

291K

Deep Learning for Machine Translation Multilingual Neural Machine Translation

Lei Li

UCSB

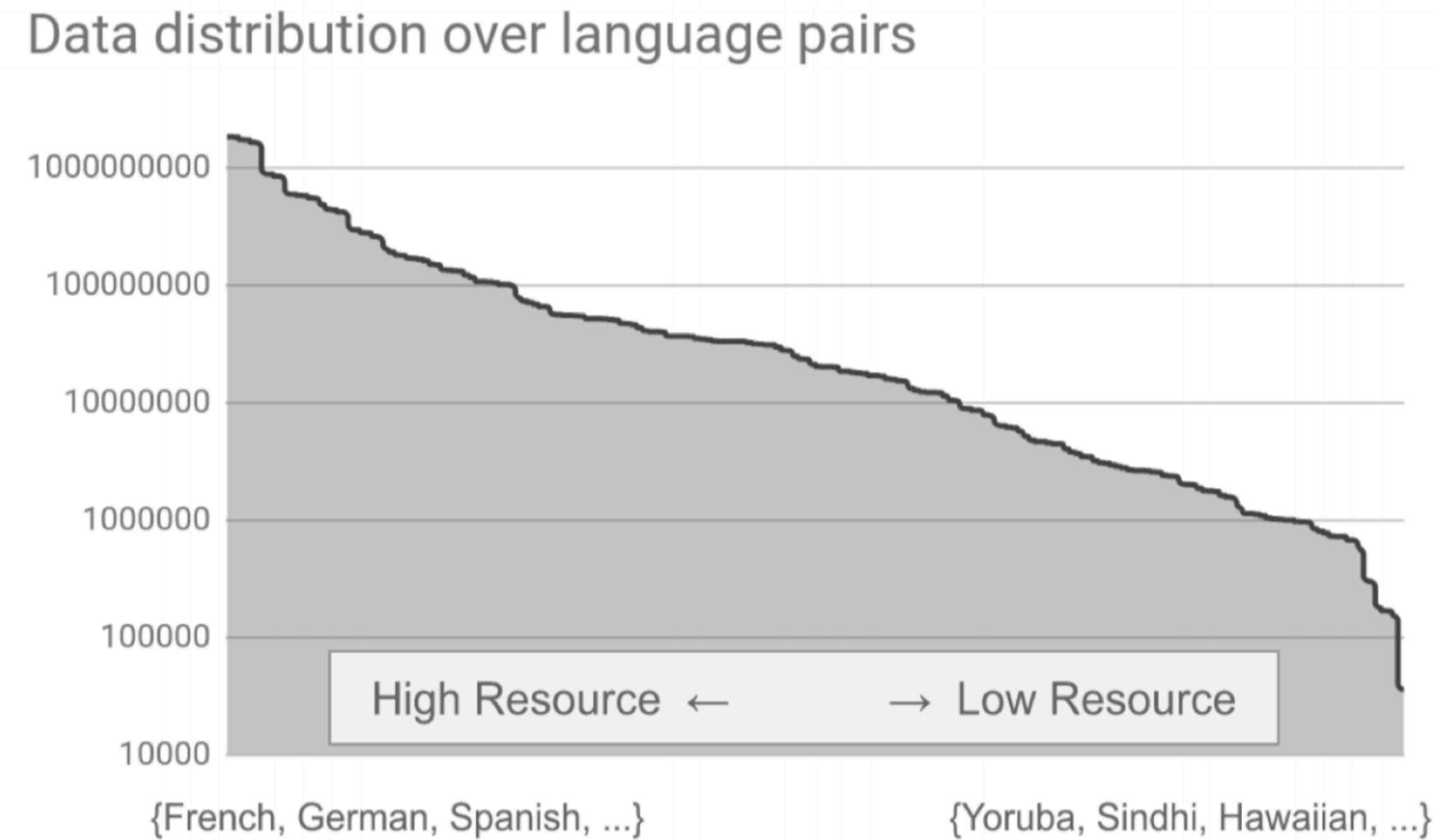
11/3/2021

Outline

- Multilingual NMT
- Architecture for MNMT
 - Multilingual Vocabulary
- Reducing multilingual interference
 - adapters for MNMT

Corpus Size in Languages

- NMT requires large amount of parallel bilingual data
- Parallel data, However, very expensive/non-trivial to obtain
 - Low resource language pairs (e.g., English-to-Tamil)
 - Low resource domains (e.g., social network)
 - but additional monolingual data on source side and/or target side. can we do reasonably well?
- Rich resource setting: in addition to parallel data (~10s millions), much larger monolingual data, can we further improve?



[Arivazhagan et al., 2019]

Multilingual Neural Machine Translation

- Bilingual NMT: one model for each translation direction
- Multilingual NMT: Develop one model to translate between all language pairs.
- Why?
 - 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

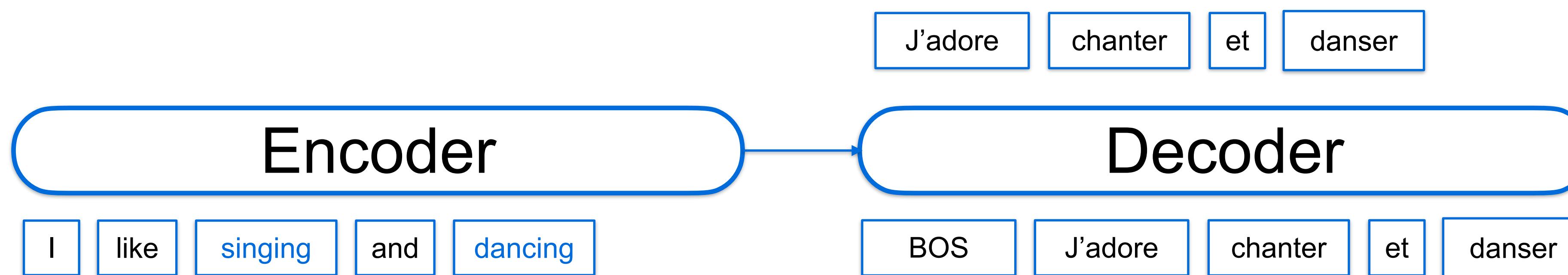
- Many-to-one:
 - Many source language to a target language
 - Usually the target is English
- One-to-Many:
 - One source language to many target languages
 - Usually the source is English
- Many-to-many:
 - Many source language to many target languages
 - Should include non-English pairs (often low-resource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

MNMT at Testing Time

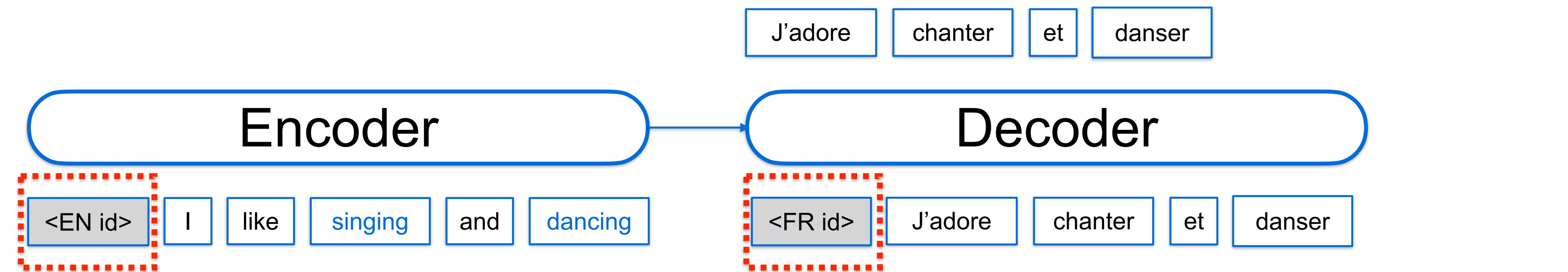
- Regular:
 - Testing language appeared during training (but not the sentence)
- Exotic (Unseen) pair
 - Both the testing source language and target language appeared in the training, but the source-target pair never appeared in the training
 - Also known as zero-shot MNMT
- Exotic (Unseen) source
 - 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.

A single model for Multilingual NMT

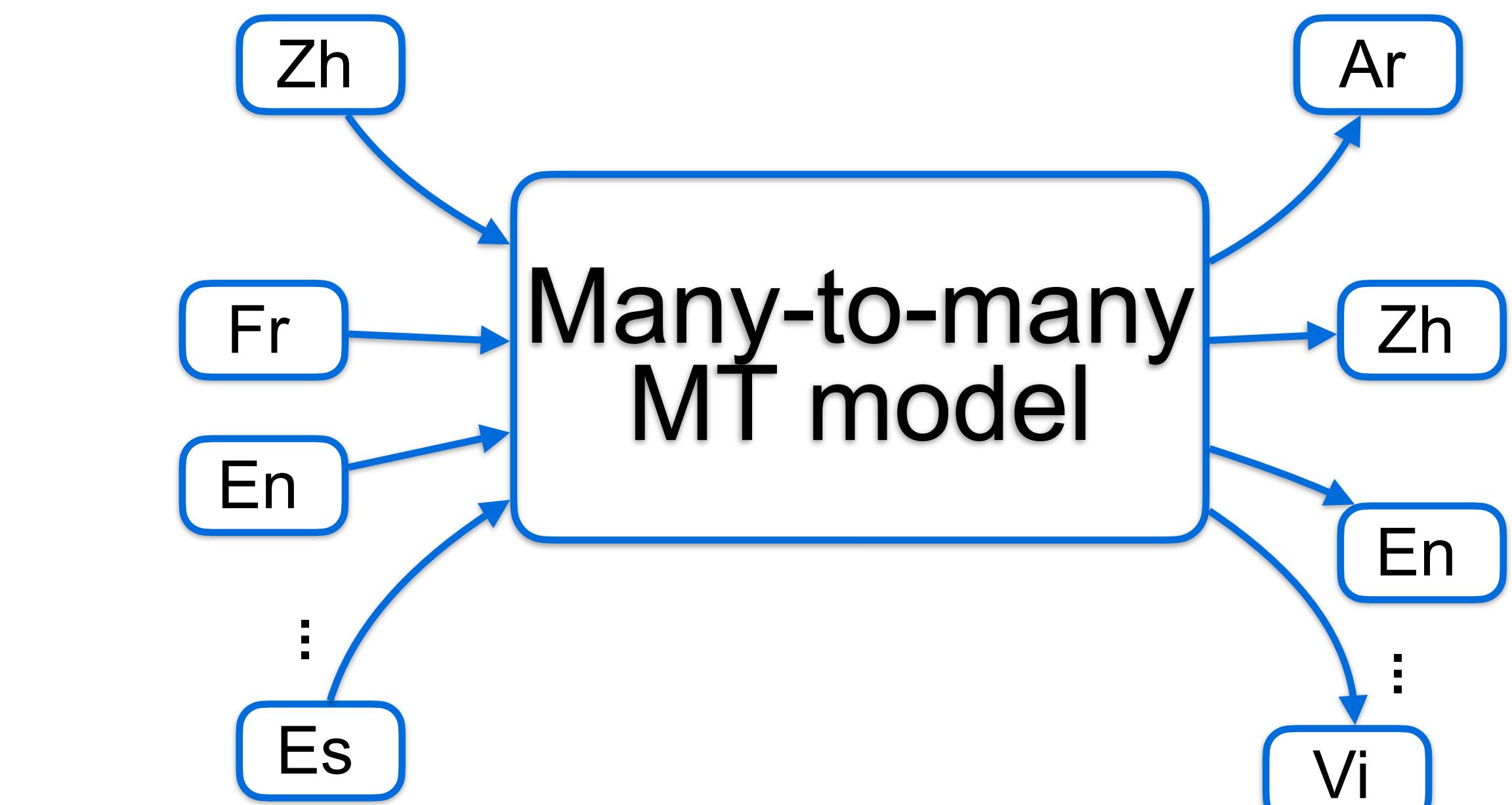
- Language-specific encoding (@en@car, @de@automobile)
- But hard to learn a joint embedding.
- Challenge:
 - large vocabulary (twice many)
 - how does the model know it is to translate into German or French?



Multilingual Machine Translation - Language Tag

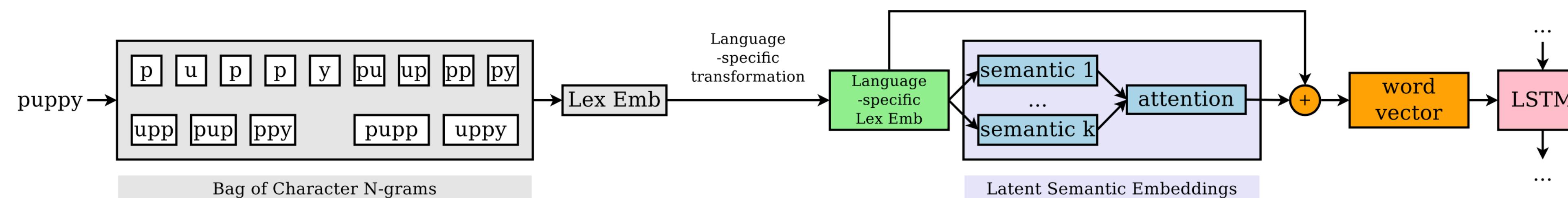


- One model can translate between many languages.
- Language Tag is used to indicate the source and target language.
- Vocabulary is built jointly



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 System

- LSTM-s2s:
 - 8 layer LSTM encoder, 1st layer bidirectional
 - 8 layer LSTM decoder with attention
- Combine De-En and Fr-En to train a joint NMT
- One model to translate two directions

Table 1: Many to One: BLEU scores on for single language pair and multilingual models. *: no oversampling

Model	Single	Multi	Diff
WMT De→En	30.43	30.59	+0.16
WMT Fr→En	35.50	35.73	+0.23
WMT De→En*	30.43	30.54	+0.11
WMT Fr→En*	35.50	36.77	+1.27
Prod Ja→En	23.41	23.87	+0.46
Prod Ko→En	25.42	25.47	+0.05
Prod Es→En	38.00	38.73	+0.73
Prod Pt→En	44.40	45.19	+0.79

Google's MNMT System

- One-to-many is more difficult than many-to-one MNMT
-

Table 2: One to Many: BLEU scores for single language pair and multilingual models. *: no oversampling

Model	Single	Multi	Diff
WMT En→De	24.67	24.97	+0.30
WMT En→Fr	38.95	36.84	-2.11
WMT En→De*	24.67	22.61	-2.06
WMT En→Fr*	38.95	38.16	-0.79
Prod En→Ja	23.66	23.73	+0.07
Prod En→Ko	19.75	19.58	-0.17
Prod En→Es	34.50	35.40	+0.90
Prod En→Pt	38.40	38.63	+0.23

Google's MNMT

- Combining multiple source languages and multiple target languages together will degrade the performance a bit, but still surprising to see one model work as well for many-to-many English-centric pairs. ß

English-centric Many-to-Many

Model	Single	Multi	Diff
WMT En→De	24.67	24.49	-0.18
WMT En→Fr	38.95	36.23	-2.72
WMT De→En	30.43	29.84	-0.59
WMT Fr→En	35.50	34.89	-0.61
WMT En→De*	24.67	21.92	-2.75
WMT En→Fr*	38.95	37.45	-1.50
WMT De→En*	30.43	29.22	-1.21
WMT Fr→En*	35.50	35.93	+0.43
Prod En→Ja	23.66	23.12	-0.54
Prod En→Ko	19.75	19.73	-0.02
Prod Ja→En	23.41	22.86	-0.55
Prod Ko→En	25.42	24.76	-0.66
Prod En→Es	34.50	34.69	+0.19
Prod En→Pt	38.40	37.25	-1.15
Prod Es→En	38.00	37.65	-0.35
Prod Pt→En	44.40	44.02	-0.38

Google's MNMT

- Training 12 language pairs together

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 vs single	-	-1.72	-1.43	-0.95	-0.76
	-	-5.6%	-4.7%	-3.1%	-2.5%

Google's MNMT Zero-shot

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

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



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

Google's MNMT Zero-shot

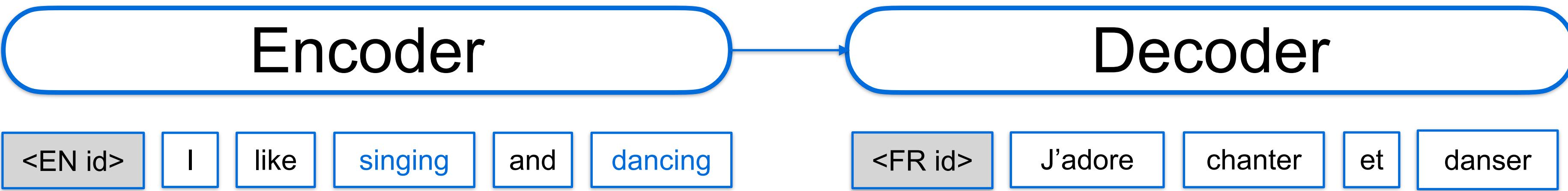
- MNMT is worse than pivot on zero-shot directions

zero-shot

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
NMT Es→Ja implicitly bridged	9.14

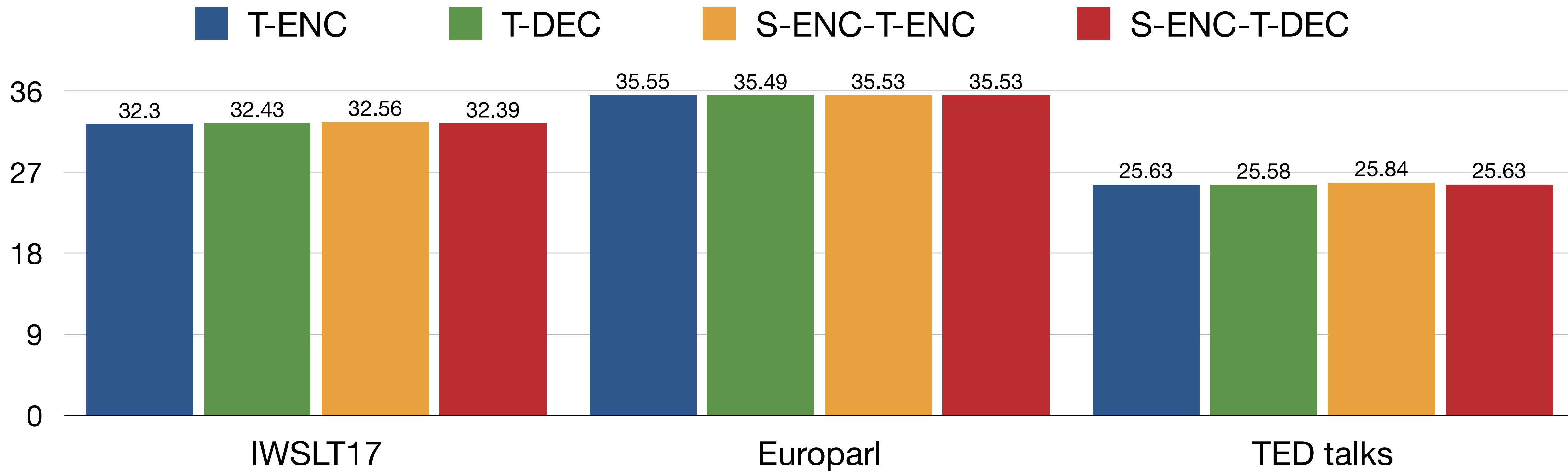
Source Language Tag or target Language Tag?



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

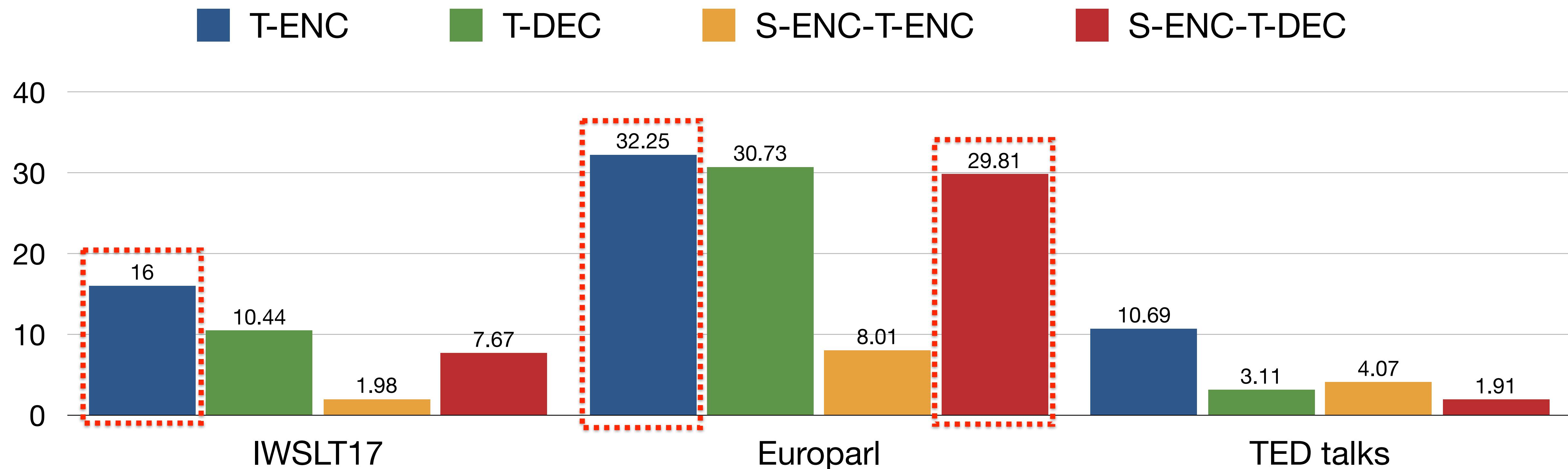
Language Tag Does not Affect Performance on Supervised Directions

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

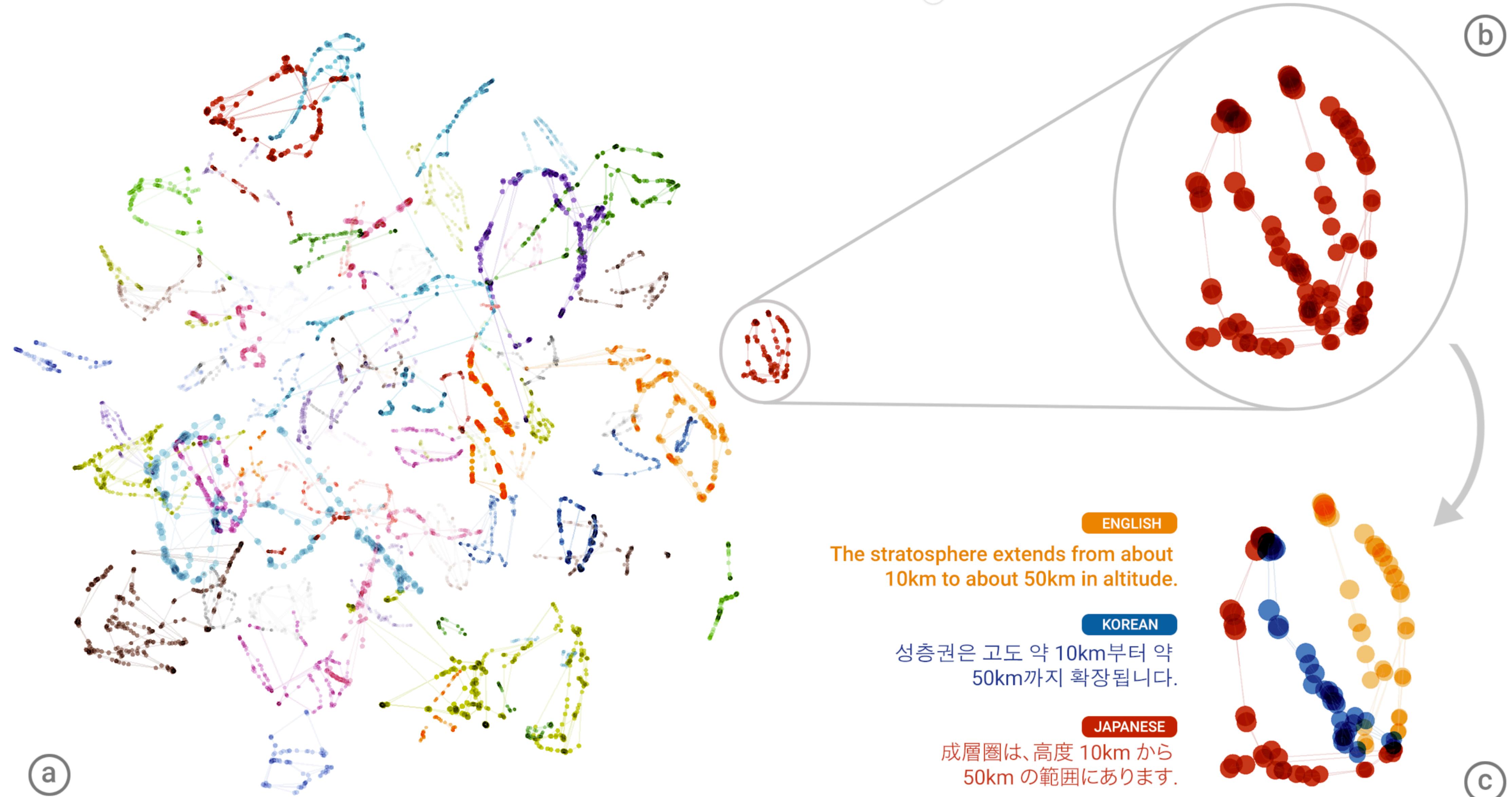


Target Language Tag on Encoder Strategy Gets Best Zero-Shot Performance

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



Does sentence have similar emb. representation?



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: 나 는 東 京 大 学 的 生 이 니 다 . → I am a student at Tokyo University.
- Japanese/Korean: 私は東京大学 ト き ジ う 大 学 の 学 生 で す 。 → 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지구의 중심의 가까이에 어딘가에도착하고 있어야한다.
0.60	나는지구의 센터의 가까이에 어딘가에도착하고 있어야한다.
0.70	나는지구의 중심근처어딘가에도착해야합니다.
0.90	나는어딘가지구의 중심근처에도착해야합니다.
1.00	나는어딘가지구의 중심근처에도착해야합니다.

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

Unfortunate 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

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

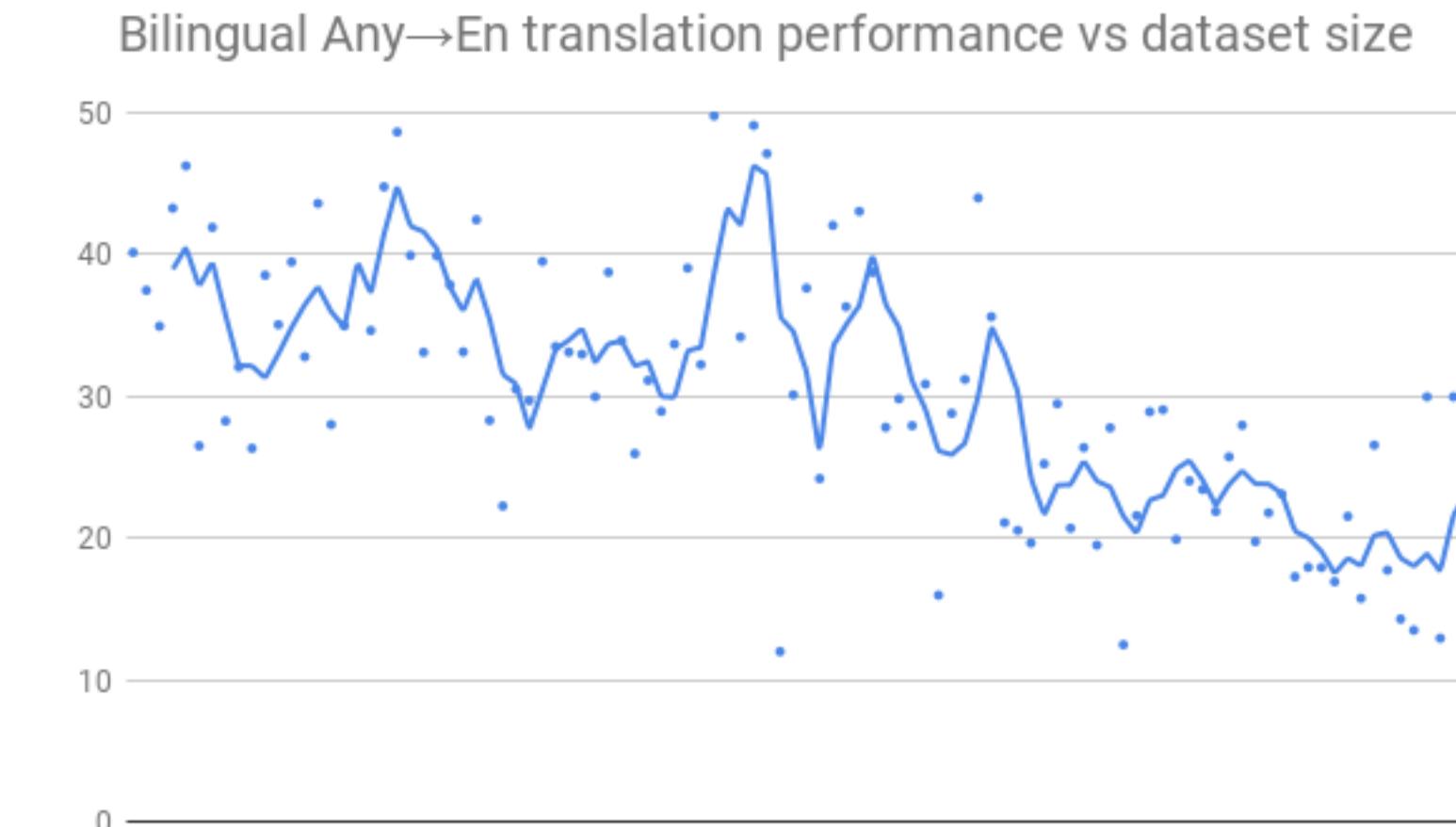
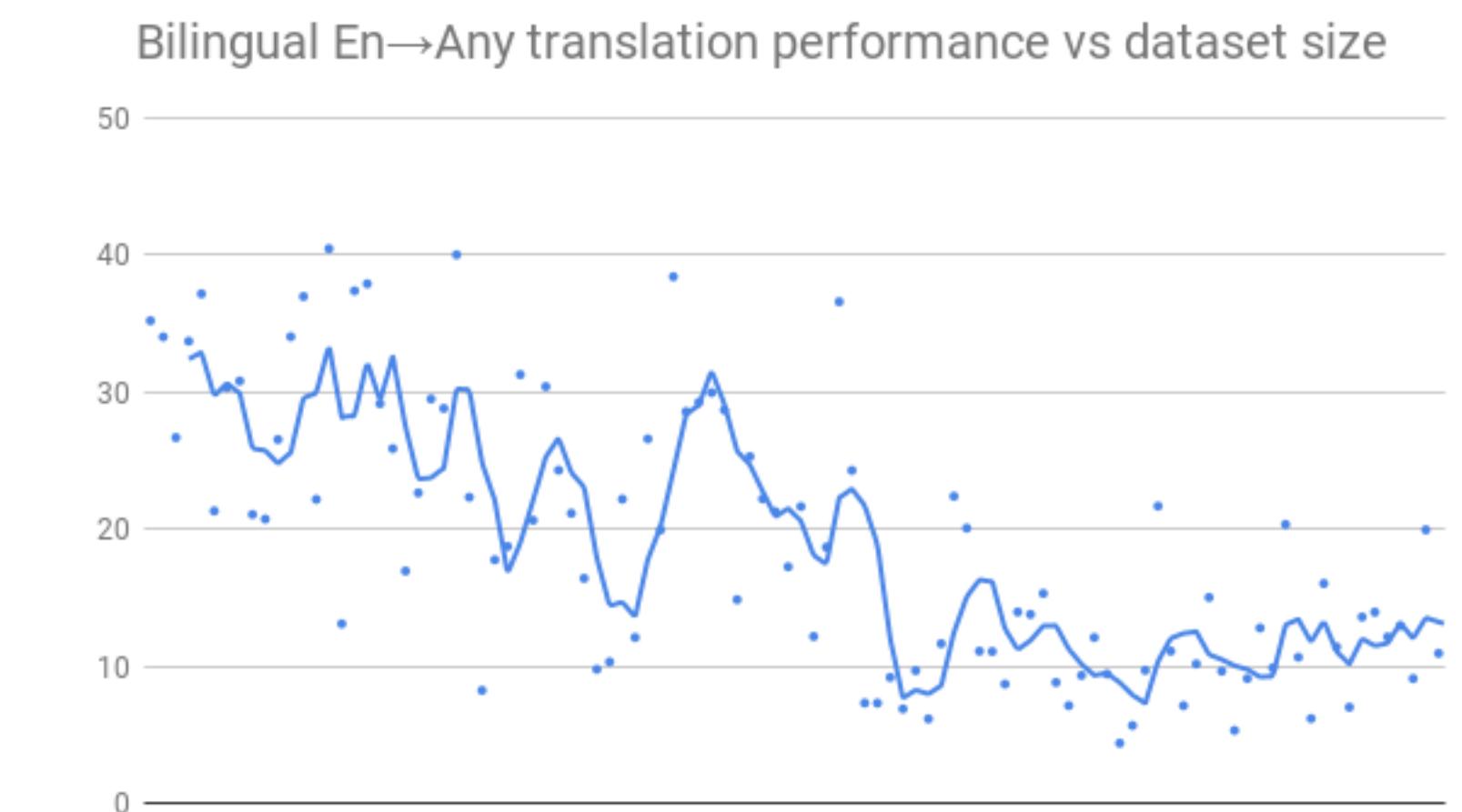
	Ar-Fr	Fr-Ar	Ru-Uk	Uk-Ru	Avg.
5-to-5	1.66	4.49	3.7	3.02	3.21
25-to-25	1.83	5.52	16.67	4.31	7.08
50-to-50	4.34	4.72	15.14	20.23	11.1
75-to-75	1.85	4.26	11.2	15.88	8.3
103-to-103	2.87	3.05	12.3	18.49	9.17

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)

<i>En</i> → <i>Any</i>	High 25	Med. 52	Low 25
Bilingual	29.34	17.50	11.72
<i>All</i> → <i>All</i>	28.03	16.91	12.75
<i>En</i> → <i>Any</i>	28.75	17.32	12.98
<i>Any</i> → <i>En</i>	High 25	Med. 52	Low 25
Bilingual	37.61	31.41	21.63
<i>All</i> → <i>All</i>	33.85	30.25	26.96
<i>Any</i> → <i>En</i>	36.61	33.66	30.56



Sampling of Data

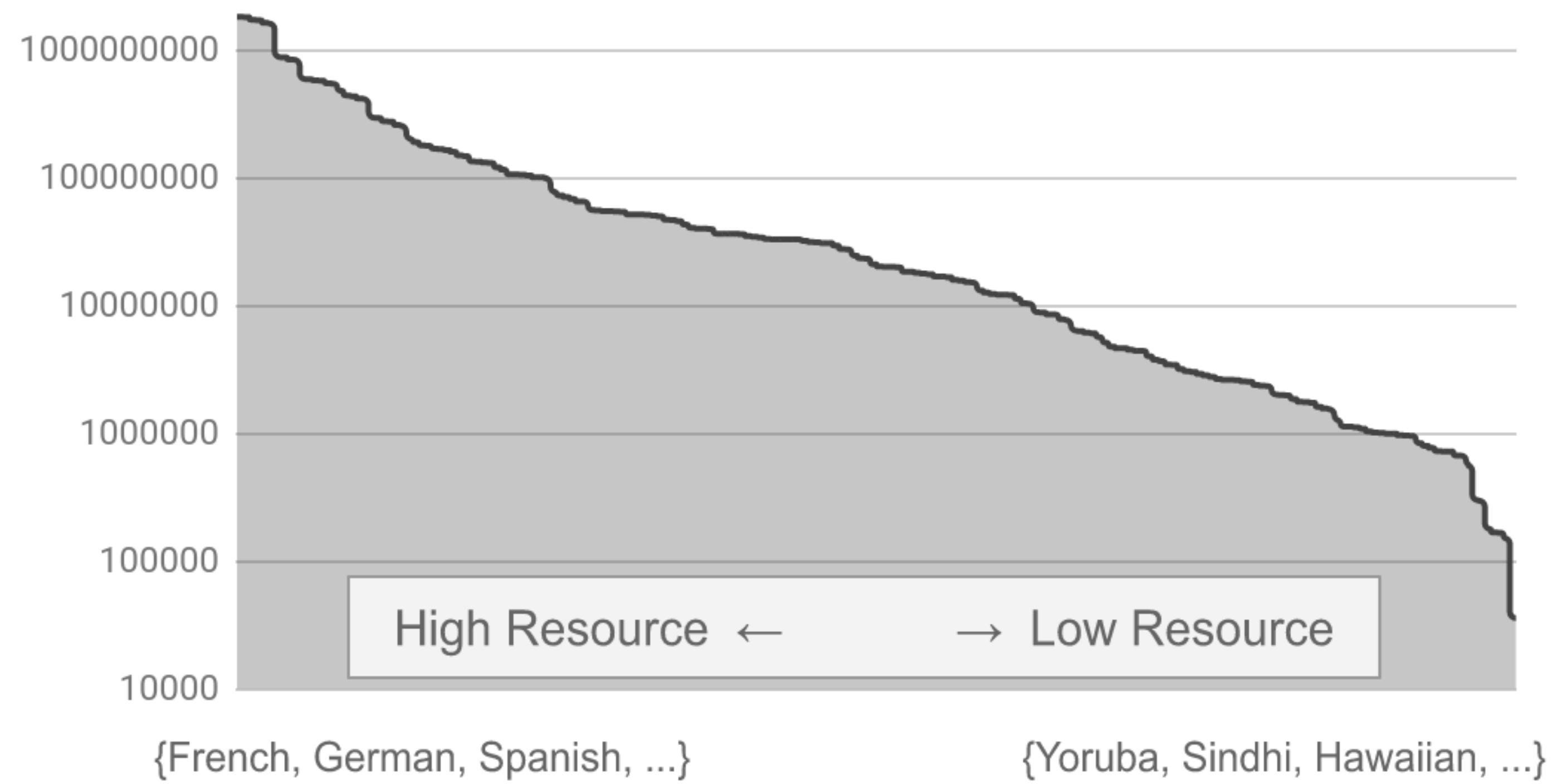
- sample data prob w.r.t

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

-

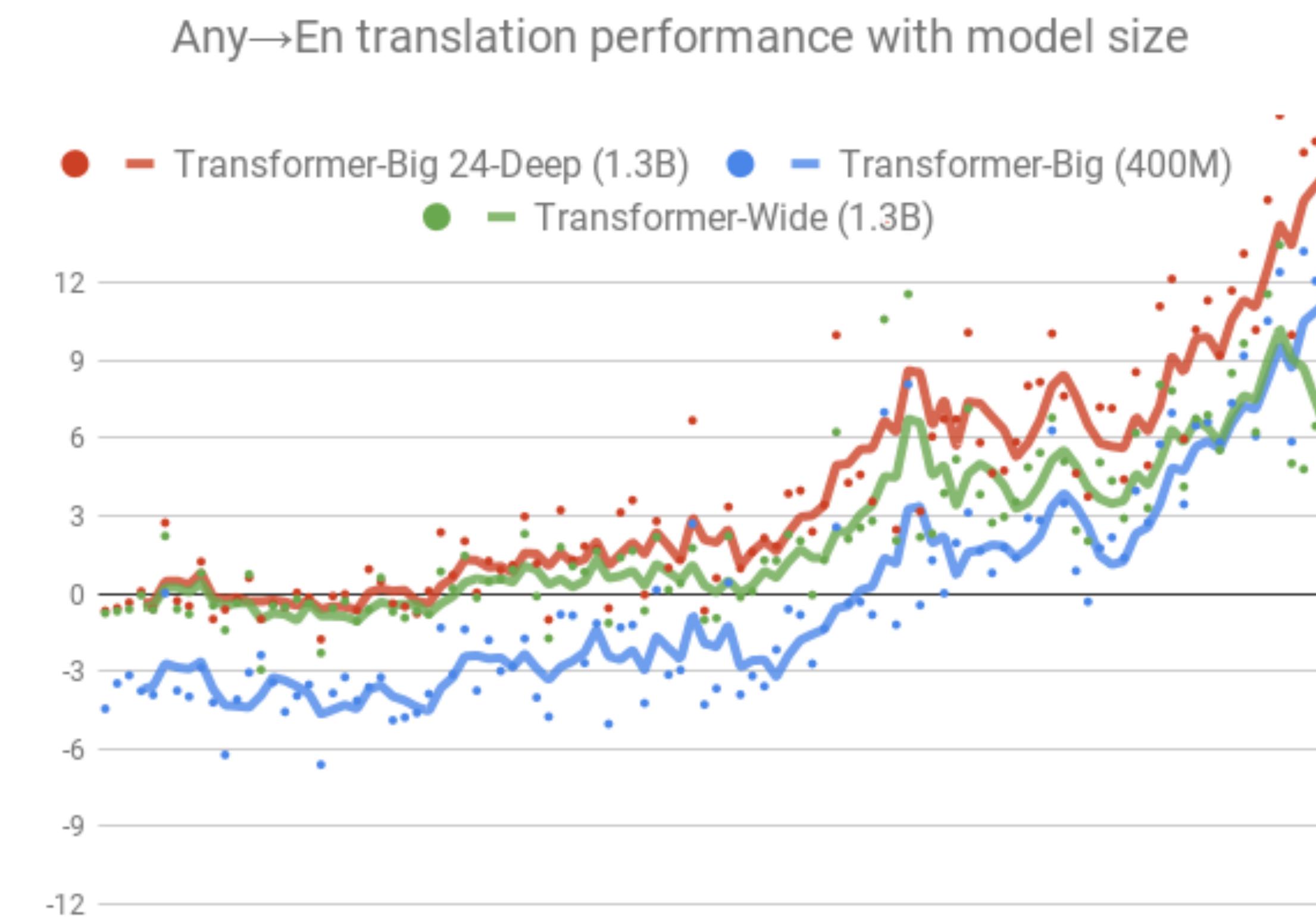
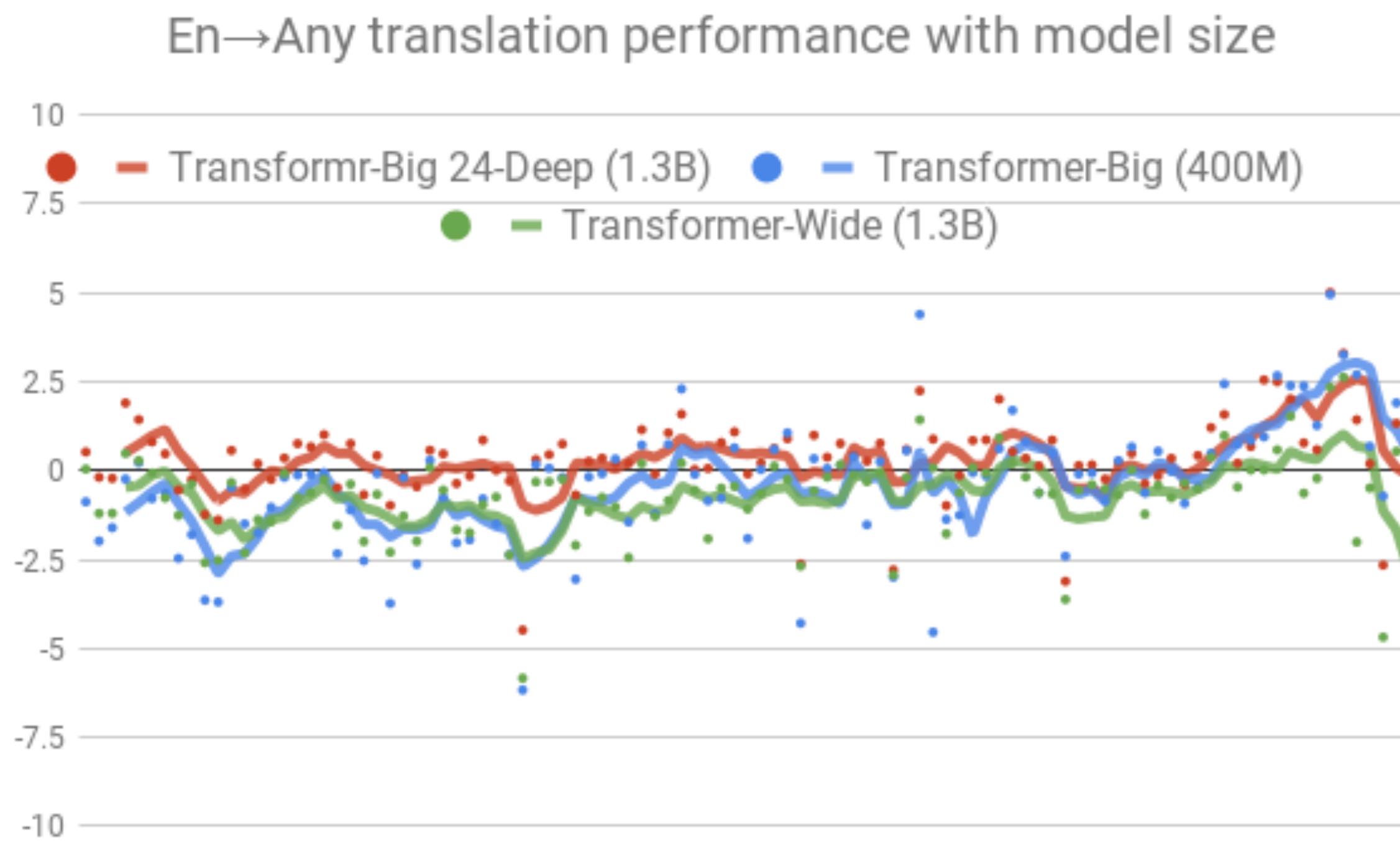
$En \rightarrow Any$	High 25	Med. 52	Low 25
$T_V = 1$	27.81	16.72	12.73
$T_V = 100$	27.83	16.86	12.78
$T_V = 5$	28.03	16.91	12.75
$Any \rightarrow En$	High 25	Med. 52	Low 25
$T_V = 1$	33.82	29.78	26.27
$T_V = 100$	33.70	30.15	26.91
$T_V = 5$	33.85	30.25	26.96

Data distribution over language pairs



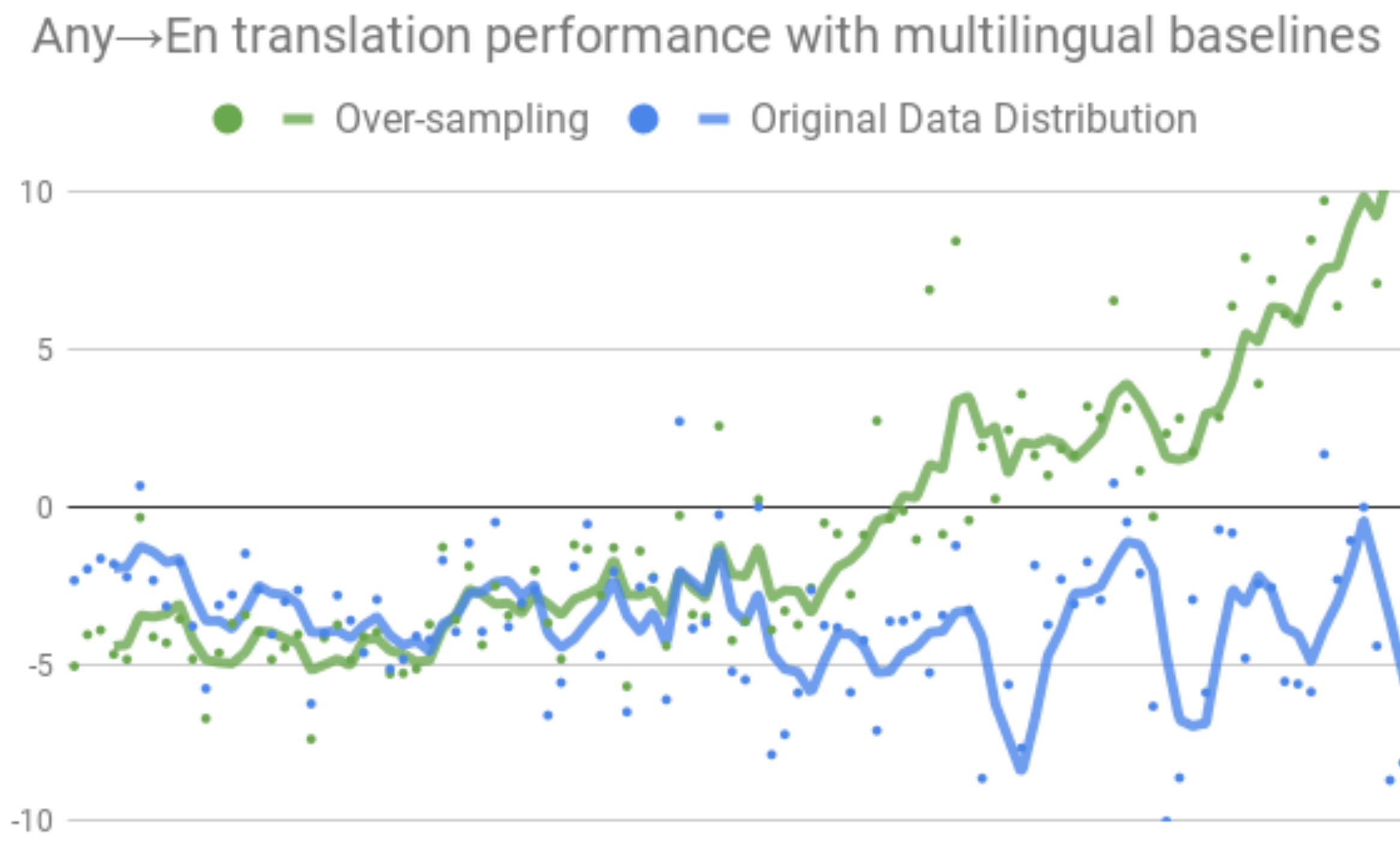
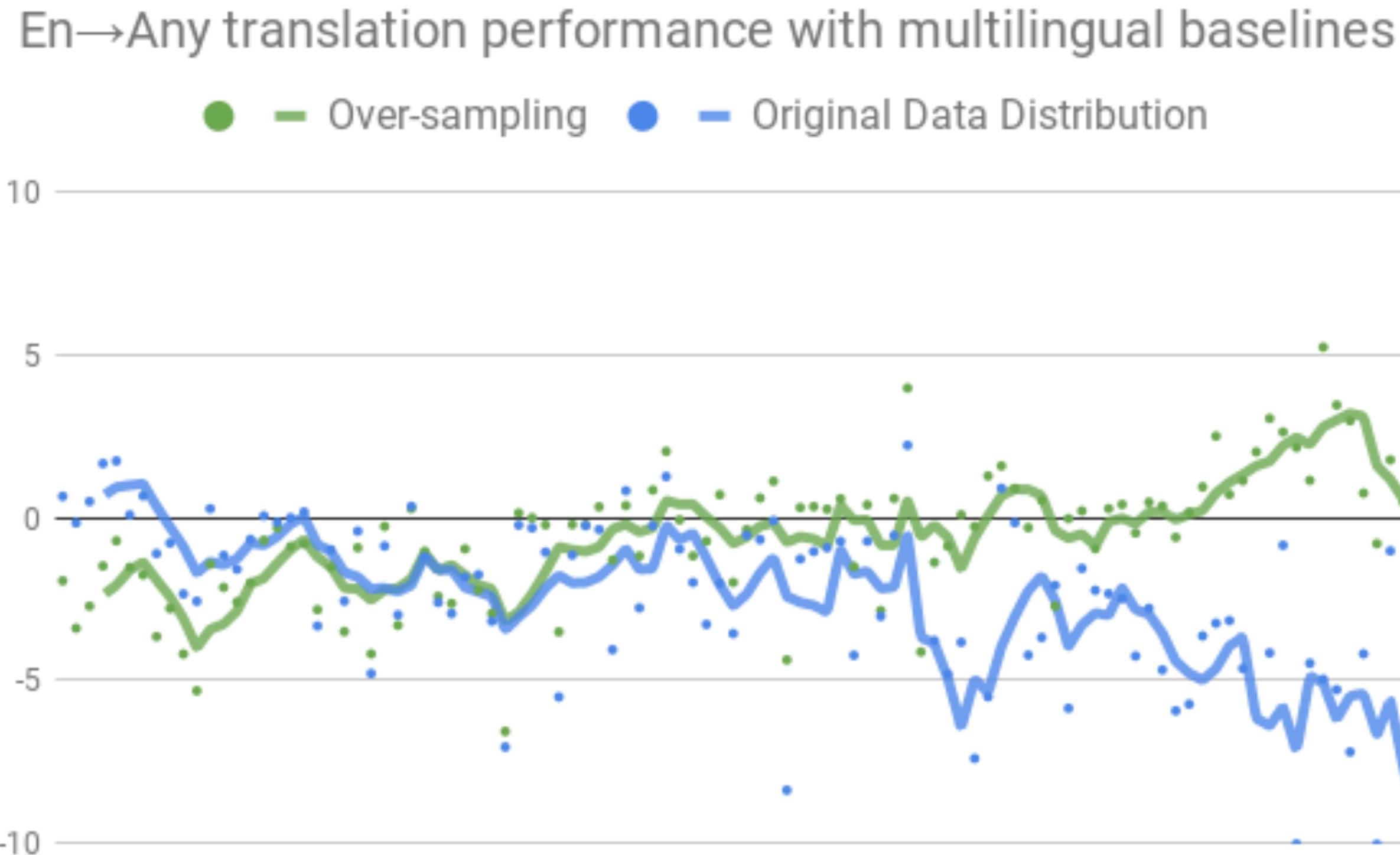
Bigger Model

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



Limitation

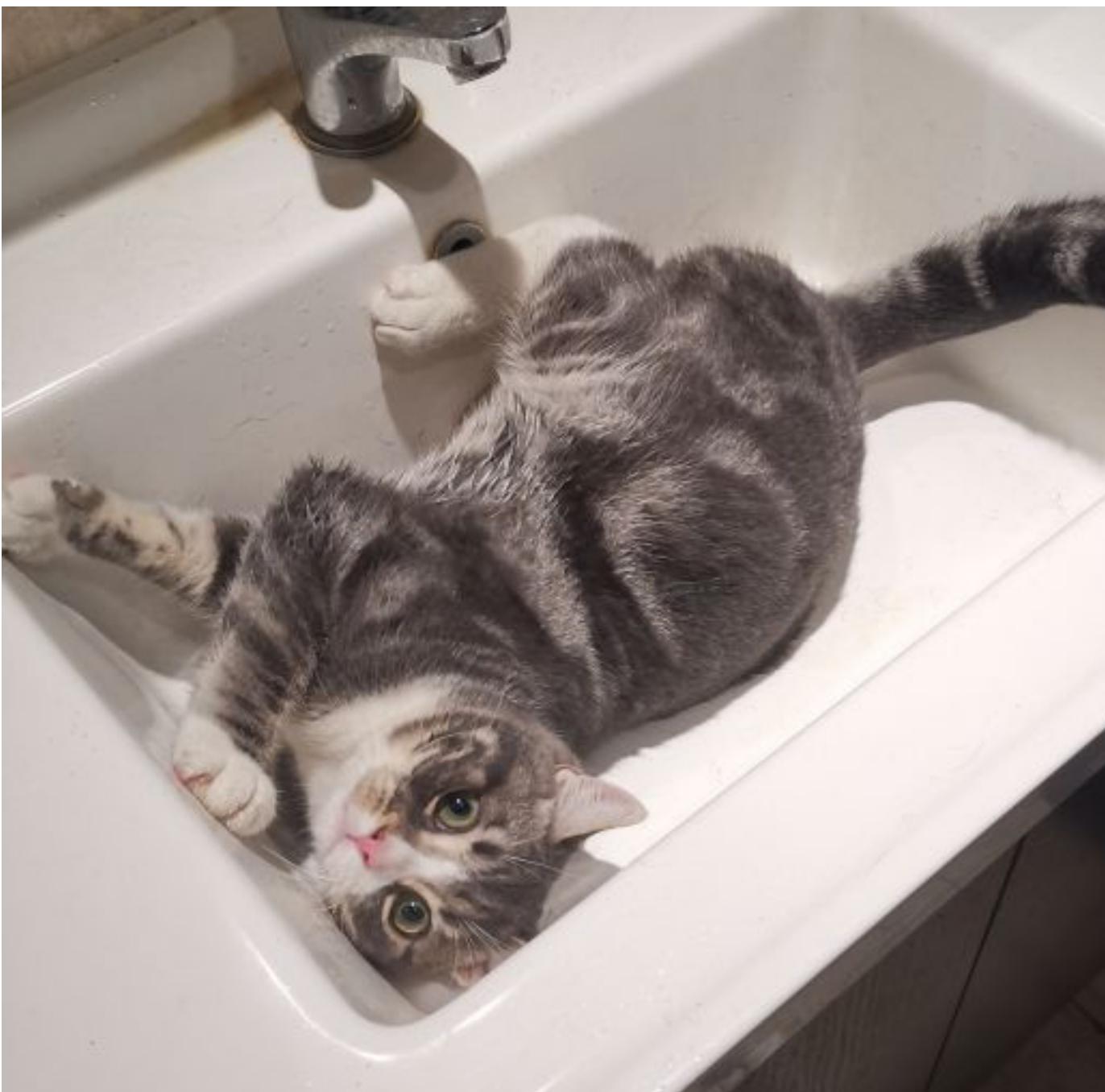
- mTransformer boosts performance on low-resource languages but not high-resource
- Zero-shot directions are not usable yet.



MT w/ Adapter

Parameter Interference issue for MNMT

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



Bilingual

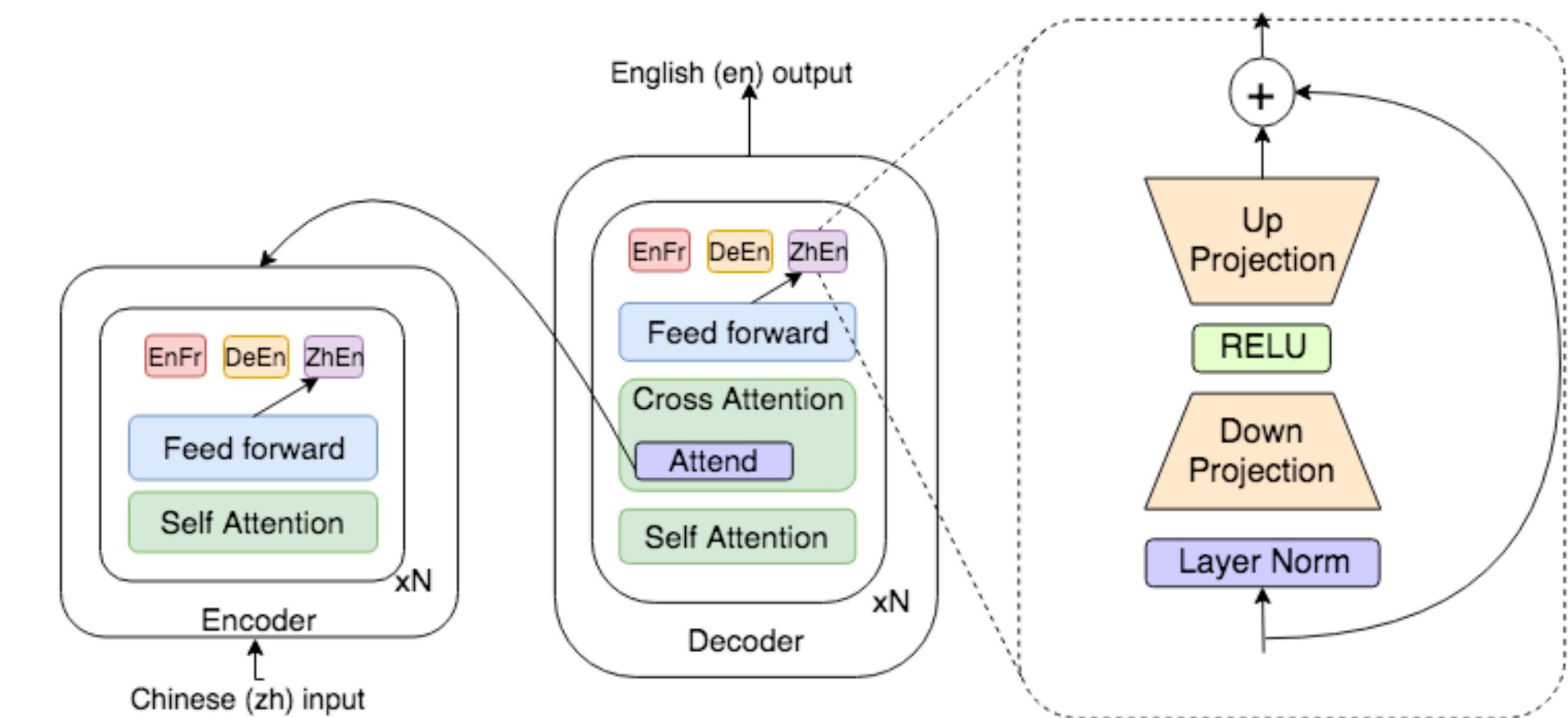
Multilingual

Language Aware mTransformer

- deep mTransformer
 - 12e12d **+2BLEU**
- language-aware layer normalization **+2~3BLEU**
 - each language has its own normalization
- language-aware linear transformation
 - the output of encoder is transformed with a language-specific matrix
- Online back translation (+up to **10BLEU**)
- Evaluated on OPUS100: 55M sentence pairs

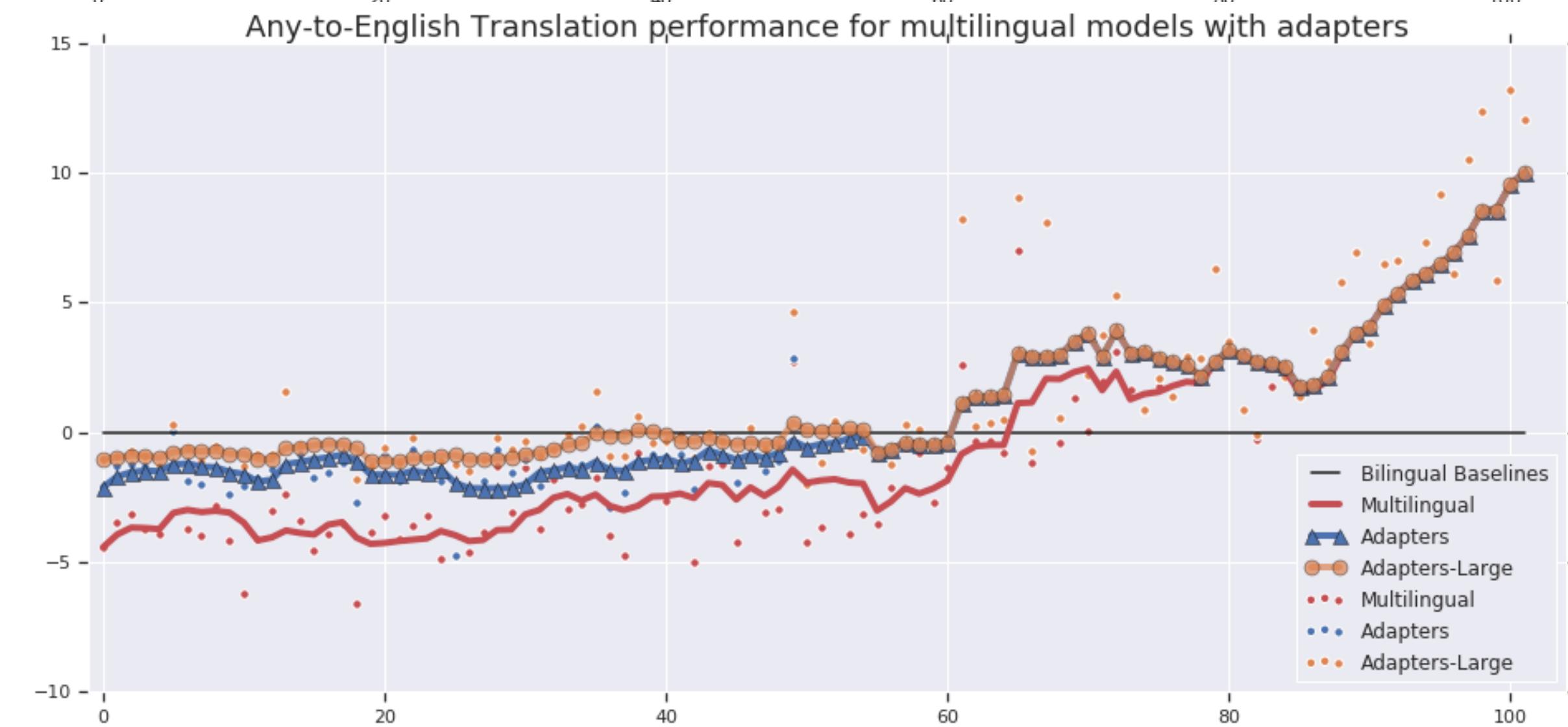
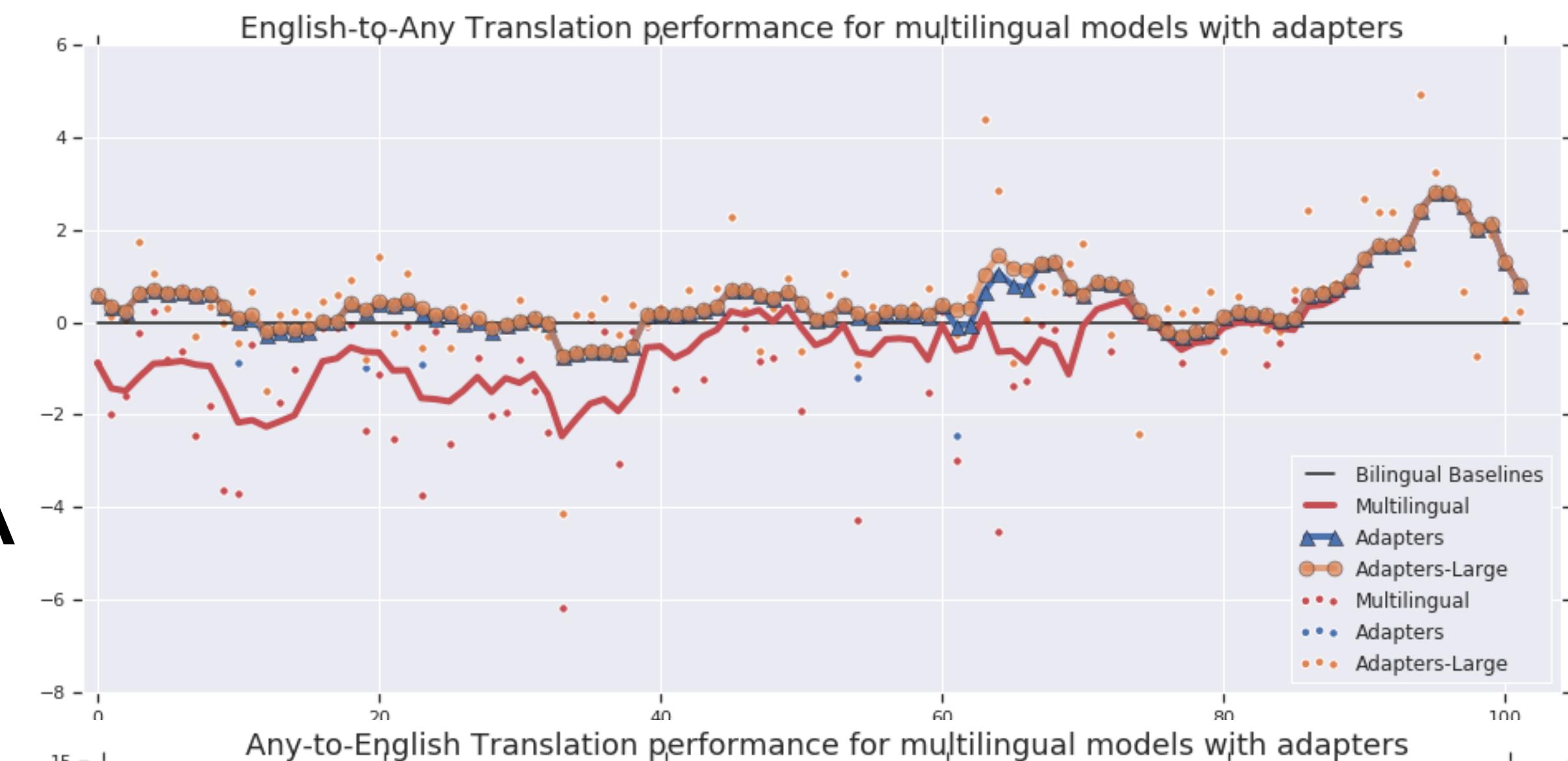
Multilingual NMT with Serial Adapter

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



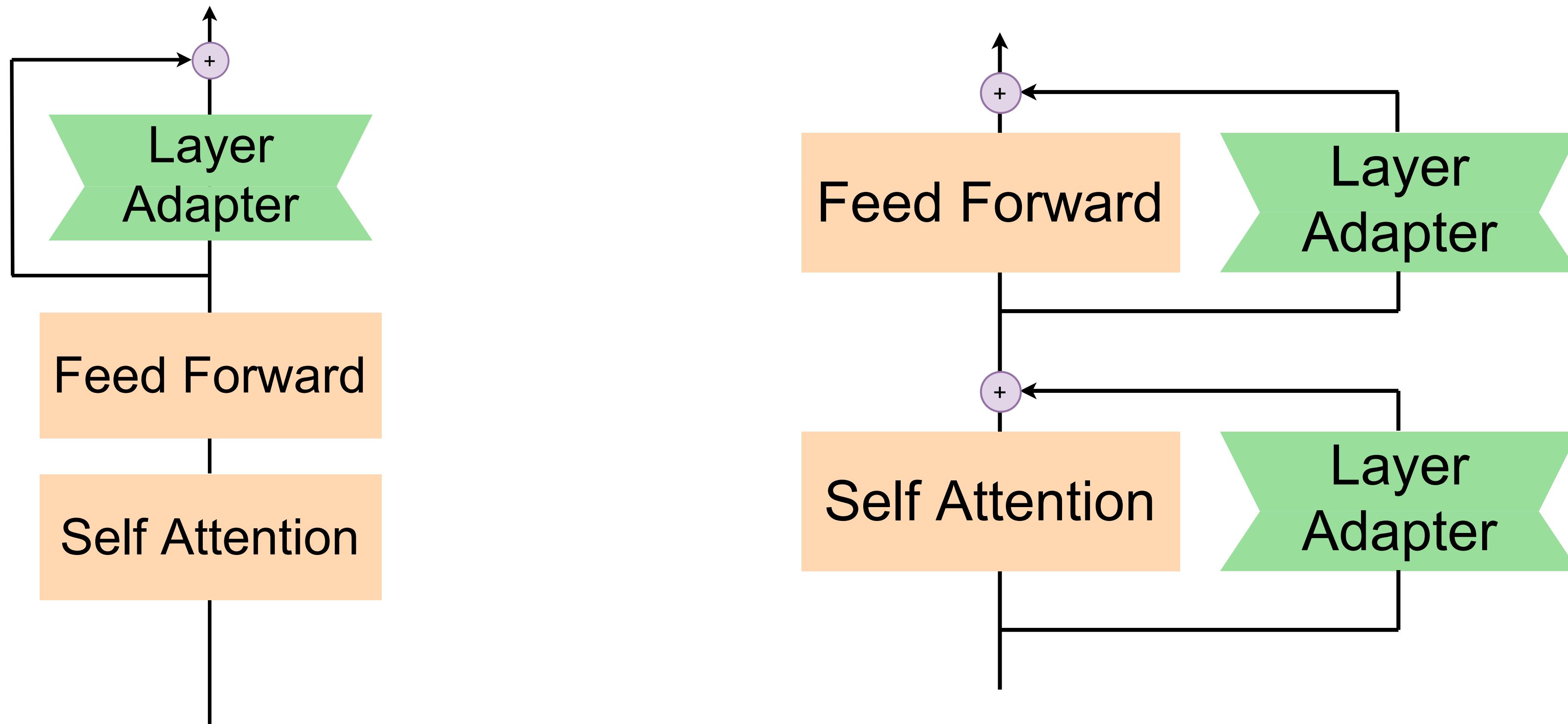
Serial Adapter improves Multilingual Translation

- 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.



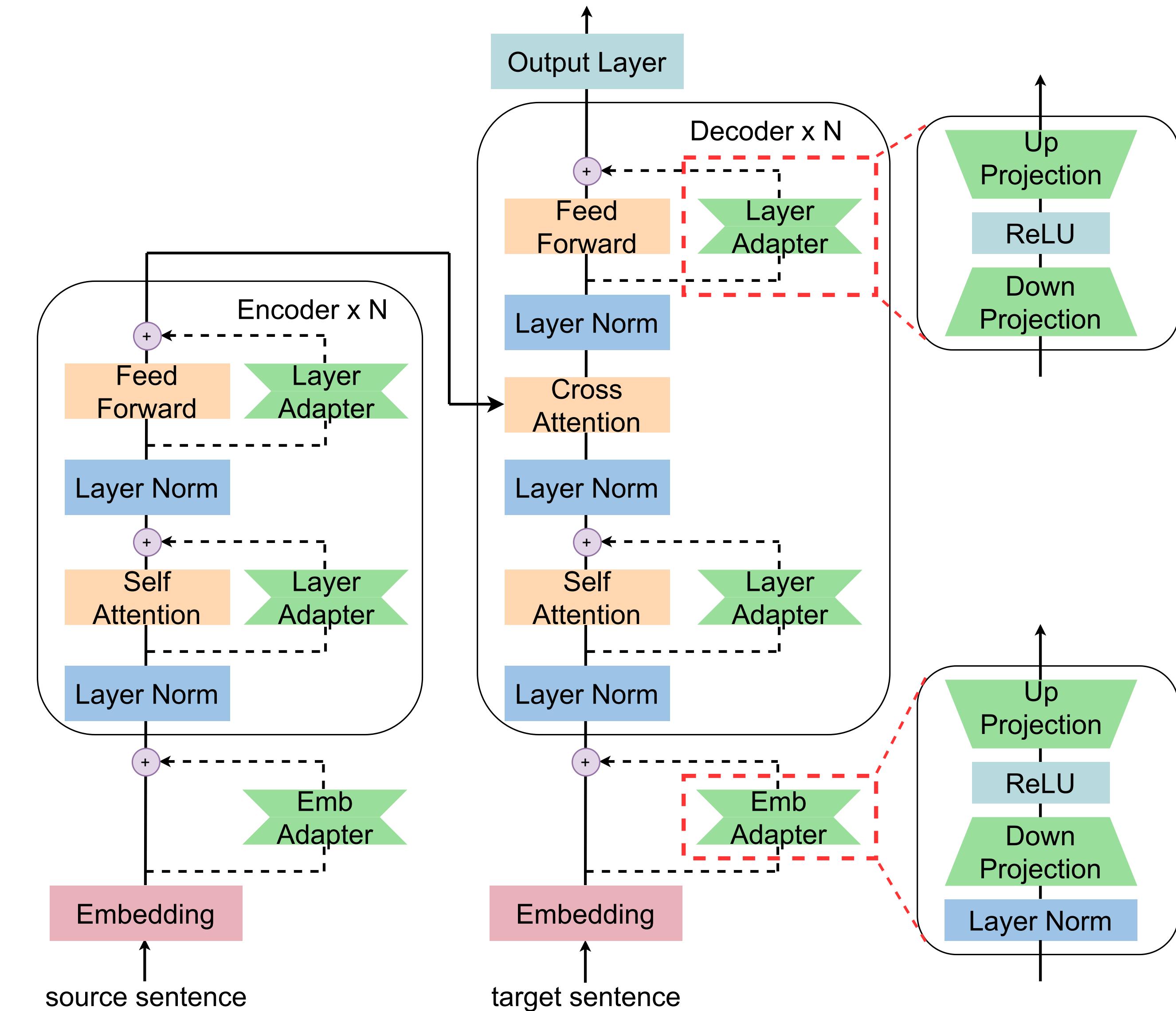
Counter Interference

- Which adapter will remove noise?



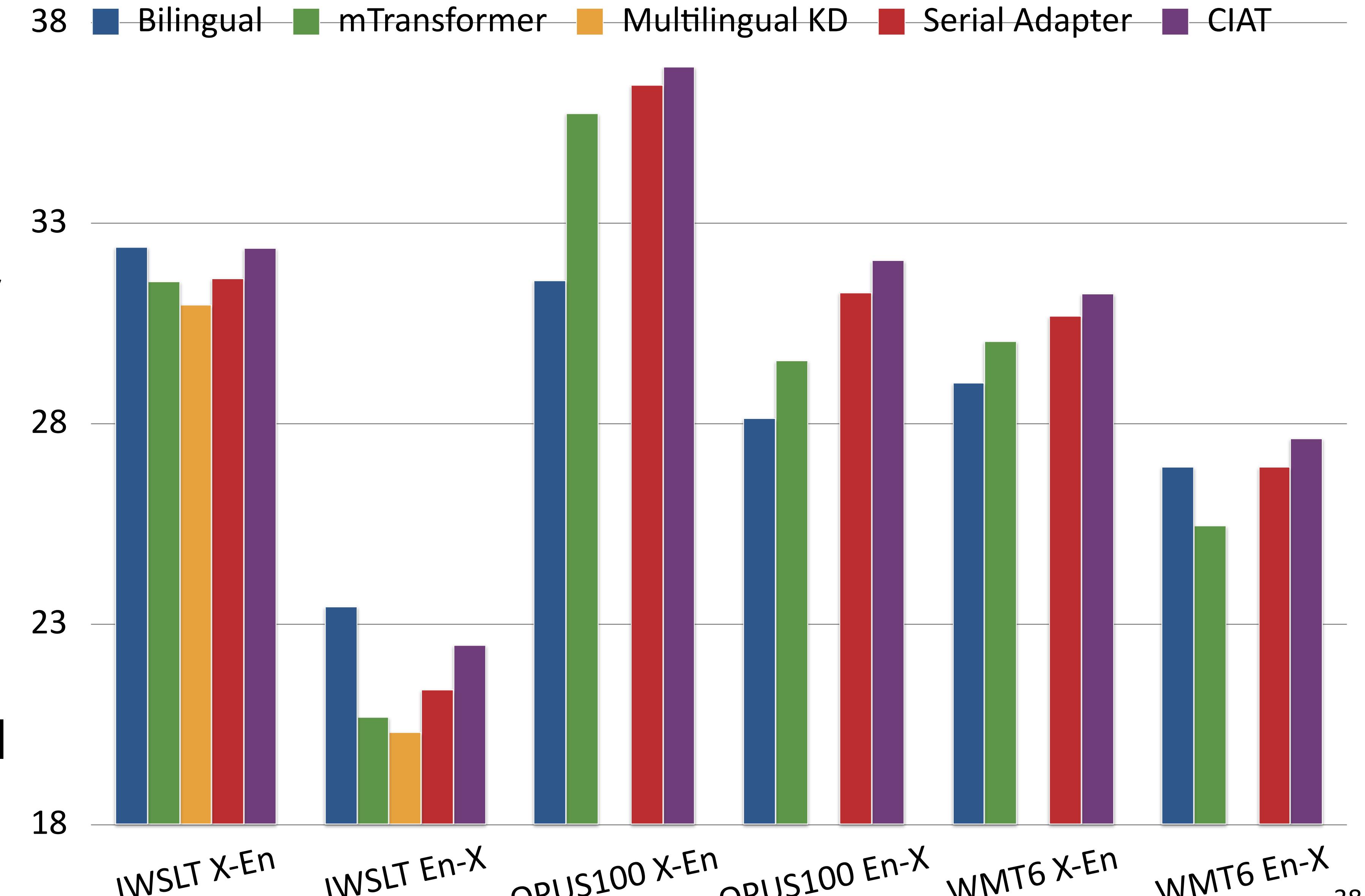
Parallel Adapter - CIAT

- Design rationale:
 - process before multilingual interference is introduced in each layer
- Embedding adapter
- Parallel layer adapter
- Training:
 - Pretrain mTransformer on multilingual data
 - Fix mTransformer and train parallel adapters on specific language pairs



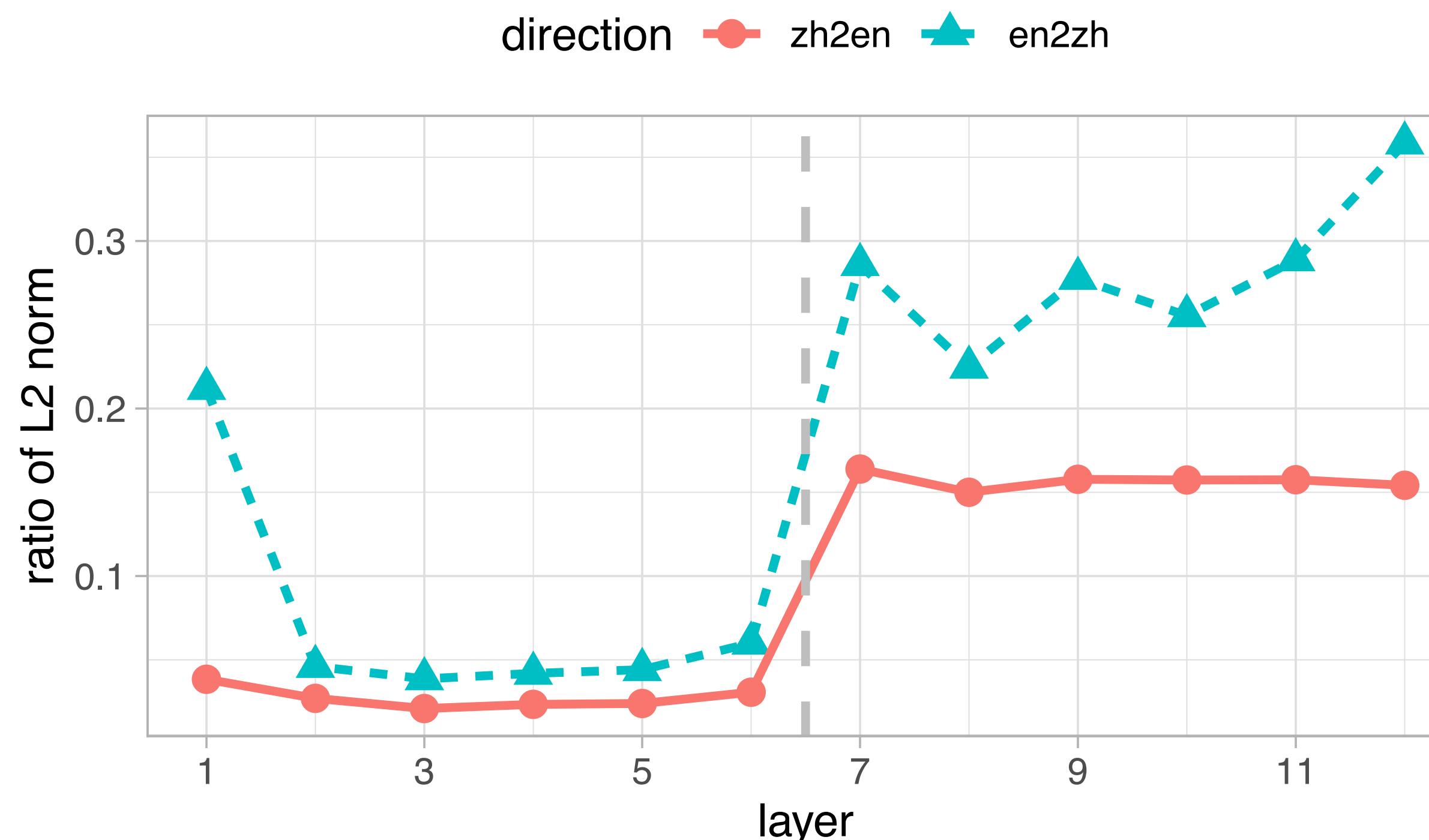
Comparing MNMT w/ Adapters

- 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.

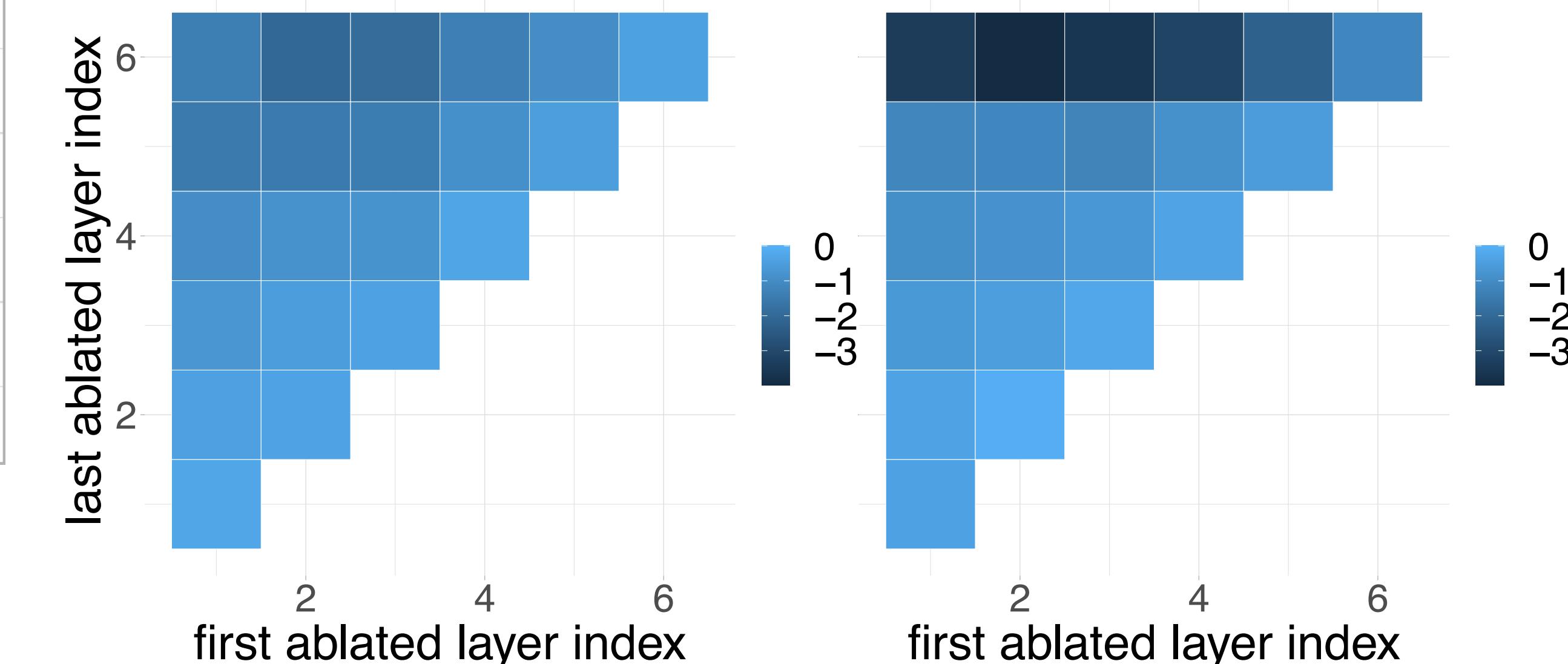


Which layer-adapter are more important?

- Upper decoder layer adapter is more important

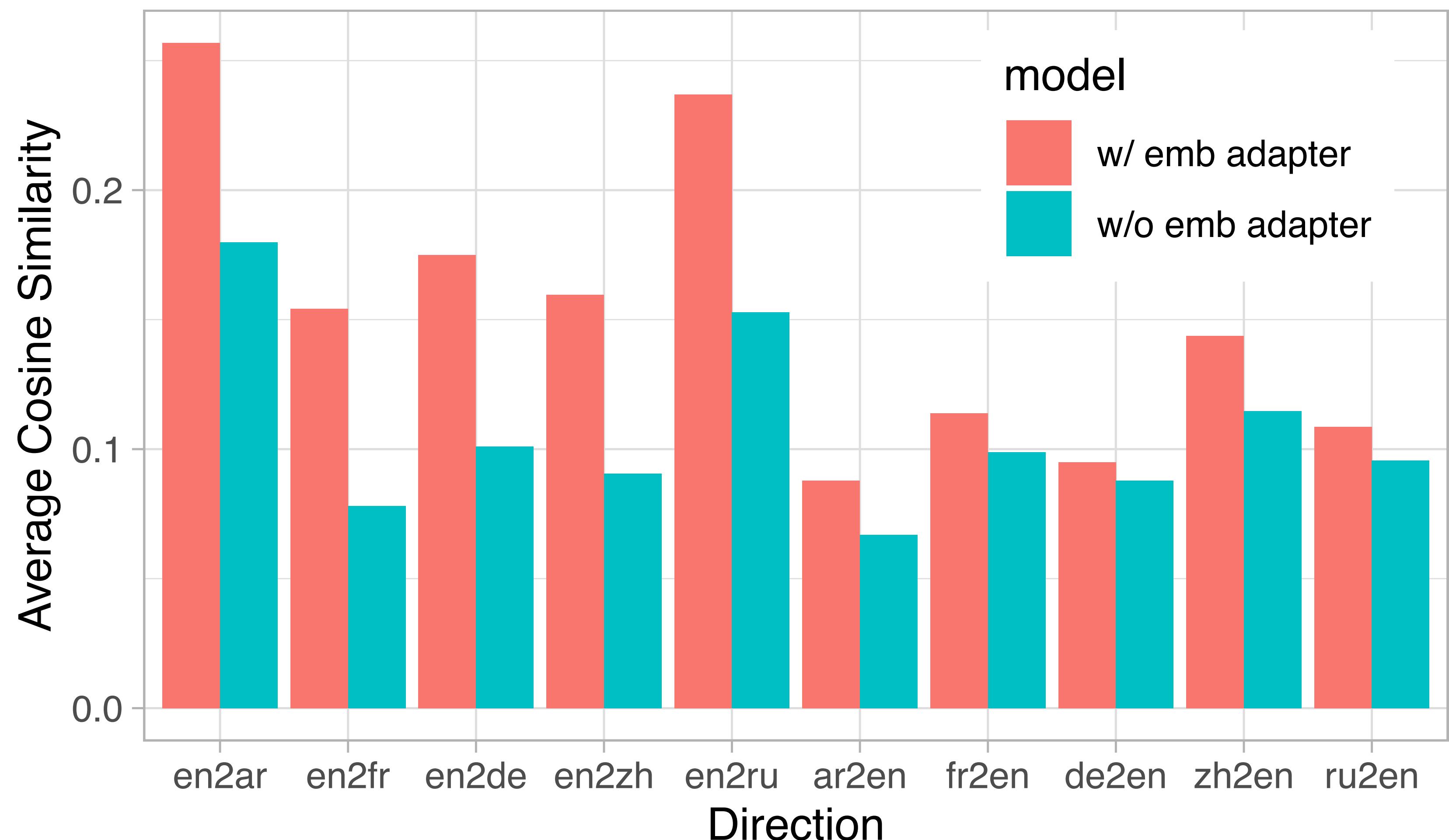


Drop in perf. removing
enc/dec adapters



Embedding Adapter is also important!

- Embedding adapter enhance the word embedding similarity between language pairs



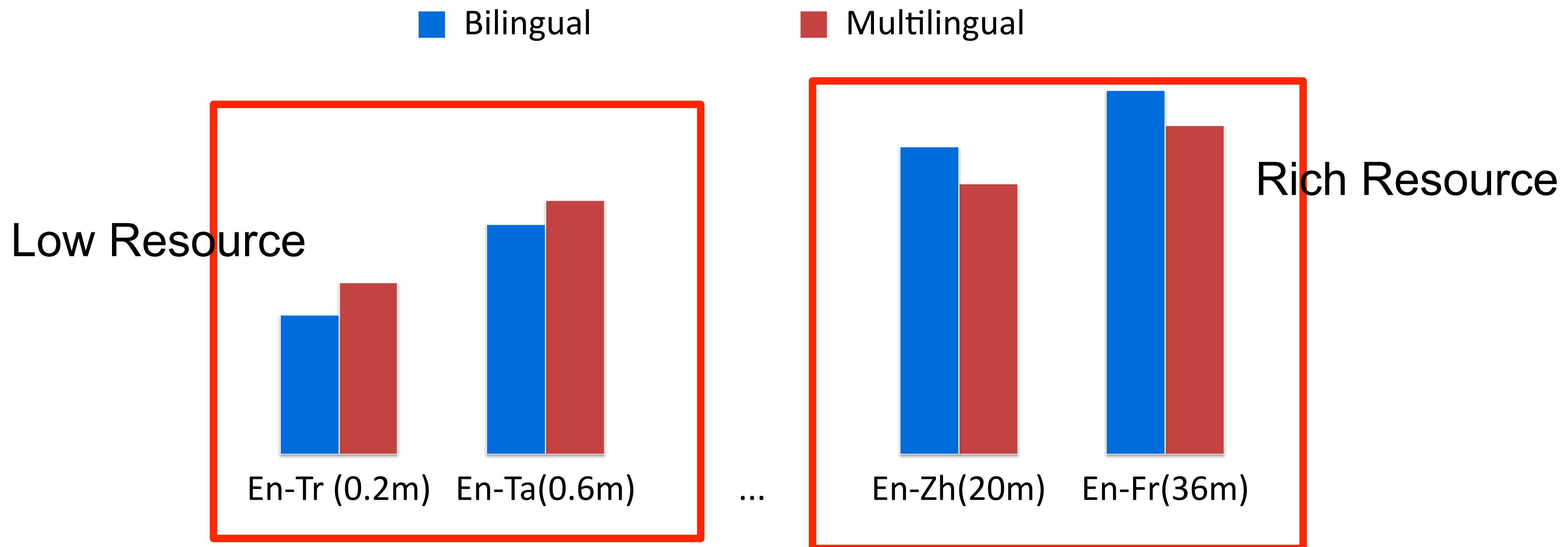
Benefit of MNMT w/ Adapter

- 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: $N \times 242m$
 - mTransformer: $242m$
 - mTransformer+Serial Adapter: $242m + N \times 12.6m$
 - mTransformer+parallel adapter (CIAT): $242m + N \times 12.6\sim27.3m$
- Plug-and-play: CIAT only needs to finetune adapter

Exploiting Model Capacity with Language-specific Subnet

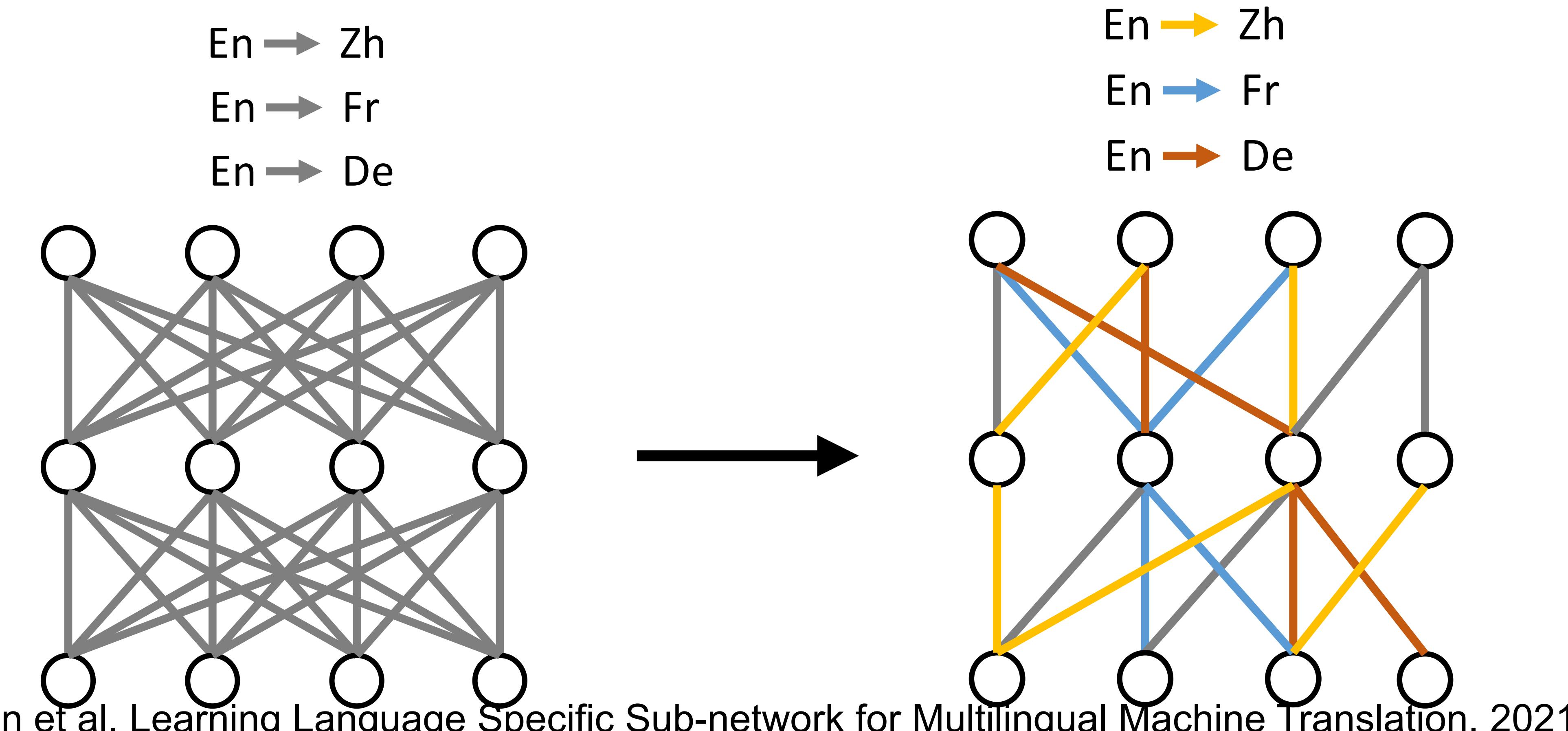
Challenge of Multilingual NMT

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



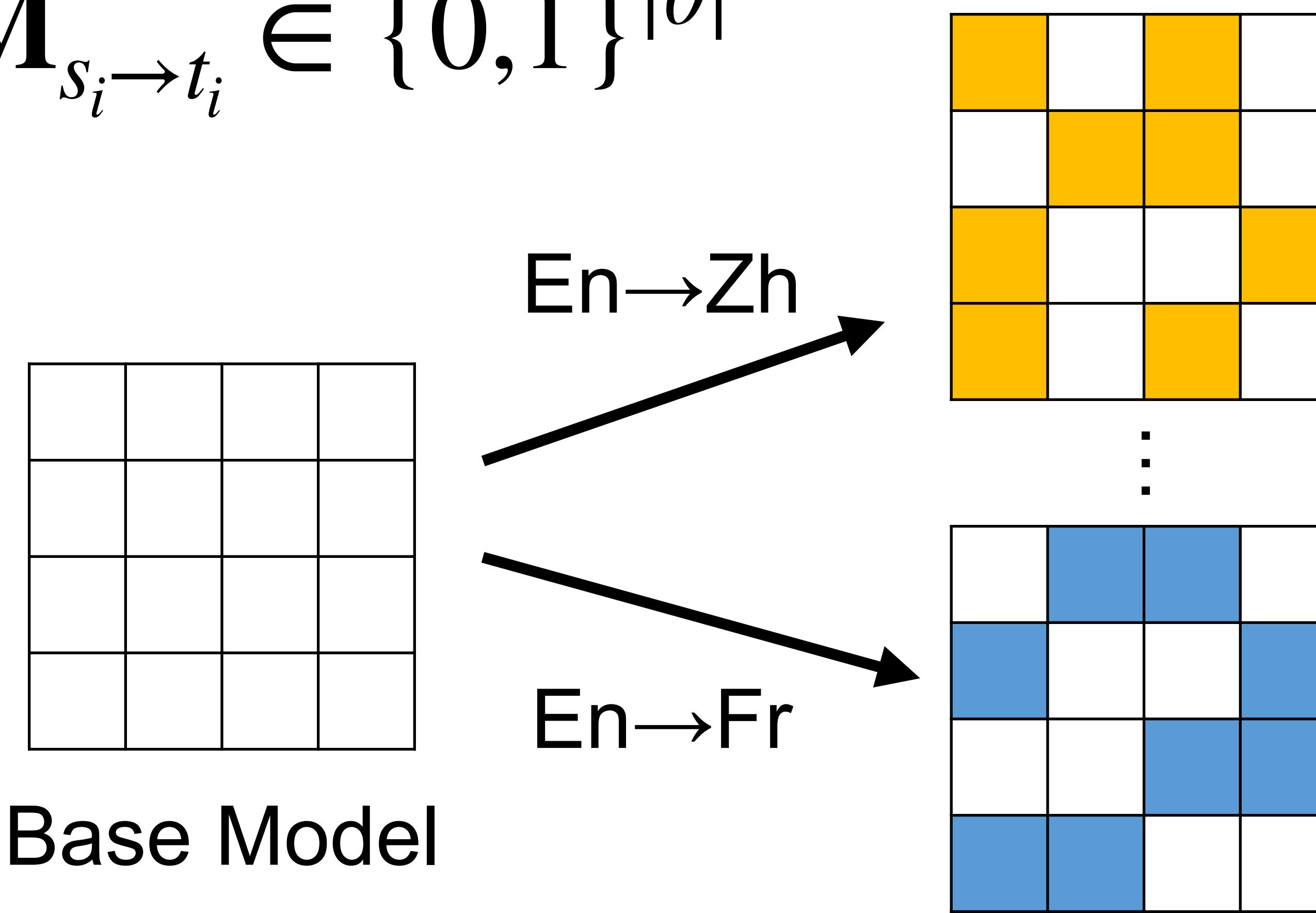
Language-Specific Sub-network (LaSS)

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



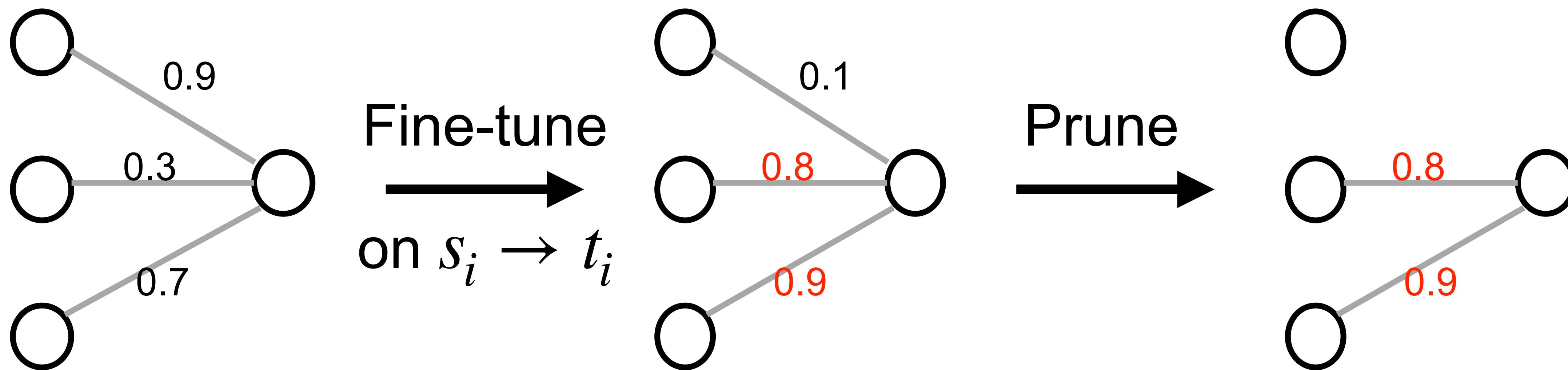
LaSS overall framework

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



How to find language-specific sub-network: Intuition

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



How to find language-specific masks

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

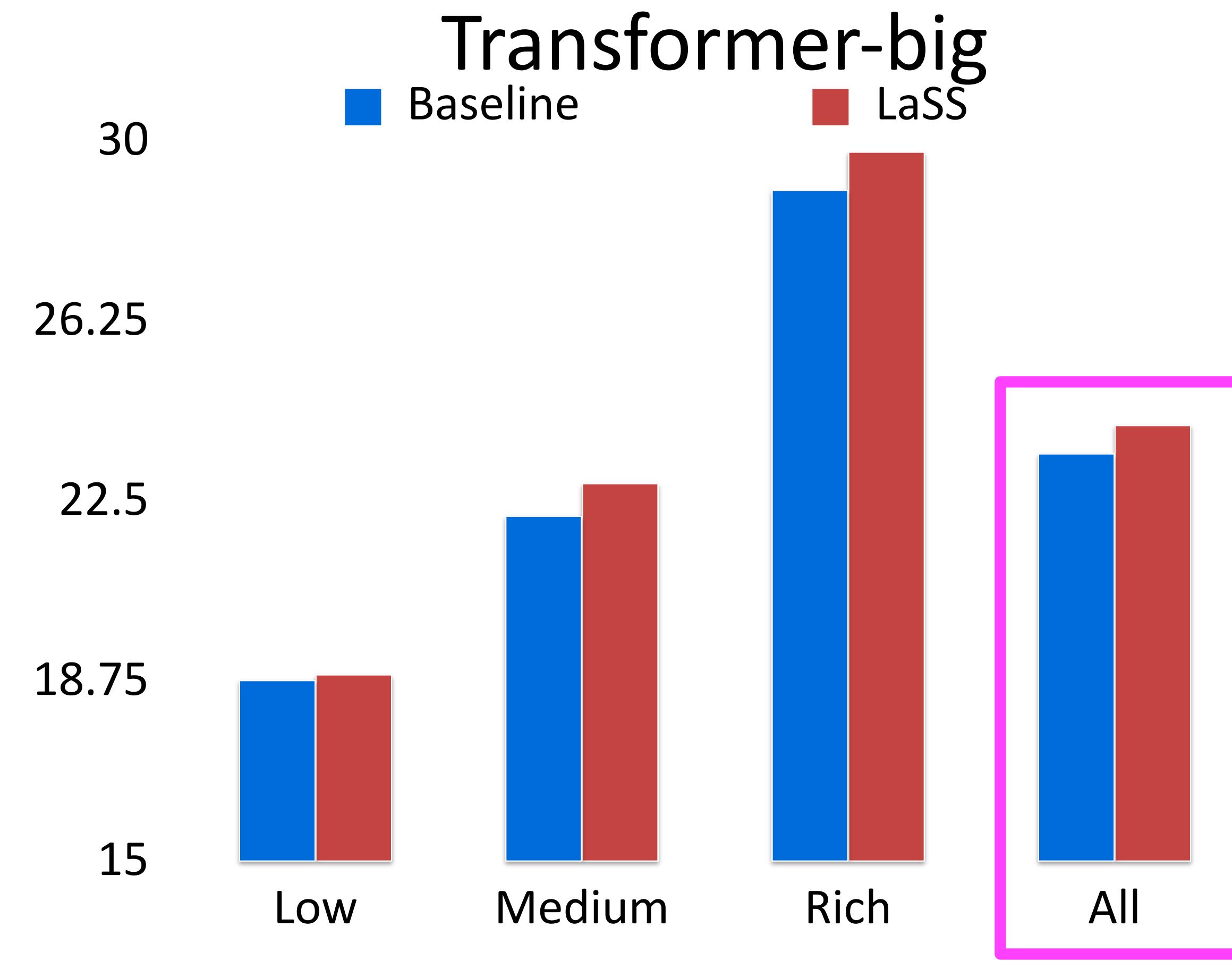
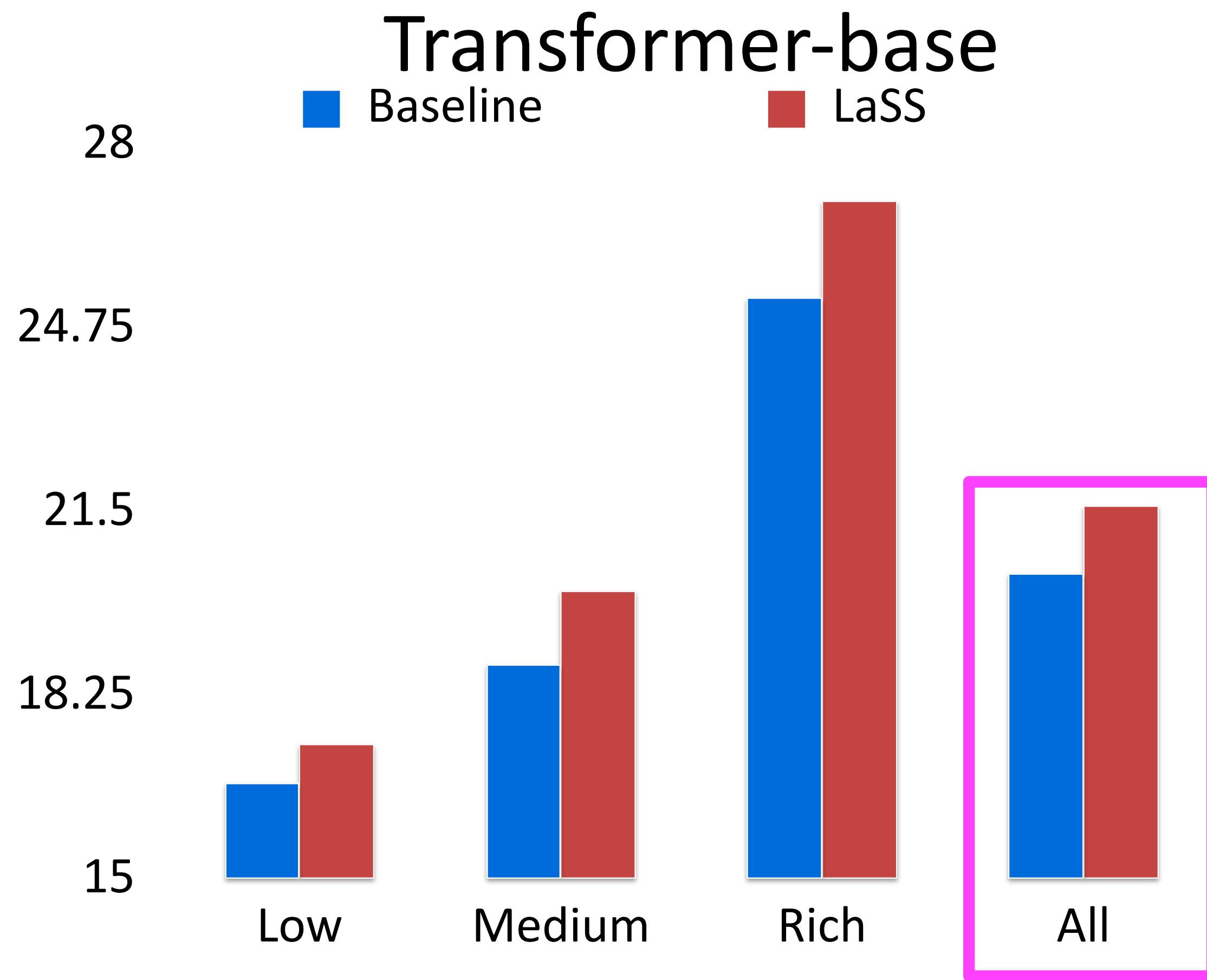
Structure-aware Joint Training

- Further continue to train θ_0 through structure-aware updating after obtaining $\mathbf{M}_{s_i \rightarrow t_i}$
 - Create batch $\mathcal{B}_{s_i \rightarrow t_i}$ full of samples from $s_i \rightarrow t_i$
 - Forward and backward with sub-network

$$\theta_{s_i \rightarrow t_i} = \left\{ \theta_0^j \mid \mathbf{M}_{s_i \rightarrow t_i}^j = 1 \right\}$$

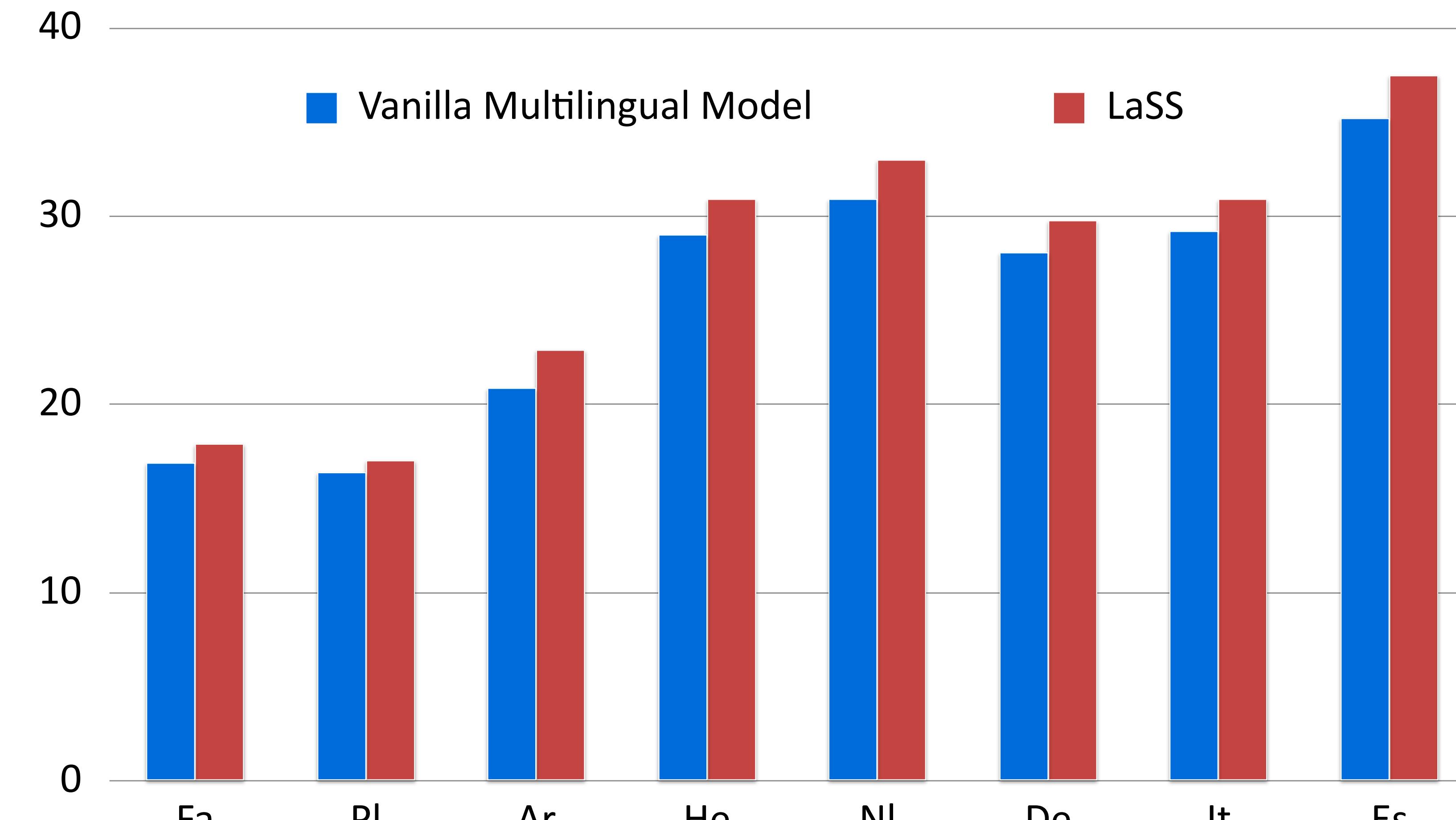
Efficacy in alleviating Parameter Interference

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



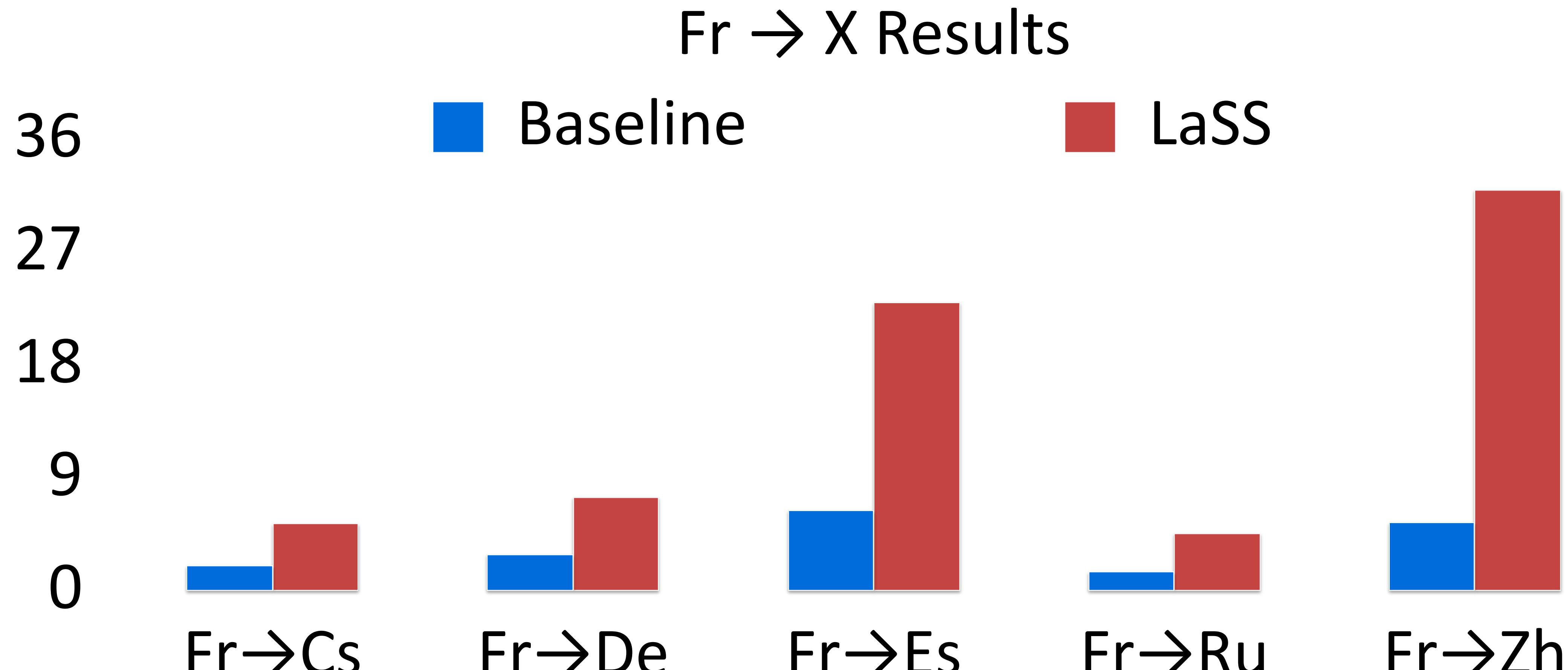
Efficacy in alleviating Parameter Interference

- LaSS obtains consistent performance gains.
 - IWSLT



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



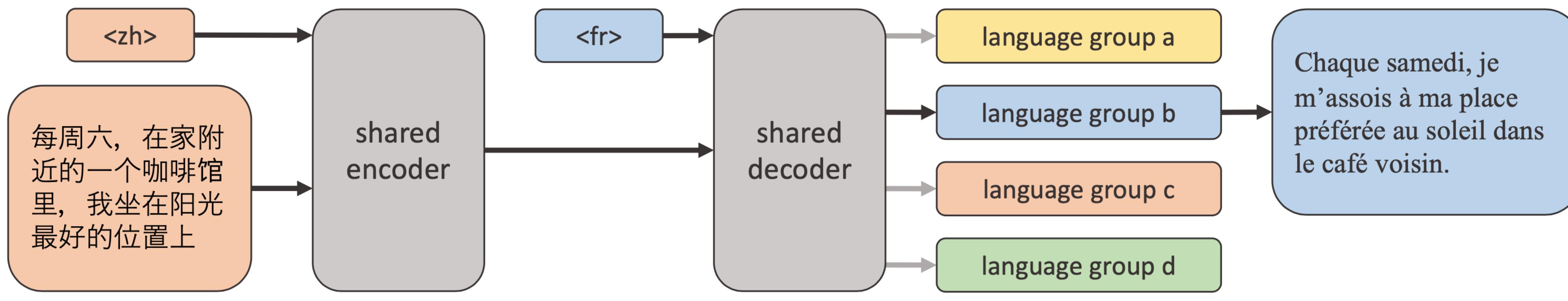
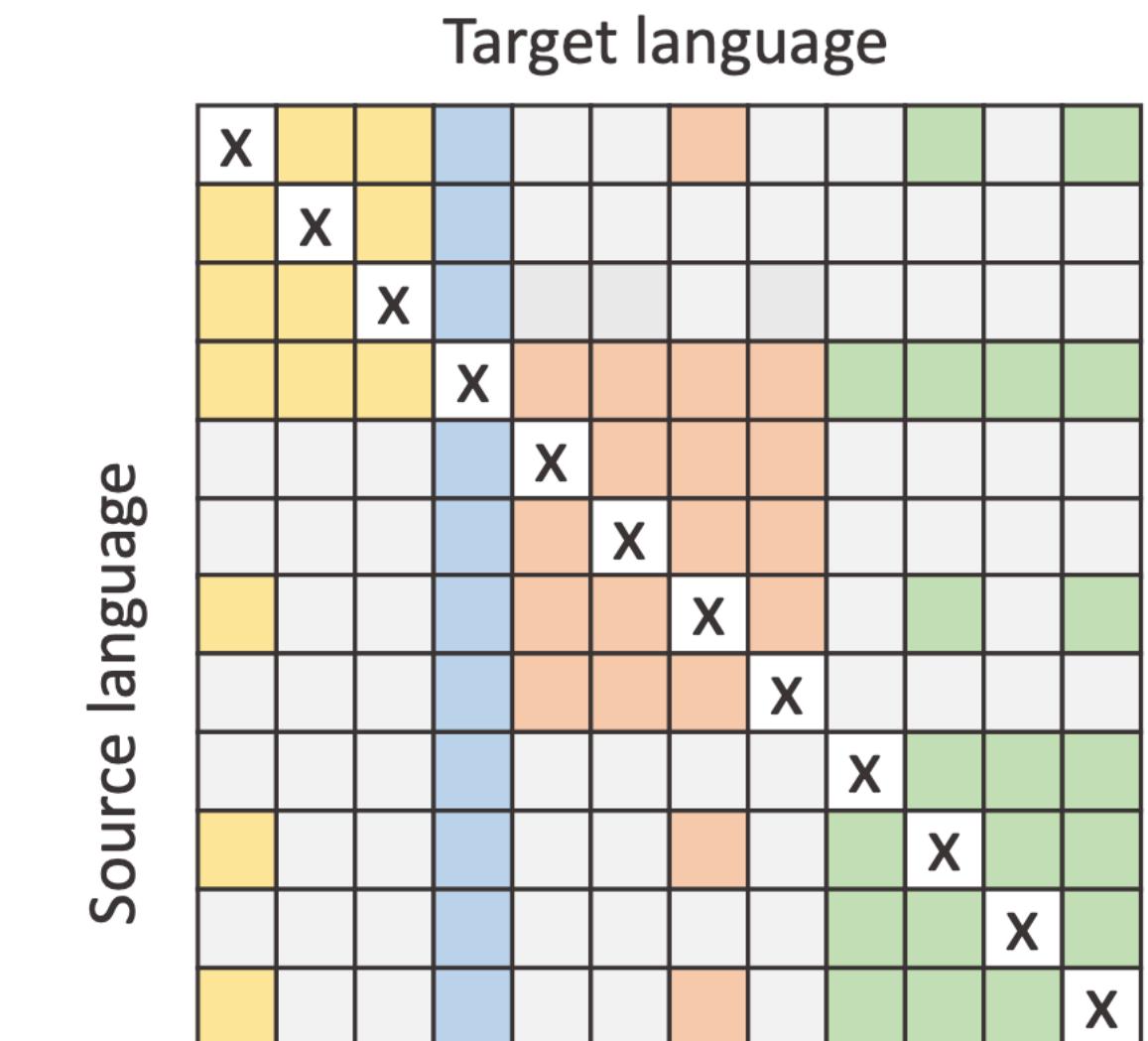
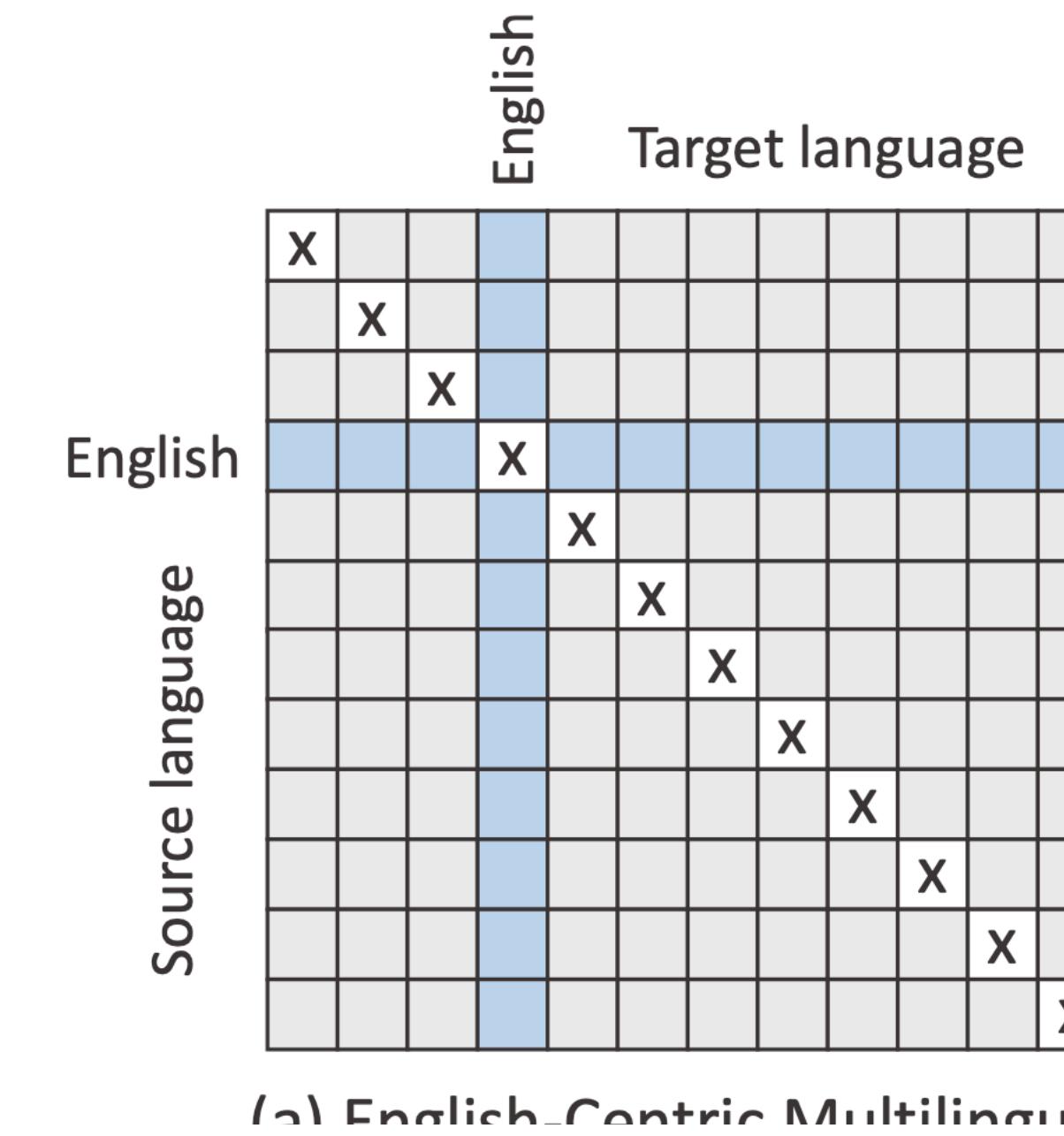
Benefits of Language-specific Subnet

- 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



100 Language 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

- CCAigned [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.

The power of non-English parallel data

- Not necessarily fair performance.

Setting	To English	From English	Non-English
Bilingual baselines	27.9	24.5	8.3
English-Centric	31.0	24.2	5.7
English-Centric with Pivot	—	—	10.4
Many-to-Many	31.2	24.1	15.9

	Source	Target	Test Set	BLEU		
				English-Centric	M2M-100	Δ
India	Hindi	Bengali	TED	3.9	8.7	+4.8
	Hindi	Marathi	TED	0.4	8.4	+8.0
	Hindi	Tamil	TED	1.1	7.5	+6.4
South Africa	Afrikaans	Xhosa	Autshumato	0.1	3.6	+3.5
	Afrikaans	Zulu	Autshumato	0.3	3.6	+3.3
	Afrikaans	Sesotho	Autshumato	0.0	2.1	+2.1
	Xhosa	Zulu	Autshumato	0.1	3.6	+3.5
	Sesotho	Zulu	Autshumato	0.1	1.2	+1.1
	Arabic	French	TED	5.3	20.8	+15.5
Chad						
DR Congo	French	Swahili	Tatoeba	1.8	5.7	+3.9
Kazakhstan	Kazakh	Russian	TED	0.5	4.5	+4.0
Singapore	Chinese	Tamil	TED	0.2	8.0	+7.8

	Source	Target	Test Set	BLEU		
				English-Centric	M2M-100	Δ
Austria	German	Croatian	TED	9.6	21.3	+11.7
	German	Hungarian	TED	11.3	17.4	+6.1
Belgium	Dutch	French	TED	16.4	25.8	+9.4
	Dutch	German	TED	18.1	26.3	+8.2
Belarus	Belarusian	Russian	TED	10.0	18.5	+8.5
Croatia	Croatian	Serbian	TED	22.4	29.8	+7.4
	Croatian	Hungarian	TED	12.4	17.5	+5.1
	Croatian	Czech	TED	15.2	22.5	+7.3
Cyprus	Croatian	Slovak	TED	13.8	24.6	+10.8
	Greek	Turkish	TED	4.8	12.6	+7.8
	Czech	Slovak	TED	9.5	28.1	+18.6
Finland	Finnish	Swedish	TED	7.9	19.2	+11.3
Italy	Italian	French	TED	18.9	28.8	+9.9
	Italian	German	TED	18.4	25.6	+7.2
Moldova	Romanian	Russian	TED	8.0	19.0	+11.0
	Romanian	Ukrainian	TED	8.7	17.3	+8.6
Montenegro	Albanian	Croatian	TED	3.0	20.7	+17.7
	Albanian	Serbian	TED	7.8	20.6	+12.8
Romania	Romanian	German	TED	15.0	24.7	+9.7
	Romanian	Hungarian	TED	11.0	16.3	+4.3
	Romanian	Turkish	TED	5.1	12.0	+6.9
	Romanian	Armenian	TED	0.4	8.2	+7.8
Russia	Bashkir	Russian	Tatoeba	0.1	4.3	+4.2

Direction	Test Set	BLEU			Δ
		Published	M2M-100	Average	
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

Language Presentation

Reading

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