CS11-737 Multilingual NLP

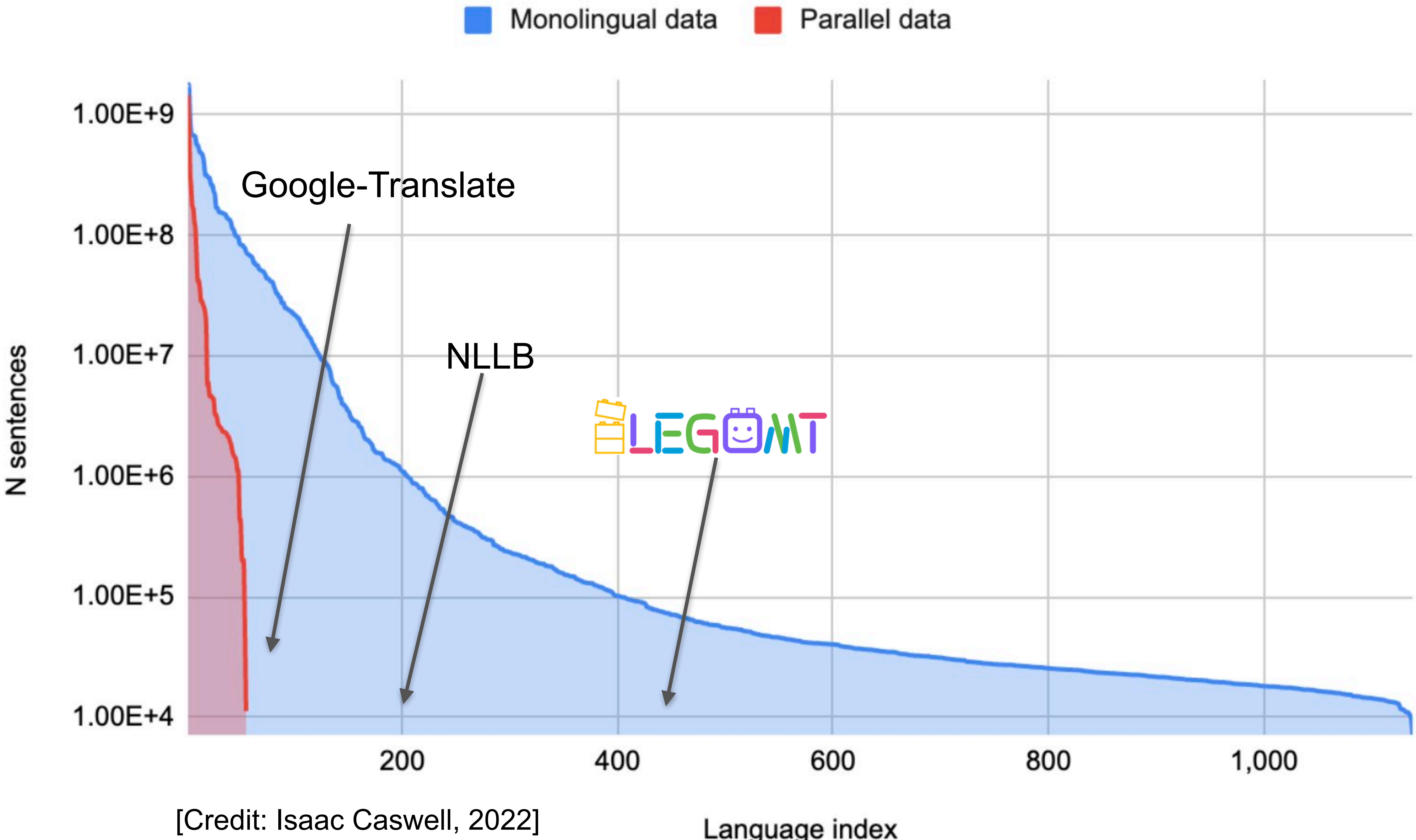
Multilingual Neural Machine Translation Model Architecture

Lei Li

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

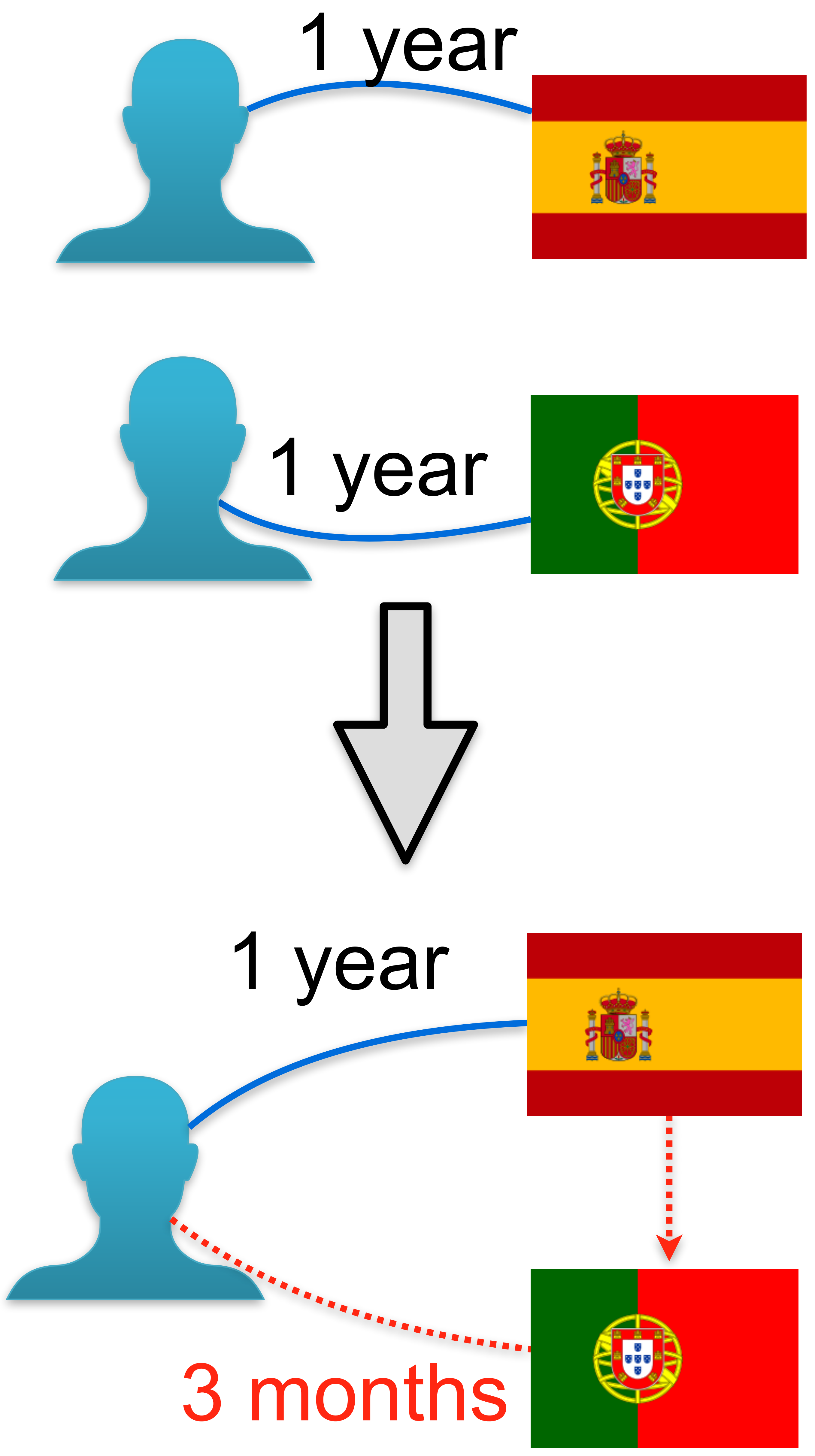
Carnegie Mellon University
Language Technologies Institute

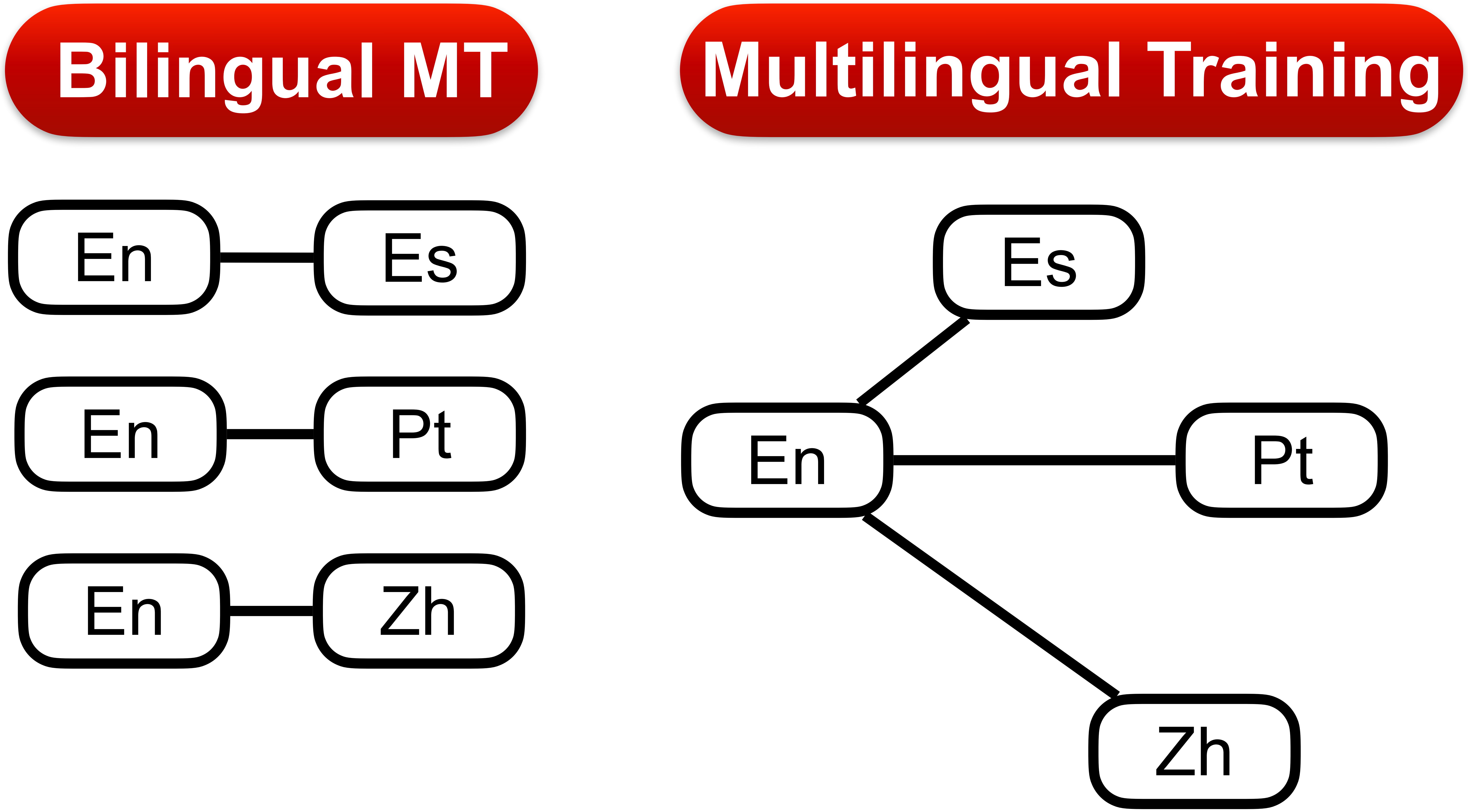
Language Data



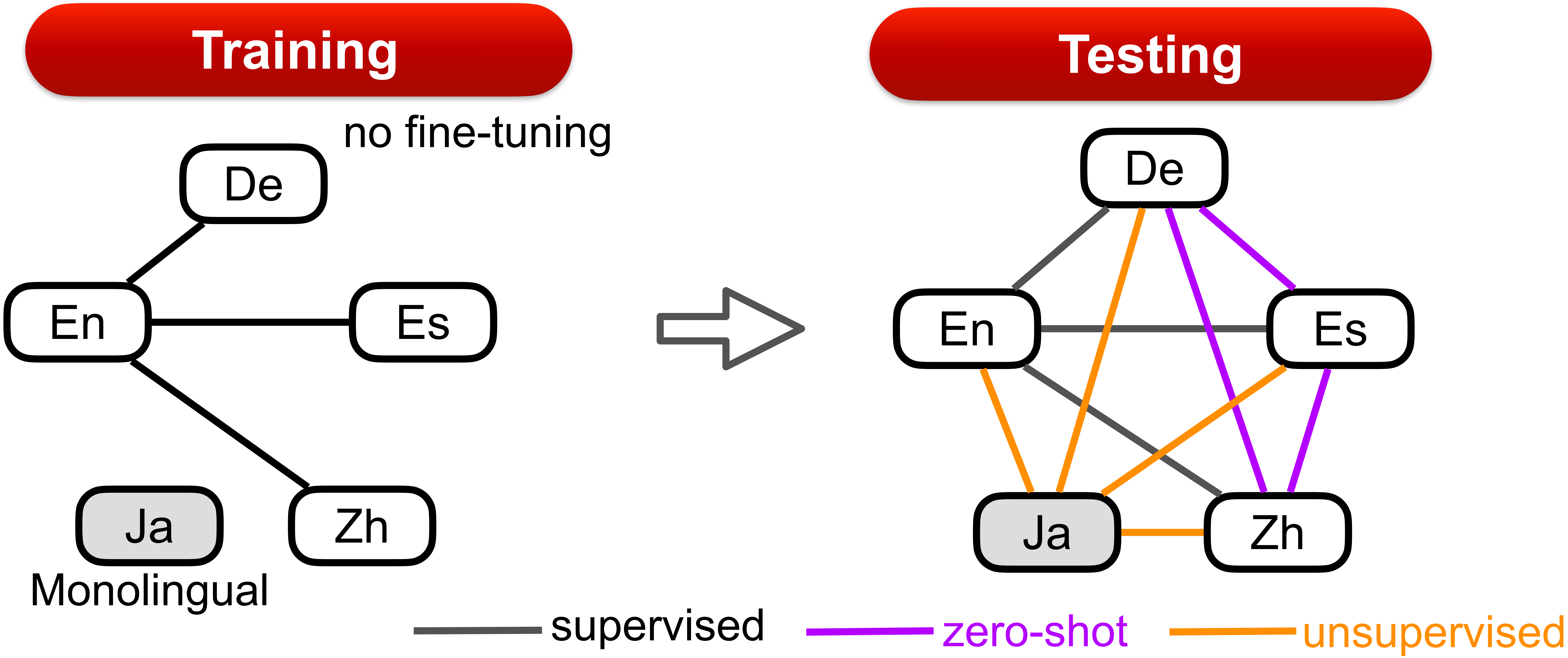
[Credit: Isaac Caswell, 2022]

Training Multilingual MT Jointly





Many-to-Many Multilingual NMT



Why Multilingual NMT?

- Develop one model to translate between all language pairs.
- Model-side: Languages with rich resource could benefit those with low resource
  - Similar languages share tokens
- Serving-side: only one model deployment versus of many deployments. Simpler workload and job management and scheduling.
  - Many languages would have much few requests but still need to occupy the servers.

MNMT Categorization

- Many-to-one:
  - Many source language to a target language
  - Usually the target is English
- One-to-Many:
  - One source language to many target languages
  - Usually the source is English
- Many-to-many:
  - Many source language to many target languages
  - Should include non-English pairs (often low-resource or zero-resource setting), very challenging!
- Which is simpler?

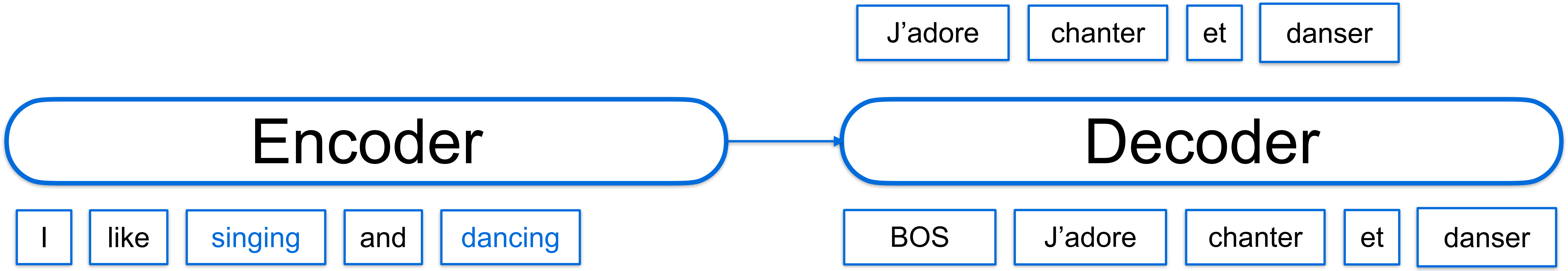
MNMT Fine-tuning Testing

- Exotic (Unseen) pair
  - Both the testing source language and target language appeared in the training, but the source-target pair never appeared in the training
  - Also known as zero-shot MNMT
- Exotic (Unseen) source
  - Testing source language never occur in the training
- Exotic (Unseen) target
  - Testing target language never occur in the training
- Exotic (Unseen) full
  - Neither the source language nor the target language for testing occur in the training
  - Is it even possible? Yes, for the pre-train fine-tuning paradigm.

MNMT with Language Tags

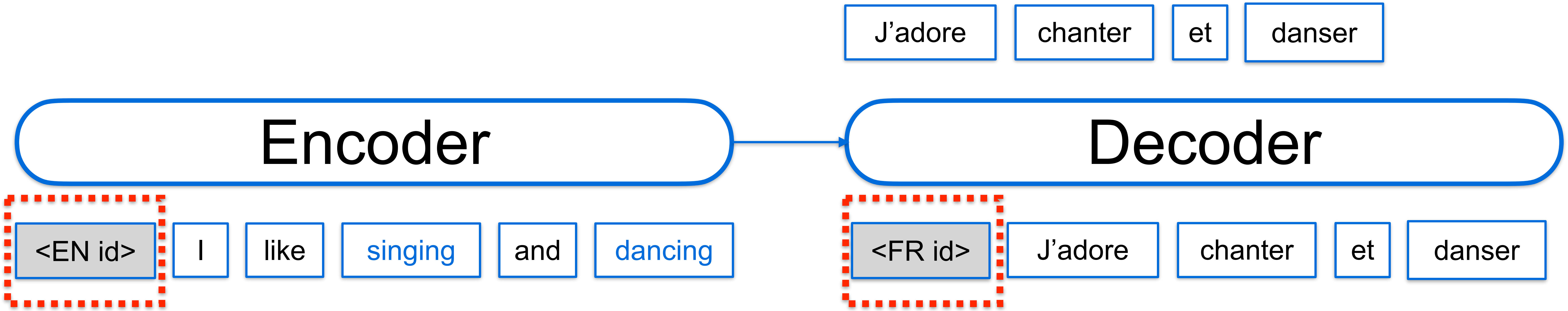
A single model for Multilingual NMT

- Language-specific encoding (@en@car, @de@automobile)
- But hard to learn a joint embedding.
- Challenge:
  - large vocabulary (twice many)
  - how does the model know it is to translate into German or French?

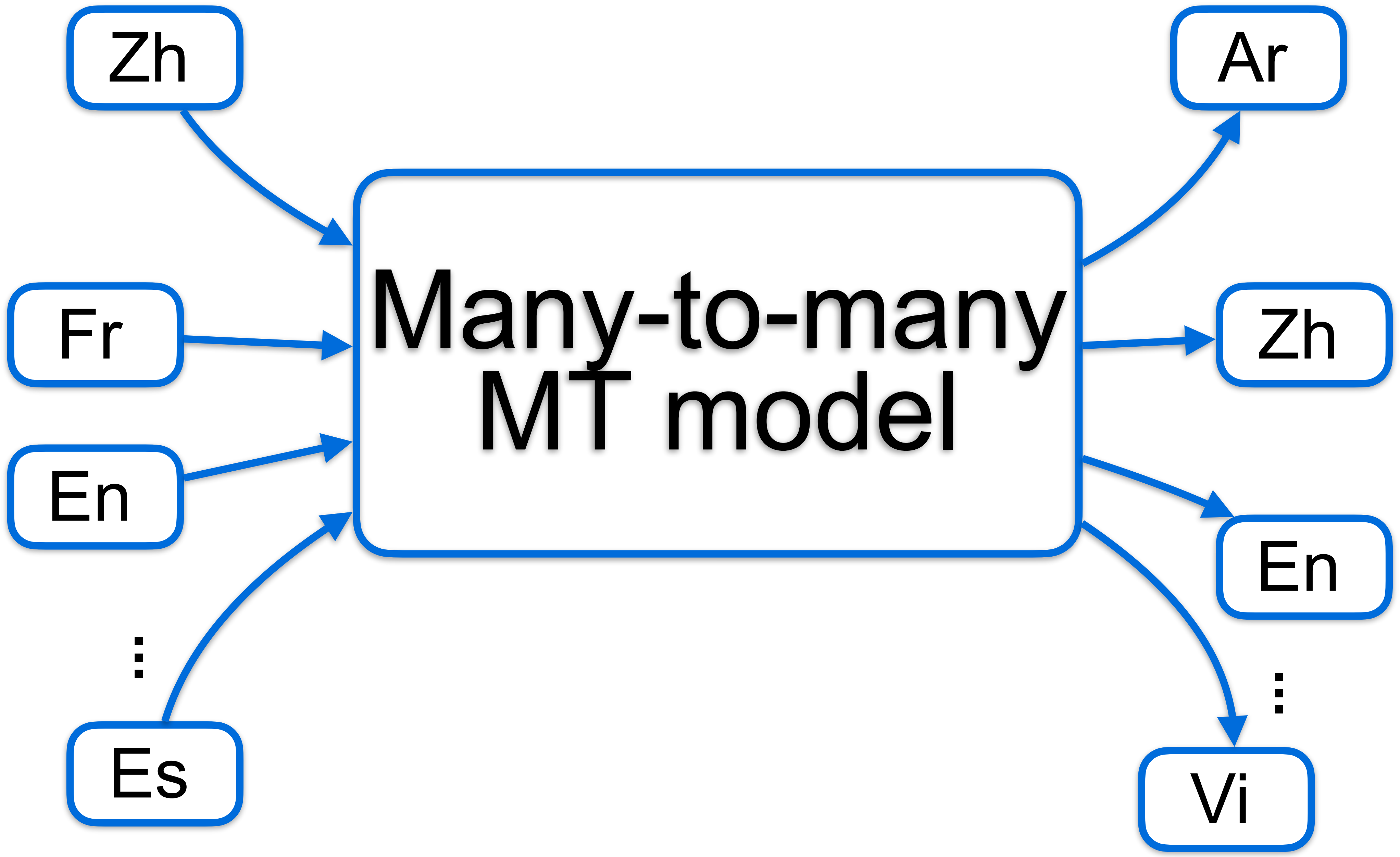


Ha et al. Toward Multilingual Neural Machine Translation with Universal Encoder and Decoder. 2016

Multilingual Machine Translation - Language Tag



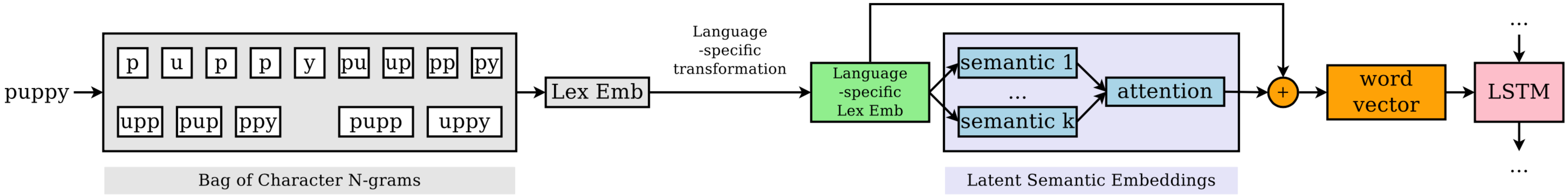
- One model can translate between many languages.
- Language Tag is used to indicate the source and target language.
- Vocabulary is built jointly



Johnson et al. Google’s Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. 2017

Vocabulary

- Single joint vocabulary [Johnson 2017]
  - combine all corpus together, and apply BPE
- Soft-decoupled encoding [Wang et al 2019]
- Even better: learned vocabulary [Xu 2021], (later in class)



Google's MNMT

- Training 12 language pairs together
- LSTM-s2s:
  - 8 layer LSTM encoder, 1st layer bidirectional
  - 8 layer LSTM decoder with attention

Table 4: Large-scale experiments: BLEU scores for single language pair and multilingual models.

Model	Single	Multi	Multi	Multi	Multi
#nodes	1024	1024	1280	1536	1792
#params	3B	255M	367M	499M	650M
En→Ja	23.66	21.10	21.17	21.72	21.70
En→Ko	19.75	18.41	18.36	18.30	18.28
Ja→En	23.41	21.62	22.03	22.51	23.18
Ko→En	25.42	22.87	23.46	24.00	24.67
En→Es	34.50	34.25	34.40	34.77	34.70
En→Pt	38.40	37.35	37.42	37.80	37.92
Es→En	38.00	36.04	36.50	37.26	37.45
Pt→En	44.40	42.53	42.82	43.64	43.87
En→De	26.43	23.15	23.77	23.63	24.01
En→Fr	35.37	34.00	34.19	34.91	34.81
De→En	31.77	31.17	31.65	32.24	32.32
Fr→En	36.47	34.40	34.56	35.35	35.52
ave diff	-	-1.72	-1.43	-0.95	-0.76
vs single	-	-5.6%	-4.7%	-3.1%	-2.5%

Johnson et al. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. 2017

Google's MNMT Zero-shot

- Bilingual pivot
- Multilingual joint
- What is missing in the table?
  - Multilingual pivot

Table 5: Portuguese→Spanish BLEU scores using various models.

	Model	Zero-shot	BLEU
(a)	PBMT bridged	no	28.99
(b)	NMT bridged	no	30.91
(c)	NMT Pt→Es	no	31.50
(d)	Model 1 (Pt→En, En→Es)	yes	21.62
(e)	Model 2 (En↔{Es, Pt})	yes	24.75
(f)	Model 2 + incremental training	no	31.77

zero-shot

no longer zero-shot, since additional Pt-Es pairs are used.

Johnson et al. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. 2017

Google's MNMT Zero-shot

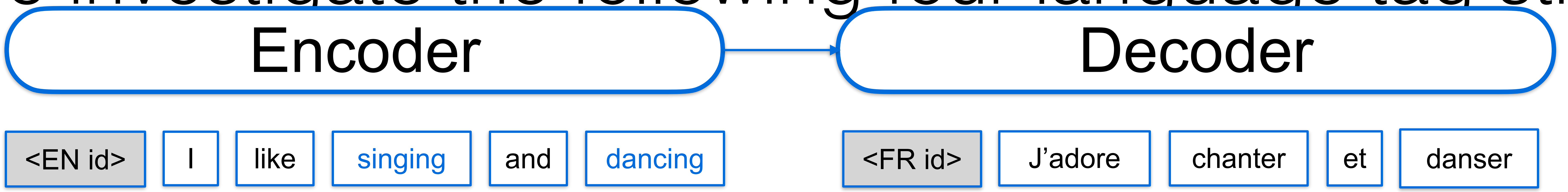
- MNMT is worse than pivot on zero-shot directions

Table 6: Spanish→Japanese BLEU scores for explicit and implicit bridging using the 12-language pair large-scale model from Table 4.

	Model	BLEU
	NMT Es→Ja explicitly bridged	18.00
zero-shot	NMT Es→Ja implicitly bridged	9.14

Source Language Tag or target Language Tag?

- We investigate the following four language tag str

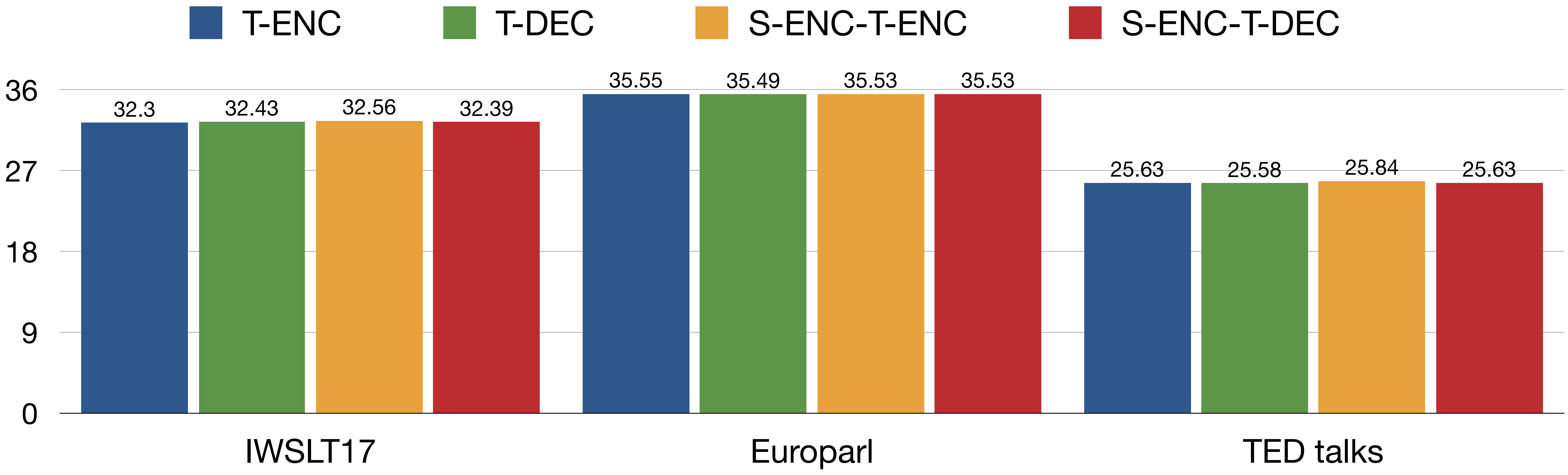


Strategy	Source sentence	Target sentence
Original	Hello World!	¡Hola Mundo
T-ENC	__es__ Hello World!	¡Hola Mundo
T-DEC	Hello World!	__es__ ¡Hola Mundo
S-ENC-T-ENC	__en__ __es__ Hello World!	¡Hola Mundo
S-ENC-T-DEC	__en__ Hello World!	__es__ ¡Hola Mundo

Wu et al. Language Tags Matter for Zero-Shot Neural Machine Translation 2021.

Language Tag Does not Affect Performance on Supervised Directions

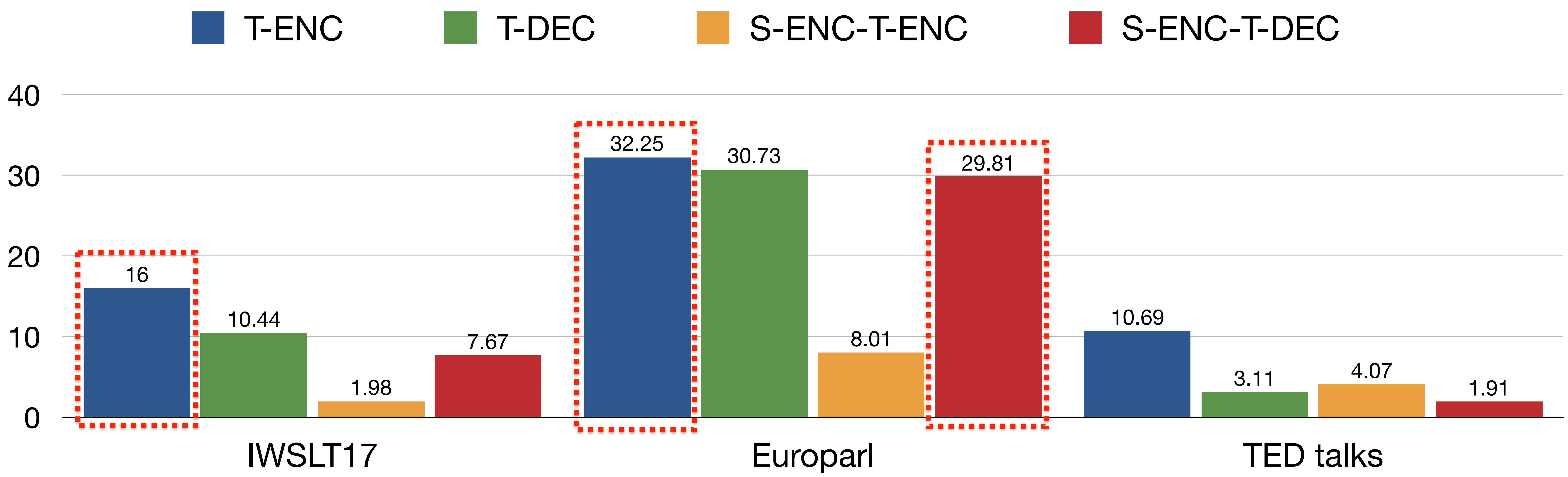
Supervised directions: The directions which has been seen together in the training time.



Wu et al. Language Tags Matter for Zero-Shot Neural Machine Translation 2021.

Target Language Tag on Encoder Strategy Gets Best Zero-Shot Performance

Zero-shot directions: The directions between known languages that the model has never seen together at training time.



Wu et al. Language Tags Matter for Zero-Shot Neural Machine Translation 2021.

Mixed Source Language can still be Translated

- {Ja, Ko} -> En
- Japanese: 私は東京大学の学生です。 → I am a student at Tokyo University.
- Korean: 나ㄴㄴㅡ ㅗㄷ쿄 ㅐㄷ학ㅢㅇ 학ㅐㅇㅅㅂ이니다. → I am a student at Tokyo University.
- Japanese/Korean: 私は東京大学ㅏㅎㄱㅇㅐㅅ입ㅣㄴ 다. → I am a student of Tokyo University.

Mixed Decoder for Target Language

- En -> {Ja, Ko}
- Either generate Japanese or Korean

Table 8: Gradually mixing target languages Ja/Ko.

w_{ko}	I must be getting somewhere near the centre of the earth.
0.00	私は地球の中心の近くにどこかに行っているに違いない。
0.40	私は地球の中心近くのどこかに着いているに違いない。
0.56	私は地球の中心の近くのどこかになっているに違いない。
0.58	私は지구の中心의가까이에어딘가에도착하고있어야한다。
0.60	나는지구의센터의가까이에어딘가에도착하고있어야한다。
0.70	나는지구의중심근처어딘가에도착해야합니다。
0.90	나는어딘가지구의중심근처에도착해야합니다。
1.00	나는어딘가지구의중심근처에도착해야합니다。

Multilingual NMT with mTransformer

- Model: Transformer-base (6e6d, 512) ==> mTransformer
- Data: TED-talk, 59 languages, 116 directions

	Az-En	Be-En	Gl-En	Sk-En	Avg.
# of examples	5.9k	4.5k	10k	61k	20.3k
Neubig & Hu 18					
baselines	2.7	2.8	16.2	24	11.42
many-to-one	11.7	18.3	29.1	28.3	21.85
Wang et al. 18	11.82	18.71	30.3	28.77	22.4
Ours					
many-to-one	11.24	18.28	28.63	26.78	21.23
many-to-many	**12.78**	**21.73**	**30.65**	**29.54**	**23.67**

	Ar-En	De-En	He-En	It-En	Avg.
# of examples	213k	167k	211k	203k	198.5k
baselines	27.84	30.5	**34.37**	33.64	31.59
many-to-one	25.93	28.87	30.19	32.42	29.35
many-to-many	**28.32**	**32.97**	33.18	**35.14**	**32.4**

Aharoni et al. Massively Multilingual Neural Machine Translation. 2019

Limitation of mTransformer: does not work for Many-to-Many En-X

	En-Az	En-Be	En-Gl	En-Sk	Avg.
# of examples	5.9k	4.5k	10k	61k	20.3k
baselines	2.16	2.47	3.26	5.8	3.42
one-to-many	**5.06**	**10.72**	**26.59**	**24.52**	**16.72**
many-to-many	3.9	7.24	23.78	21.83	14.19

	En-Ar	En-De	En-He	En-It	Avg.
# of examples	213k	167k	211k	203k	198.5k
baselines	12.95	23.31	23.66	30.33	22.56
one-to-many	**16.67**	**30.54**	**27.62**	**35.89**	**27.68**
many-to-many	14.25	27.95	24.16	33.26	24.9

Table 3: En→X test BLEU on the TED Talks corpus

Aharoni et al. Massively Multilingual Neural Machine Translation. 2019

Even More Languages

- mTransformer
  - 6e6d, 1024 -> 8192
  - 473m parameters
- 103 Languages (inc. En)
  - 64k vocab

# of language pairs	102
examples per pair	
min	63,879
max	1,000,000
average	940,087
std. deviation	188,194
total # of examples	95,888,938

	Ar	Az	Be	De	He	It	Nl	Ro	Sk	Tr	Avg.
baselines	23.34	16.3	21.93	30.18	31.83	**36.47**	36.12	34.59	25.39	27.13	28.33
many-to-one	**26.04**	**23.68**	**25.36**	35.05	**33.61**	35.69	**36.28**	36.33	28.35	**29.75**	**31.01**
many-to-many	22.17	21.45	23.03	**37.06**	30.71	35.0	36.18	**36.57**	**29.87**	27.64	29.97

Table 5: X→En test BLEU on the 103-language corpus

	Ar	Az	Be	De	He	It	Nl	Ro	Sk	Tr	Avg.
baselines	10.57	8.07	15.3	23.24	19.47	31.42	28.68	27.92	11.08	15.54	19.13
one-to-many	**12.08**	**9.92**	**15.6**	**31.39**	**20.01**	**33**	**31.06**	**28.43**	**17.67**	**17.68**	**21.68**
many-to-many	10.57	9.84	14.3	28.48	17.91	30.39	29.67	26.23	18.15	15.58	20.11

Table 6: En→X test BLEU on the 103-language corpus

Aharoni et al. Massively Multilingual Neural Machine Translation. 2019

More language trained together, but

	Ar-En	En-Ar	Fr-En	En-Fr	Ru-En	En-Ru	Uk-En	En-Uk	Avg.
5-to-5	**23.87**	**12.42**	**38.99**	**37.3**	29.07	**24.86**	**26.17**	16.48	**26.14**
25-to-25	23.43	11.77	38.87	36.79	**29.36**	23.24	25.81	**17.17**	25.8
50-to-50	23.7	11.65	37.81	35.83	29.22	21.95	26.02	15.32	25.18
75-to-75	22.23	10.69	37.97	34.35	28.55	20.7	25.89	14.59	24.37
103-to-103	21.16	10.25	35.91	34.42	27.25	19.9	24.53	13.89	23.41

mTransformer Zero-shot Performance

	Ar-Fr	Fr-Ar	Ru-Uk	Uk-Ru	Avg.
5-to-5	1.66	4.49	3.7	3.02	3.21
25-to-25	1.83	**5.52**	**16.67**	4.31	7.08
50-to-50	**4.34**	4.72	15.14	**20.23**	**11.1**
75-to-75	1.85	4.26	11.2	15.88	8.3
103-to-103	2.87	3.05	12.3	18.49	9.17

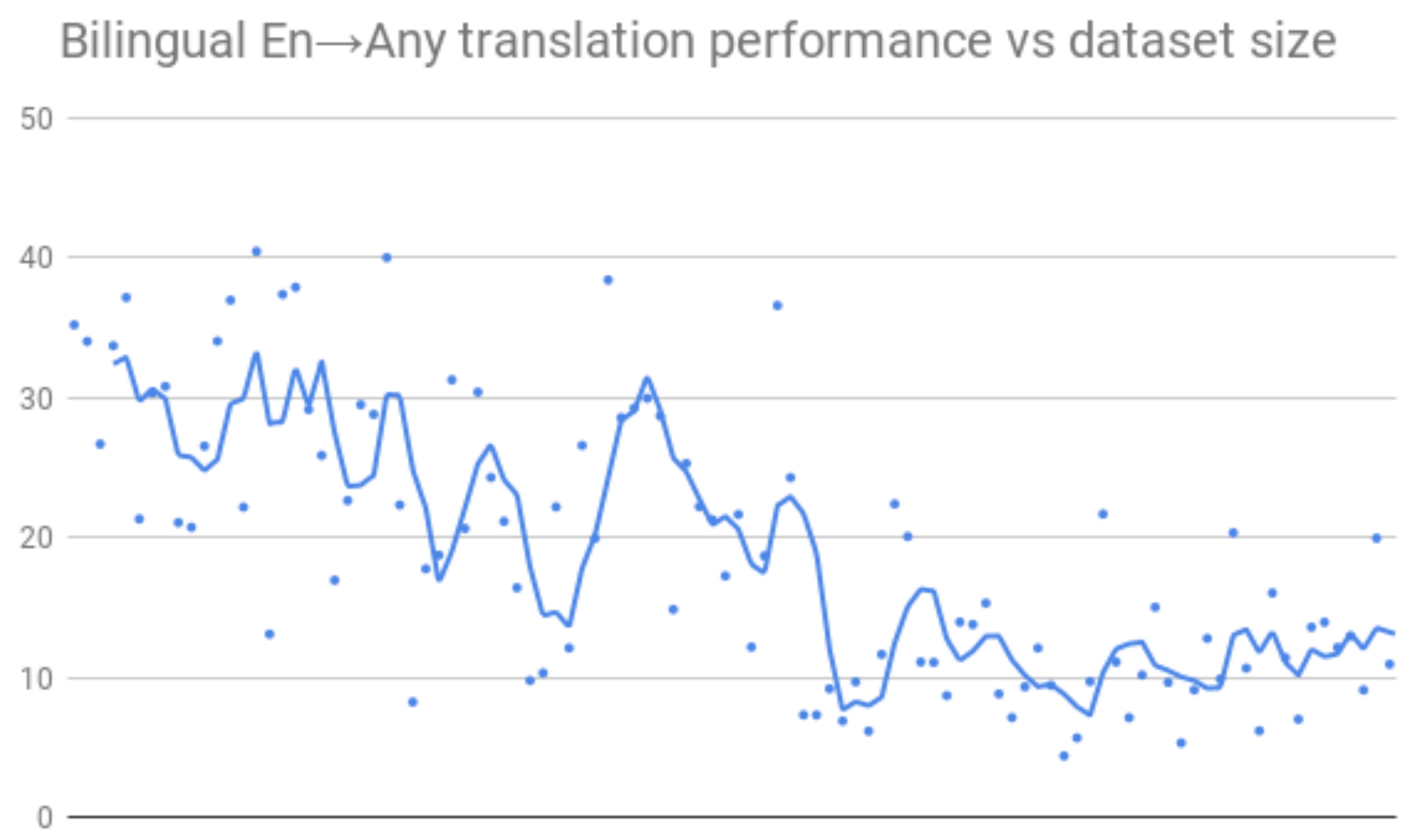
Table 8: Zero-Shot performance while varying the number of languages involved

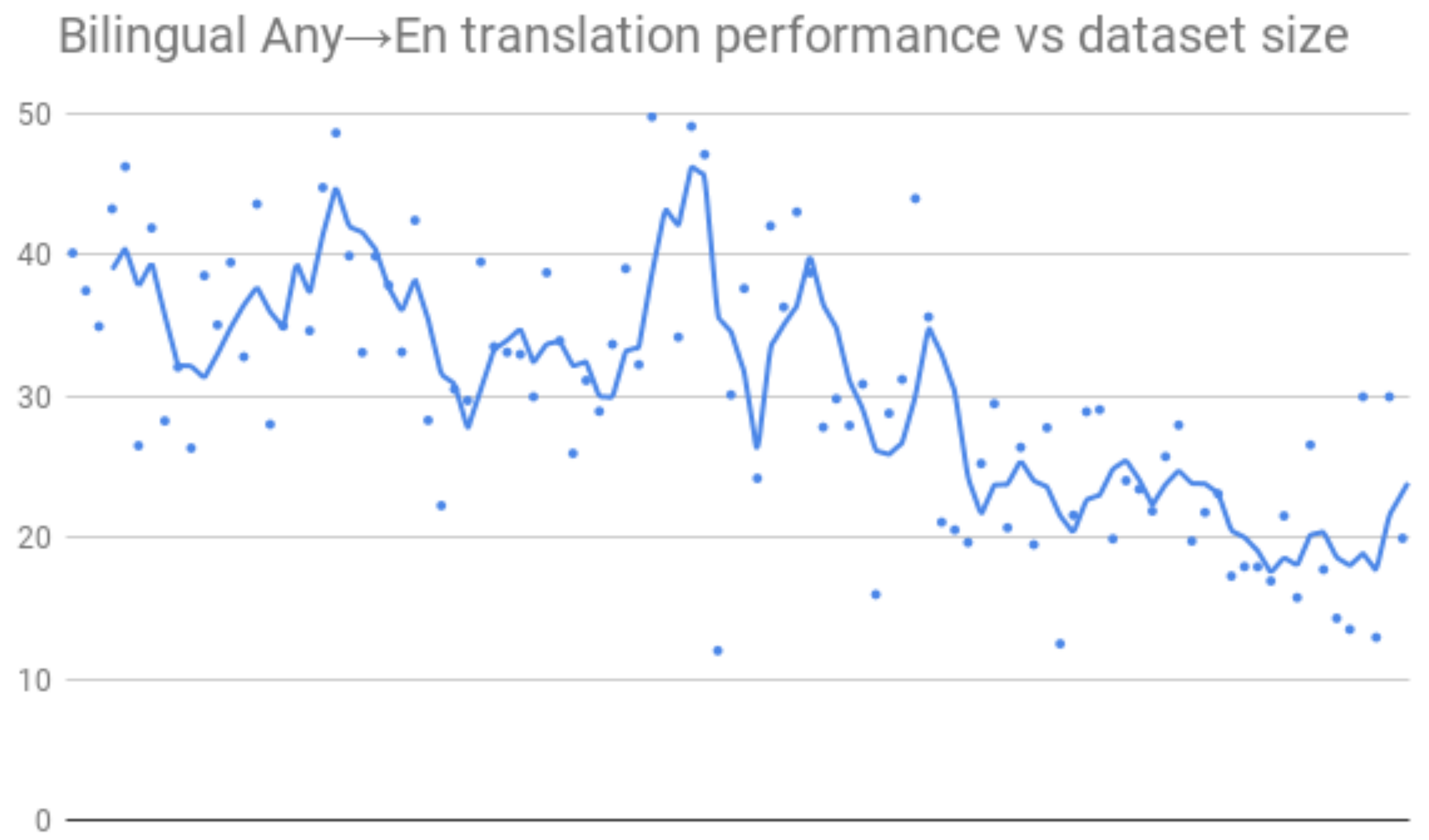
Bigger Data

- Data: 25 billion sentence pairs in 103 languages
- Model: mTransformer with 375million params (larger than Transformer-big)

En→Any	High 25	Med. 52	Low 25
Bilingual	29.34	17.50	11.72
All→All	28.03	16.91	12.75
En→Any	28.75	17.32	12.98

Any→En	High 25	Med. 52	Low 25
Bilingual	37.61	31.41	21.63
All→All	33.85	30.25	26.96
Any→En	36.61	33.66	30.56





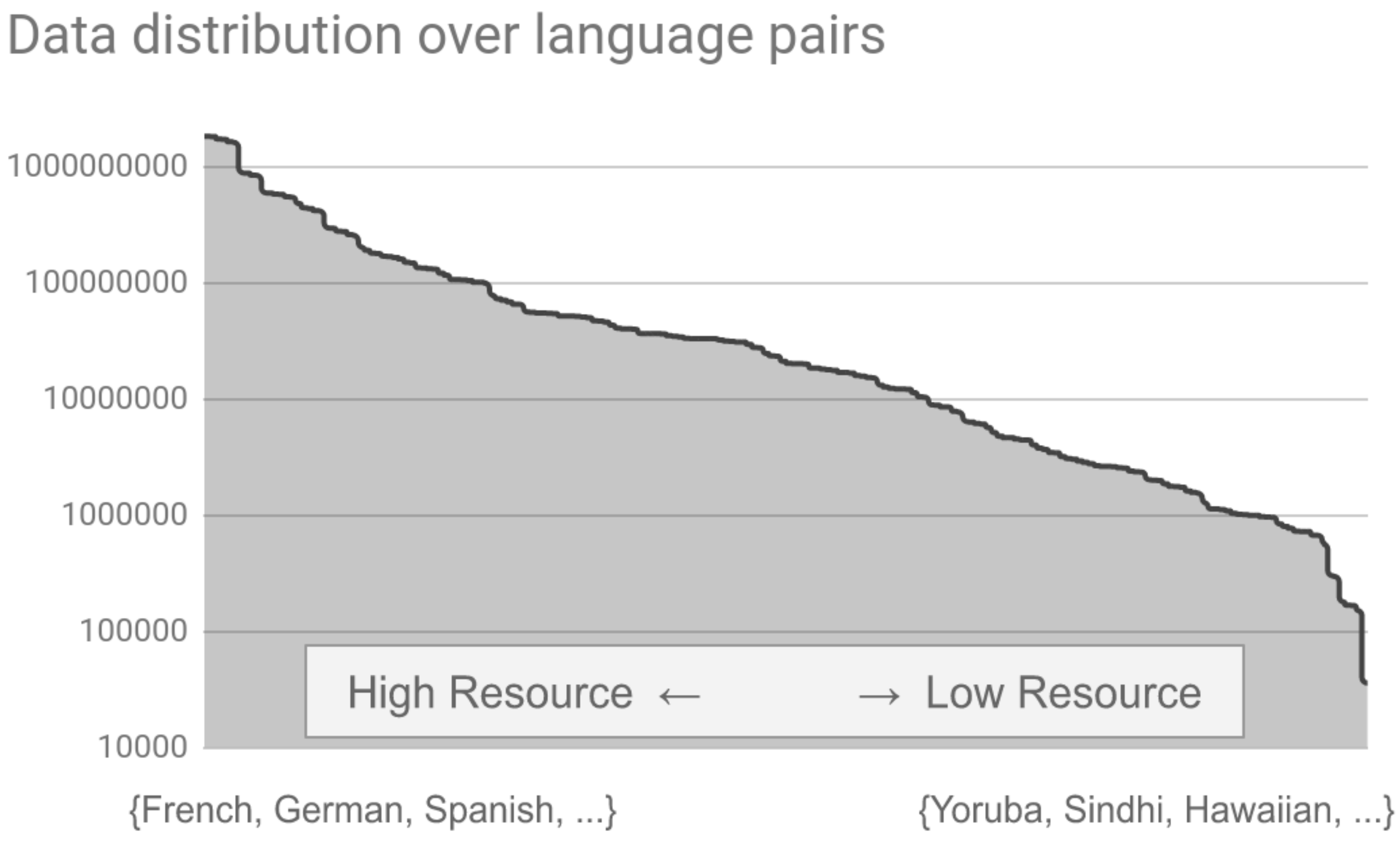
Arivazhagan et al. Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges. 2019

Sampling of Data

- sample data prob w.r.t

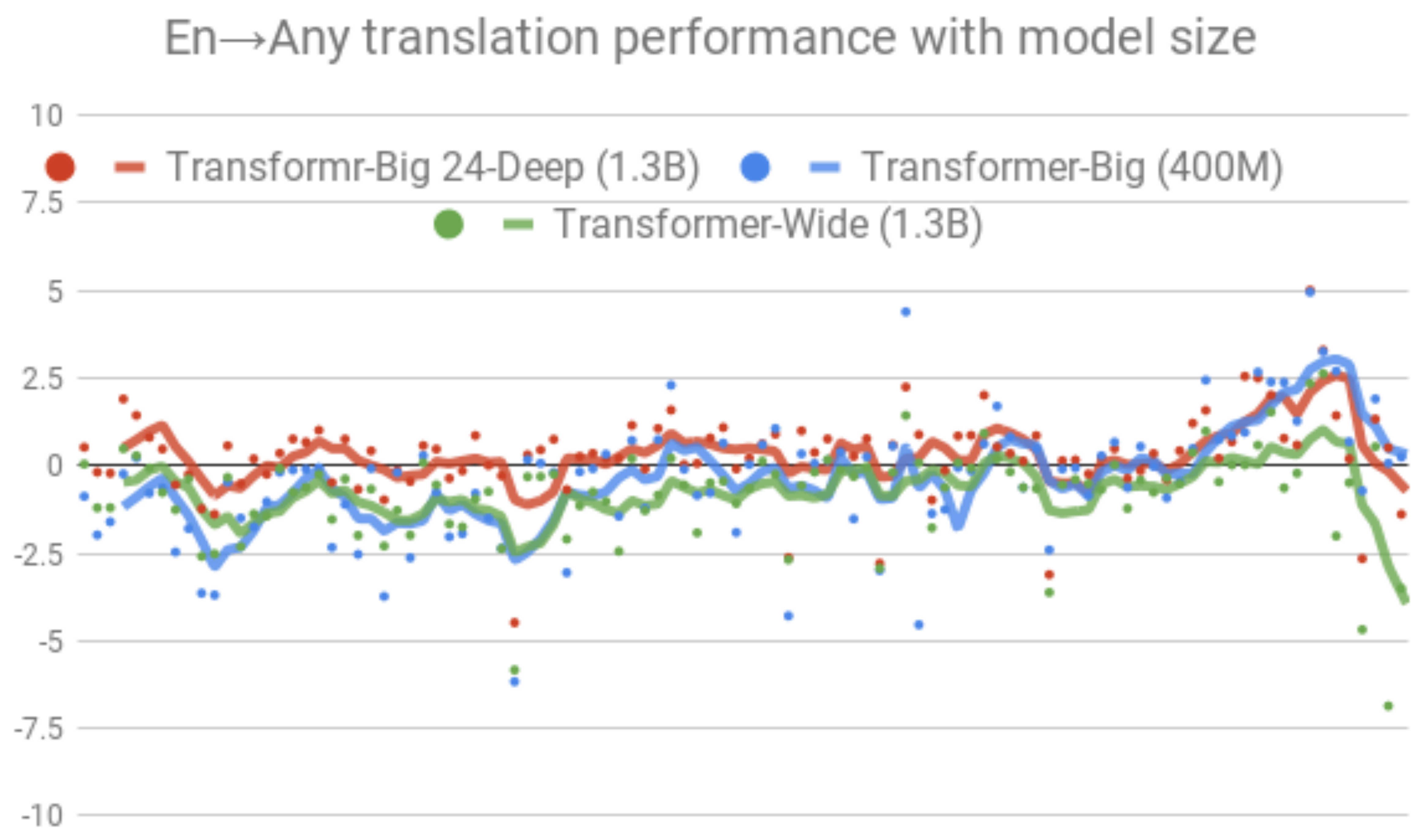
$$p^{\frac{1}{T}}$$

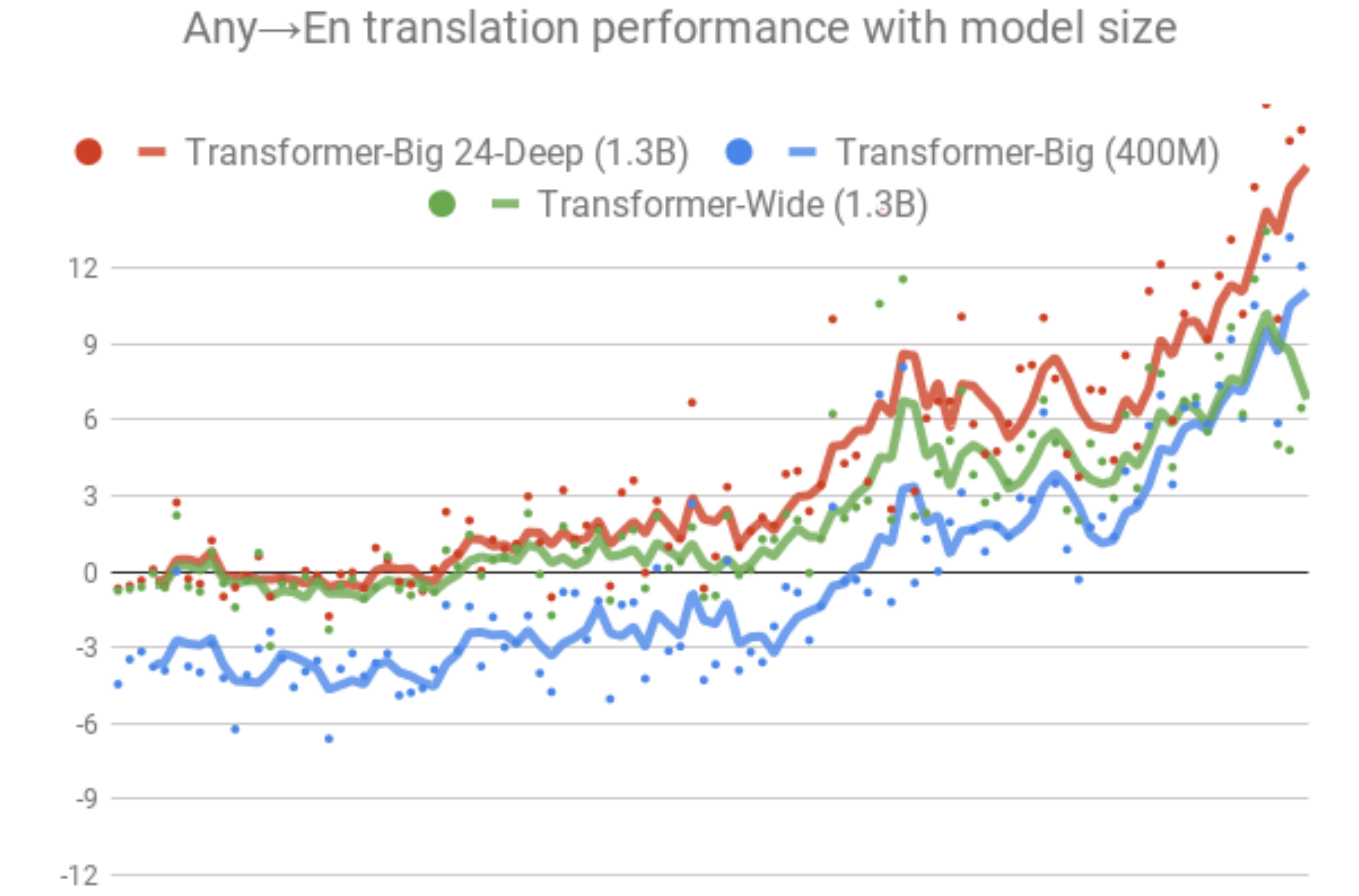
$En \to Any$	High 25	Med. 52	Low 25
$T_V = 1$	27.81	16.72	12.73
$T_V = 100$	27.83	16.86	12.78
$T_V = 5$	28.03	16.91	12.75
$Any \to En$	High 25	Med. 52	Low 25
$T_V = 1$	33.82	29.78	26.27
$T_V = 100$	33.70	30.15	26.91
$T_V = 5$	33.85	30.25	26.96



Bigger Model

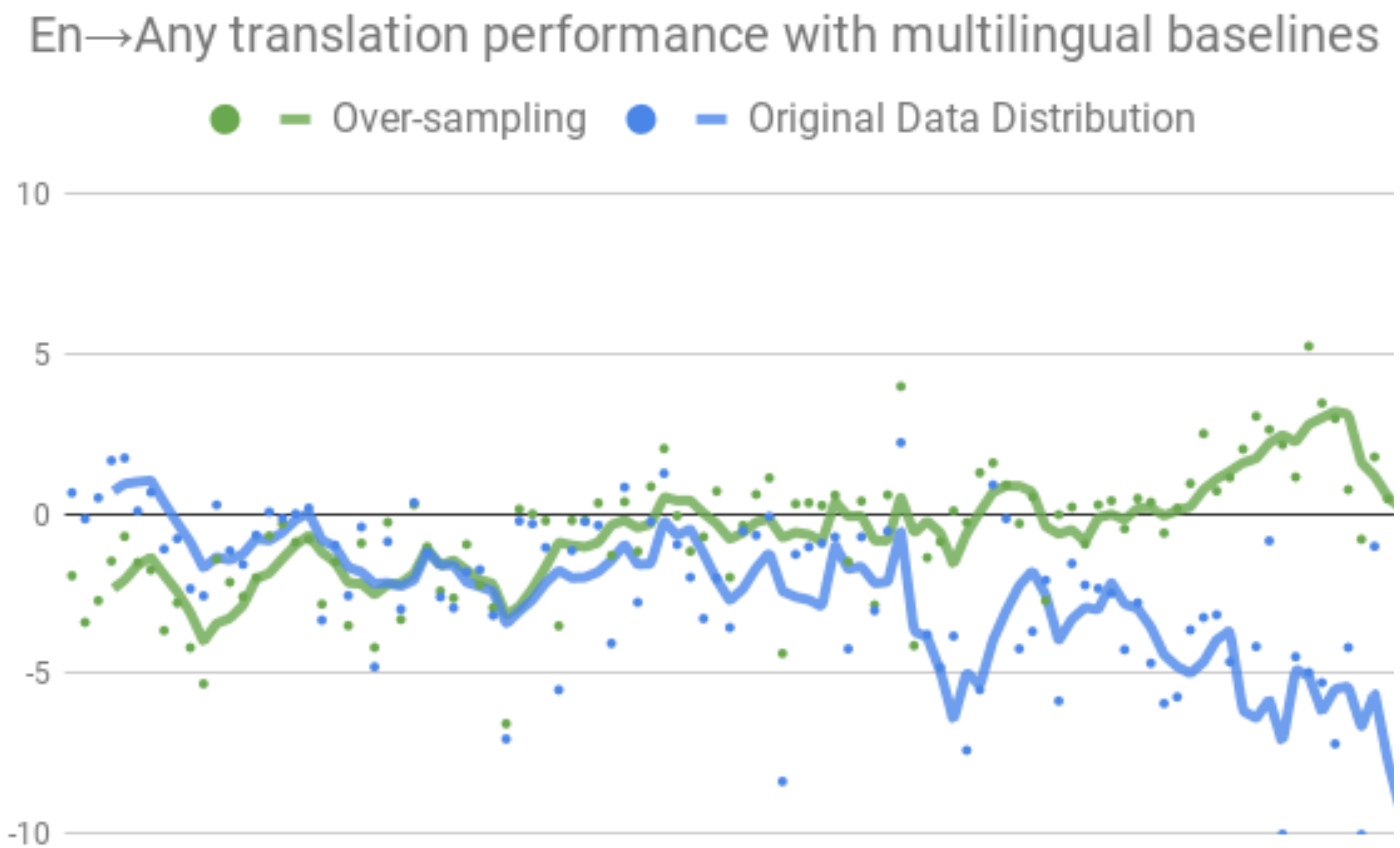
- mTransformer:
  - 400m, 1.3B wide (12e12d), 1.3B deep (24e24d)
  - Deep is better than wide!

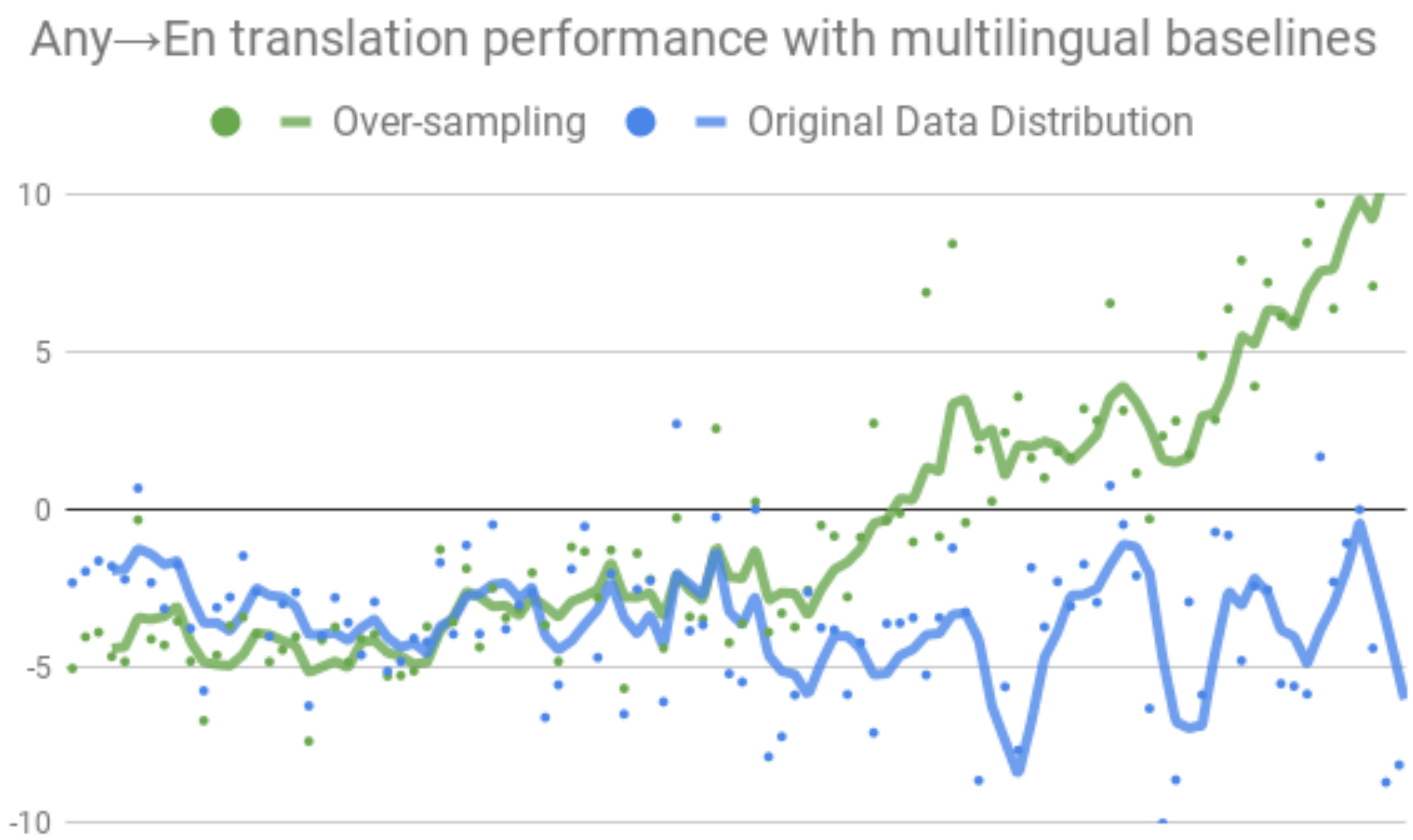




Limitation

- mTransformer boosts performance on low-resource languages but not high-resource
- Zero-shot directions are not usable yet.





Arivazhagan et al. Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges. 2019

MT w/ Adapter

Parameter Interference issue for MNMT

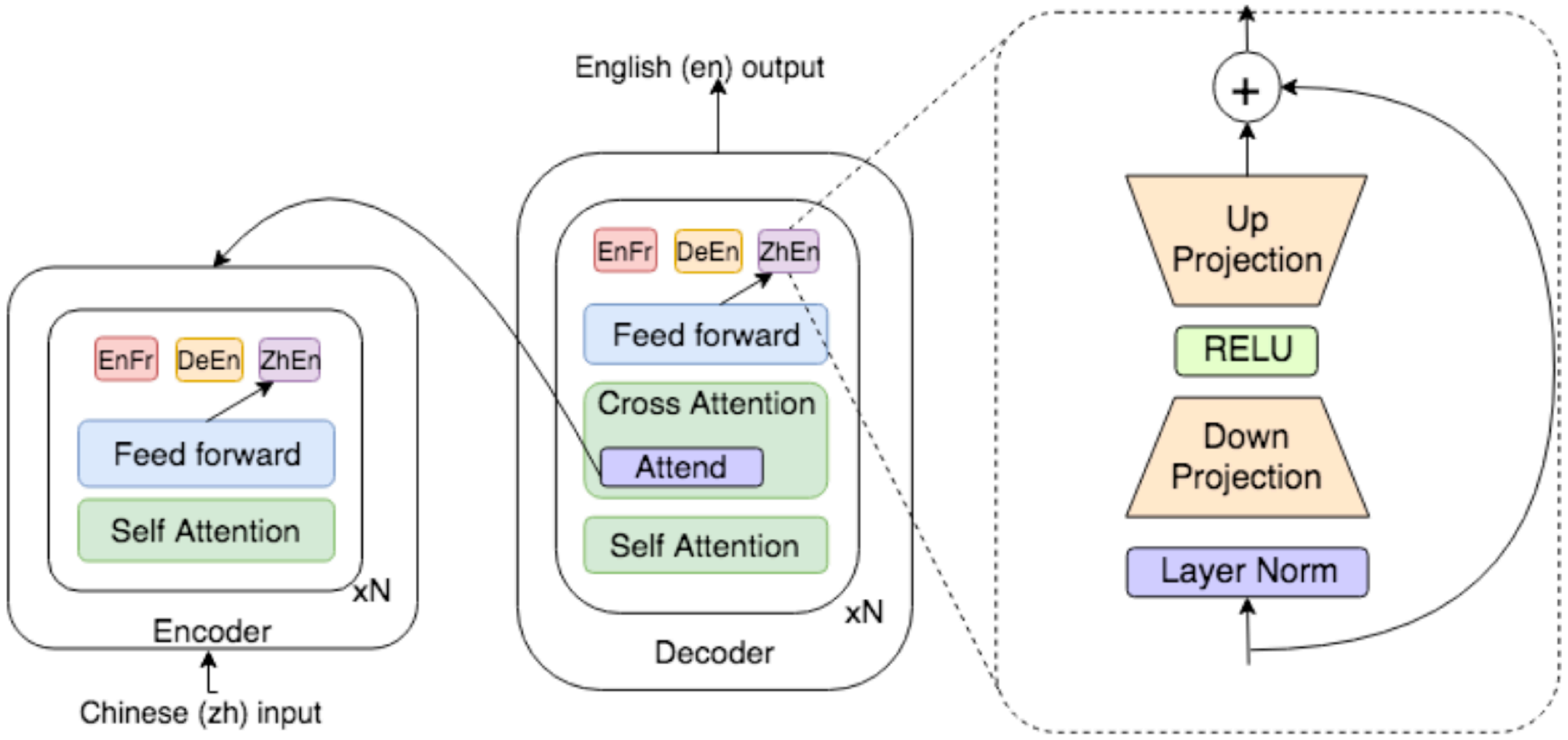
- Insufficient model capacity
  - the sharing model capacity has to be split for different translation directions

Bilingual

Multilingual

Multilingual NMT with Serial Adapter

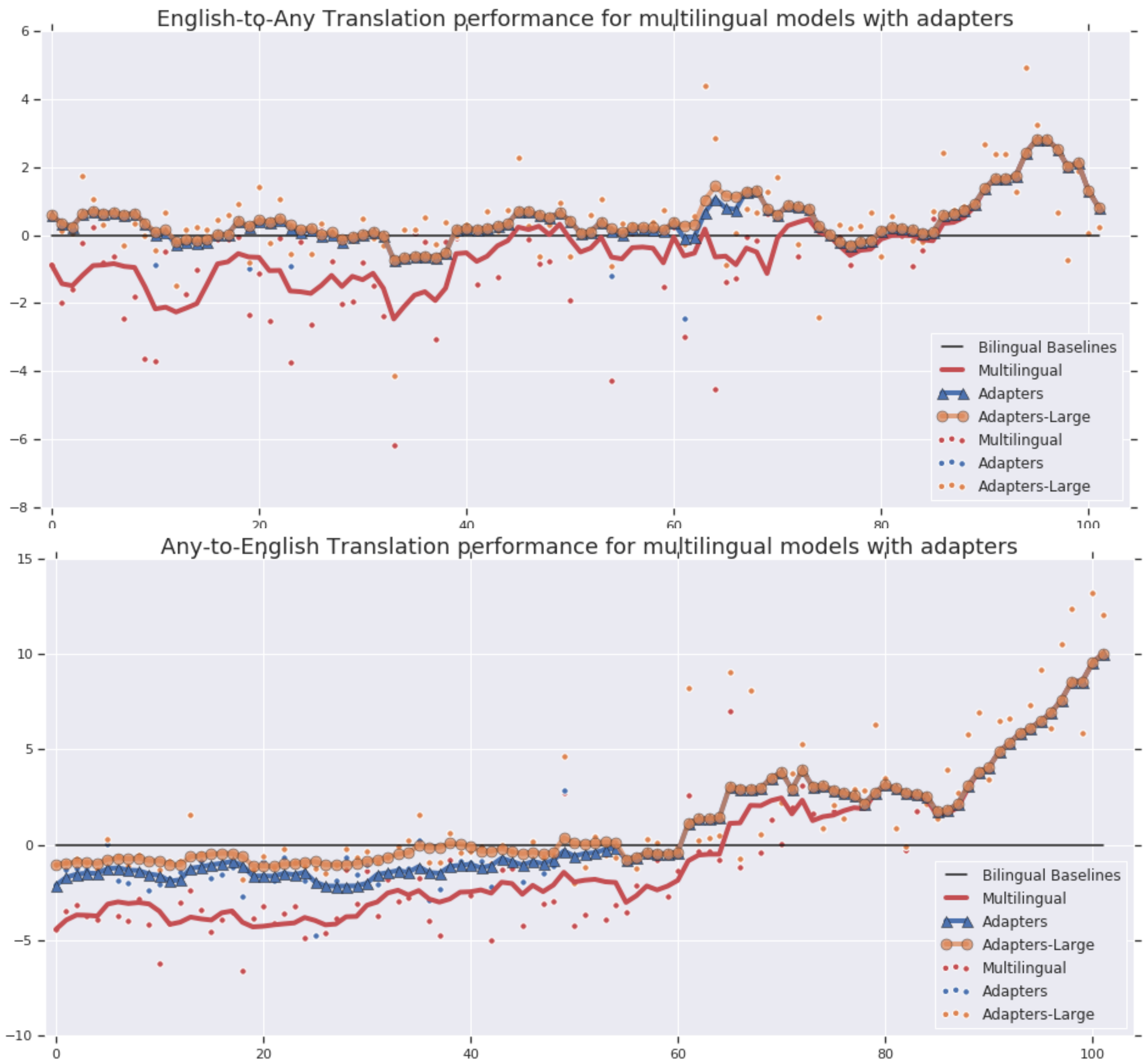
- For each layer, adding language-specific module
- $\tilde{z}$ =LNT(z_i).
- h =relu($W \tilde{z}$)
- x =$Wh + z$
- Could be used for both domain adaptation and MNMT
- Joint training the whole architecture



Bapna & Firat, Simple, Scalable Adaptation for Neural Machine Translation, 2019

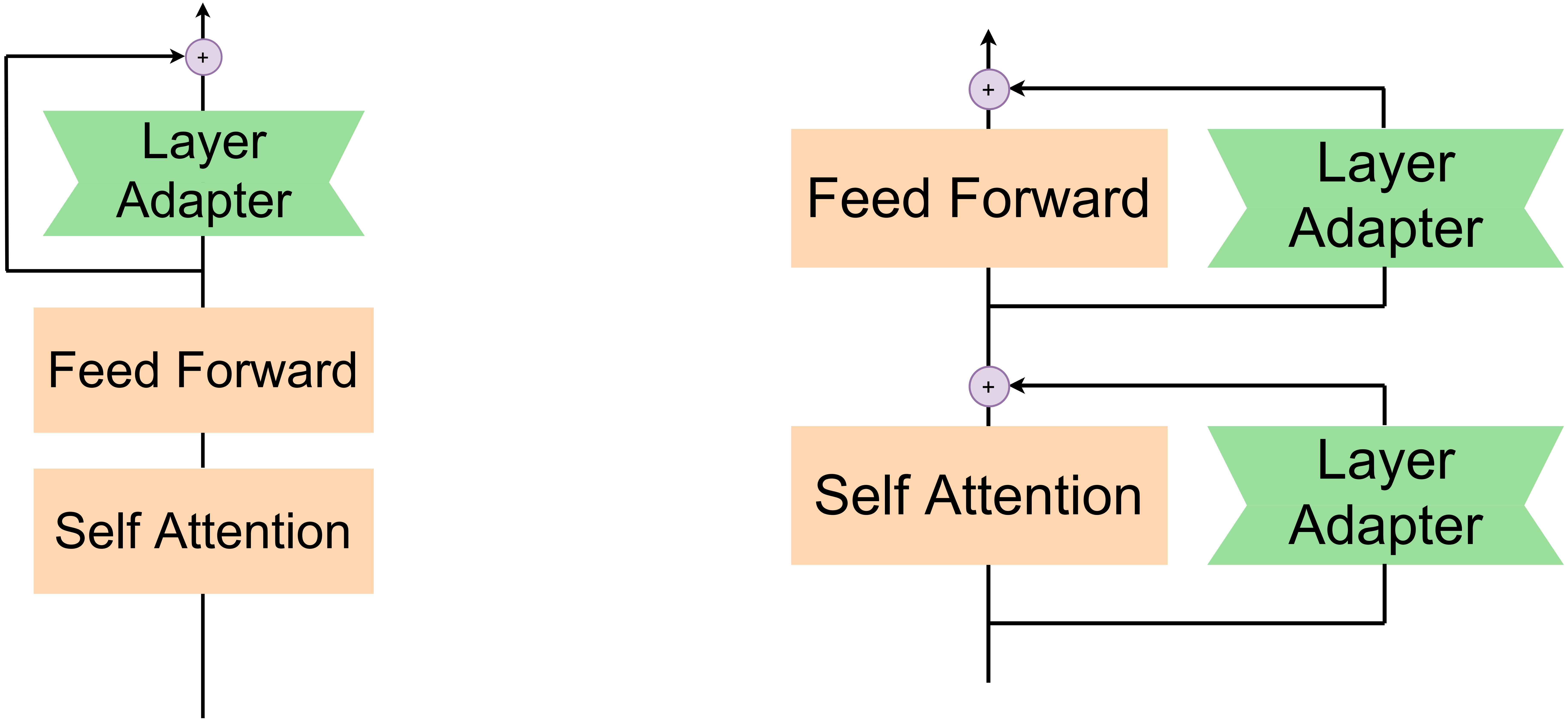
Serial Adapter improves Multilingual Translation

- on rich-resource lang.
- But serial-adapter is not plug-and-play
    - Joint training mTransformer+SA will be better than training mTransformer, fix, and train adapter.
    - Adapter has tight integration with the main architecture.



Counter Interference

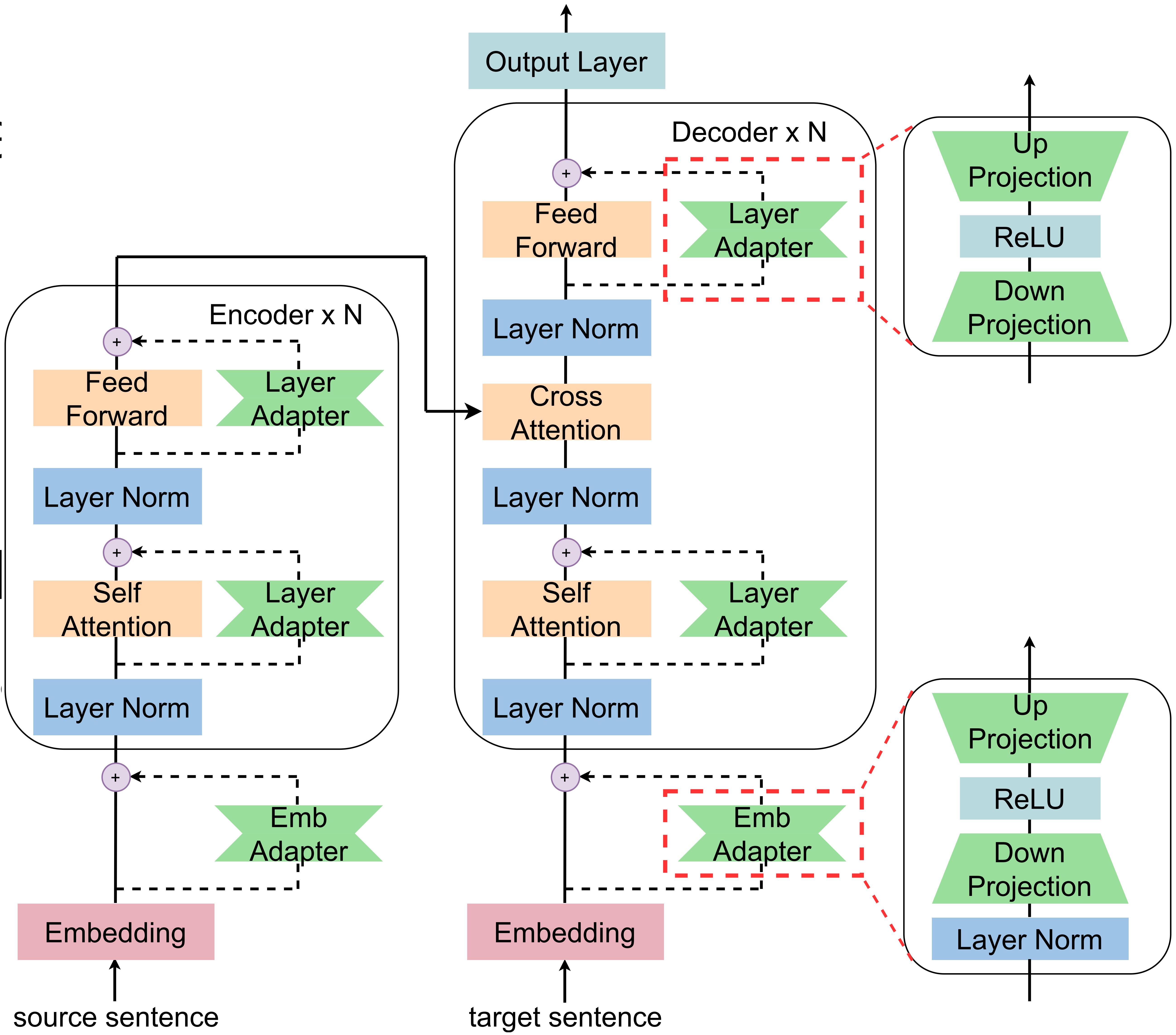
- Which adapter will remove noise?



Zhu et al. Counter-Interference Adapter for Multilingual Machine Translation. 2021

Parallel Adapter - CIAT

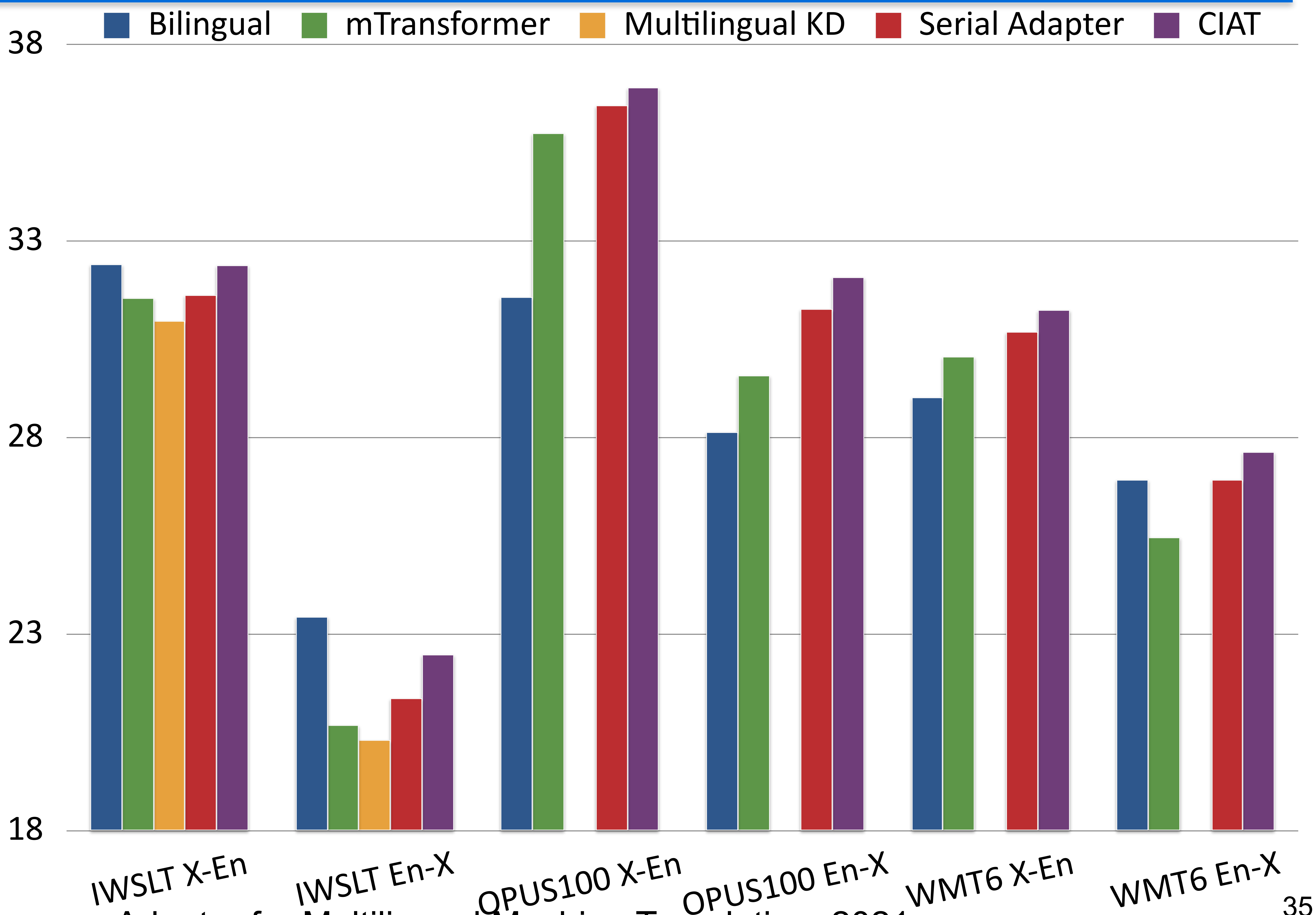
- Design rationale:
  - process before multilingual inte
- Embedding adapter
- Parallel layer adapter
- Training:
  - Pretrain mTransformer on multil
  - Fix mTransformer and train par pairs



Zhu et al. Counter-Interference Adapter for Multilingual Machine Translation. 2021

Comparing MNMT w/ Adapters

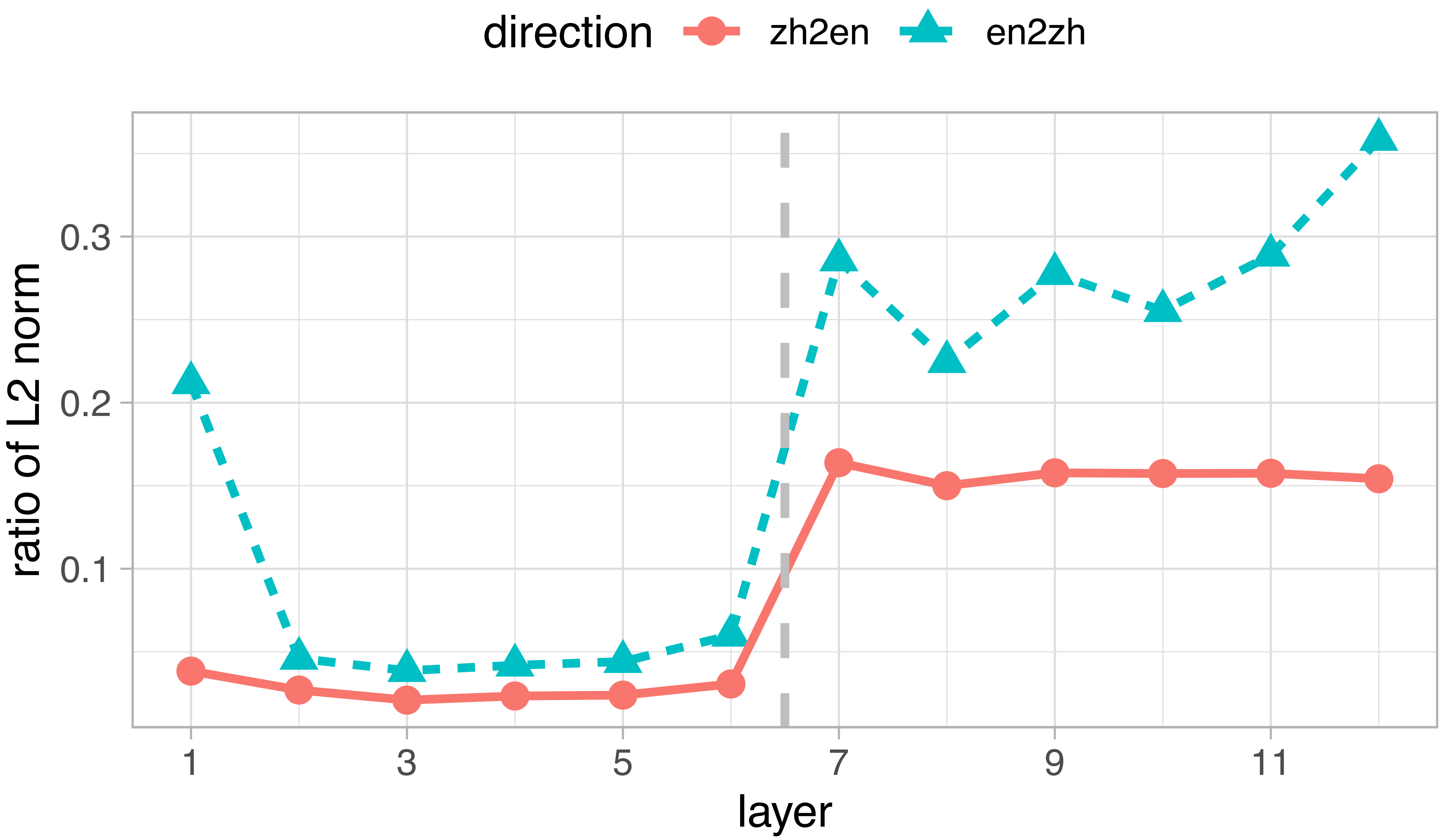
- mTransformer could be worse than bilingual Transformer
- Both serial adapter and parallel adapter (CIAT) improves mTransformer
- Parallel even beat bilingual Transformer! Serial adapter does not.



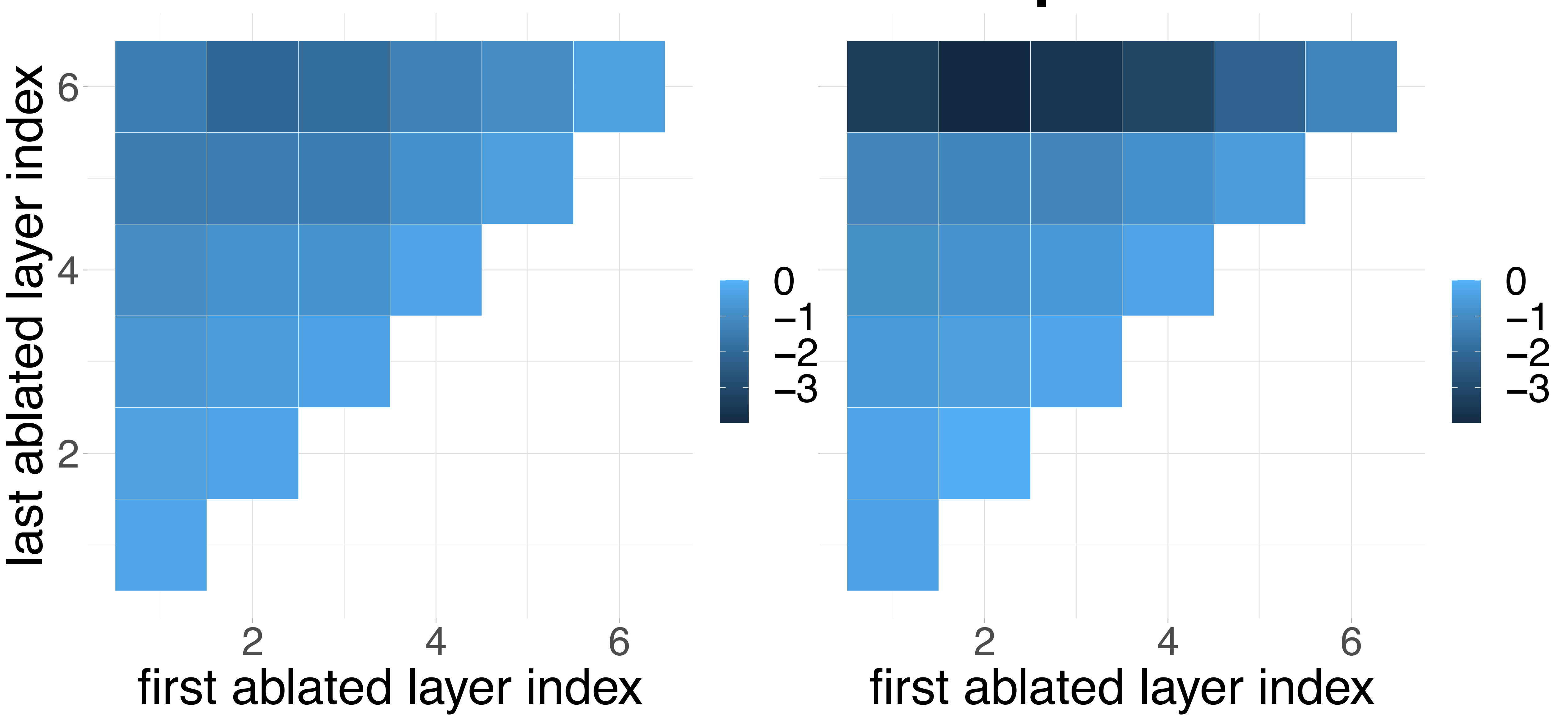
Zhu et al. Counter-Interference Adapter for Multilingual Machine Translation. 2021

Which layer-adapter are more important?

- Upper decoder layer adapter is more important

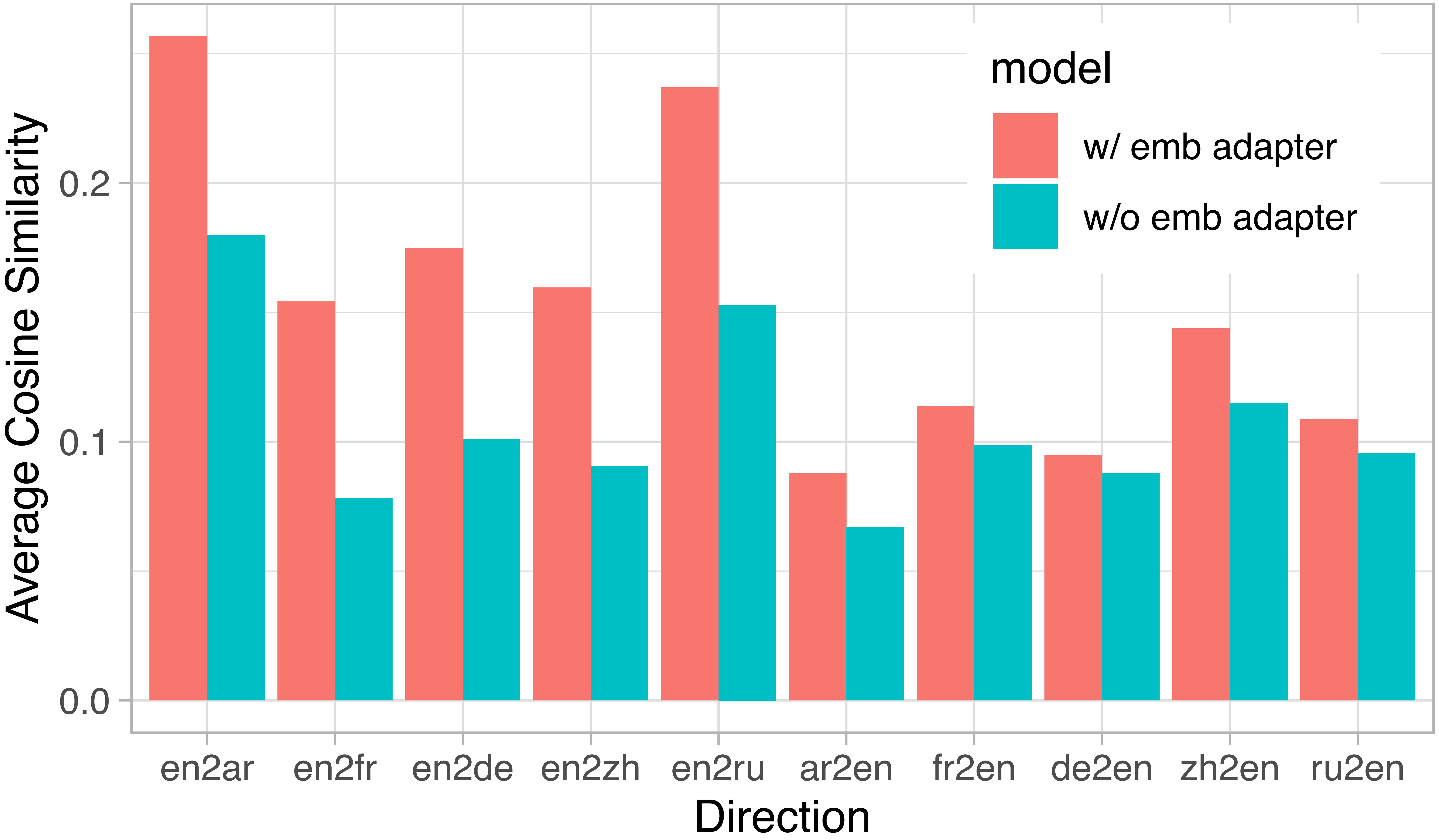


Drop in perf. removing enc/dec adapters



Embedding Adapter is also important!

- Embedding adapter enhance the word embedding similarity between language pairs



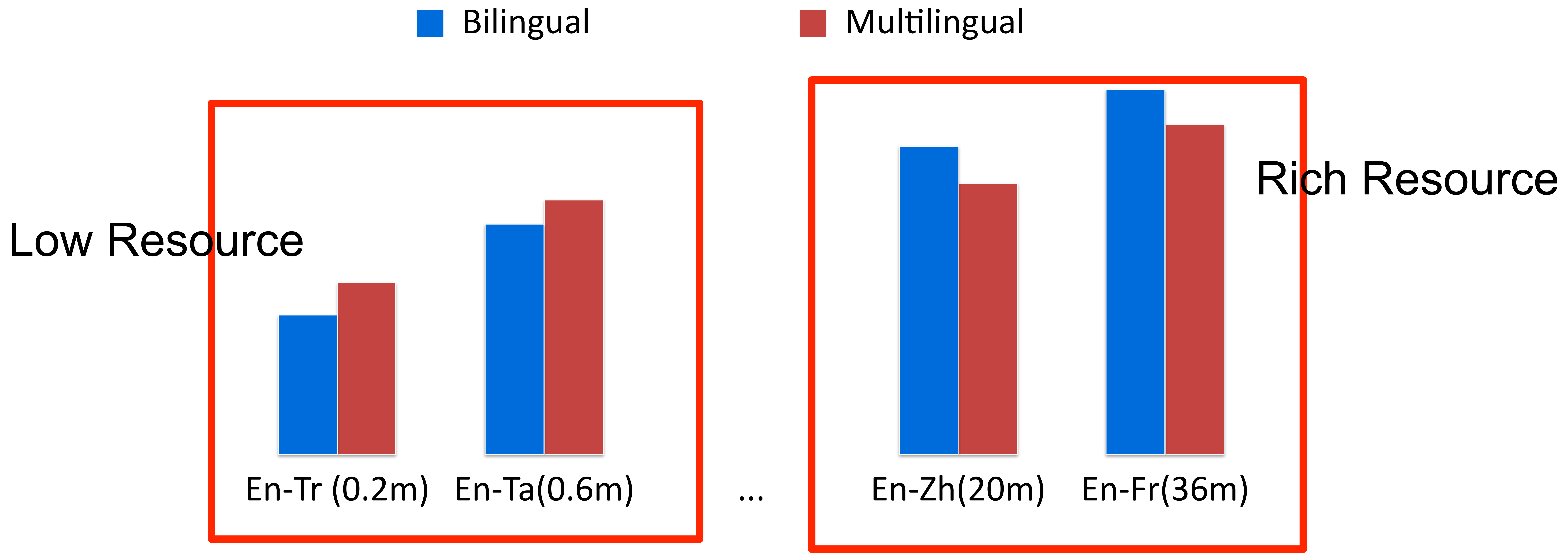
Benefit of MNMT w/ Adapter

- Improve the performance on MNMT, even beat Bilingual NMT
  - Reducing interference among large languages
  - Boost performance on zero-shot setting
- With a fraction of overhead
  - Bilingual Transformer-big: N x 242m
  - mTransformer: 242m
  - mTransformer+Serial Adapter: 242m + N x 12.6m
  - mTransformer+parallel adapter (CIAT): 242m + N x 12.6~27.3m
- Plug-and-play: CIAT only needs to finetune adapter

Exploiting Model Capacity with Language-specific Subnet

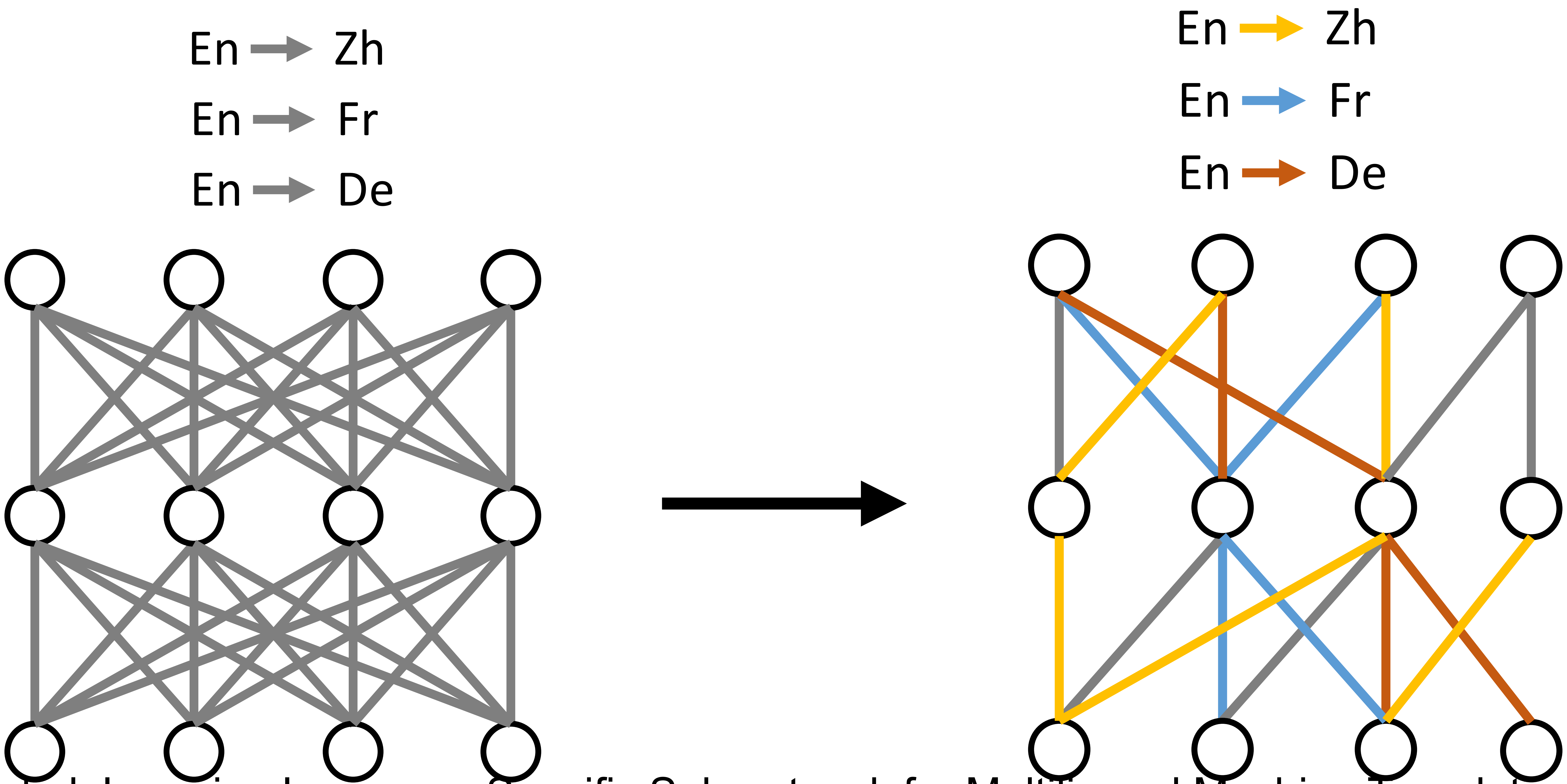
Challenge of Multilingual NMT

- Challenge: Performance degradation for rich-resource
  - caused by Parameter Interference



Language-Specific Sub-network (LaSS)

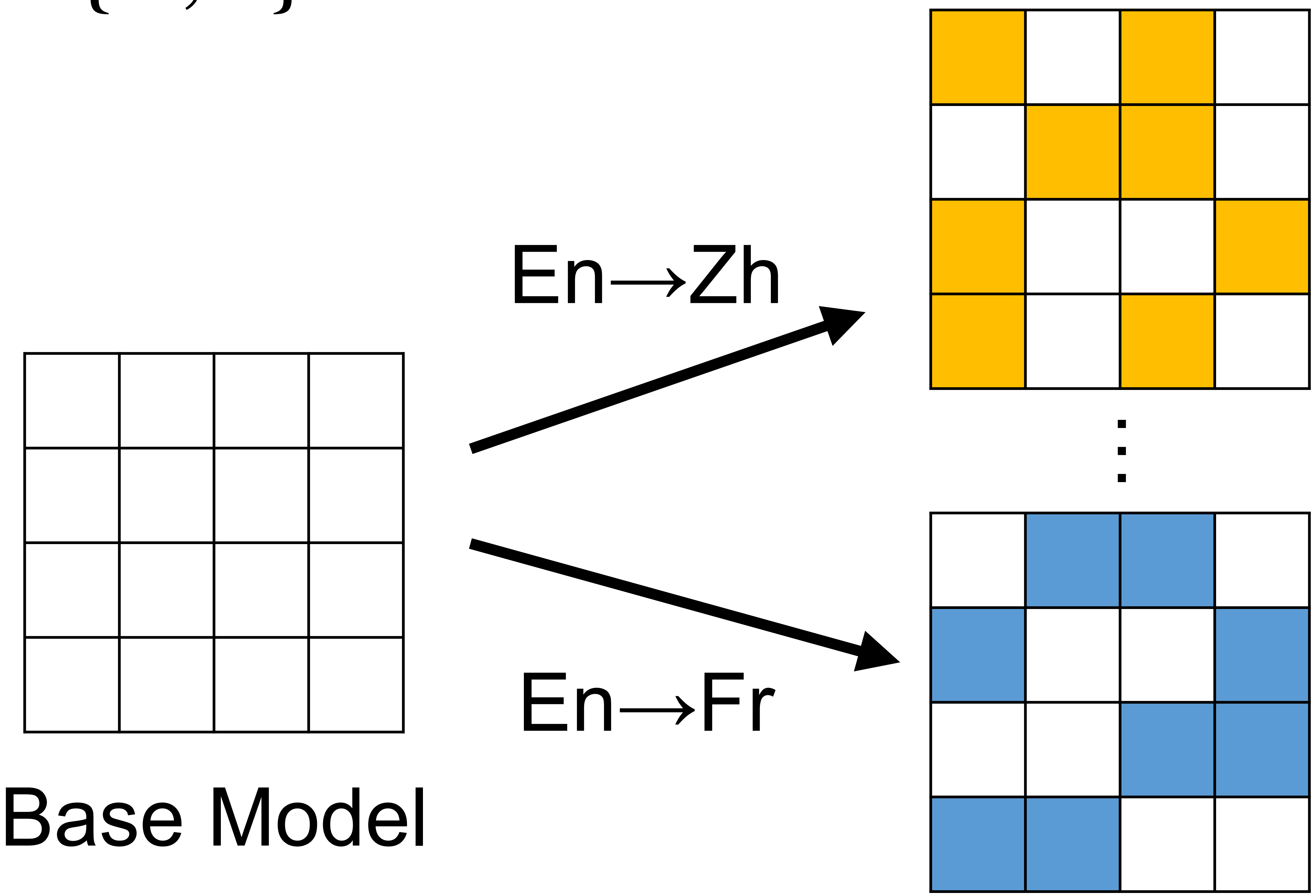
- Each direction has
  - shared parameters with other directions
  - preserves its language-specific parameters



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

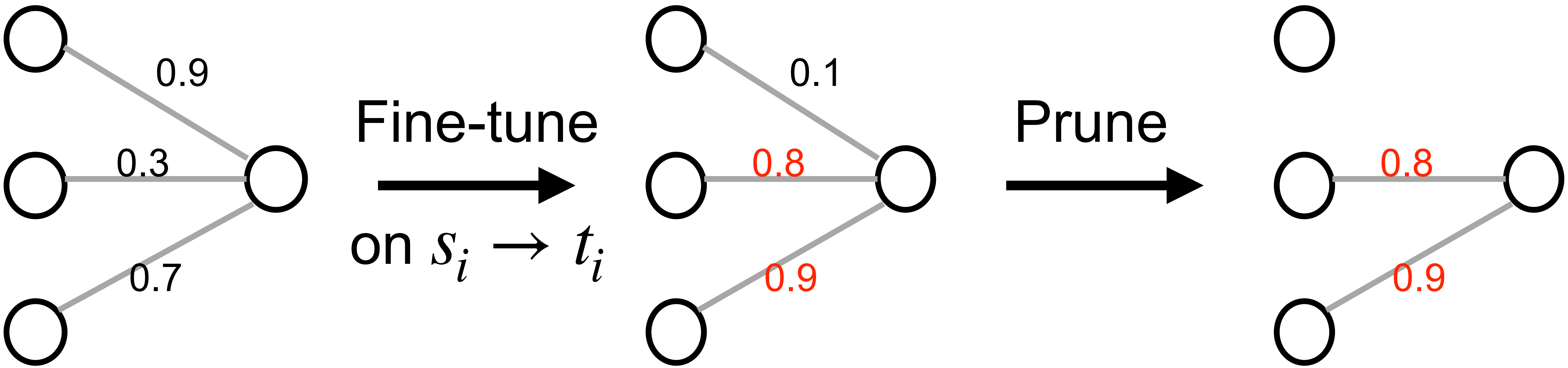
LaSS overall framework

- For each language pair $s_i \rightarrow t_i$, a sub-network is selected from base model θ_0 indicated by a binary mask $\mathbf{M}_{s_i \rightarrow t_i} \in \{0,1\}^{|\theta|}$



How to find language-specific sub-network: Intuition

- Fine-tuning and pruning
  - Fine-tuning on $s_i \to t_i$ **amplifies** important weights and **diminishes** the unimportant weights.



Lin et al. Learning Language Specific Sub-network for Multilingual Machine Translation. 2021

How to find language-specific masks

- Start with a vanilla multilingual model θ_0 jointly trained on $\left\{\mathscr{D}_{s_i \to t_i}\right\}_{i=1}^{N}$
- For each language pair $s_i \to t_i$, fine-tuning θ_0 on $\mathscr{D}_{s_i \to t_i}$, respectively
- Rank the weights in fine-tuned model and prune the lowest α percent to obtain $\mathbf{M}_{s_i \to t_i}$

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

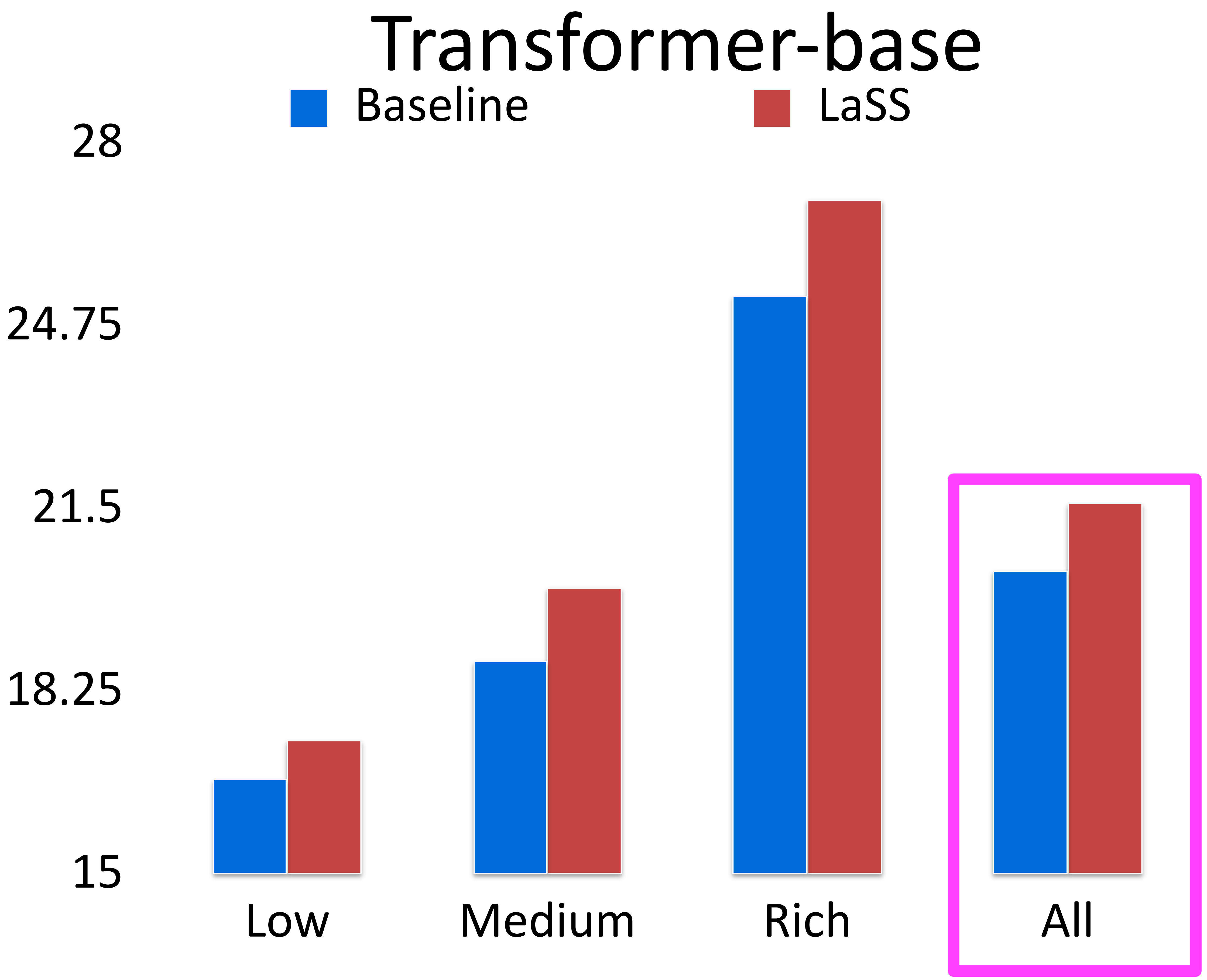
Structure-aware Joint Training

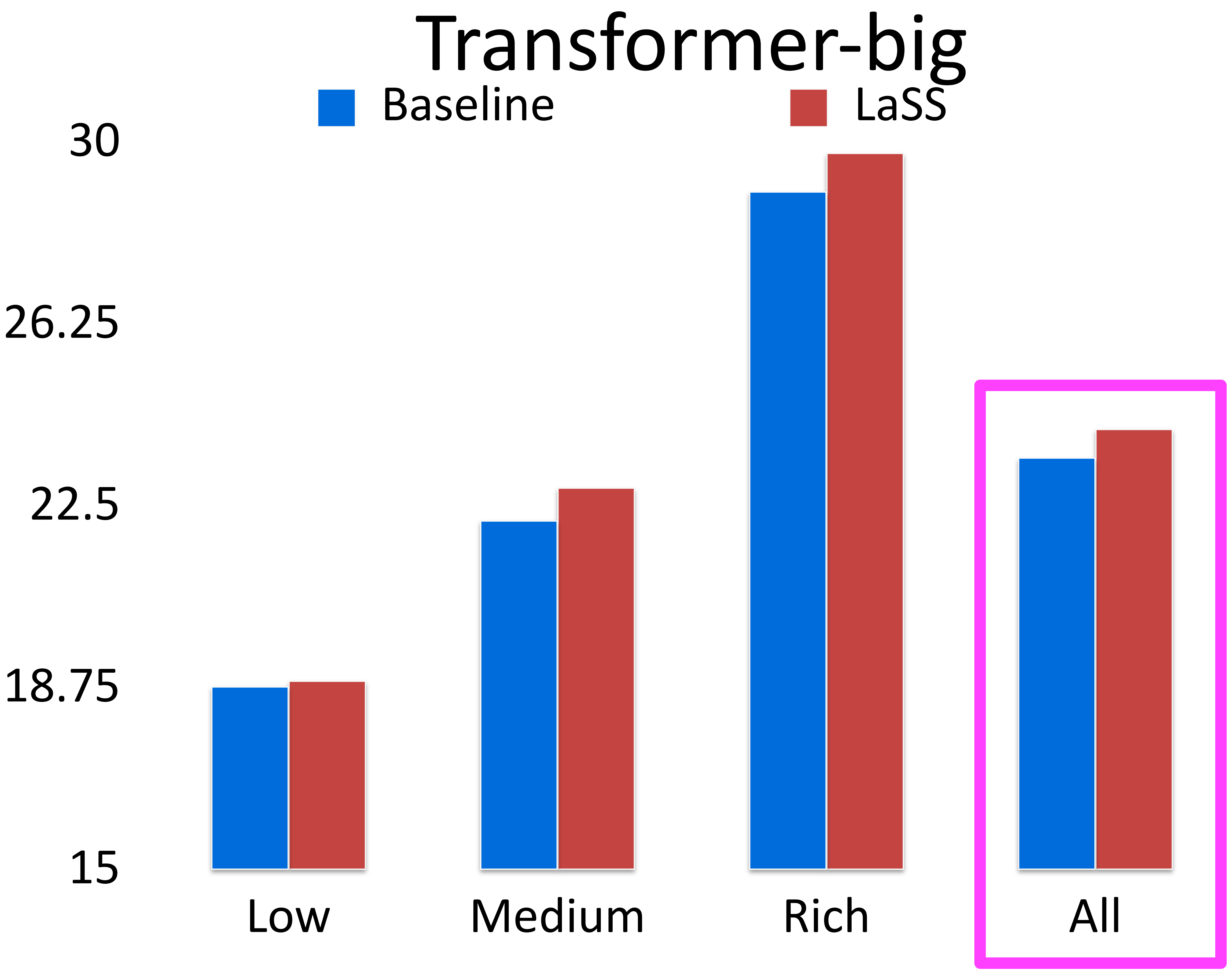
- Further continue to train θ_0 through structure-aware updating after obtaining $\mathbf{M}_{s_i \to t_i}$
  - Create batch $\mathscr{B}_{s_i \to t_i}$ full of samples from $s_i \to t_i$
  - Forward and backward with sub-network $\theta_{s_i \to t_i} = \left\{ \theta_0^j \mid \mathbf{M}_{s_i \to t_i}^j = 1 \right\}$

Lin et al. Learning Language Specific Sub-network for Multilingual Machine Translation. 2021

Efficacy in alleviating Parameter Interference

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

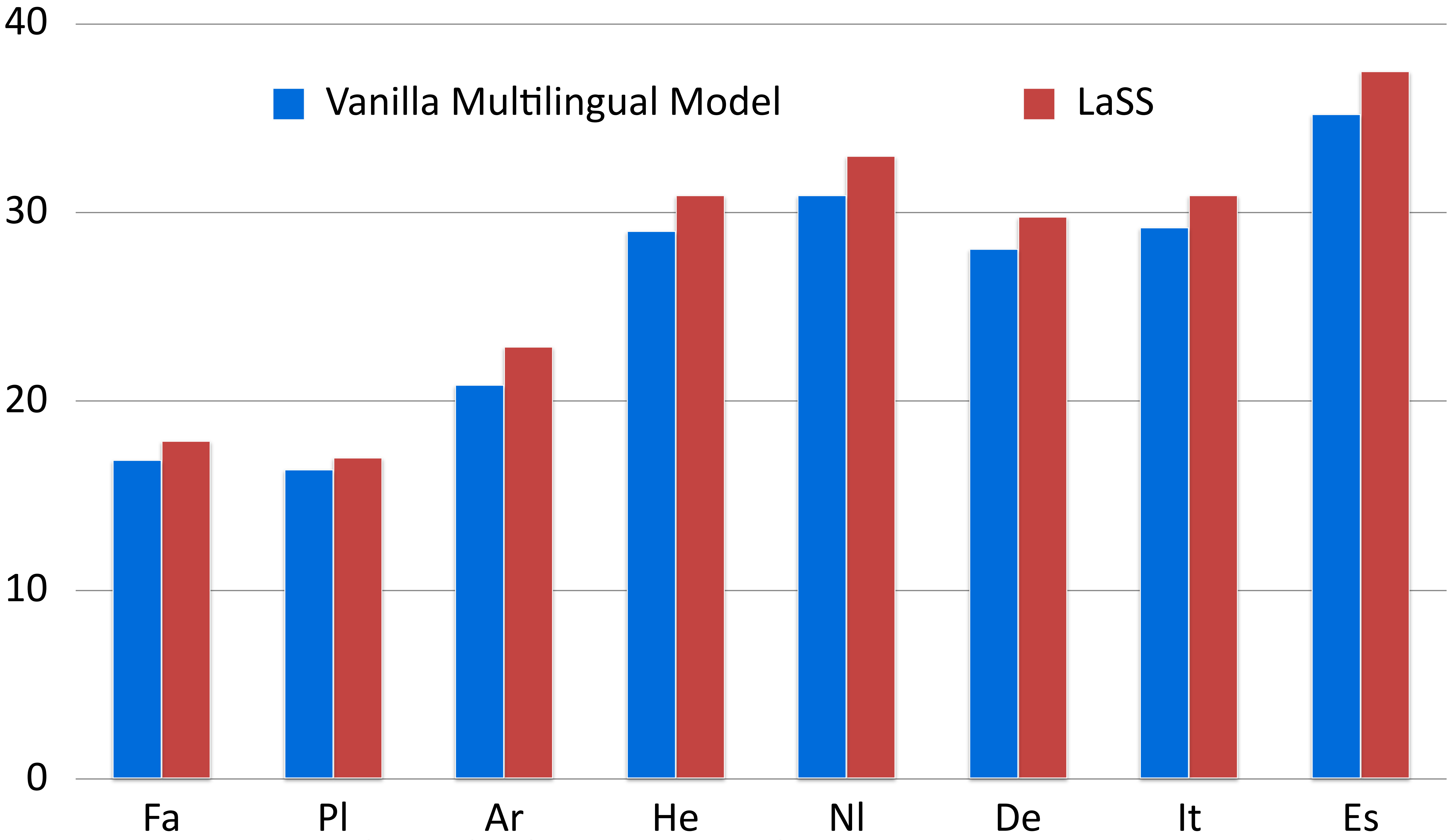




Lin et al. Learning Language Specific Sub-network for Multilingual Machine Translation, 2021

Efficacy in alleviating Parameter Interference

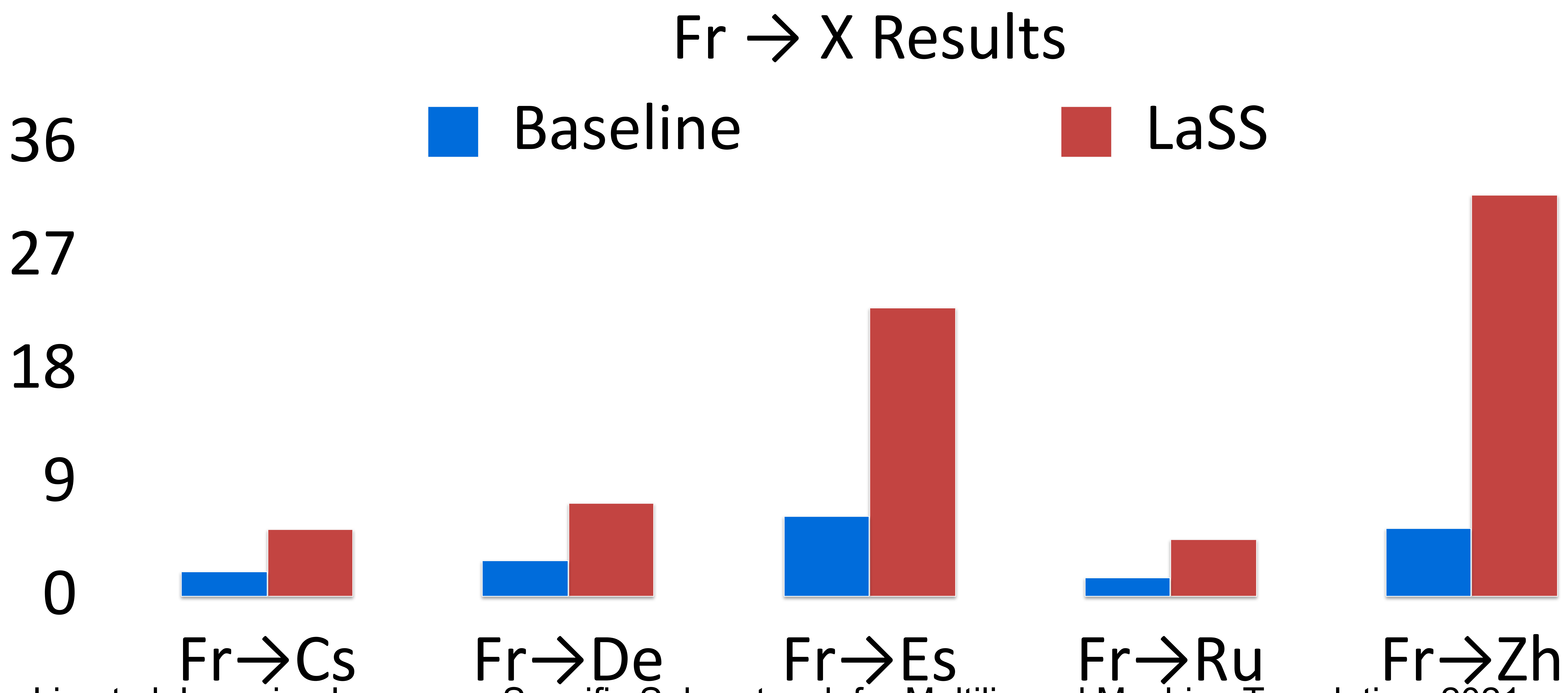
- LaSS obtains consistent performance gains.
  - IWSLT



Lin et al. Learning Language Specific Sub-network for Multilingual Machine Translation. 2021

LaSS obtains large gains in zero-shot translation

- An average of 8.3 BLEU gains on 30 language pairs
- 26.5 BLEU gains for Fr→Zh



Lin et al. Learning Language Specific Sub-network for Multilingual Machine Translation. 2021

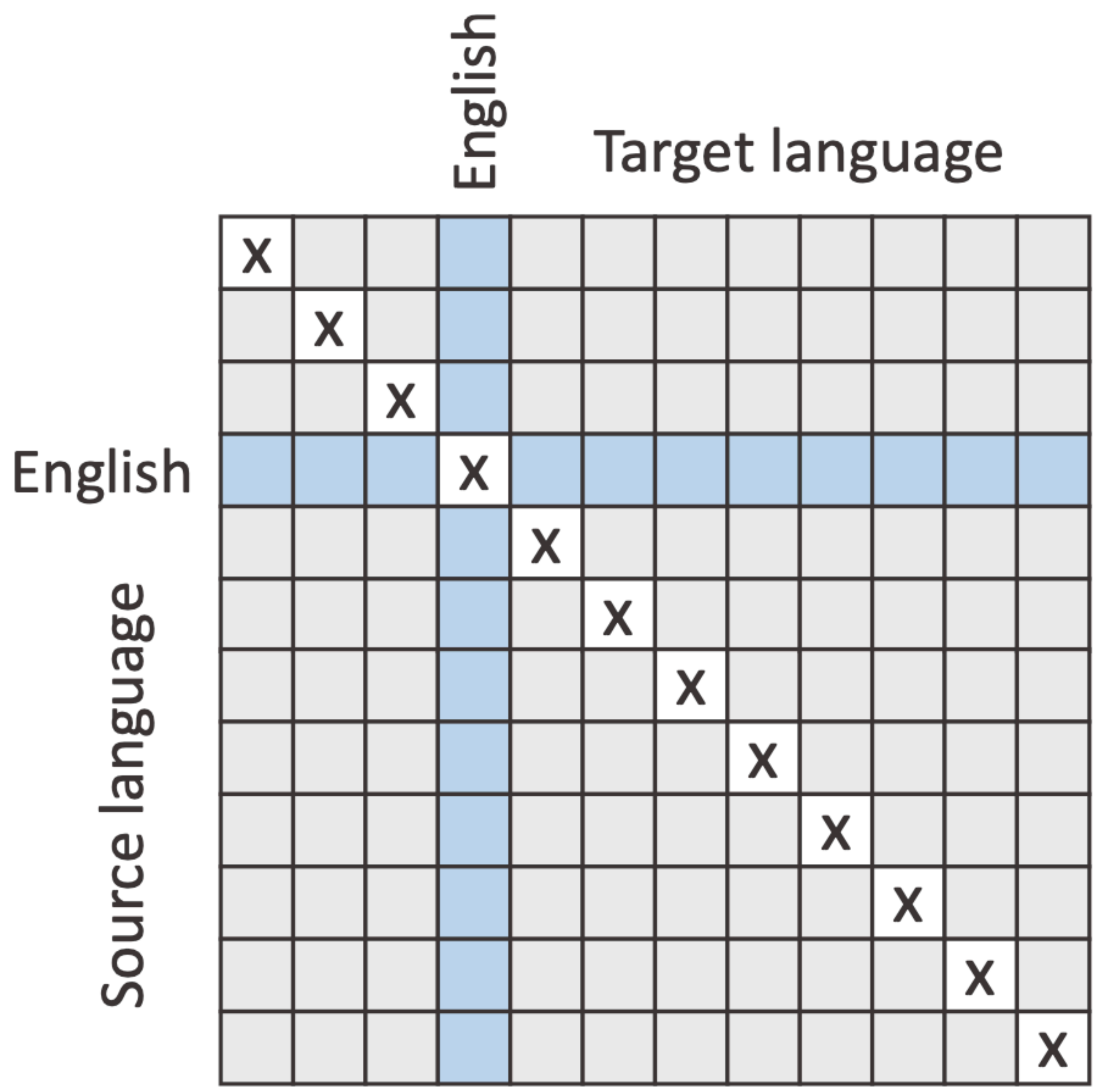
Benefits of Language-specific Subnet

- The same number of parameters, no extra parameter
- Improved performance on both rich-resource and zero-shot translation directions.

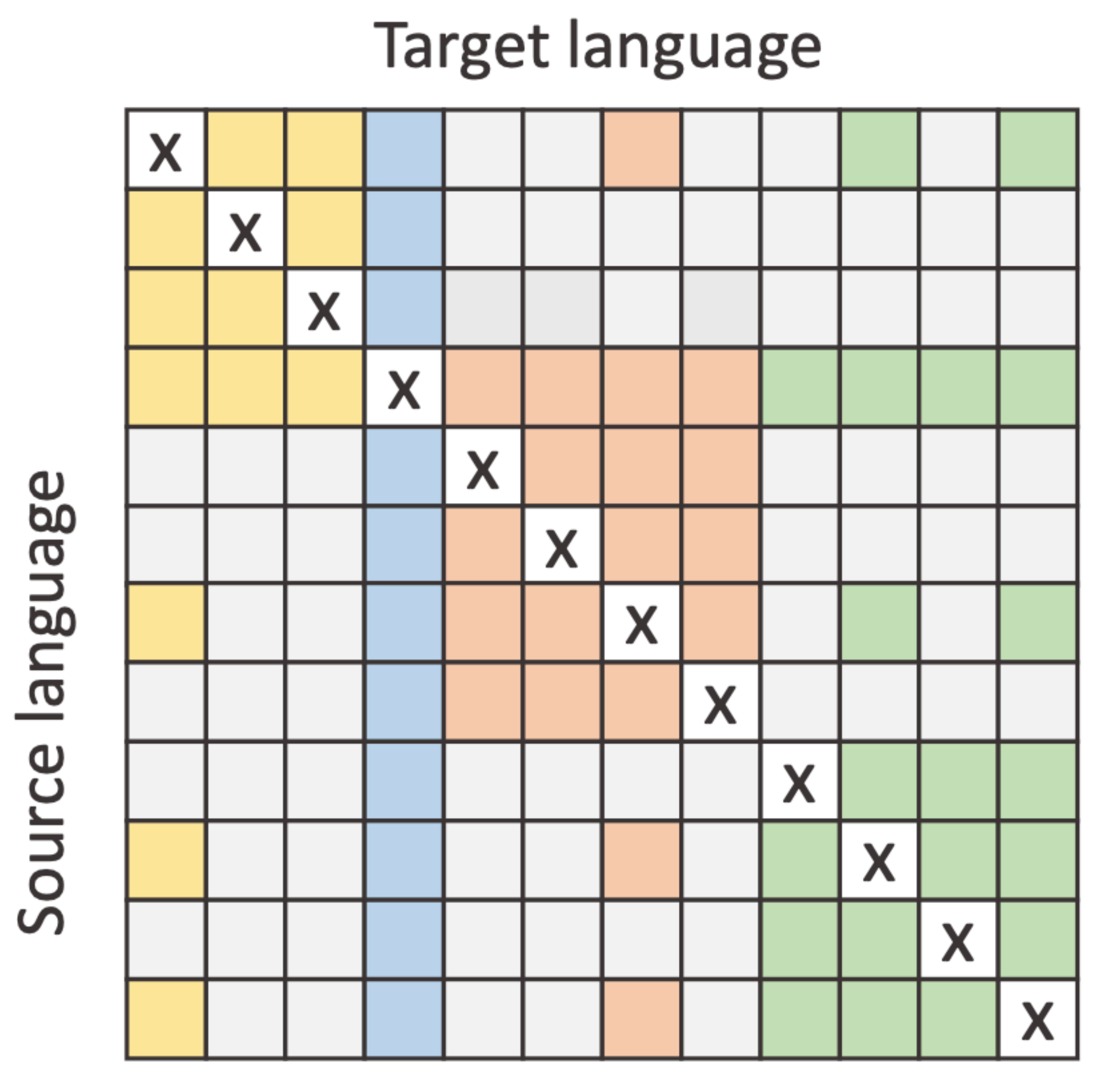
What do we need for larger scale?

Full Many-to-Many MNMT

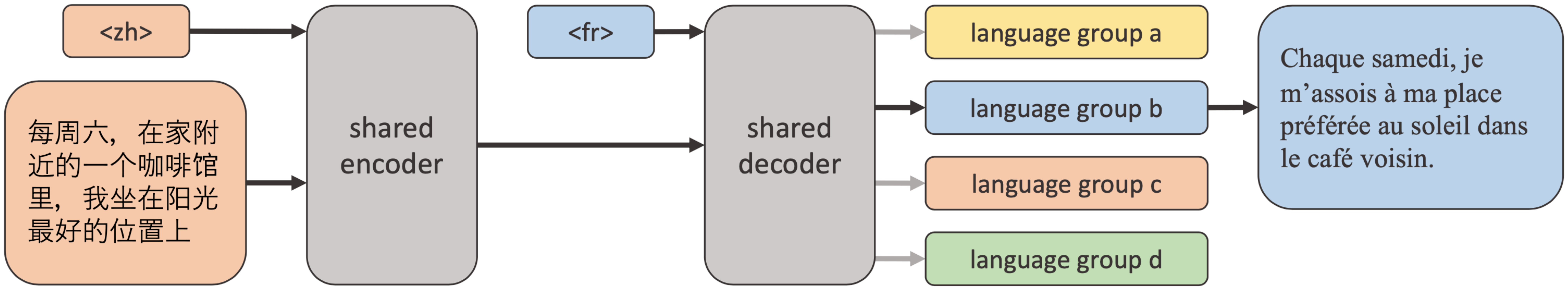
- Previous many-to-many MNMT does not work well on non-English pairs
-



(a) English-Centric Multilingual



(b) M2M-100: Many-to-Many Multilingual Model



(c) Translating from Chinese to French with Dense + Language-Specific Sparse Model

Fan et al. Beyond English-Centric Multilingual Machine Translation. 2021

100 Langauge Benchmark

- WMT — 13 languages
- WAT — Burmese-English
- IWSLT — 4 languages
- FLORES— Sinhala and Nepali <—> English
- TED—The TED Talks data set4 (Ye et al., 2018) contains translations between more than 50 languages; most of the pairs do not include English. The evaluation data is n-way parallel and contains thousands of directions.
- Autshumato— 11-way parallel data set comprising 10 African languages and English from the government domain. Half-half split.
- Tatoeba— 692 test pairs from mixed domains where sentences are contributed and translated by volunteers online. The evaluation pairs we use from Tatoeba cover 85 different languages.

Data mining for parallel corpus

- CCAligned [El-Kishky et al 2020]
  - use LASER encoder to produce sentence embedding
  - for every Eng sentence, use vector search engine (e.g. FAISS) to search candidate aligned sentence by comparing sentence embedding
  - parallel or comparable web-document pairs in 137 languages aligned with English.
- Use language family as bridge to mine
  - non-English pairs
- Total Training Data: 7.5B parallel sentences, corresponding to 2200 directions.

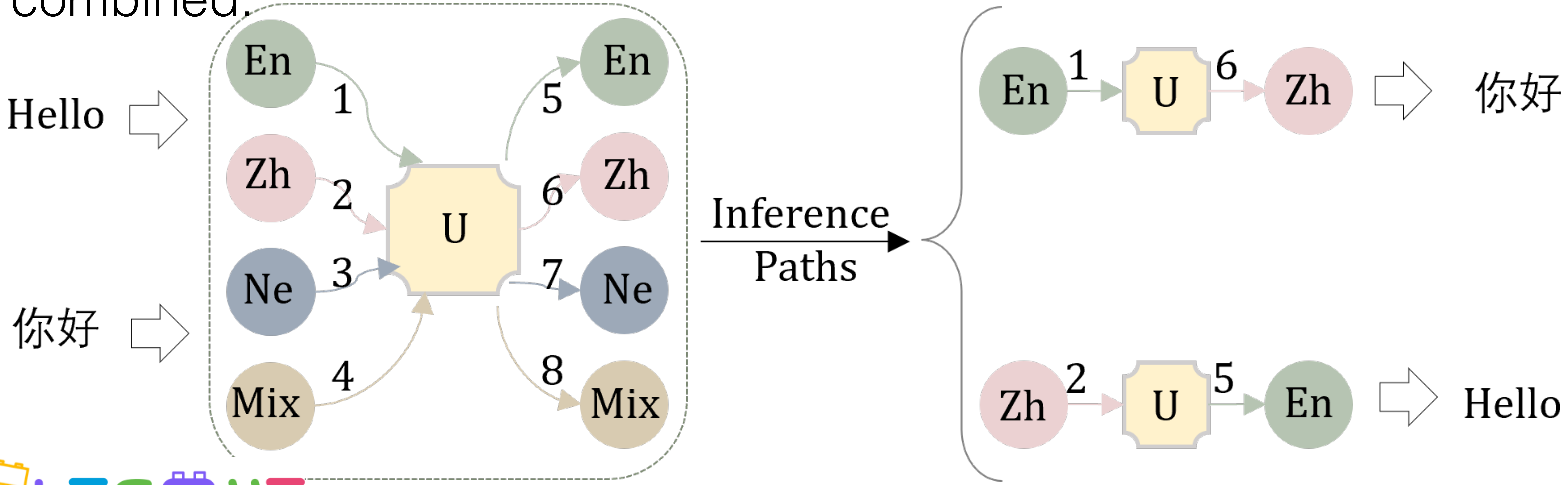
Direction	Test Set	BLEU Published	BLEU M2M-100	BLEU Δ
Without Improvement				
English-Chinese (Li et al., 2019)	WMT'19	38.2	33.2	-5.0
English-Finnish (Talman et al., 2019)	WMT'17	28.6	28.2	-0.4
English-Estonian (Pinnis et al., 2018)	WMT'18	24.4	24.1	-0.3
Chinese-English (Li et al., 2019)	WMT'19	29.1	29.0	-0.1
With Improvement				
English-French (Edunov et al., 2018)	WMT'14	43.8	43.8	0
English-Latvian (Pinnis et al., 2017)	WMT'17	20.0	20.5	+0.5
German-English (Ng et al., 2019)	WMT'19	39.2	40.1	+0.9
Lithuanian-English (Pinnis et al., 2019)	WMT'19	31.7	32.9	+1.2
English-Russian (Ng et al., 2019)	WMT'19	31.9	33.3	+1.4
English-Lithuanian (Pinnis et al., 2019)	WMT'19	19.1	20.7	+1.6
Finnish-English (Talman et al., 2019)	WMT'17	32.7	34.3	+1.6
Estonian-English (Pinnis et al., 2018)	WMT'18	30.9	33.4	+2.5
Latvian-English (Pinnis et al., 2017)	WMT'17	21.9	24.5	+2.6
Russian-English (Ng et al., 2019)	WMT'19	37.2	40.5	+3.3
French-English (Edunov et al., 2018)	WMT'14	36.8	40.4	+3.6
English-German (Ng et al., 2019)	WMT'19	38.1	43.2	+5.1
English-Turkish (Sennrich et al., 2017)	WMT'17	16.2	23.7	+7.5
Turkish-English (Sennrich et al., 2017)	WMT'17	20.6	28.2	+7.6
	Average	30.0	31.9	**+1.9**

LegoMT

Lego-MT: Detachable Architecture

Each branch contains a complete encoder-decoder for a language/ language group.

7 branches for central languages, and 1 branch for all languages combined.



Lego-MT: Learning Detachable Models for Massively Multilingual Machine Translation [Yuan, Lu, Zhu, Kong, **Lei Li**, Xu, ACL 2023]

Data Flow in Lego-MT

Lego-MT: Learning Detachable Models for Massively Multilingual Machine Translation [Yuan, Lu, Zhu, Kong, **Lei Li**, Xu, ACL 2023]

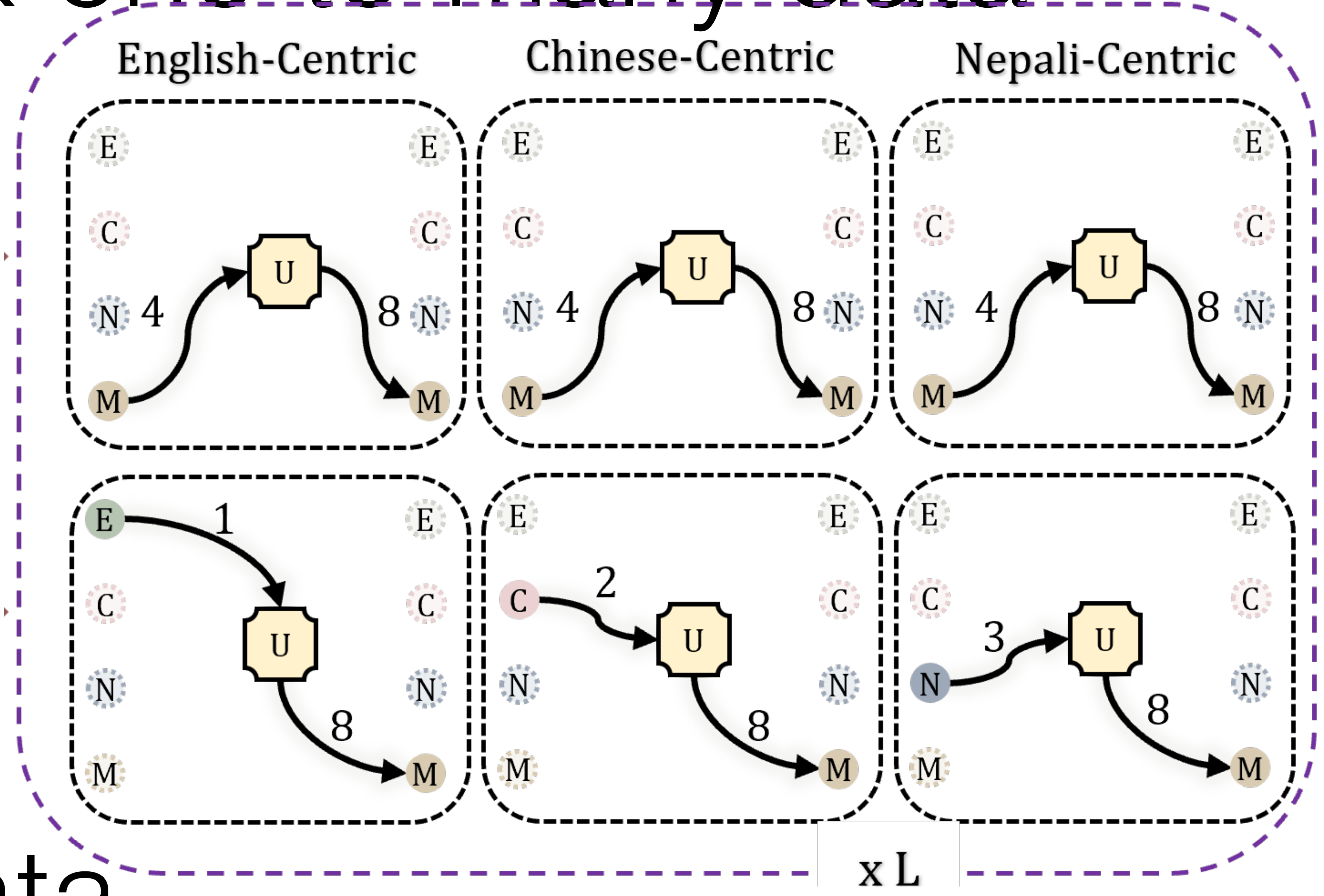
Lego-MT Two-stage Training

- 1st stage: training on many-to-many & one-to-many data

$$\min L_{mix} + L_{enc}$$

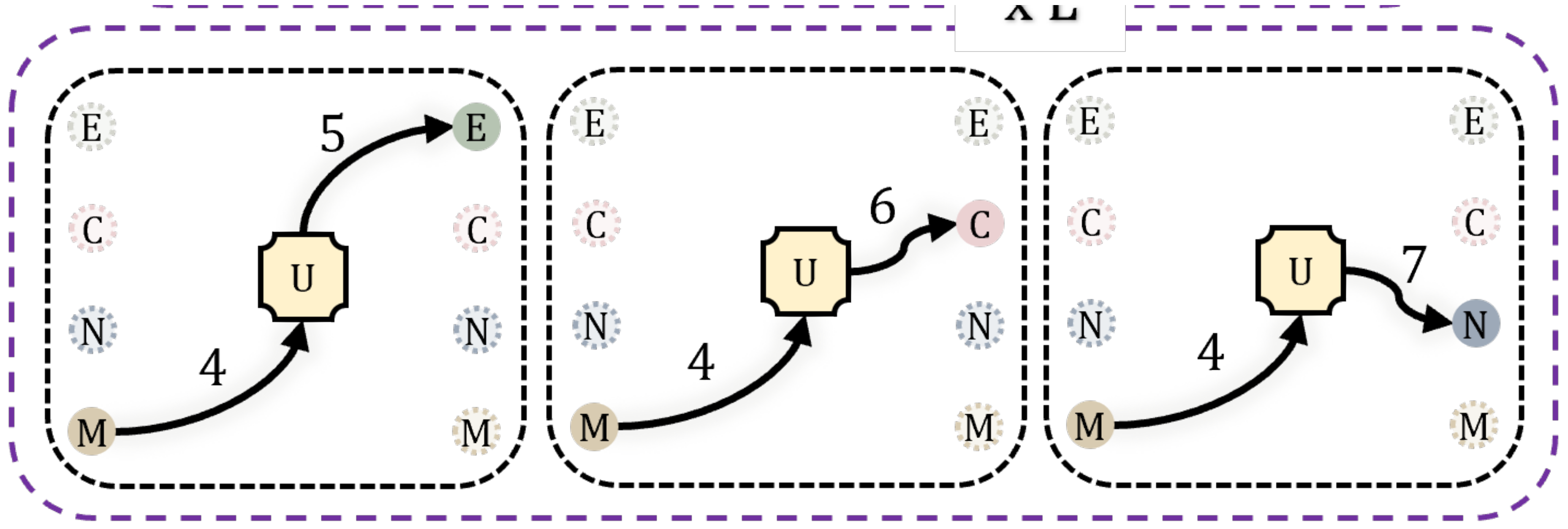
$$L_{mix} = -\sum_{x,y \sim D_{\text{multi}}} \log P_{\theta_{mix}}(y|x)$$

$$L_{enc} = -\sum_{x,y \sim D_{\text{lg}\to\cdot}} \log P_{\theta_{enc}}(y|x)$$



- 2nd stage: training on many-to-one data

$$L_{dec} = -\sum_{x,y \sim D_{\cdot\to\text{lg}}} \log P_{\theta_{dec}}(y|x)$$

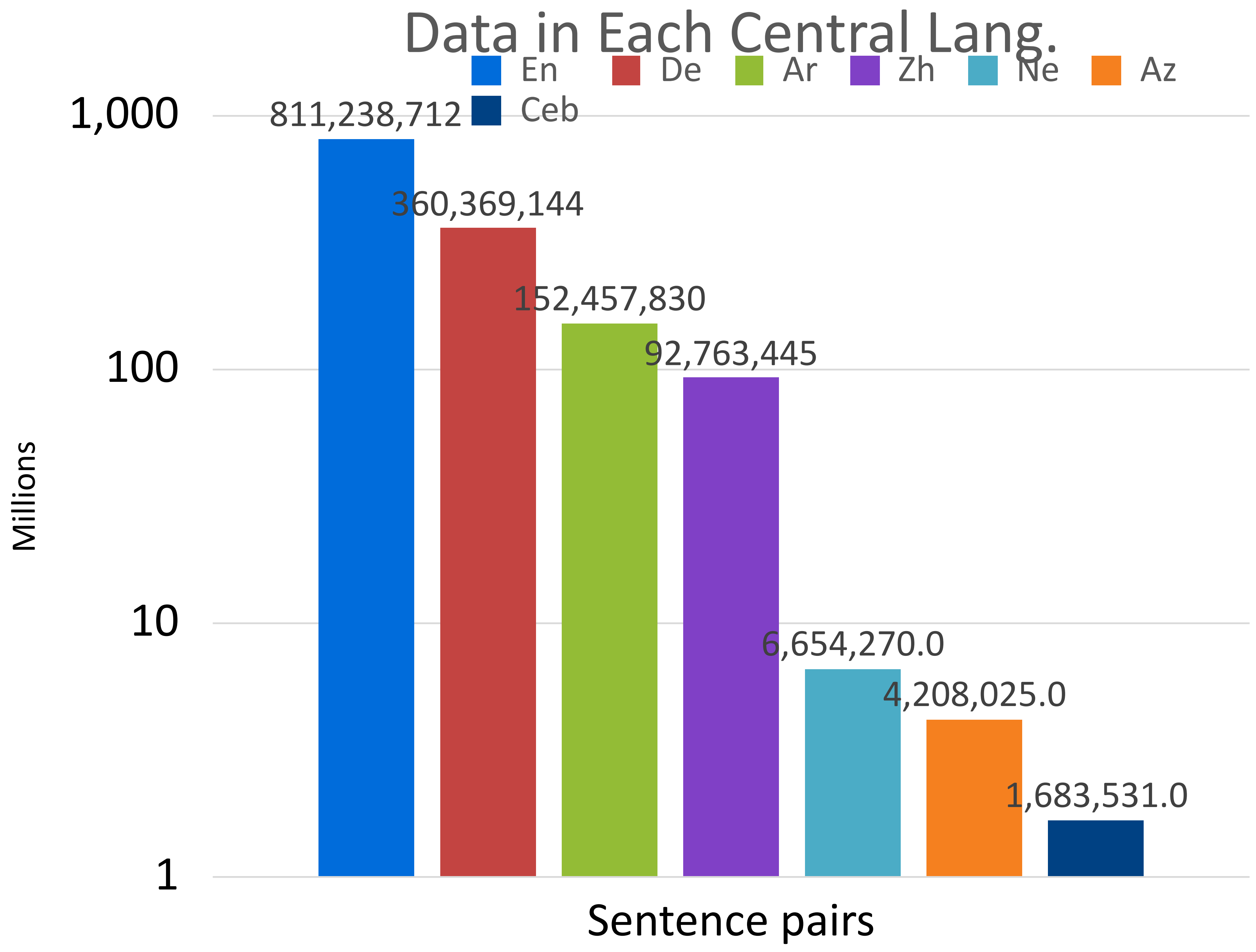


Fix the encoder of mix-flow branch

Lego-MT: Learning Detachable Models for Massively Multilingual Machine Translation [Yuan, Lu, Zhu, Kong, **Lei Li**, Xu, ACL 2023]

Multi-centric Data for 433 Languages

- Training Data
  - 1.3B sentence pairs collected from OPUS
  - 433 languages including 7 central languages
- Testing:
  - Flores-101 Devtest, human written translation pairs covering 101 languages.
  - 7×85 translation directions
- Evaluation Metric:
  - spBLEU, same in Flores-101



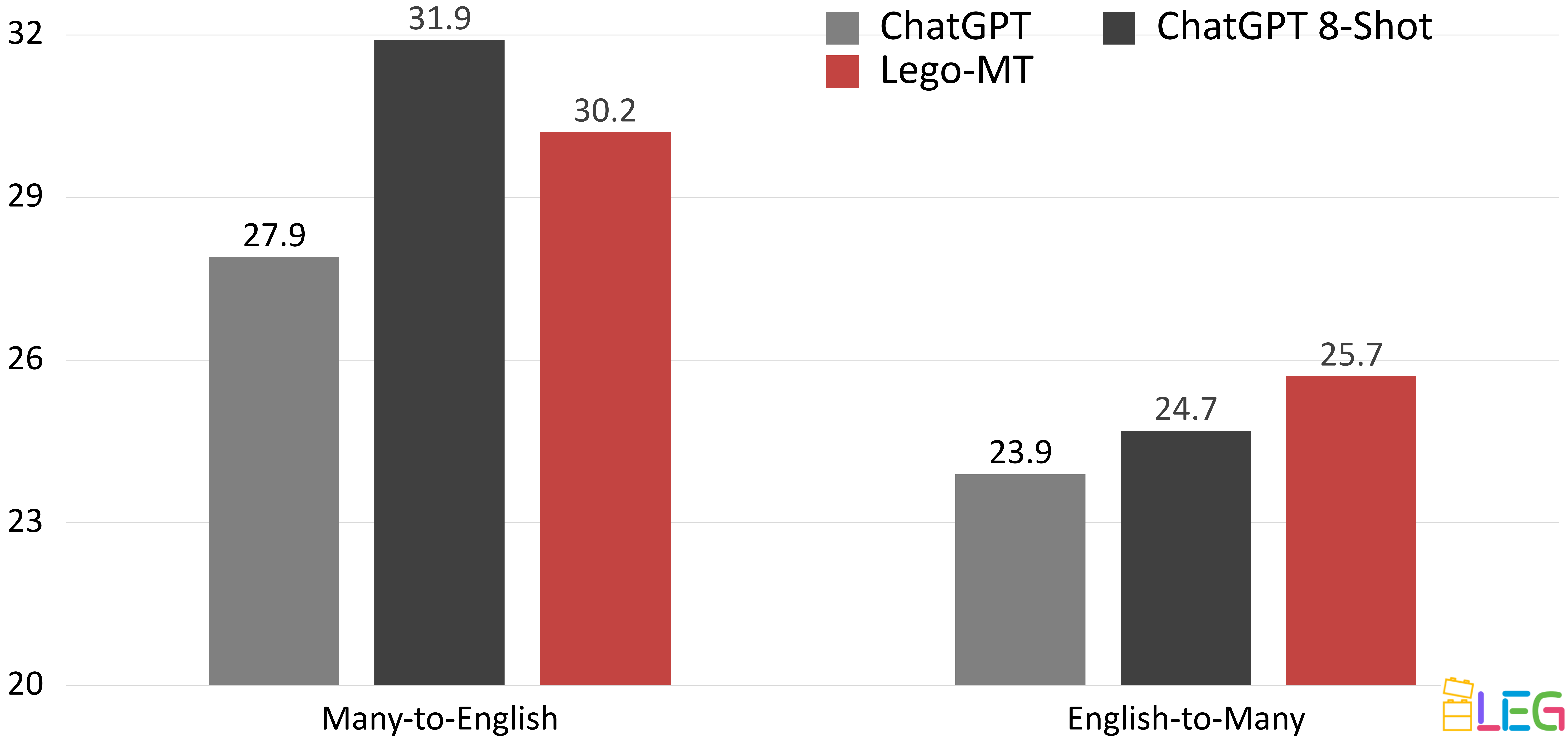
Lego-MT Model Configuration

- Model Parameters
  - Each Flow: 0.6B parameters
  - Total Training Parameters:
    - 9.6B = 1.2B (Mix-Flow) + 0.6 * 7 (Enc-Flow) + 0.6 * 7 (Dec-Flow)
  - Inference Parameter：
    - 1.2B (Each branch can be independently loaded during inference)
    - We use Mix-flow for multilingual evaluation
- Training Setting
  - Max token 8000
  - The training of all centric languages is conducted in random order
  - Training duration:15 days on 32 A100 GPUs.

LEGO-MT 1.2B outperforms M2M-100 12B!

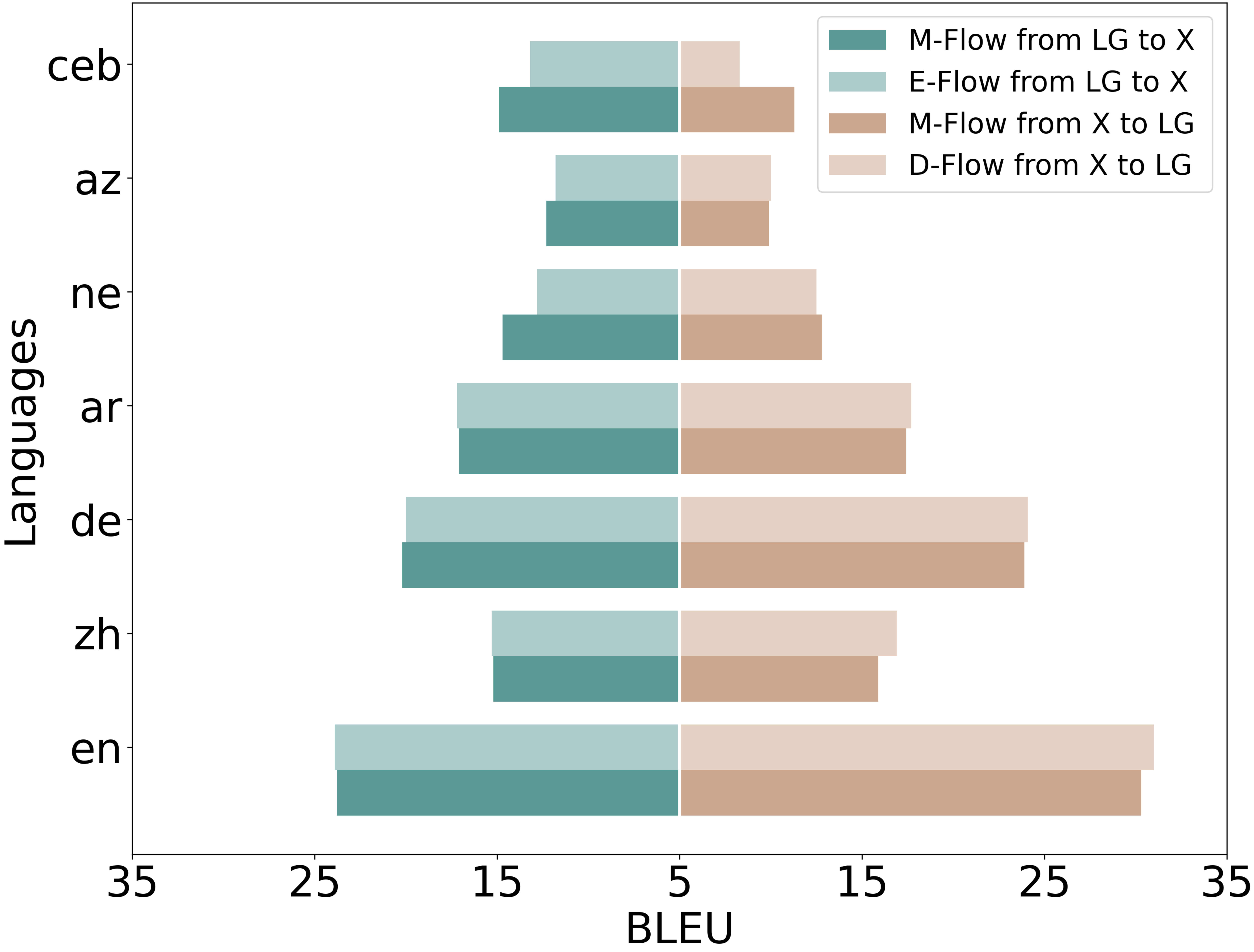
Lego-MT: Learning Detachable Models for Massively Multilingual Machine Translation [Yuan, Lu, Zhu, Kong, **Lei Li**, Xu, ACL 2023]

Lego-MT surpasses plain ChatGPT



Lego-MT: Learning Detachable Models for Massively Multilingual Machine Translation [Yuan, Lu, Zhu, Kong, **Lei Li**, Xu, ACL 2023]

w?



Lego-MT: Learning Detachable Models for Massively Multilingual Machine Translation [Yuan, Lu, Zhu, Kong, **Lei Li**, Xu, ACL 2023]

Language Presentation

Reading

- Yuan et al. LegoMT: Learning Detachable Models for Massively Multilingual Machine Translation, 2023
- Johnson et al. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. 2017
- Aharoni et al. Massively Multilingual Neural Machine Translation. 2019
- Arivazhagan et al. Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges. 2019
- Bapna & Firat, Simple, Scalable Adaptation for Neural Machine Translation, 2019
- Zhu et al. Counter-Interference Adapter for Multilingual Machine Translation. 2021
- Lin et al. Learning Language Specific Sub-network for Multilingual Machine Translation. 2021

Reading

- Sennrich et al. Improving Neural Machine Translation Models with Monolingual Data. ACL 2016.
- Cheng et al. Semi-Supervised Learning for Neural Machine Translation. ACL 2016.
- Artetxe et al. Unsupervised Neural Machine Translation. 2018
- Lample et al. Unsupervised Machine Translation Using Monolingual Corpora Only. 2018
- He et al. Dual Learning for Machine Translation. 2016.
- Gulcehre et al. On Using Monolingual Corpora in Neural Machine Translation. 2015
- Edunov et al. Understanding Back-translation at Scale. 2018.