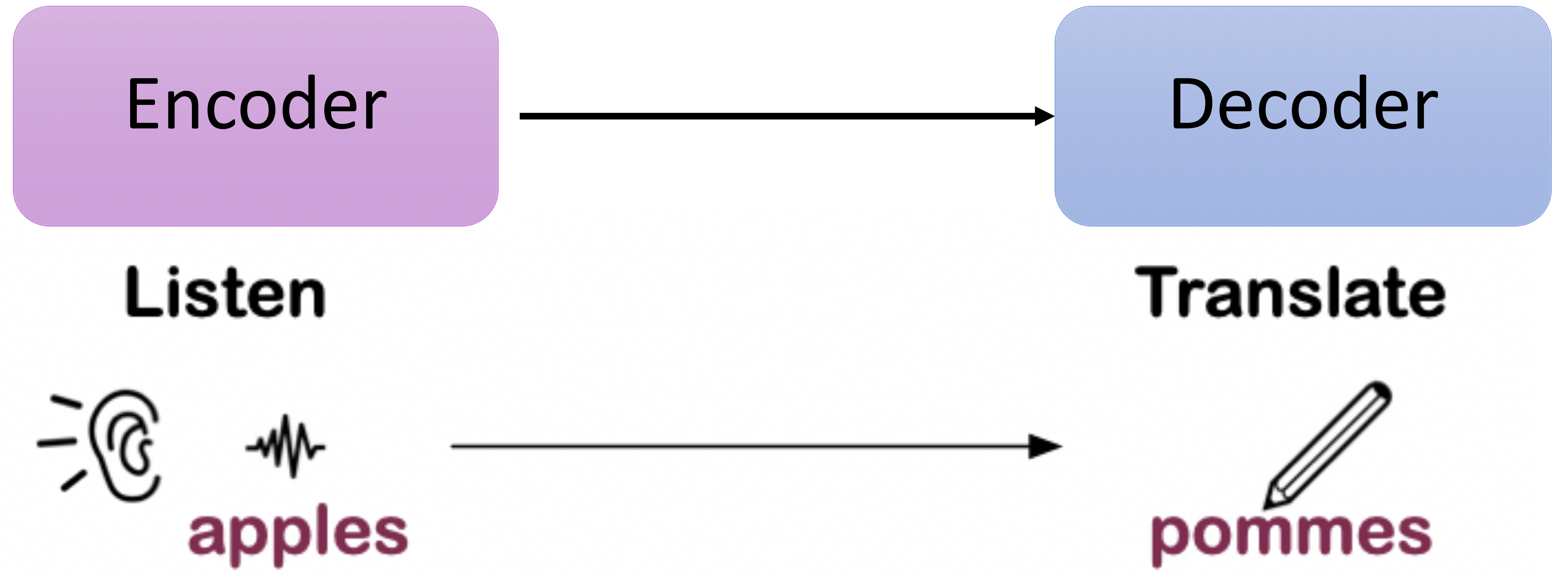


Using pre-trained LM in decoding weighting is easy!

But

🤔 How to use pre-trained BERT to improve ST performance?

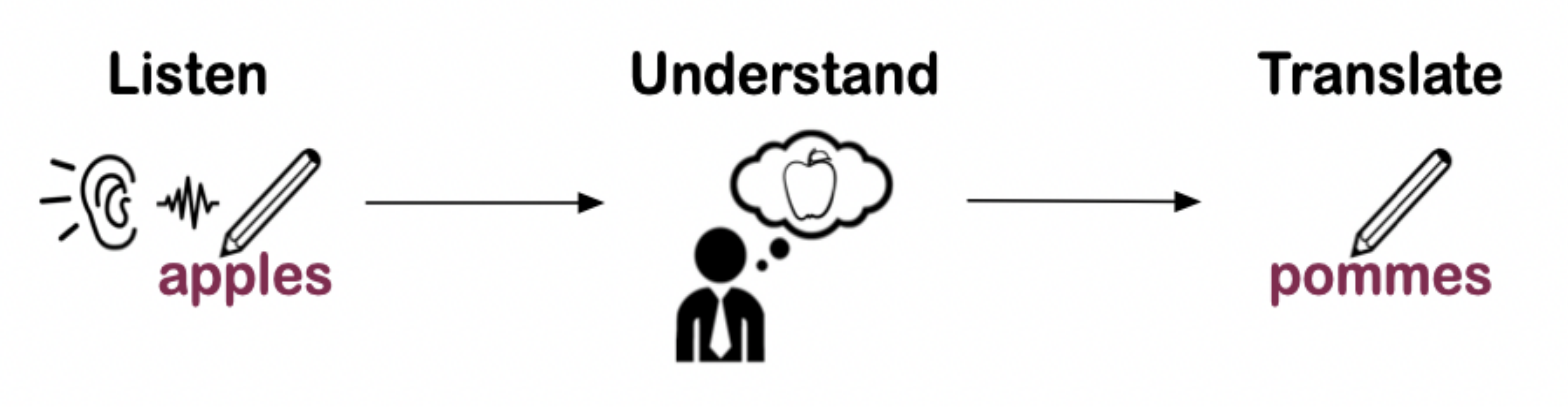
Drawbacks of the Encoder-Decoder Structure



1. A single encoder is hard to capture the representation of audio for the translation.
2. Limited in utilizing the information of *"transcription"* in the training.

Motivation: Mimic human's behavior

Question: How human translate?

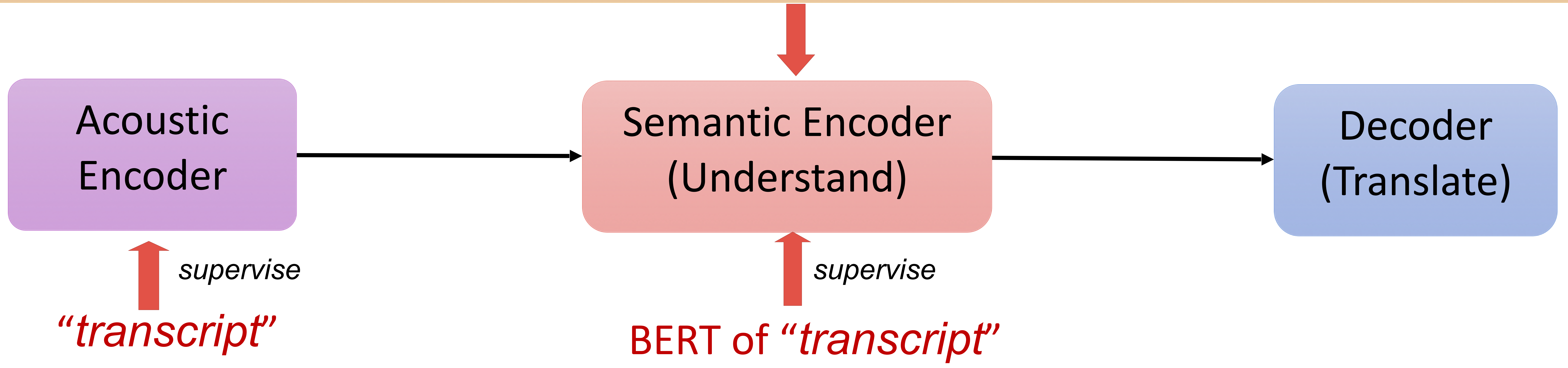


“Listen-Understand-Translate”(LUT) model based motivated by human's behavior

Motivation of Better Encoding

Drawback 1: A single encoder is not enough.

Idea 1: Introduce a semantic encoder



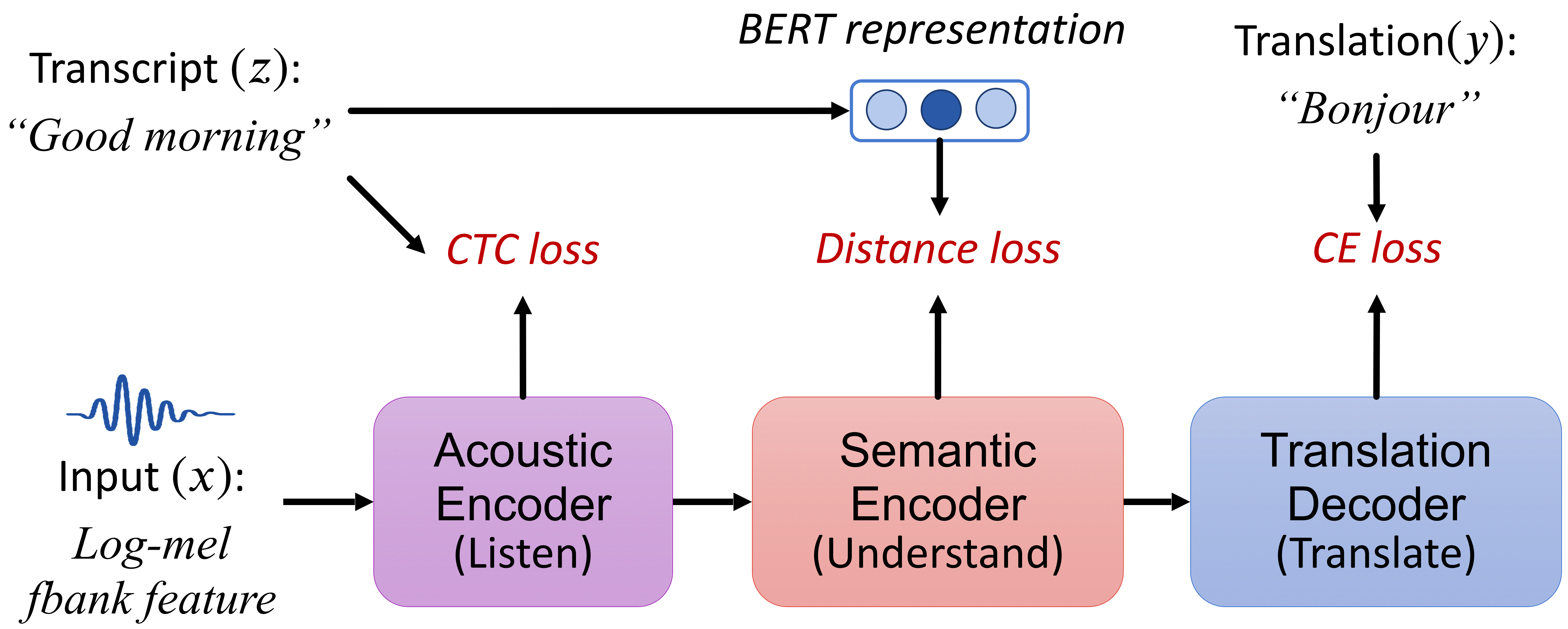
Drawback 2: Limit in using "transcript" info.

Idea 2: Utilizing the pre-trained representation (e.g. BERT) of the "transcript" to learn the semantic feature.

LUT: Utilizing Pre-trained Model on Raw Text

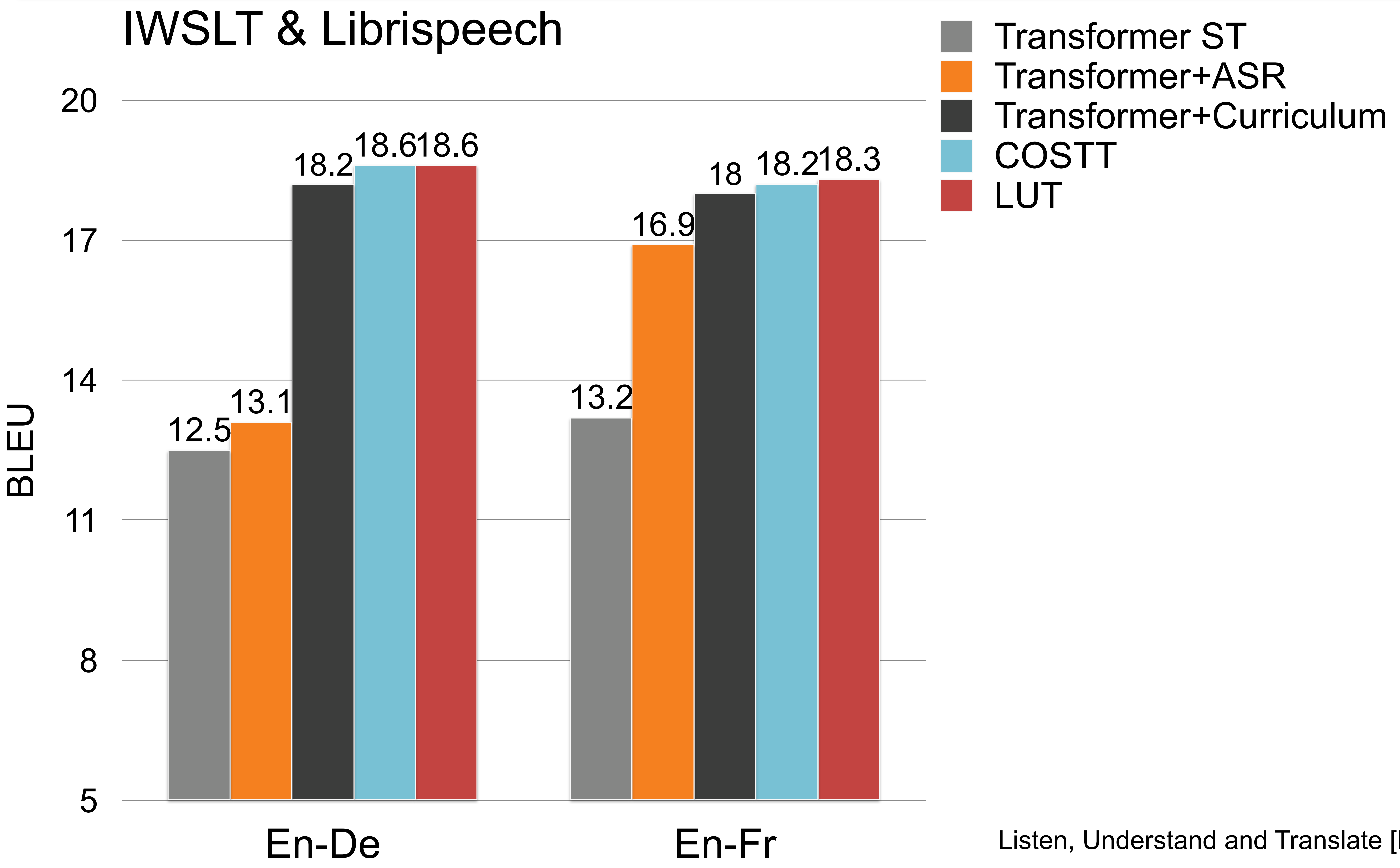
Training data: triples of

<speech, transcript_text, translate_text>



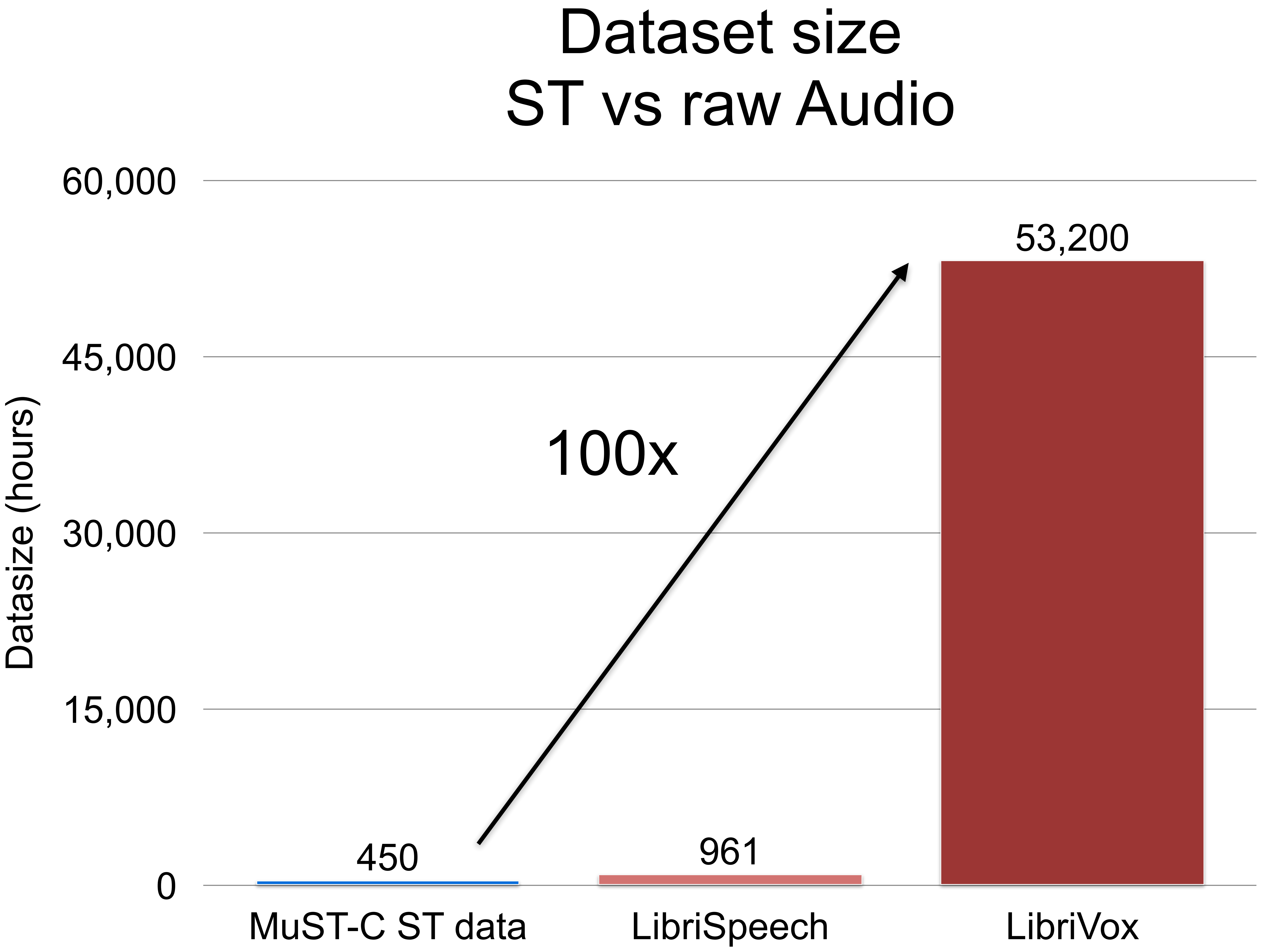
Listen, Understand and Translate [Q. Dong, R. Ye, M. Wang, H. Zhou, S. Xu, B. Xu, Lei Li, AAAI 2021]

ST Benefits from BERT, with Raw Text Pre-training



Listen, Understand and Translate [Dong et al, AAAI 2021]

Audio Pre-training

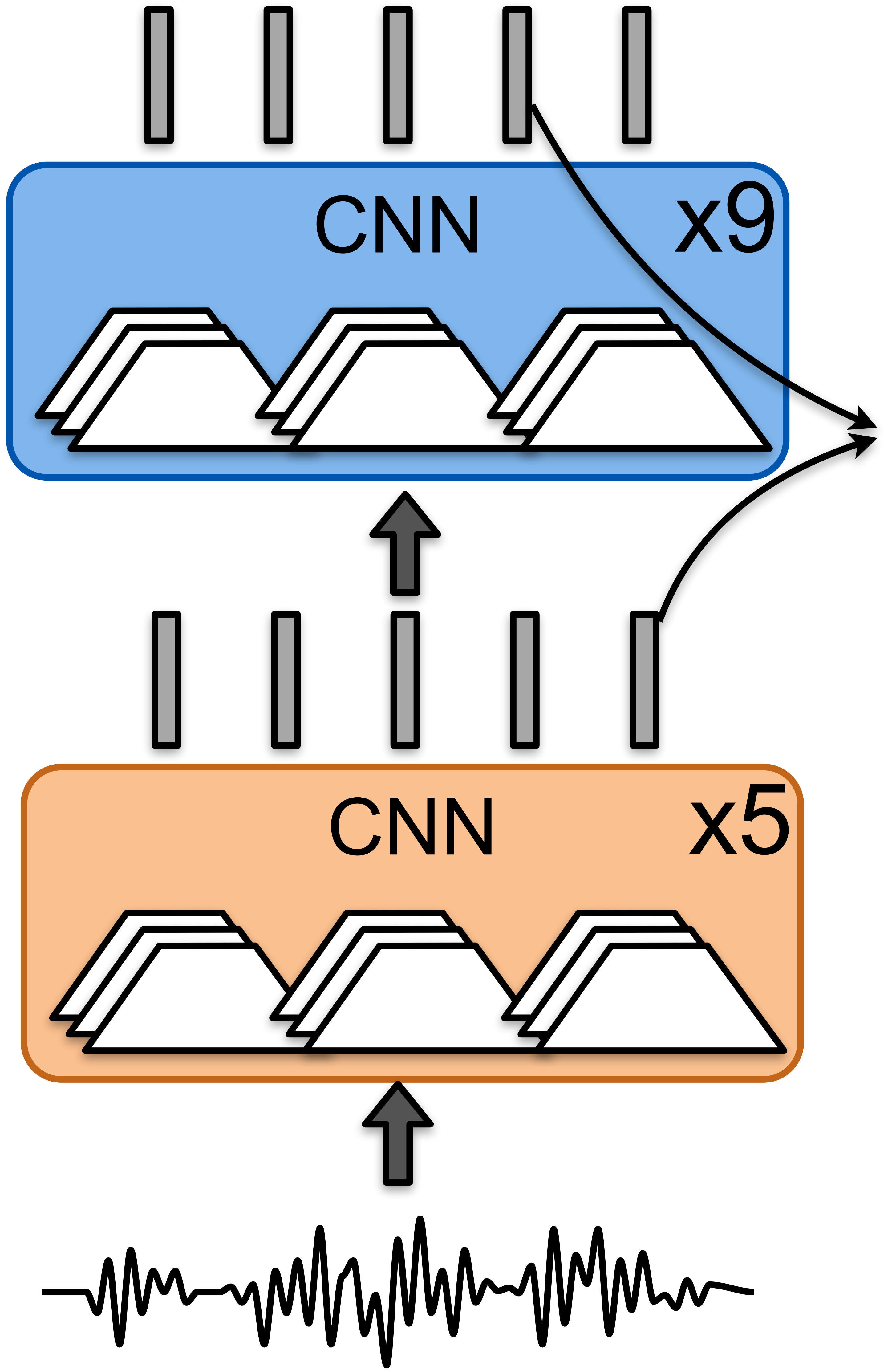


🤔 How to use larger <u>raw audio data</u> to improve ST performance?

Wav2Vec: Self-supervised Speech Representation Learning

high-level context state c, each frame ~ 210ms, stride10ms

Low level acoustic state h, each frame ~ 30ms, stride10ms



Training data: LibriSpeech 960 hrs audio only

Minimize contrastive loss

$$L = -\sum \left(\log \sigma(z_{t+1} \cdot h_t) + \sum \log \sigma(-z_{\_} \cdot h_t) \right)$$

Bring closer context and acoustic state

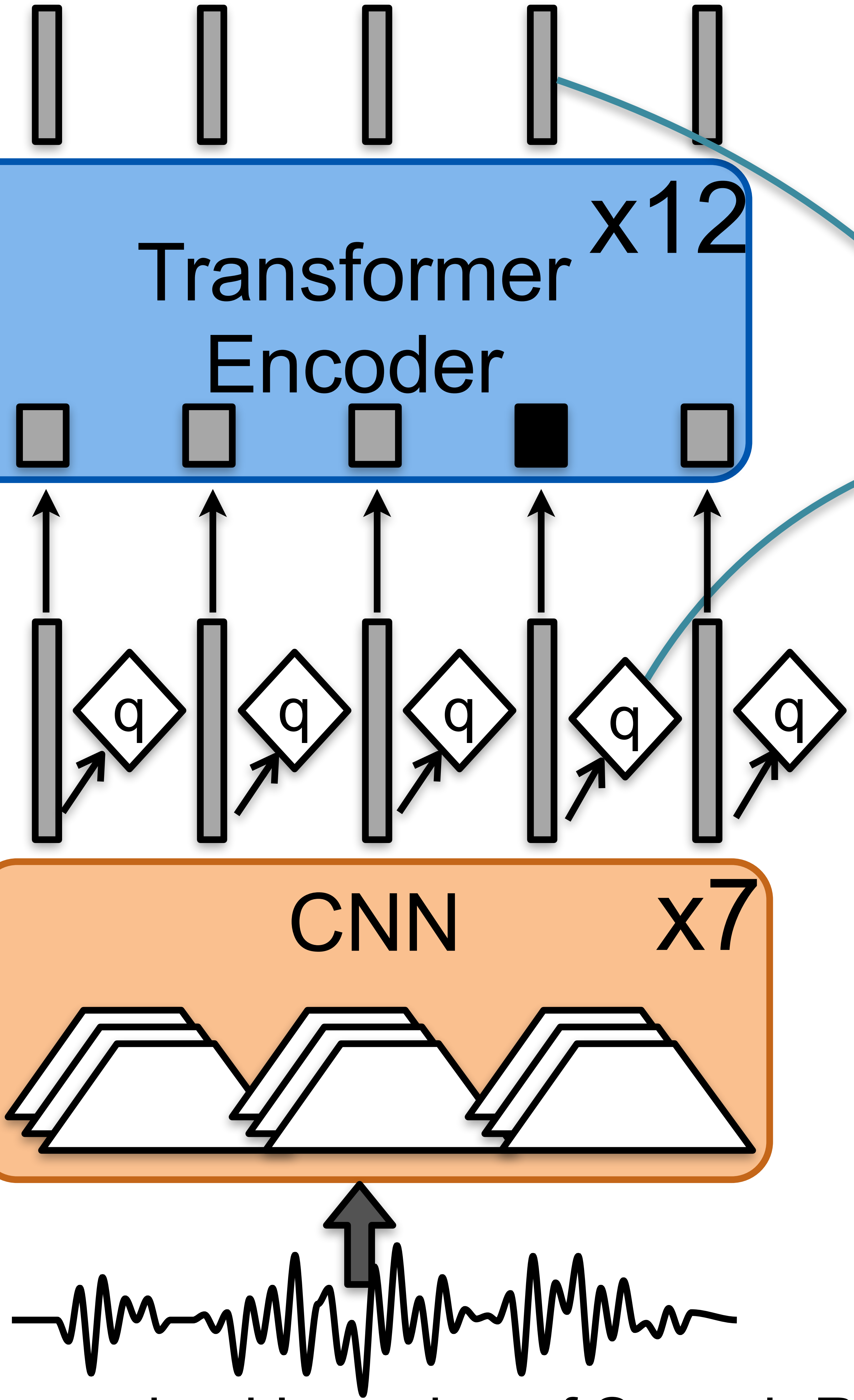
Bring further context and negative sampled acoustic state

wav2vec: Unsupervised Pre-training for Speech Recognition [Schneider et al, Interspeech 2010]

Wav2Vec2.0: Contrastive on quantized acoustic state

Masked context during training

Quantized low-level acoustic state, each frame ~ 25ms, stride 20ms



Training data: (audio only)
LibriSpeech 960 hrs
LibriVox 53k hrs

Minimize contrastive loss

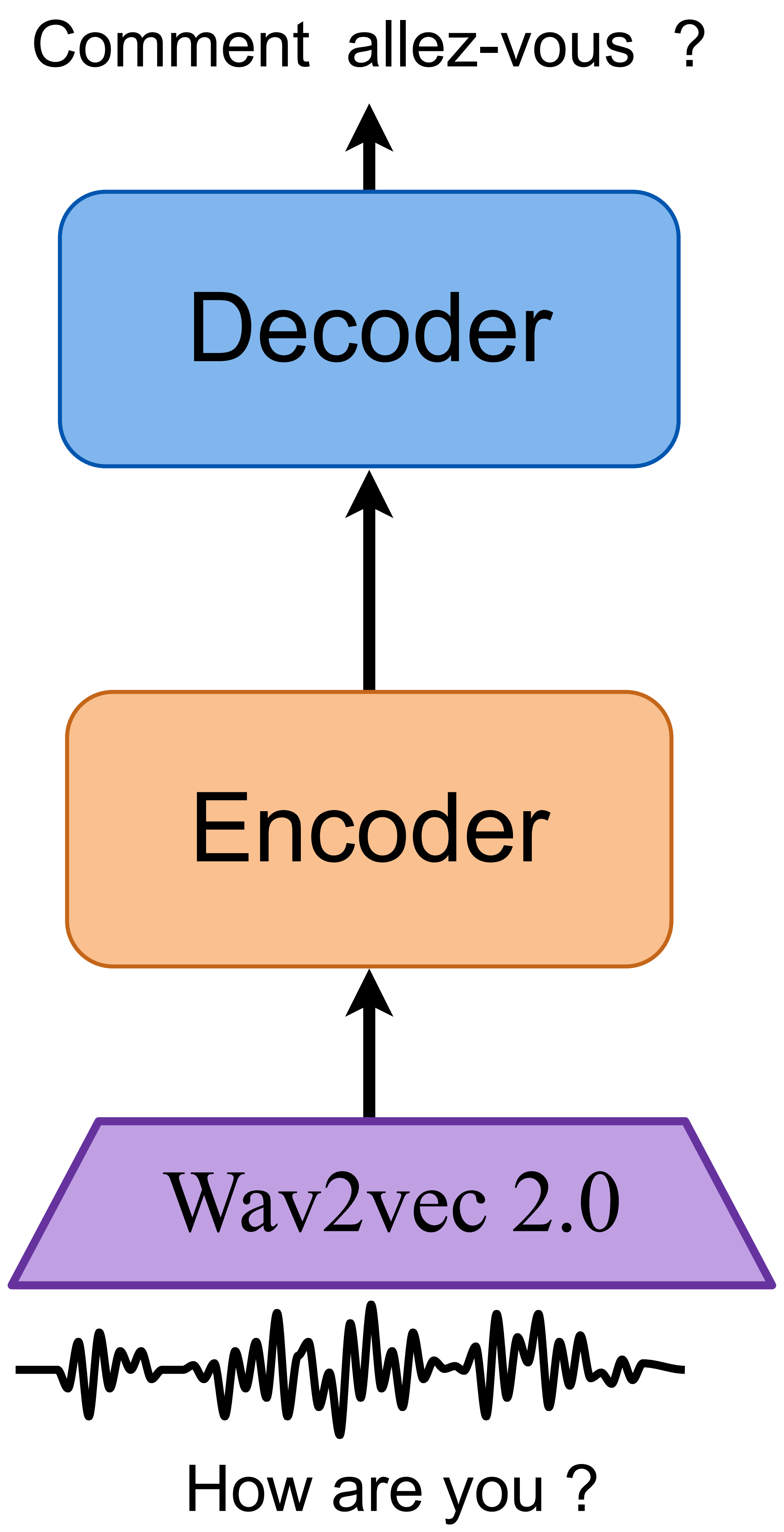
$$L = -\sum \log \frac{\exp Sim(c_t, q_t)}{\sum \exp Sim(c_t, q_-)} + \text{penalty}$$

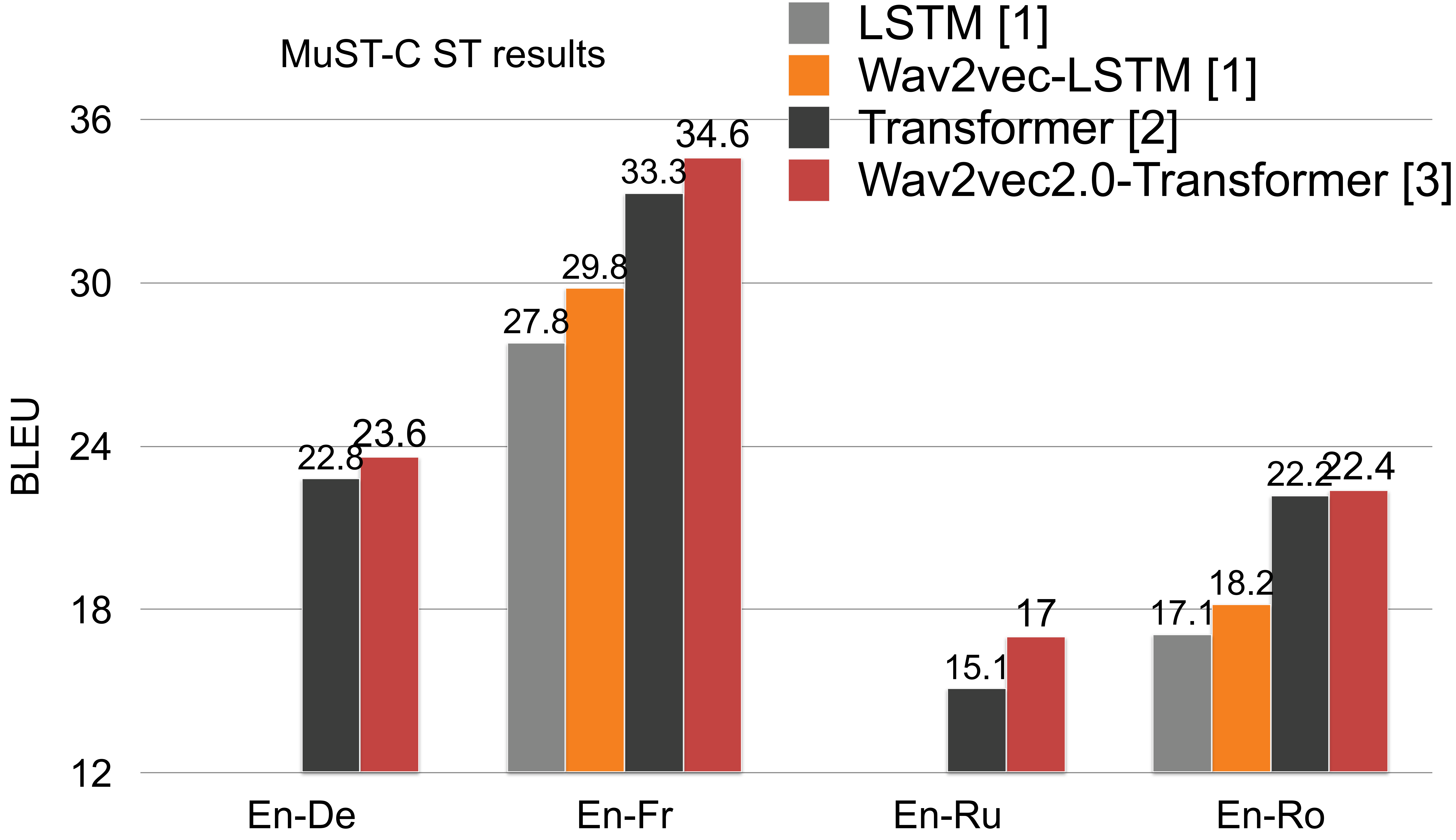
Bring closer masked context and quantized acoustic state

Wav2vec2.0: a Framework for Self-Supervised Learning of Speech Representations [Baevski et al, NeurIPS 2020]

Speech Translation with Audio-Pretrain

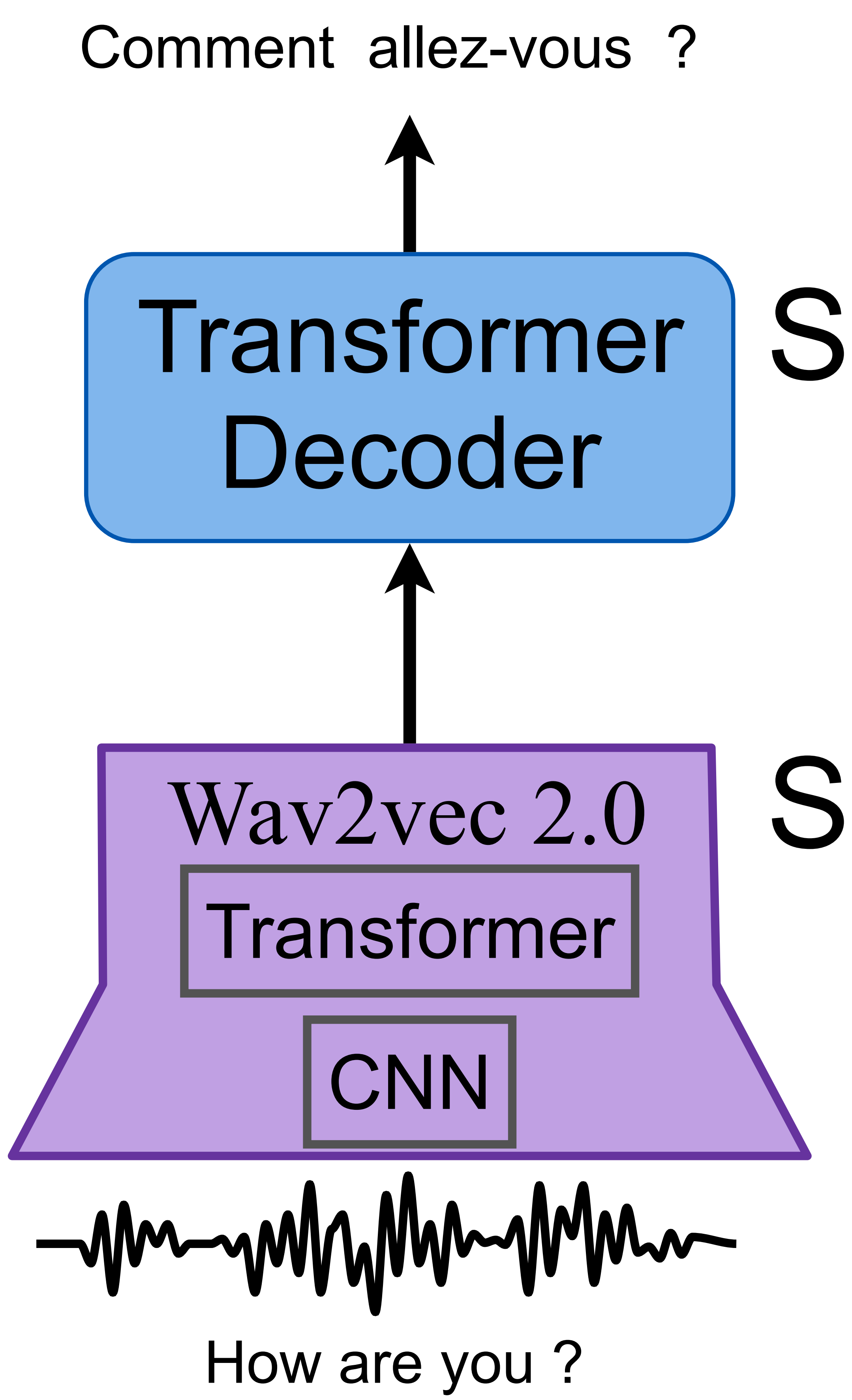
Wav2vec Pretrain + Fine-tune on ST





[1] Self-supervised Representations improve end-to-end speech translation [Wu et al. InterSpeech 2020]
[2] NeurST toolkit [Zhao et al ACL2021 demo] [3] End-to-end Speech Translation [Ye et al. InterSpeech 2021]

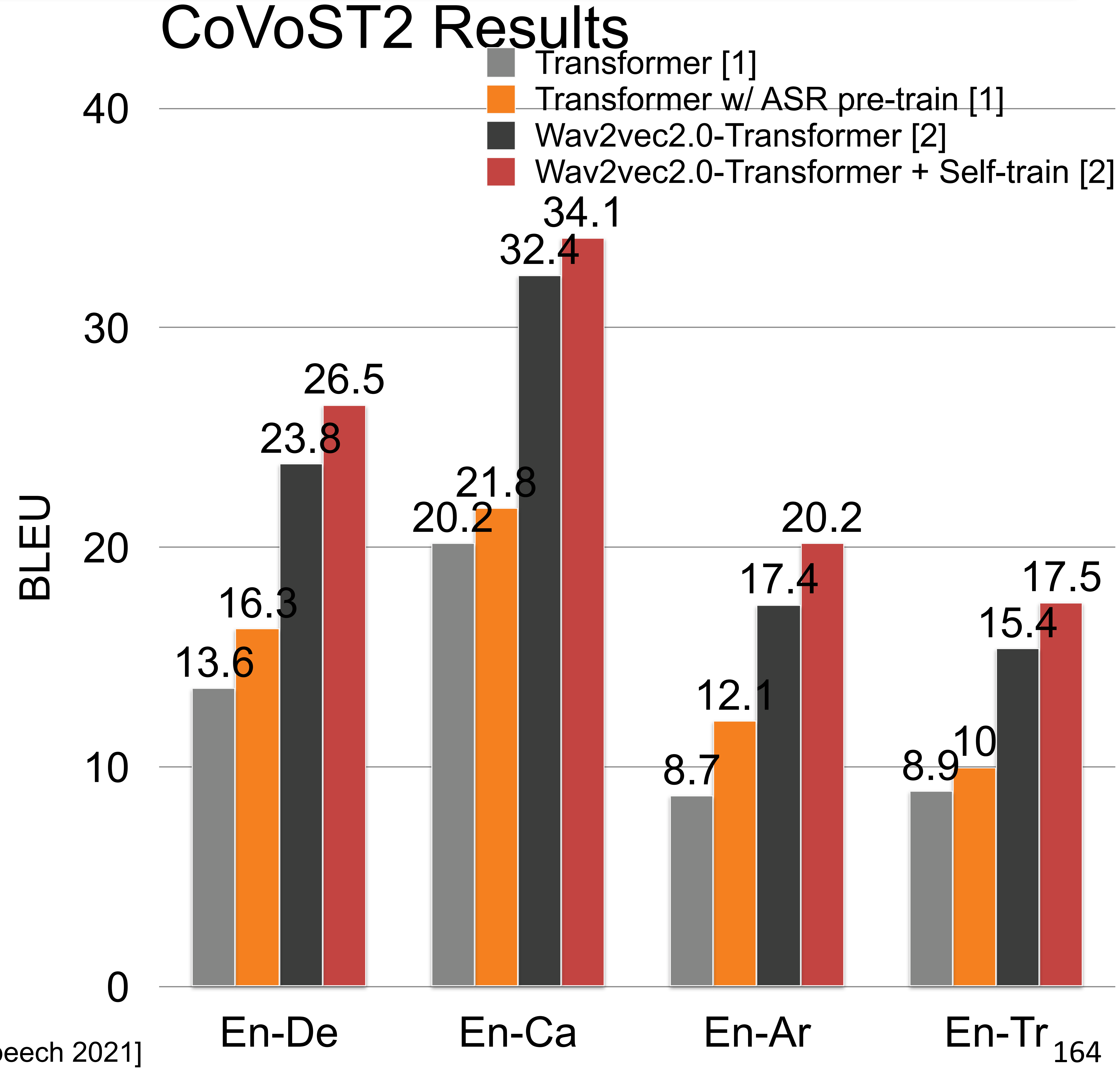
Self-training with Audio data



Step 0. Audio-only pre-training for Wav2vec2.0

Step 1. Freeze Wav2vec2.0, train on ST

Step 2. Self-train on generated pseudo-translation with LibriVox audio

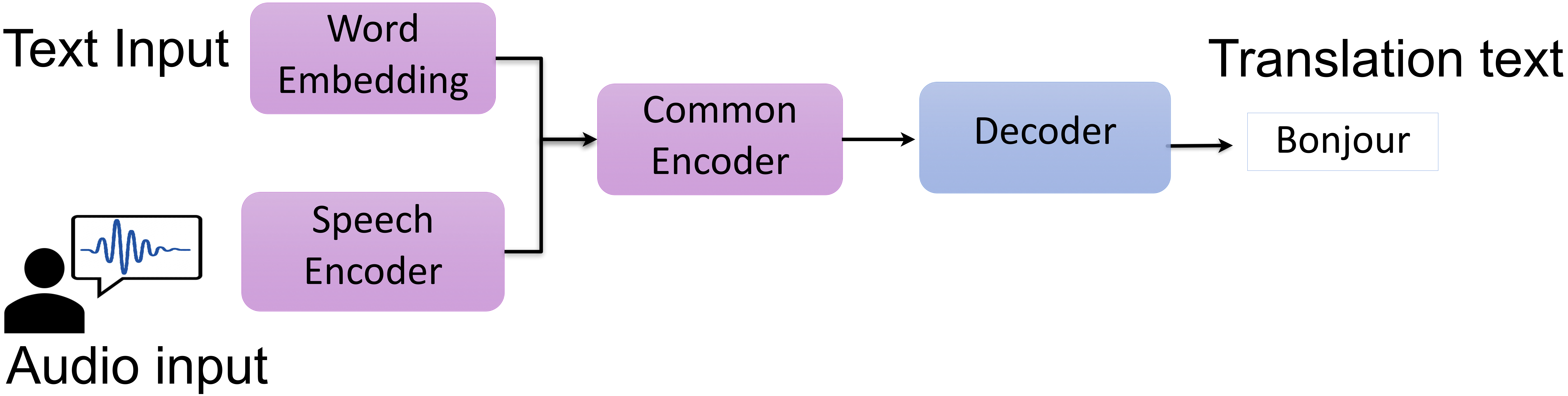


[1] CoVoST 2 and Massively Multilingual Speech-to-Text Translation, [Wang et al InterSpeech 2021]
[2] Large-Scale Self- and Semi-Supervised Learning for Speech Translation [Wang et al. 2021]

Bimodal Pre-training with Audio & MT data

- Chimera: Learning Fixed-size Shared Space for both audio and text, audio+MT pretraining [Han et al. 2021]
- XSTNet: Bring speech sequence to roughly similar length to text, then Pre-training & progressive multi-task fine-tuning [Ye et al. 2021]
- Wav2vec2.0-mTransformer LNA: Use both audio pertaining + multilingual pertained language model, and selective efficient fine-tuning [Li et al. ACL 2021]
- FAT-ST: Masked pre-training for fused audio and text [Zheng et al. ICML 2021]

Bi-modal Encoding Architecture for ST

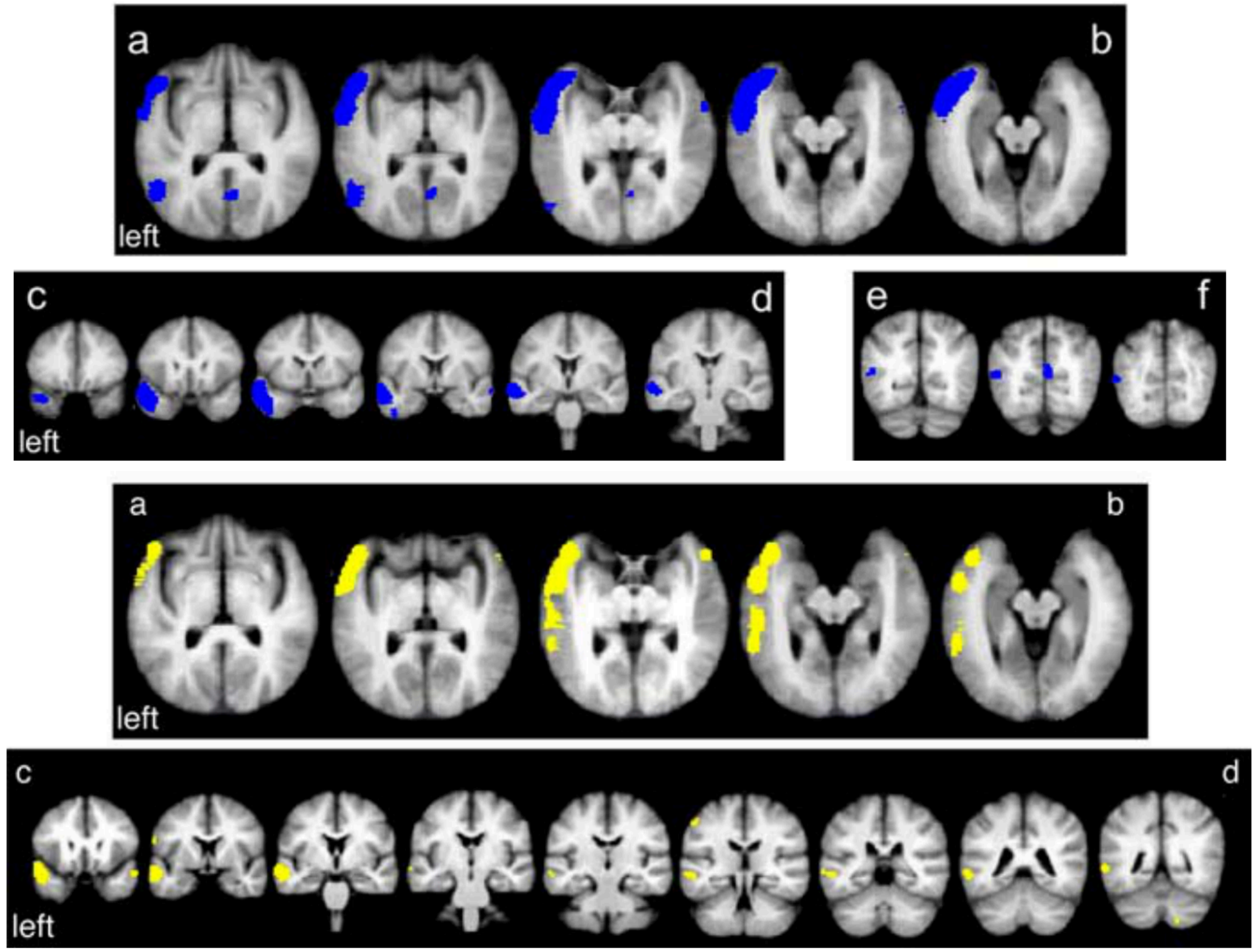


Challenges: gap between text and audio

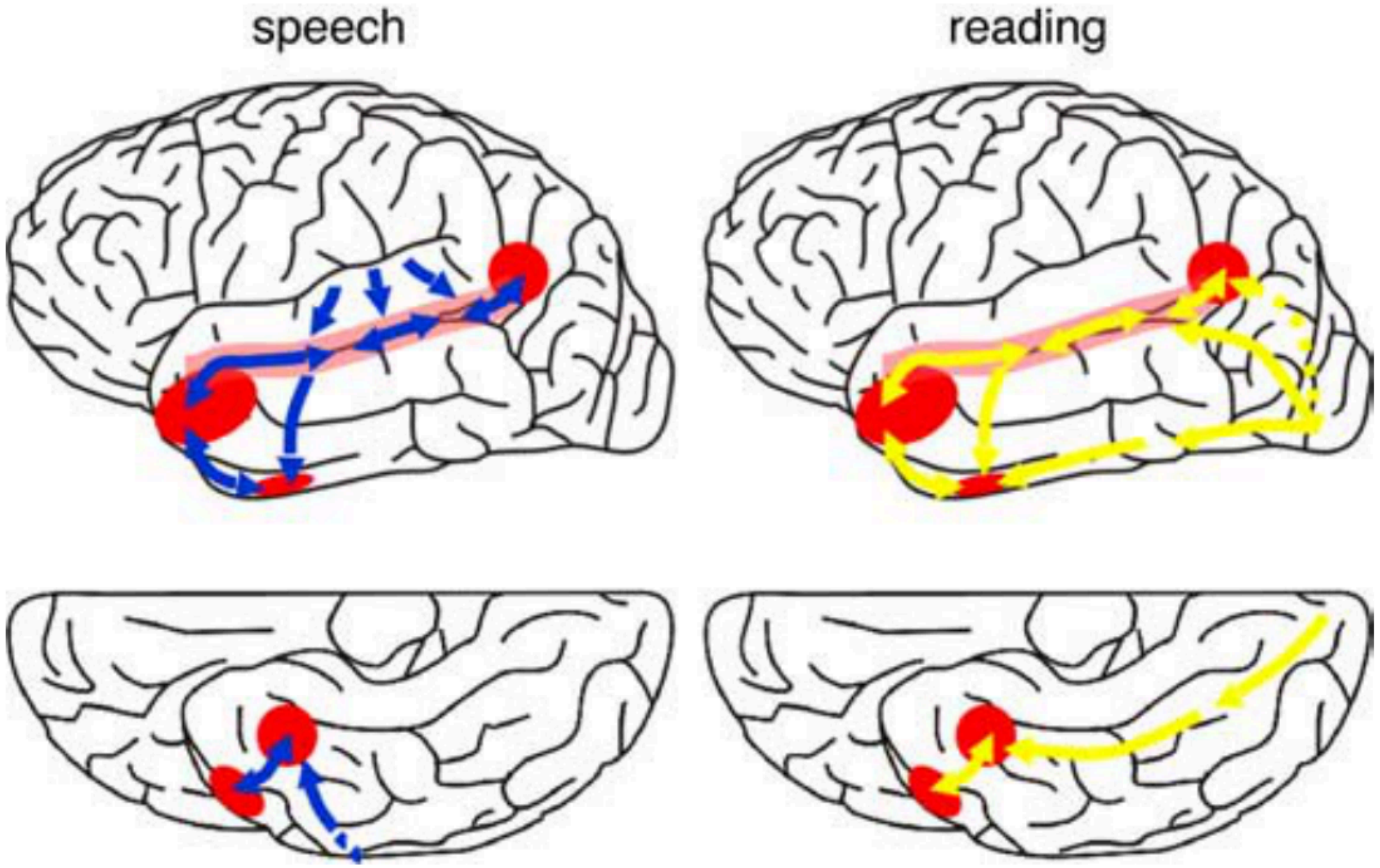
1. Length: ~20 (text) vs. ~ 1k-10k (audio)
2. Embedding space disparity

Insights from Cognitive Neuroscience

Speech and text interfere with each other in brain[1]



activation map



processing paths

<u>Convergence sites</u> of *speech* (blue) and *text* (yellow)

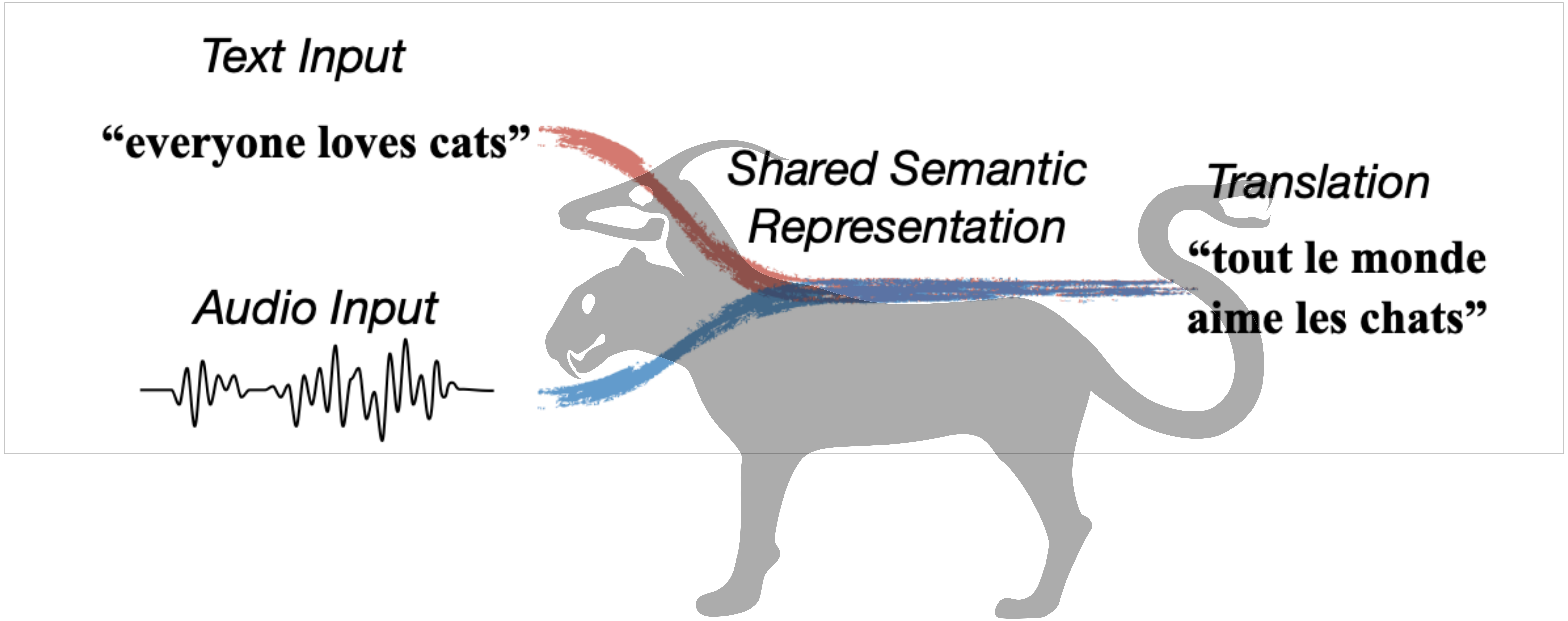
[1] Van Atteveldt, Nienke, et al. "Integration of letters and speech sounds in the human brain." *Neuron* 43.2 (2004): 271-282.

[2] Spitsyna, Galina, et al. "Converging language streams in the human temporal lobe." *Journal of Neuroscience* 26.28 (2006): 7328-7336.

Idea: Bridging the Speech-Text modality gap with Shared Semantic Representation

ST triple data:

<speech, transcript_text, translate_text>

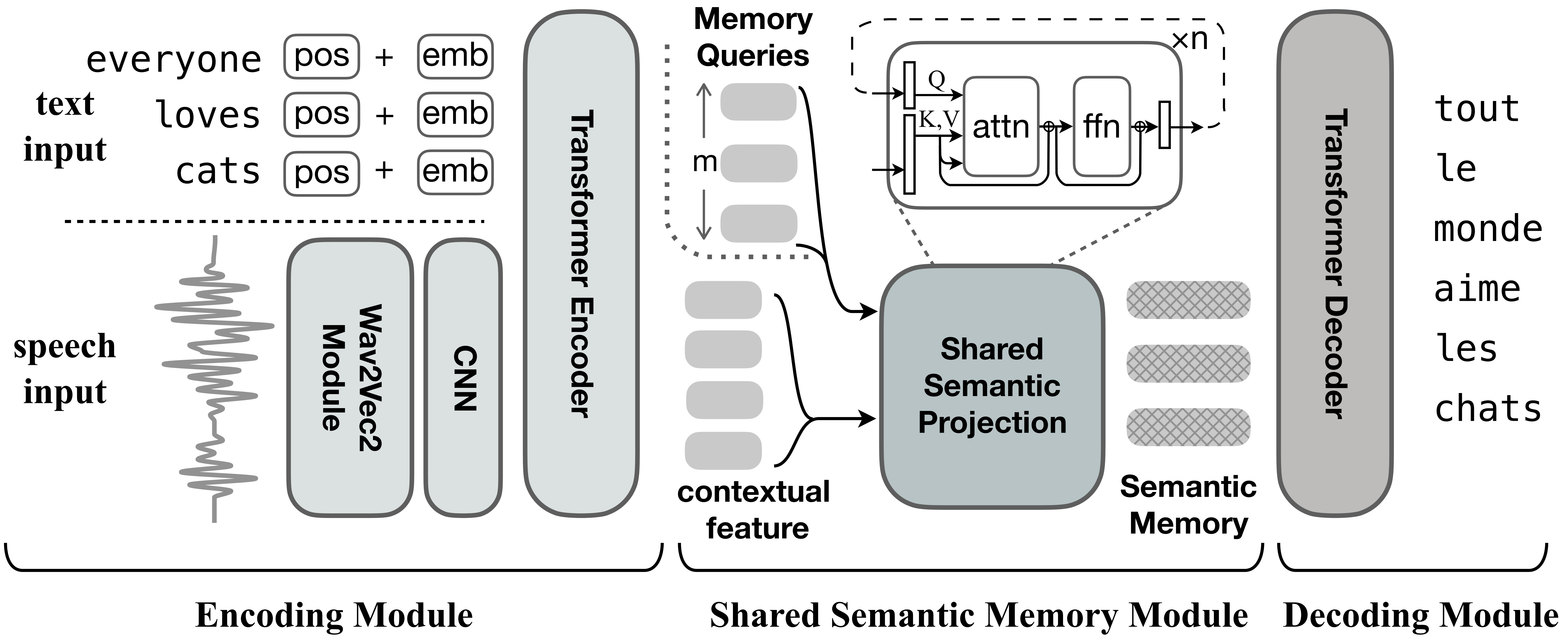


Learning Shared Semantic Space for Speech-to-Text Translation Listen [Chi Han, Mingxuan Wang, Heng Ji, Lei Li, Findings of ACL 2021]

Chimera Model for ST

Training with auxiliary objectives: ST + MT + Contrastive loss

Benefit: able to exploit large external MT data



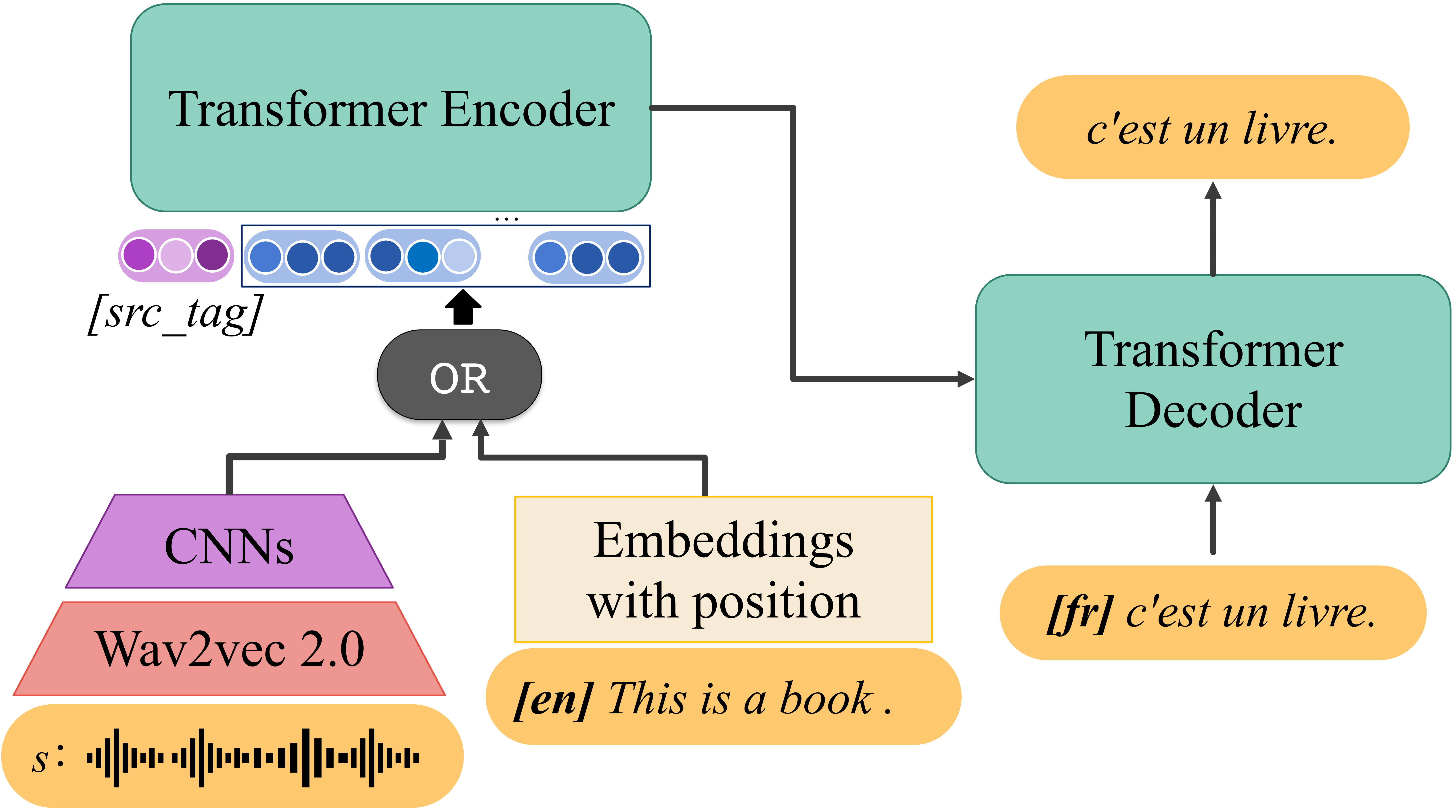
Learning Shared Semantic Space for Speech-to-Text Translation Listen [Chi Han, Mingxuan Wang, Heng Ji, Lei Li, Findings of ACL 2021]

Chimera achieves the best (so far) BLEU on all languages in MuST-C

Model	External Data			MuST-C EN-X							
	Speech	ASR	MT	EN-DE	EN-FR	EN-RU	EN-ES	EN-IT	EN-RO	EN-PT	EN-NL
FairSeq ST †	×	×	×	22.7	32.9	15.3	27.2	22.7	21.9	28.1	27.3
Espnet ST ‡	×	×	×	22.9	32.8	15.8	28.0	23.8	21.9	28.0	27.4
AFS ⋆	×	×	×	22.4	31.6	14.7	26.9	23.0	21.0	26.3	24.9
Dual-Decoder ◇	×	×	×	23.6	33.5	15.2	28.1	24.2	22.9	**30.0**	27.6
STATST ♯	×	×	×	23.1	-	-	-	-	-	-	-
MAML ♭	×	×	✓	22.1	34.1	-	-	-	-	-	-
Self-Training ◦	✓	✓	×	25.2	34.5	-	-	-	-	-	-
W2V2-Transformer *	✓	×	×	22.3	34.3	15.8	28.7	24.2	22.4	29.3	28.2
Chimera Mem-16	✓	×	✓	25.6	35.0	16.7	30.2	24.0	23.2	29.7	28.5
Chimera	✓	×	✓	**27.1** •	**35.6**	**17.4**	**30.6**	**25.0**	**24.0**	**30.2**	**29.2**

Learning Shared Semantic Space for Speech-to-Text Translation [Chi Han, Mingxuan Wang, Heng Ji, Lei Li, Findings of ACL 2021]

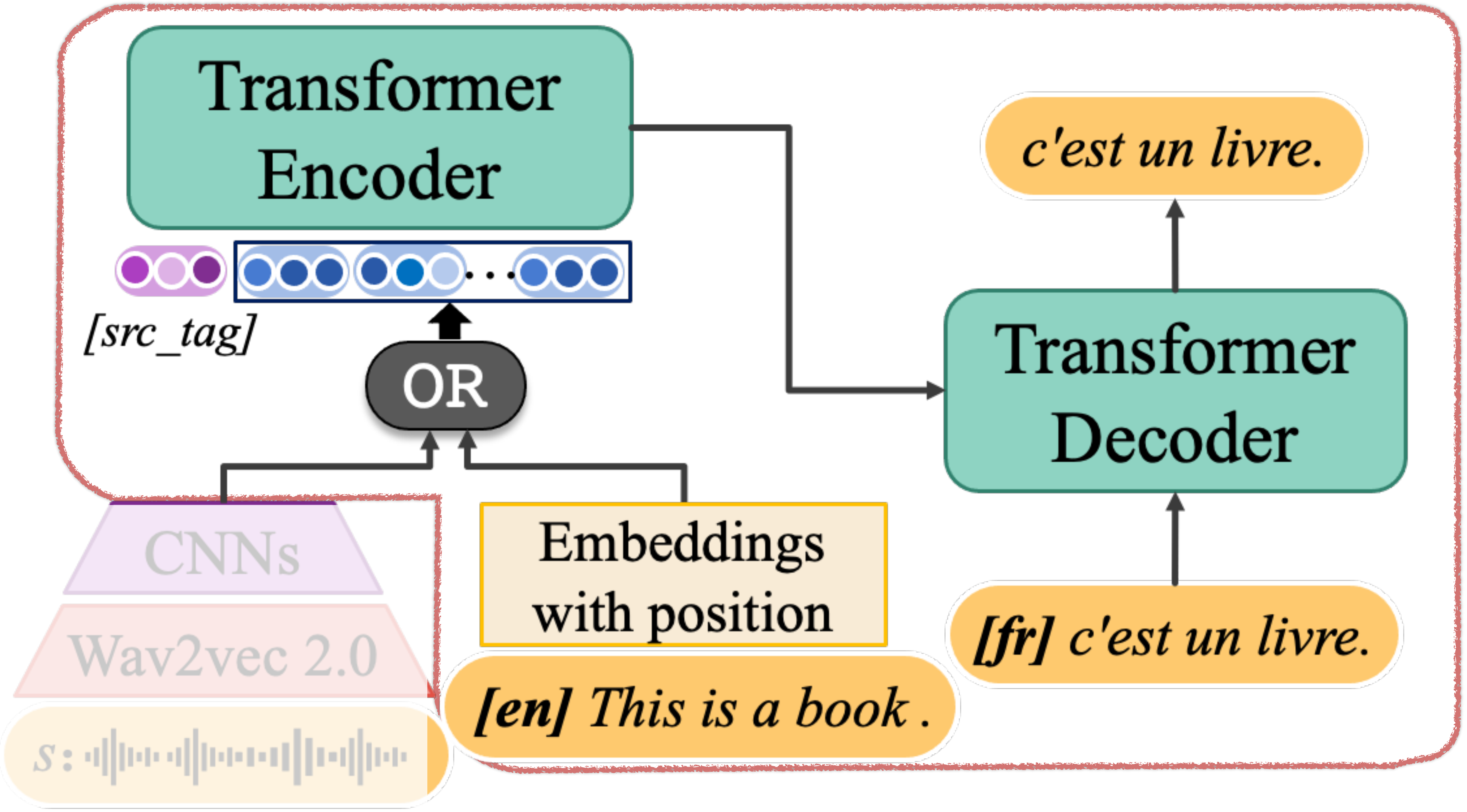
Cross Speech-Text Network (XSTNet)



End-to-end Speech Translation via Cross-modal Progressive Training [Rong Ye, Mingxuan Wang, Lei Li, Interspeech 2021]

Supports to train MT data

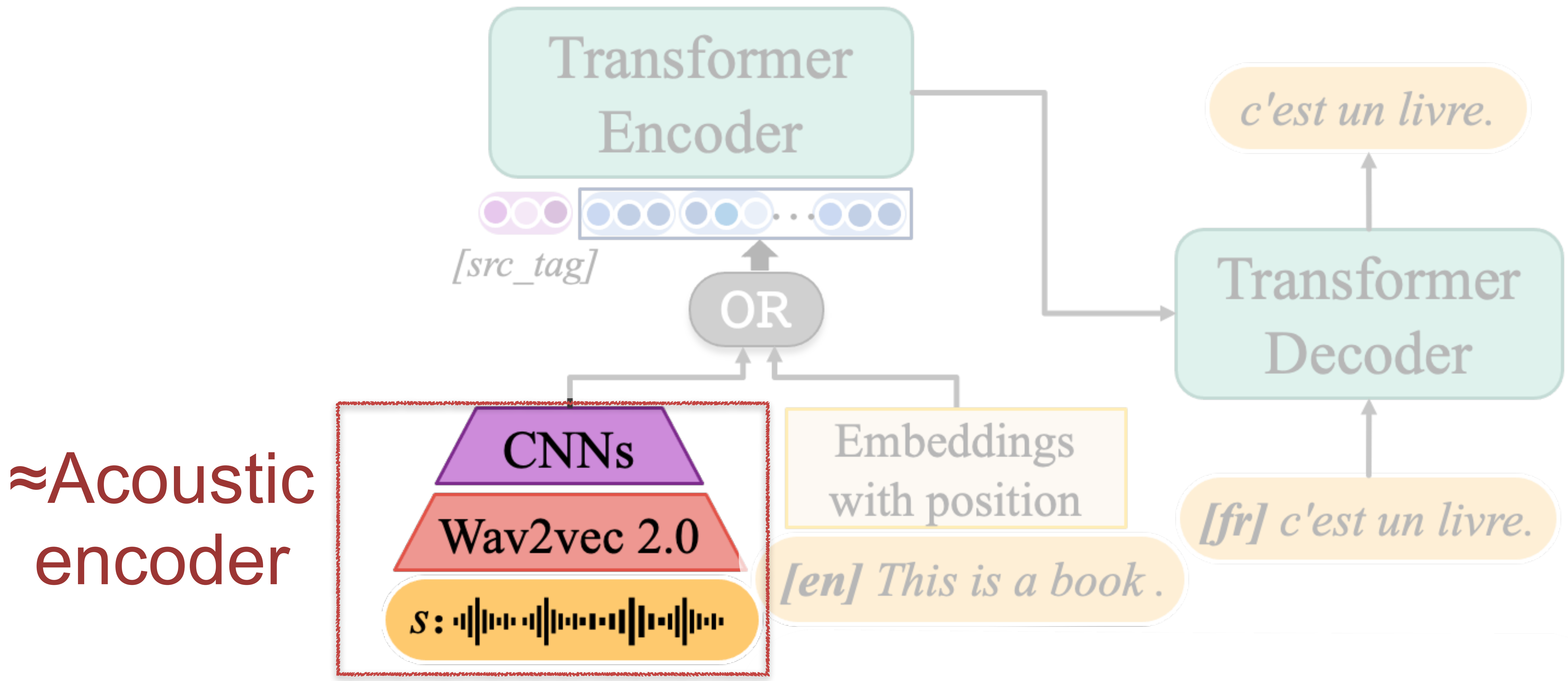
- ☑ Transformer MT model
- ☑ We can add **more external MT data** to train Transformer encoder & decoder



End-to-end Speech Translation via Cross-modal Progressive Training [Rong Ye, Mingxuan Wang, Lei Li, Interspeech 2021]

Supports inputs of two modalities

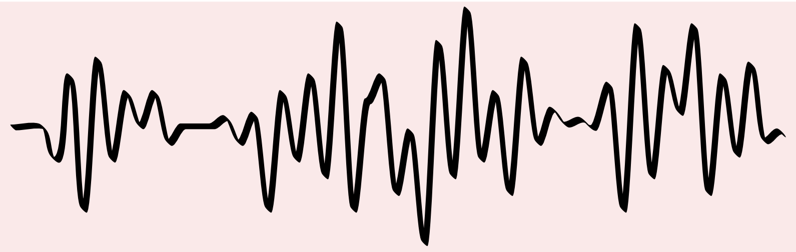
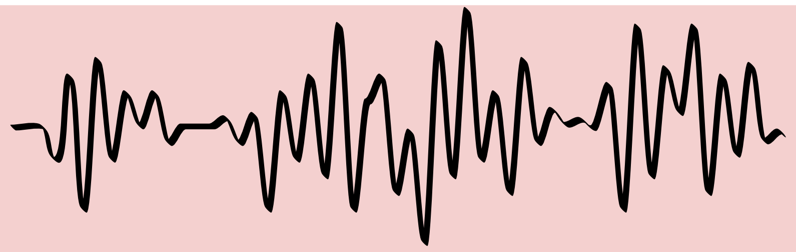
- ☑ Wav2vec2.0[1] as the acoustic encoder
- ☑ We add two convolution layers with 2-stride to shrink the length.



[1] wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020

Language indicator strategy

- We use language indicators to distinguish different tasks.

Tasks	Source input	Target output
MT	**<en>** This is a book.	**<fr>** c'est un livre.
ASR	**<audio>** 	**<en>** This is a book.
ST	**<audio>** 	**<fr>** c'est un livre.

End-to-end Speech Translation via Cross-modal Progressive Training [Rong Ye, Mingxuan Wang, Lei Li, Interspeech 2021]

Progressive Multi-task Training

Large-scale MT pre-training

Using **external MT** D_{MT-ext}

Multi-task Finetune

Using **(1) external MT** D_{MT-ext}

(2) D_{ST} with *<speech, translation>*

(3) D_{ASR} with *<speech, transcript>*

Progressive:

Don't stop

training D_{MT-ext}

End-to-end Speech Translation via Cross-modal Progressive Training [Rong Ye, Mingxuan Wang, Lei Li, Interspeech 2021]

XSTNet achieves State-of-the-art Performance

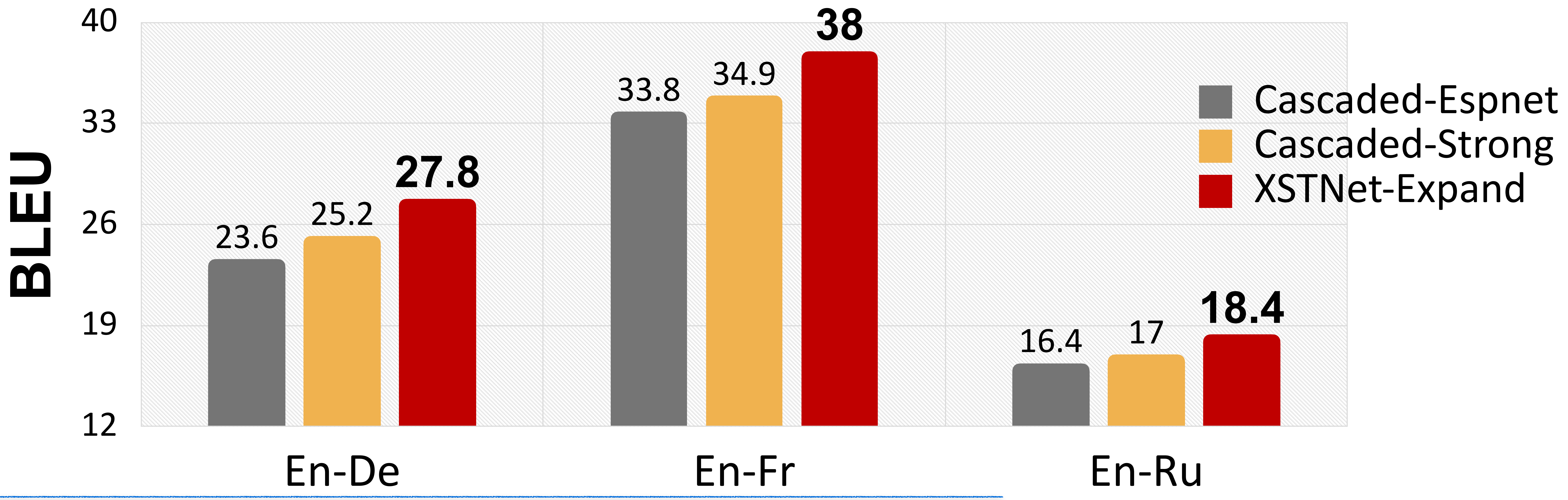
Models	External Data	Pre-train Tasks	De	Es	Fr	It	Nl	Pt	Ro	Ru	Avg.
Transformer ST [13]	×	ASR	22.8	27.4	33.3	22.9	27.2	28.7	22.2	15.1	24.9
AFS [31]	×	×	22.4	26.9	31.6	23.0	24.9	26.3	21.0	14.7	23.9
Dual-Decoder Transf. [15]	×	×	23.6	28.1	33.5	24.2	27.6	30.0	22.9	15.2	25.6
Tang et al. [2]	MT	ASR, MT	23.9	28.6	33.1	-	-	-	-	-	-
FAT-ST (Big) [6]	ASR, MT, mono-data†	FAT-MLM	25.5	30.8	-	-	30.1	-	-	-	-
W-Transf.	audio-only*	SSL*	23.6	28.4	34.6	24.0	29.0	29.6	22.4	14.4	25.7
XSTNet (Base)	audio-only*	SSL*	25.5	29.6	36.0	25.5	30.0	31.3	25.1	16.9	27.5
XSTNet (Expand)	MT, audio-only*	SSL*, MT	**27.8**§	**30.8**	**38.0**	**26.4**	**31.2**	**32.4**	**25.7**	**18.5**	**28.8**

Table 1: *Performance (case-sensitive detokenized BLEU) on MuST-C test sets.* †: *"Mono-data" means audio-only data from Librispeech, Libri-Light, and text-only data from Europarl/Wiki Text; *: "Audio-only" data from LibriSpeech is used in the pre-training of wav2vec2.0-base module, and "SSL" means the self-supervised learning from unlabeled audio data.* § *uses OpenSubtitles as external MT data.*

XSTNet-Base: Achieves the SOTA in the restricted setup

XSTNet-Expand: Goes better by using extra MT data

XSTNet better than cascaded ST! a gain of 2.6 BLEU

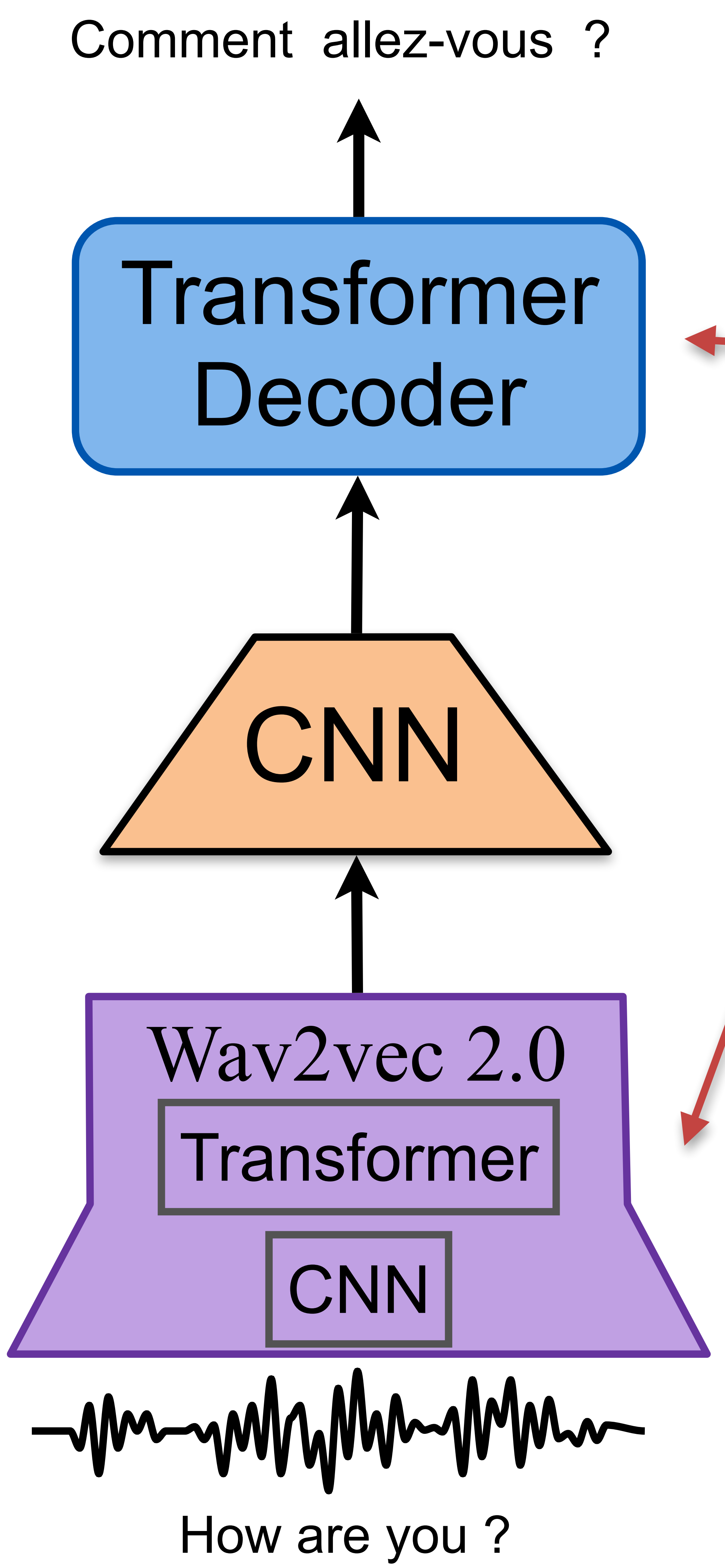


What is "Cascaded-Strong" system?

Strong ASR model + Large-scale MT data

Cascaded - Strong	Model	Training data	Performance (En-De)
ASR	W2V2+ Transformer	MuST-C D_{ASR}	WER=13.0
MT	Transformer-base	WMT + MuST-C D_{MT}	BLEU=31.7

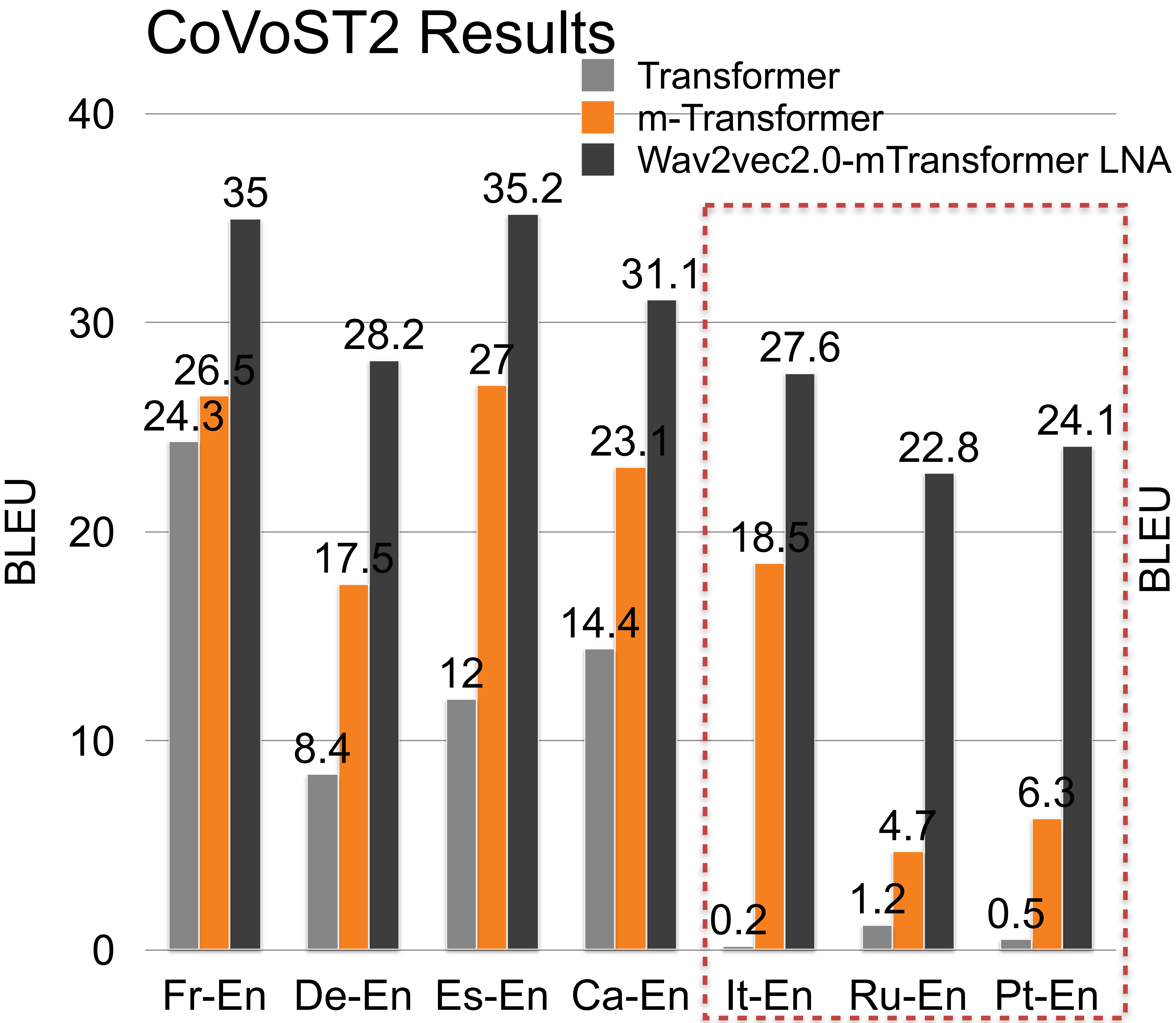
Audio and Multilingual Text Pretrain for Multilingual ST

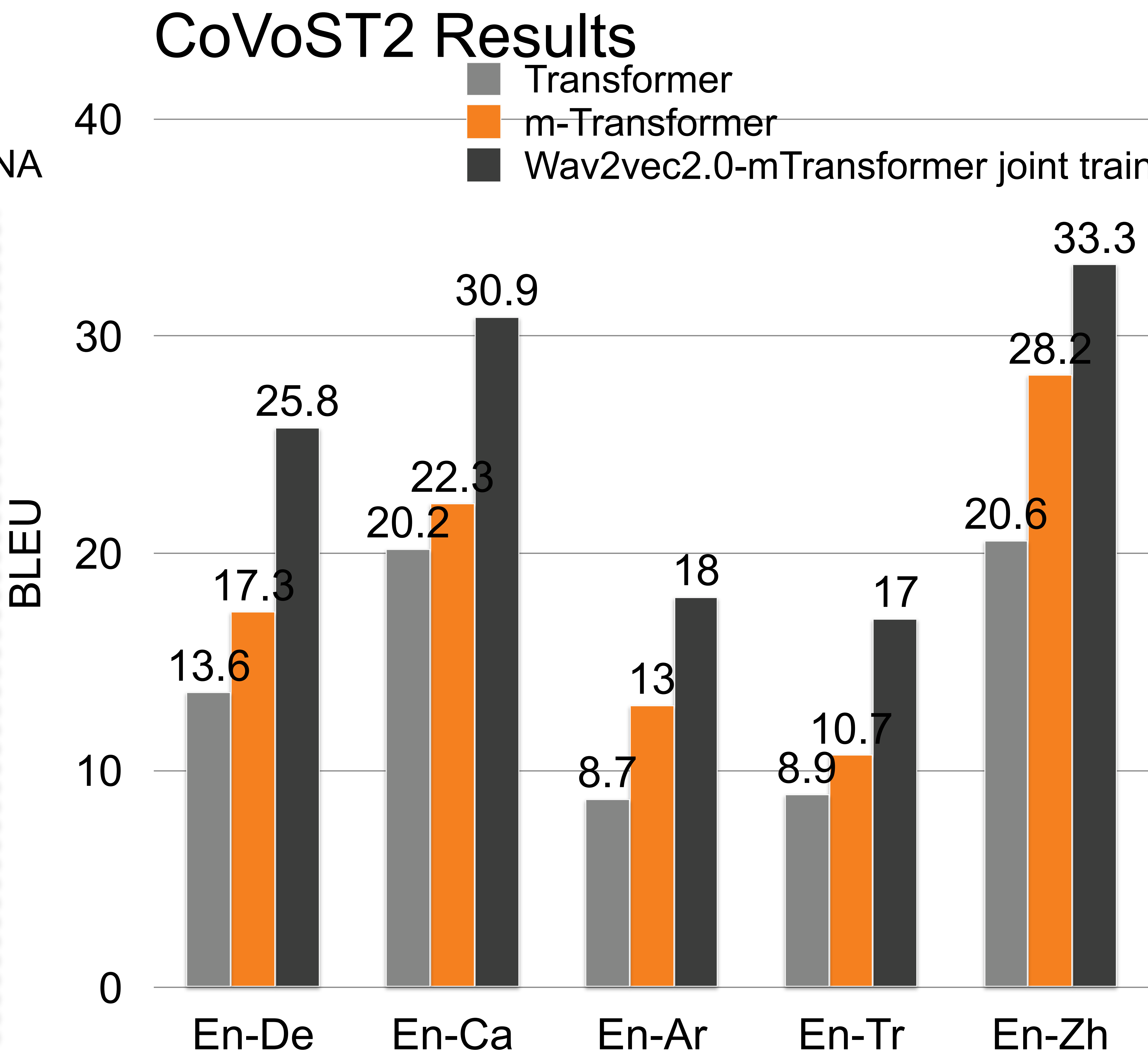


- Encoder uses Wav2vec2.0 pre-trained on LibriVox-60k audio
- Decoder: mBart pre-trained on 50 monolingual text and 49 bitext
- ST finetune strategy (LNA):
  - Only fine-tune layer-norm and attention layers
- MT+ST multitask joint train with further parallel bitext data

Multilingual Speech Translation with Efficient Finetuning of Pretrained Models [Li et al, ACL 2021]

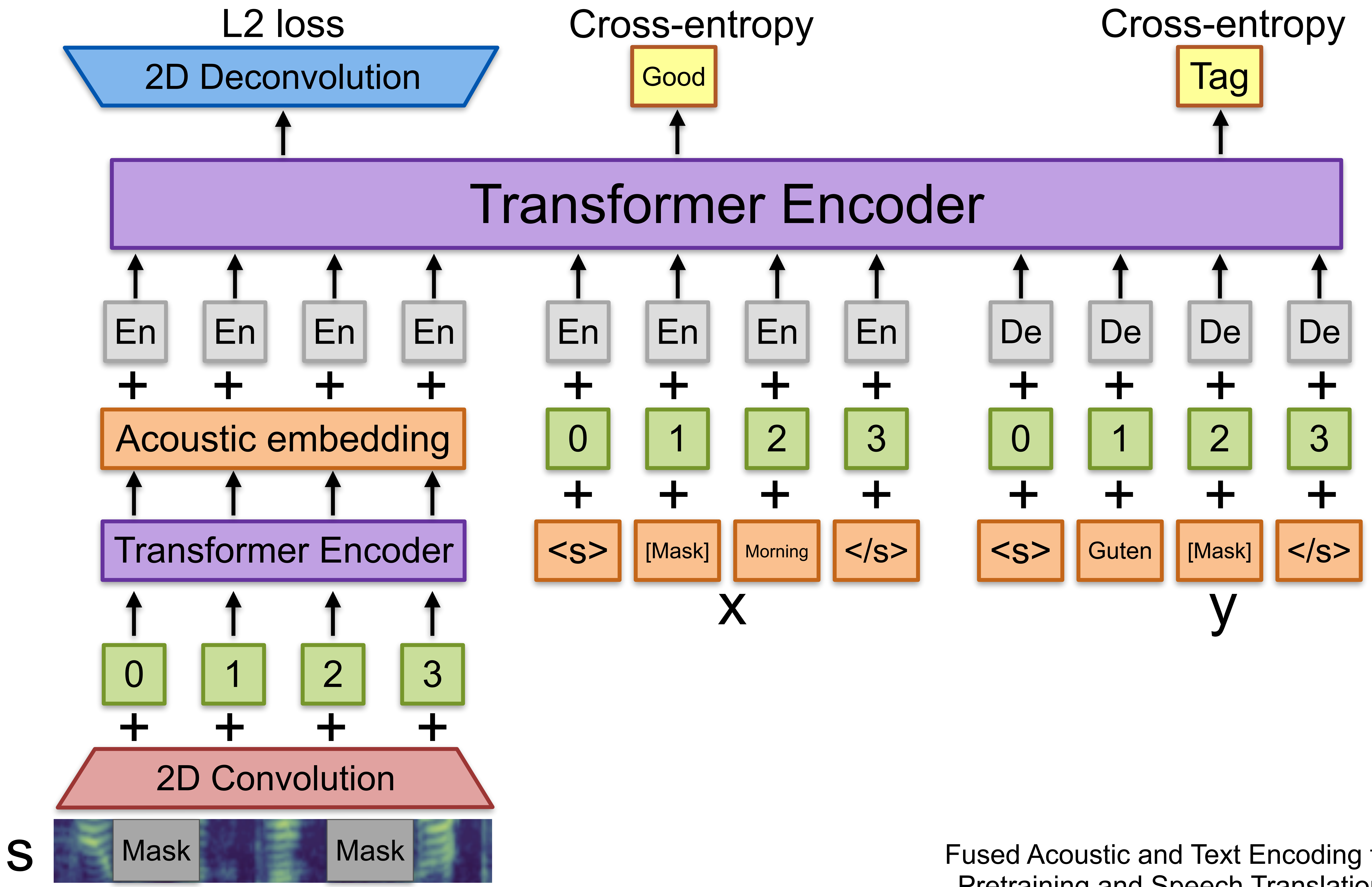
Wav2vec2.0 retraining + Multilingual training effectively transfers to low resource source language





Multilingual Speech Translation with Efficient Finetuning of Pretrained Models [Li et al, ACL 2021]

Fused Acoustic and Text Masked Language Model (FAT-MLM)

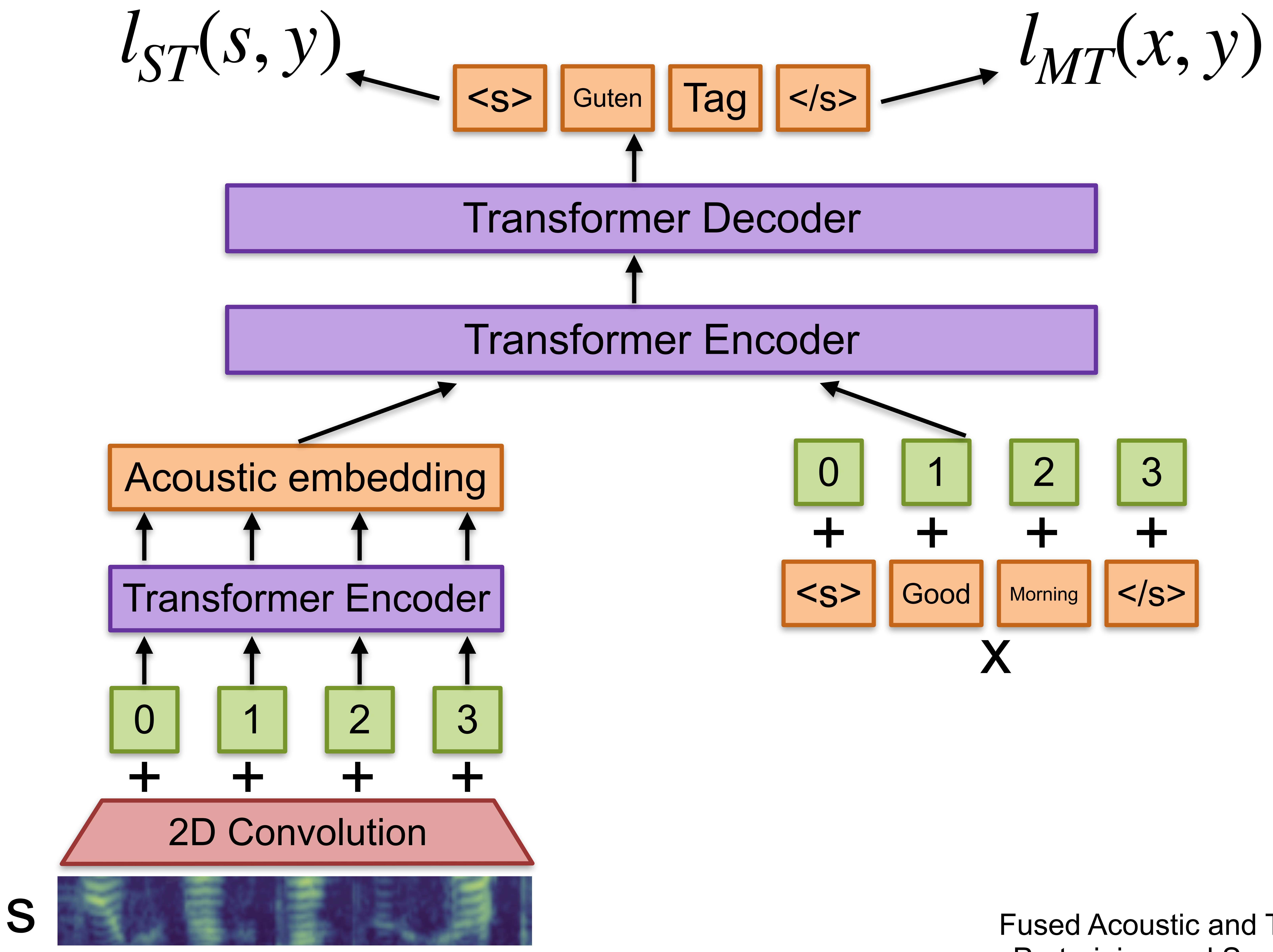


Pre-training data

1. Librispeech ASR 960h
2. Libri-light audio 3,748h
3. Europarl/wiki text 2.3M
4. MuST-C 408h
5. Europarl MT 1.9M

Fused Acoustic and Text Encoding for Multimodal Bilingual Pretraining and Speech Translation, [Zheng et al ICML 2021]

FAT-ST



Training:

- Pre-train FAT-MLM with all data
- Init FAT-ST with FAT-MLM, decoder copy encoder
- Further fine-tune on MuST-C ST data.

Fused Acoustic and Text Encoding for Multimodal Bilingual Pretraining and Speech Translation, [Zheng et al ICML 2021]

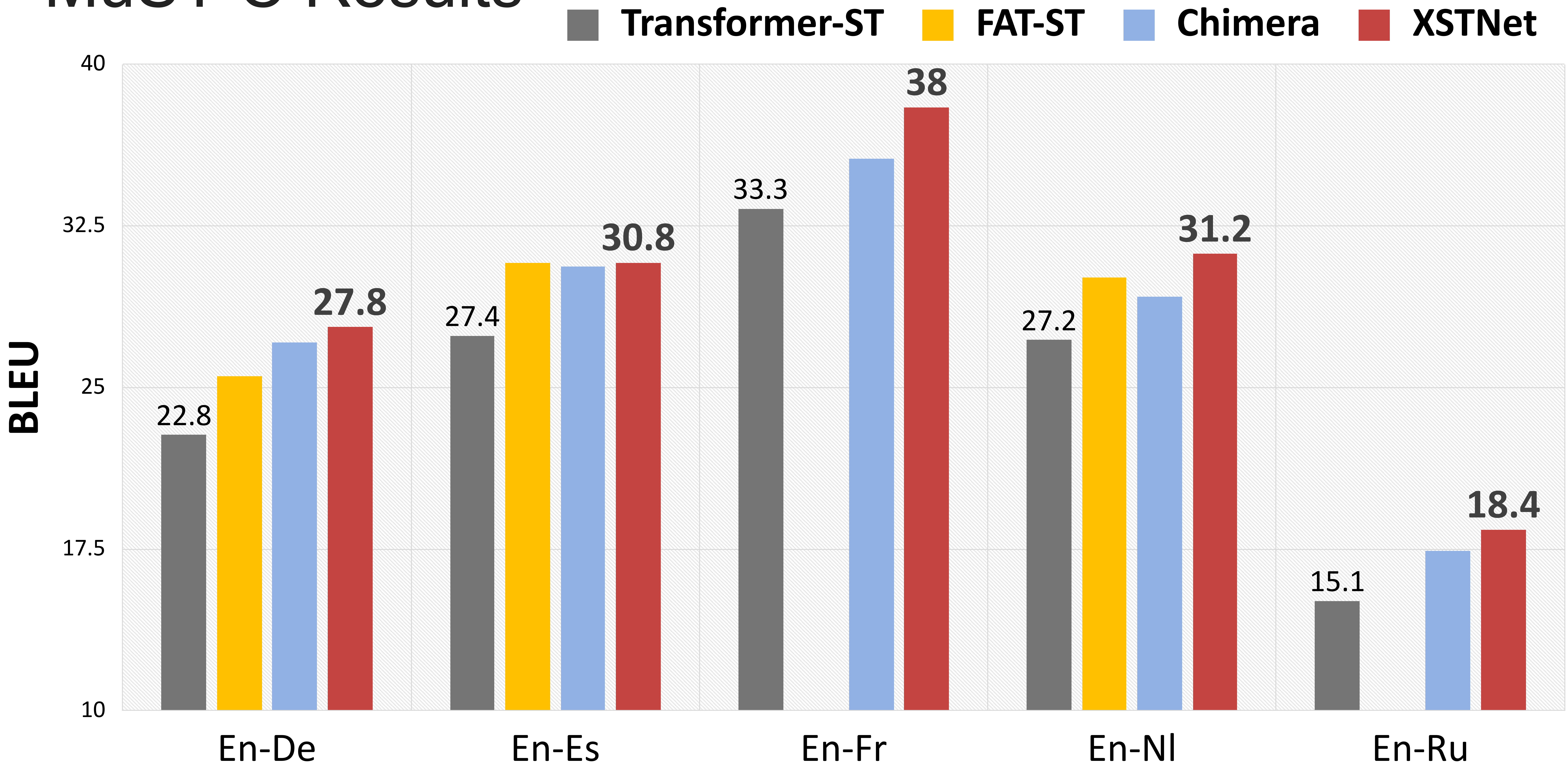
Joint audio&text Pre-training task helps ST

Pretrain Method	Models	En→De	En→Es	En→Nl	Avg.	Model Size
No Pretraining	ST	19.64	23.68	23.01	22.11	31.25M
	ST + ASR	21.70	26.83	25.44	24.66 (+2.55)	44.82M
	ST + ASR & MT	21.58	26.37	26.17	24.71 (+2.60)	56.81M
	ST + MAM	20.78	25.34	24.46	23.53 (+1.42)	33.15M
	ST + MAM + ASR	22.41	26.89	26.49	25.26 (+3.15)	46.72M
	Liu et al. (2020b)	22.55	-	-	-	-
	Le et al. (2020)	23.63	28.12	27.55	26.43 (+4.32)	51.20M
	Cascade§	23.65	28.68	27.91	26.75 (+4.64)	83.79M
	FAT-ST (base).	22.70	27.86	27.03	25.86 (+3.75)	39.34M
ASR & MT	ST	21.95	26.83	26.03	24.94 (+2.83)	31.25M
	ST + ASR & MT	22.05	26.95	26.15	25.05 (+2.94)	56.81M
MAM	FAT-ST (base)	22.29	27.21	26.26	25.25 (+3.14)	39.34M
FAT-MLM	FAT-ST (base)	**23.68**	28.61	**27.84**	26.71 (+4.60)	39.34M
	FAT-ST (big)	23.64	**29.00**	27.64	**26.76** (+4.65)	58.25M

Fused Acoustic and Text Encoding for Multimodal Bilingual Pretraining and Speech Translation, [Zheng et al ICML 2021]

Pre-training Improves ST Performance

- MuST-C Results



Summary

	Direct Supervision	Contrastive	Masked LM	Knowledge distillation	Progressive train	Selective Fine-tune	Self-training
MT Parallel Text	COSTT			[Liu et al. 2019]	XSTNet		
ASR Speech-Transcript	LUT						
Audio-only		Wav2vec Wav2vec 2.0					[Wang et al. 2021]
Raw text				LUT			
Speech+Text		Chimera	FAT-ST		XSTNet	LNA	

Summary for Speech Translation Pre-training

- Parallel speech translation data is scarce
- Pre-training to utilize external large data
  - MT data (Parallel text)
  - ASR data (Speech-transcript)
  - Raw text (Monolingual and Multilingual)
  - Audio-only
- Network architecture to solve modality disparity
  - CNN-Transformer
  - Fixed-size shared memory module
  - Bimodal input with length shrinking for audio
- Techniques to better pre-train and better fine-tune
  - Contrastive prediction
  - Masked LM
  - Quantization of audio representation
  - Knowledge distillation
  - Progressive pre-training

Summary

- Basics
  - NMT, Transformer encoder decoder.
  - Pre-training paradigm for NLP
- Monolingual Pre-training for NMT
  - Encoder pre-training
  - Seq-to-seq pre-training
- Multilingual Pre-training for NMT
- Pre-training for Speech Translation

Thanks

- Rong Ye, Chi Han, Qianqian Dong for help on beautification of the slides.

Reference

- Monolingual Pre-training
  - When and Why are Pre-trained Word Embeddings Useful for Neural Machine Translation [Qi et al., NAACL 2018]
  - Improve Neural Machine Translation by Building Word Vector [Zhang et al., AI 2020]
  - A bag of useful tricks for practical neural machine translation: Embedding layer initialization and large batch size [Neishi et al, ACL 2017]
  - Unsupervised pretraining for sequence to sequence learning, [Ramachandran et al., EMNLP 2017]
  - Incorporate BERT into Neural Machine Translation, [Zhu et al ICLR 2020]
  - Acquiring Knowledge from Pre-trained Model to Neural Machine Translation, [Weng et al AAAI 2020]
  - Towards Making Most of BERT for NMT, [Yang et al AAAI 2020]
  - Comparison between Pre-training and Large-scale Back-translation, [Huang et al., ACL 2021]
  - 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, [Lewis et al ACL 2020]

Reference

- Multilingual Pre-training
  - Cross-lingual Language Model Pre-training [Conneau et al NeurIPS 2019]
  - Alternating Language Modeling for Cross-Lingual Pre-Training [Yang et al AAAI 2020]
  - mBART: Multilingual Denoising Pre-training for Neural Machine Translation [Liu et al., TACL 2020]
  - Pre-training Multilingual Neural Machine Translation by Leveraging Alignment Information [Lin et al., EMNLP 2020]
  - CSP: Code-Switching Pre-training for Neural Machine Translation [Yang et al., EMNLP 2020]
  - Contrastive Learning for Many-to-many Multilingual Neural Machine Translation [Pan et al., ACL 2021]
  - Learning Language Specific Sub-network for Multilingual Machine Translation [Lin et al., ACL 2021]

Reference

- Speech Translation
  - wav2vec: Unsupervised Pre-training for Speech Recognition
  - wav2vec 2.0: A framework for self-supervised learning of speech representations
  - Investigating self-supervised pre-training for end-to-end speech translation
  - Self-supervised representations improve end-to-end speech translation (wav2vec + LSTM seq2seq)
  - Large-Scale Self-and Semi-Supervised Learning for Speech Translation
  - Consecutive Decoding for Speech-to-text Translation
  - "Listen, Understand and Translate": Triple Supervision Decouples End-to-end Speech-to-text Translation
  - Learning Shared Semantic Space for Speech-to-Text Translation [ACL 21]
  - Multilingual Speech Translation with Efficient Finetuning of Pretrained Models [ACL 21]
  - Fused Acoustic and Text Encoding for Multimodal Bilingual Pretraining and Speech Translation [ICML 21]
  - End-to-end Speech Translation via Cross-modal Progressive Training [Interspeech 21]
  - Curriculum Pre-training for End-to-end Speech Translation [ACL 20]
  - End-to-End Speech Translation with Knowledge Distillation [Interspeech 19]