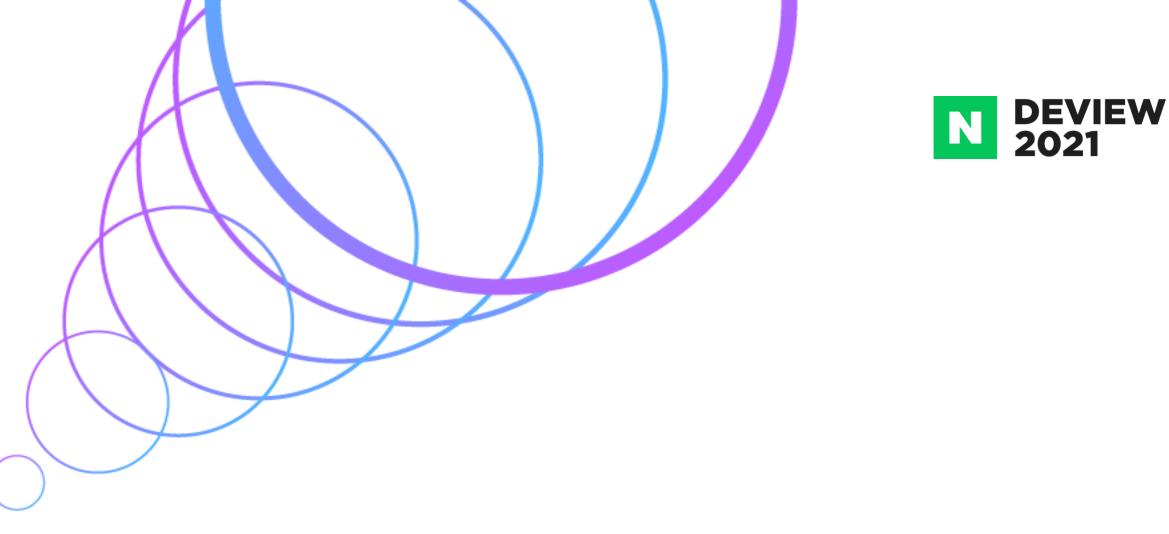
Efficient Multilingual Neural Machine Translation



papago



Alexandre Berard, Laurent Besacier, Vassilina Nikoulina NAVER LABS Europe

Other Contributors: Stephane Clinchant, Matthias Galle NAVER LABS Europe Kweonwoo Jung, Dain Lee NAVER Corp. (Papago) Asa Cooper Stickland University of Edinburgh Ahmet Ustun University of Groningen





CONTENTS

1. Introduction to multilingual neural machine translation 2. Speeding up inference 3. Efficient domain adaptation 4. Learning new languages efficiently 5. Learning new languages without parallel data



1.Introduction to multilingual neural machine translation (MNMT)

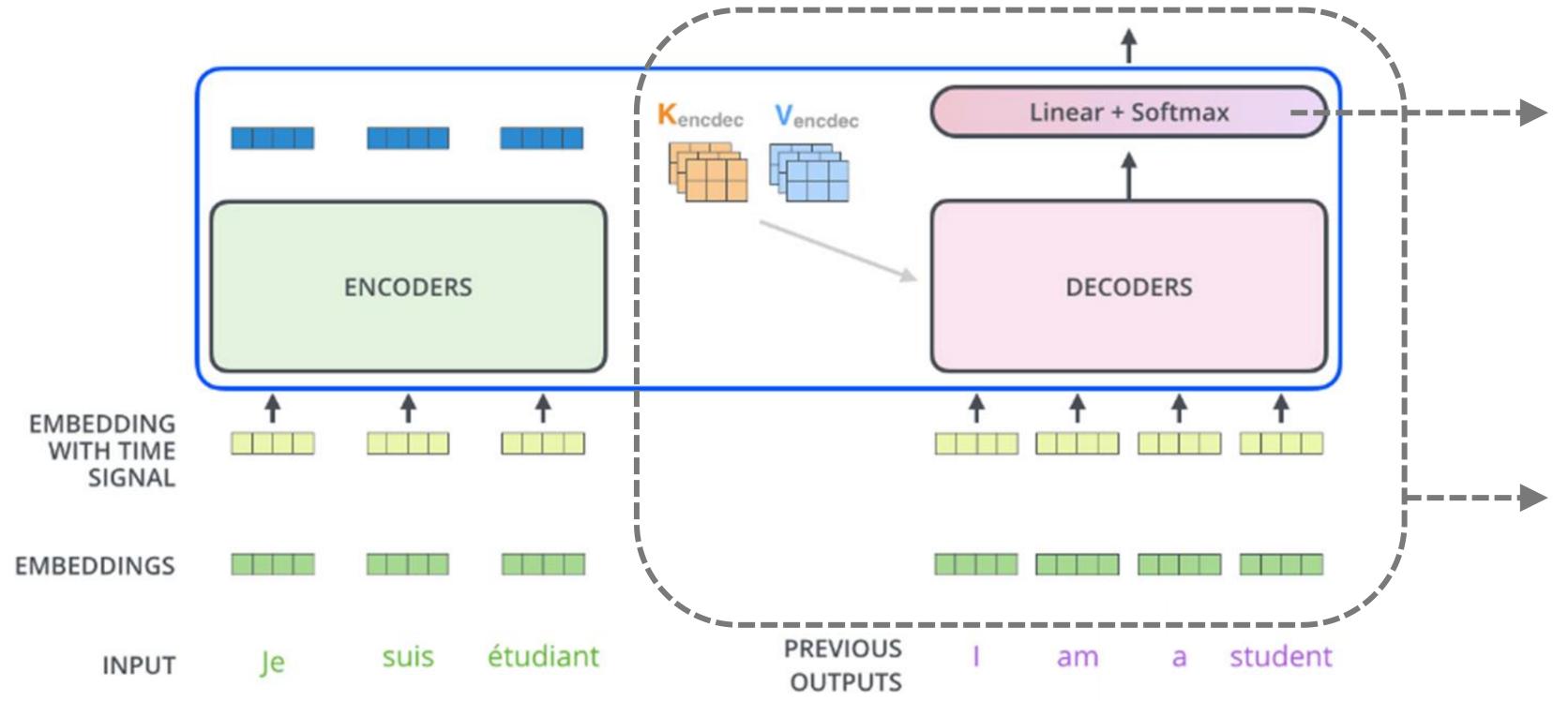




1.1 Encoder-decoder

Decoding time step: 1 2 3 4 5 6

OUTPUT



From *https://jalammar.github.io/illustrated-transformer*



am a student <end of sentence>

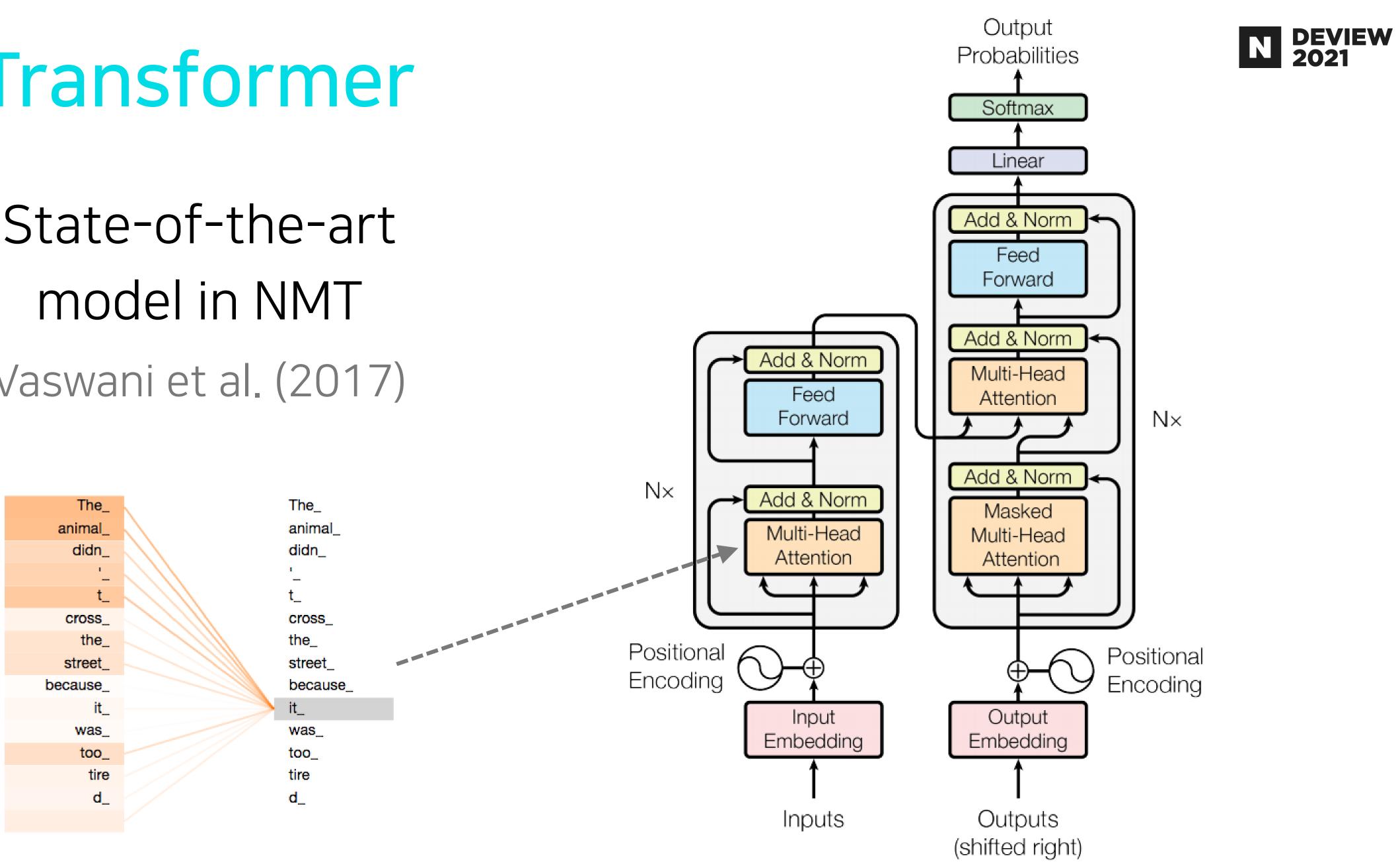
~15% of parameters Grows with vocabulary size

Most costly part Called for each new token



1.1 Transformer

State-of-the-art model in NMT Vaswani et al. (2017)





1.2 Pre-processing

Tokenize text sequences into sequences of *wordpieces* 고맙습니다. → _고 | 맙 | 습니다 | . → 147 | 1809 | 13 | 1009 Thank you. → _Thank | _you | . → 663 | 54 | 1029 Typically with the *Byte Pair Encoding* algorithm





1.2 Pre-processing

Tokenize text sequences into sequences of *wordpieces* 고맙습니다. → _고 | 맙 | 습니다 | . → 147 | 1809 | 13 | 1009 Thank you. → _Thank | _you | . → 663 | 54 | 1029 Typically with the *Byte Pair Encoding* algorithm

Vocab ID	Wordpiece
13	습니다
145	_안
147	_고
1009	
1017	하
1809	맙
1872	녕

Korean vocabulary



Vocab ID	Wordpiece	
28	_m	
38	or	
54	_you	
346	_G	
488	ood	
663	_Thank	
1029		

English vocabulary



1.3 Machine translation evaluation

BLEU (Papineni et al., 2002)

- Default evaluation metric in MT
- Based on *precision* of matched N-grams
- Not perfect: complementary evaluation is often helpful

枪手被警方击毙。

the gunman was shot to death by the police.

the gunman was police kill .
wounded police jaya of
the gunman was shot dead by the police .
the gunman arrested by police kill .
the gunmen were killed .
the gunman was shot to death by the police .



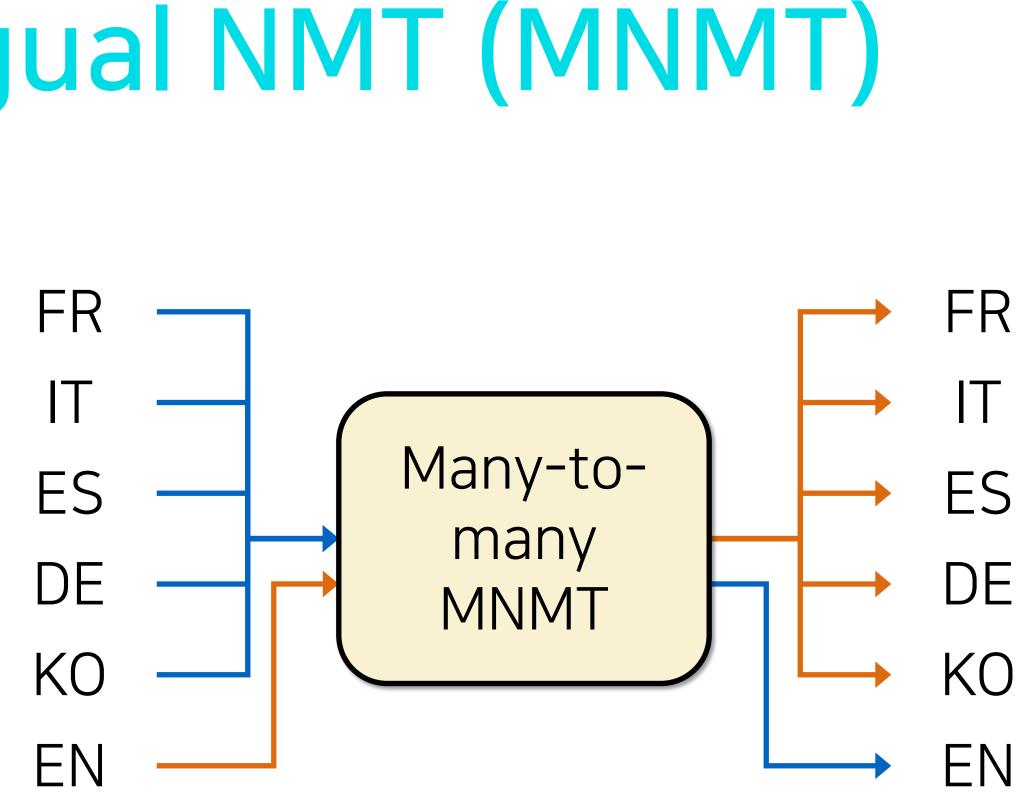
(Source Original)

(Reference Translation)

- #1
- #2
- #3
- #4
- #5
- #6 Best BLEU Score



1.4 Multilingual NMT (MNMT)

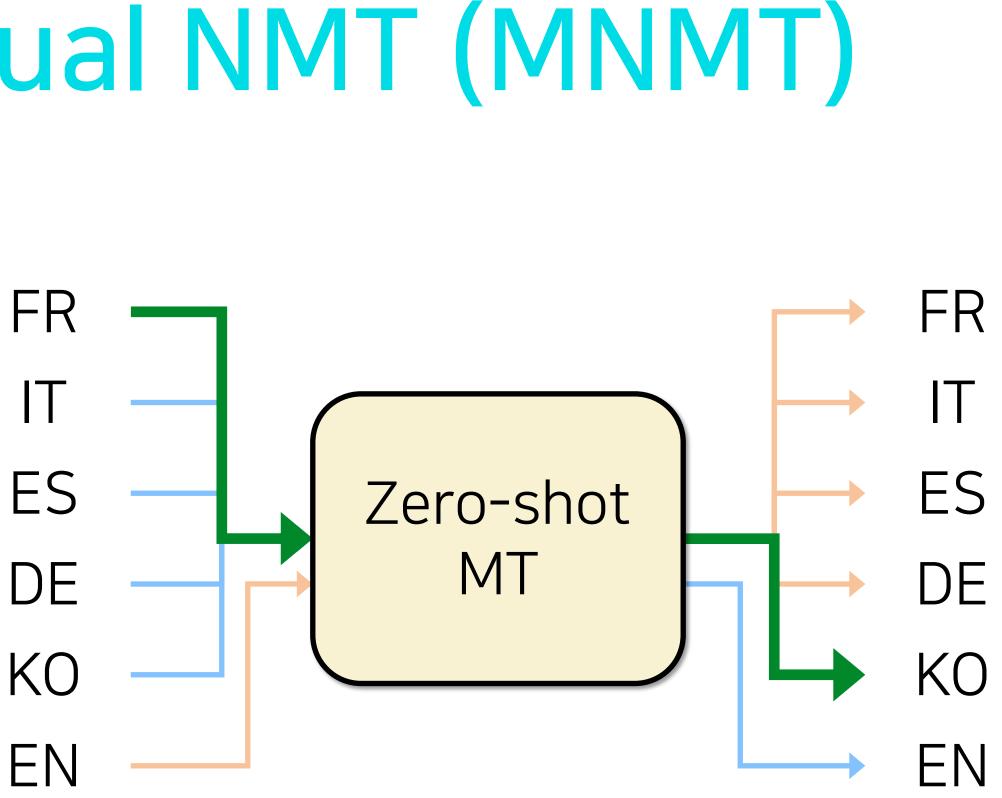


- One single model for multiple source and target languages Shared vocabulary and embeddings
- —
- Choose target language with source-side tag: <2FR>, <2IT>, etc. - Often trained on *English-centric* data: FR-EN, EN-KO, but no FR-KO





1.4 Multilingual NMT (MNMT)



Zero-shot translation

Translate in a language pair that was not seen at training But whose source and target language are known

<2KO> | _Merci | _beau | coup | . → _고 | 맙 | 습니다 | .



1.5 Challenges of MNMT

Multilingual NMT is convenient in production

- single model for all language pairs
- knowledge transfer



	 Massively multilingual with 400 million parameters 	
+5 BLEU		
Bilingual Baselines \rightarrow –	A M	A
-5 BLEU		Low Resource Language

Translation quality improvement of a single massively multilingual model as we increase the capacity (number of parameters) compared to 103 individual bilingual baselines.

> Arivazhagan et al. (2019) Google's Massively MNMT paper

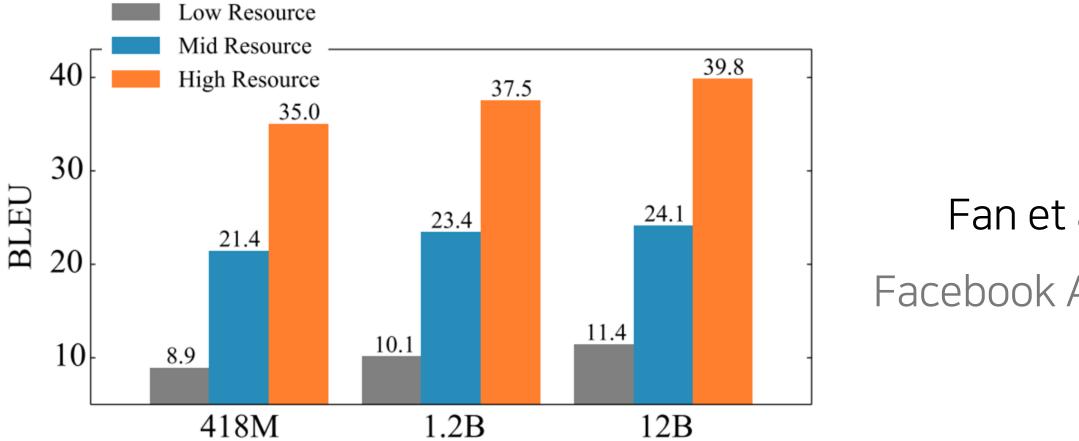




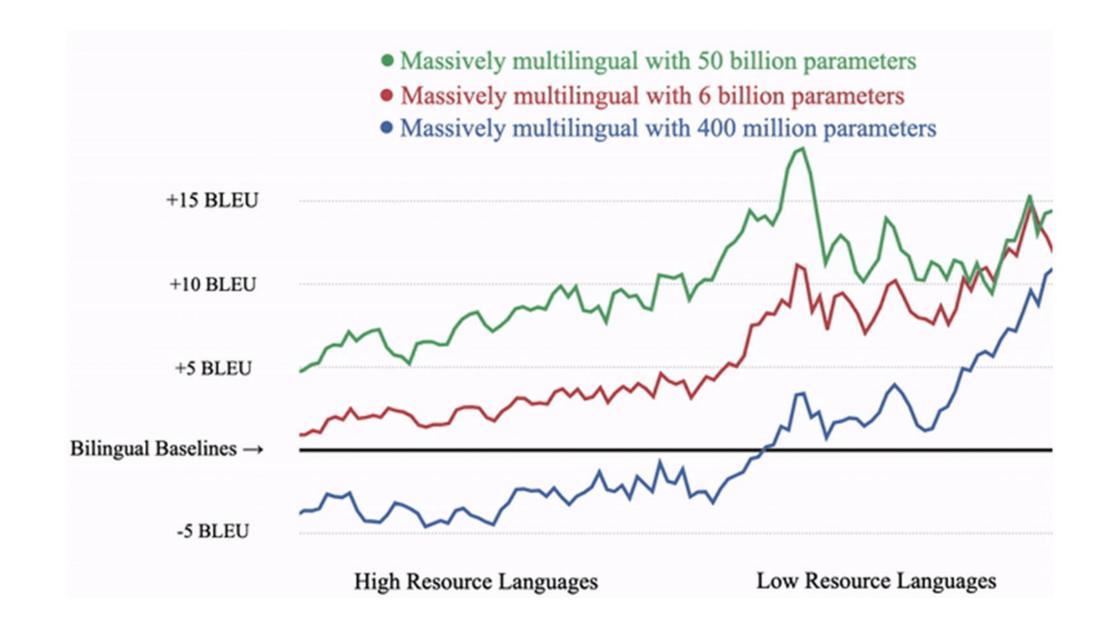
1.5 Challenges of MNMT

Multilingual NMT is convenient in production, but it requires bigger models

- slower at inference
- costly to train







Translation quality improvement of a single massively multilingual model as we increase the capacity (number of parameters) compared to 103 individual bilingual baselines.

Fan et al. (2020) Facebook Al's M2M-100

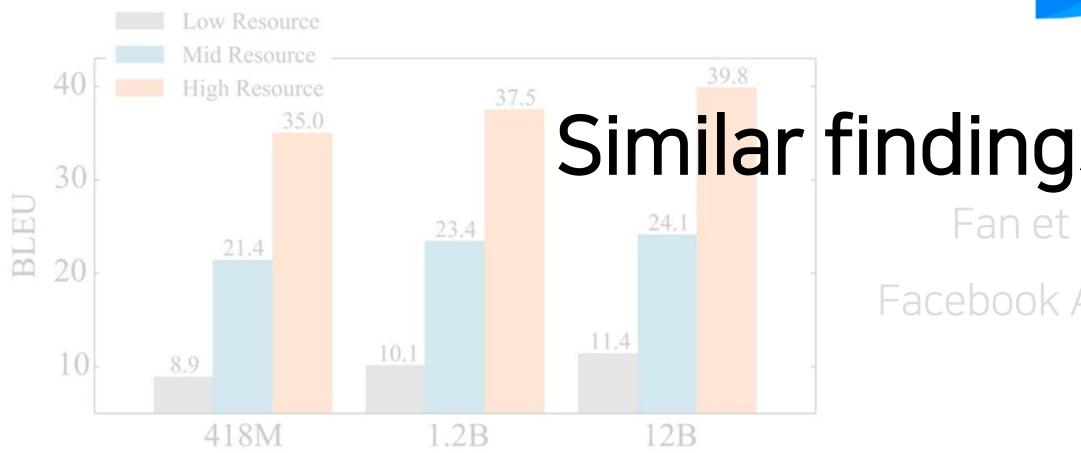
Arivazhagan et al. (2019) Google's Massively MNMT paper



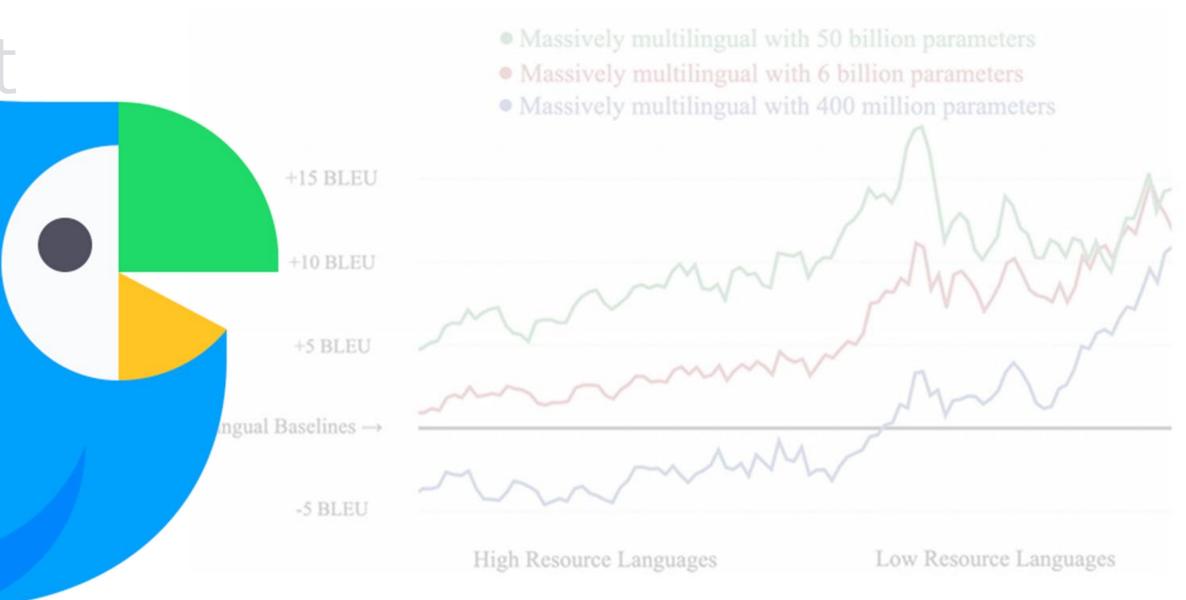
1.5 Challenges of MNMT

Multilingual NMT is convenient in production, but it requires bigger models

- slower at inference
- costly to train







Similar findings at NAVER Papago Arivazhagan et al. (2019) Fan et al. (2020) Google's Massively MNMT paper Facebook Al's M2M-100

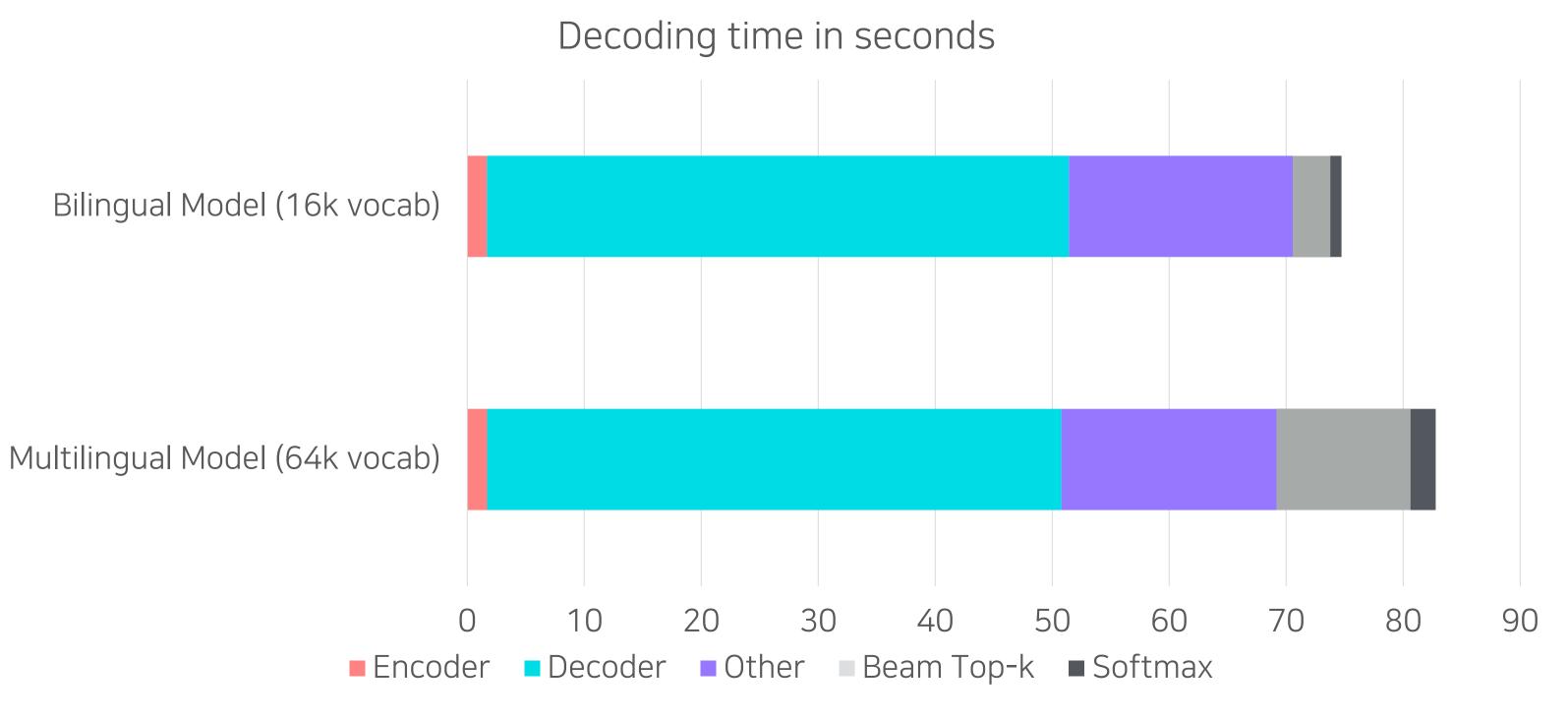


2. Speeding up inference

Efficient Inference for Multilingual Neural Machine Translation A. Berard, D. Lee, S. Clinchant, K. Jung and V. Nikoulina EMNLP 2021



2.1 Introduction



- Most time is spent in the *decoder*
- Encoder time is negligible



- Softmax and beam search times are higher for the multilingual model



2.2 Techniques

Faster decoder

- Deep encoder / shallow decoder(s)
- Hybrid model with shallow RNN decoder





2.2 Techniques

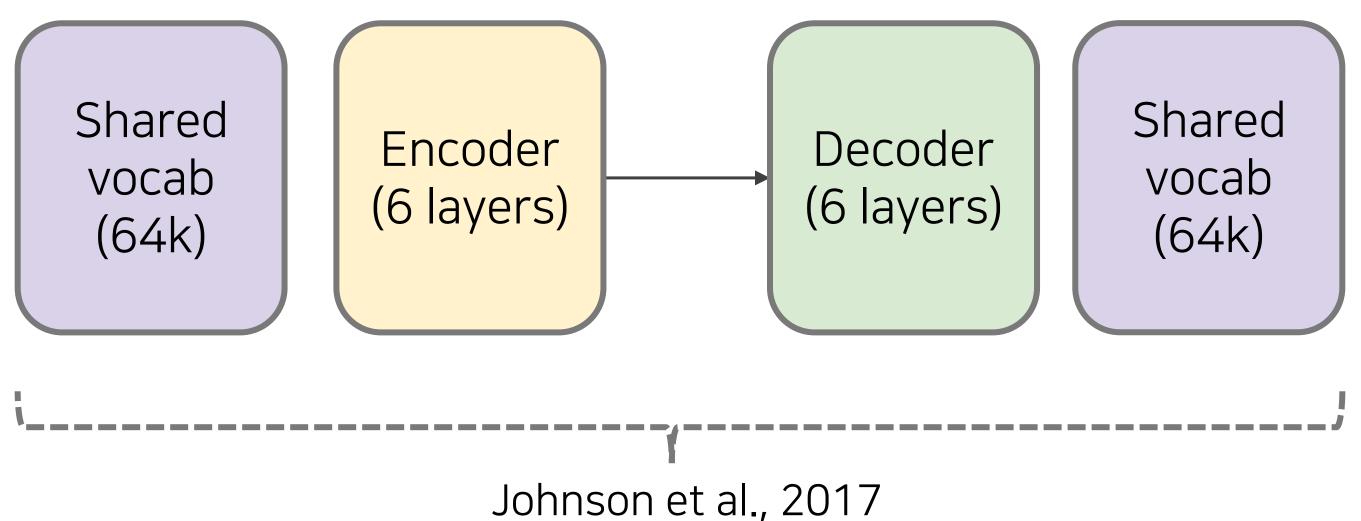
Faster decoder - Deep encoder / shallow decoder(s) - Hybrid model with shallow RNN decoder

Reducing softmax / beam search cost - Language-specific vocabulary filtering





2.2 Techniques: baseline model

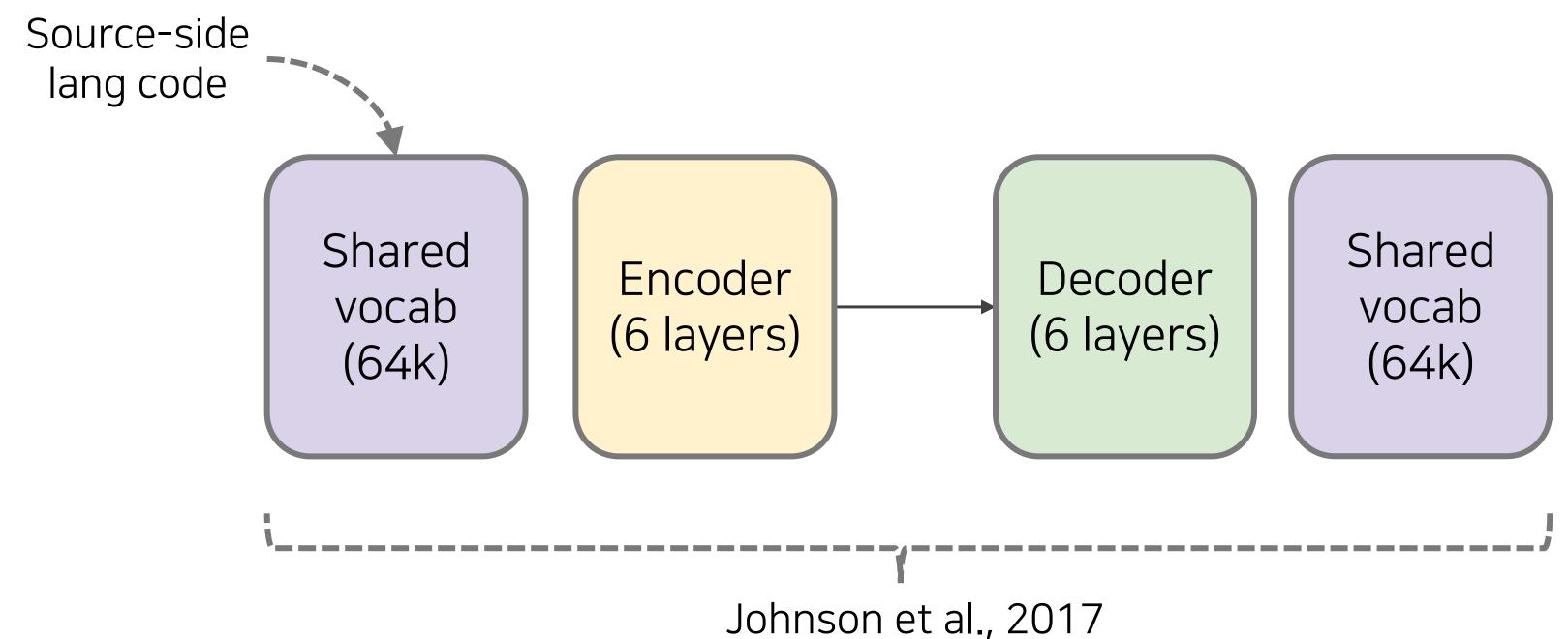








2.2 Techniques: baseline model

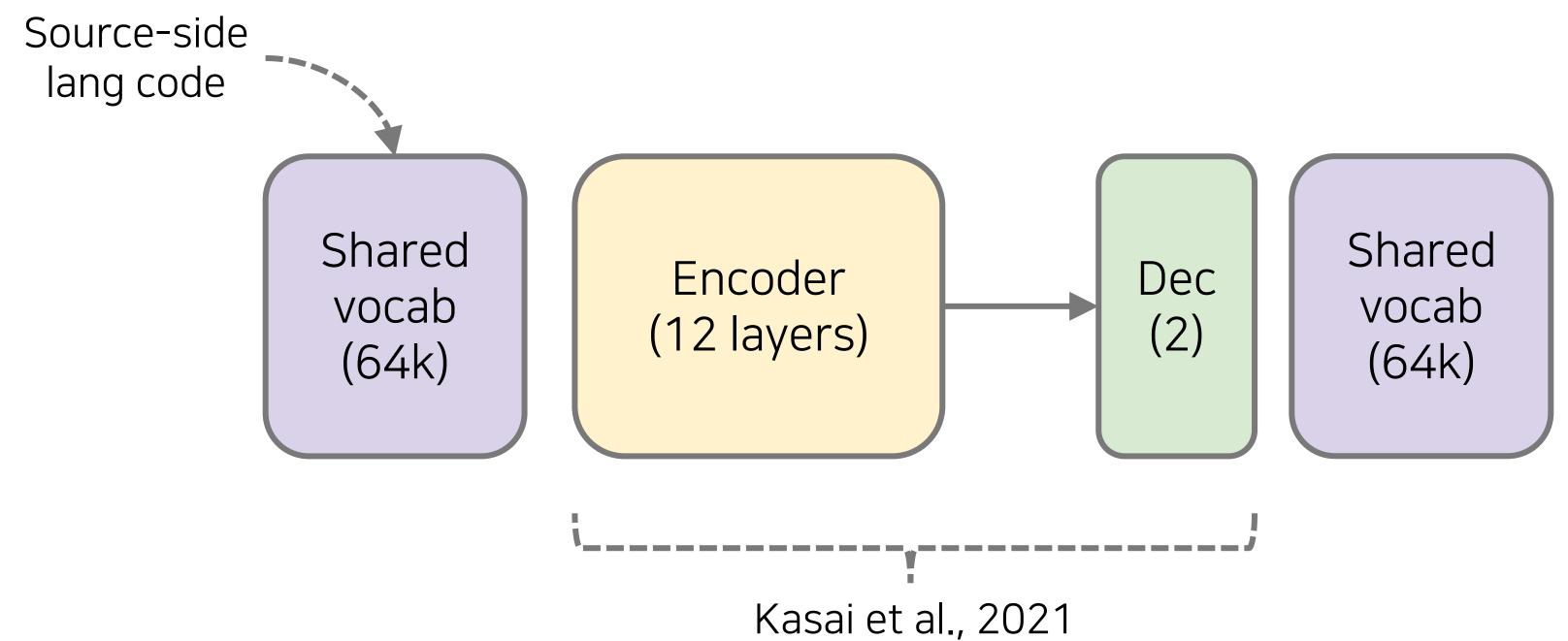








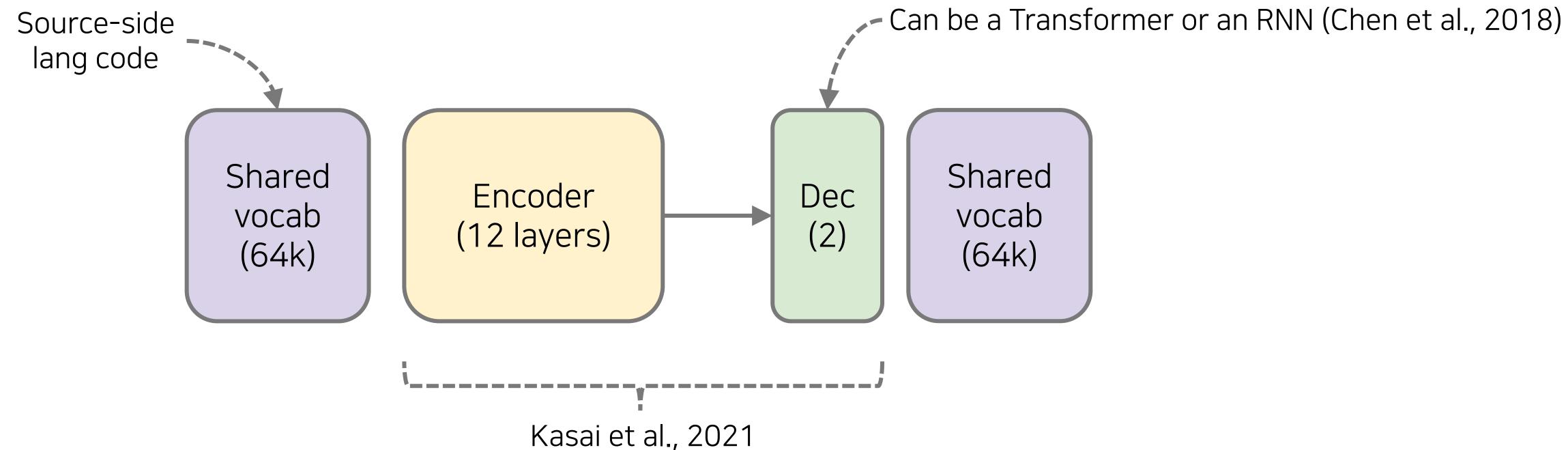
2.2 Techniques: fast MNMT model







2.2 Techniques: fast MNMT model

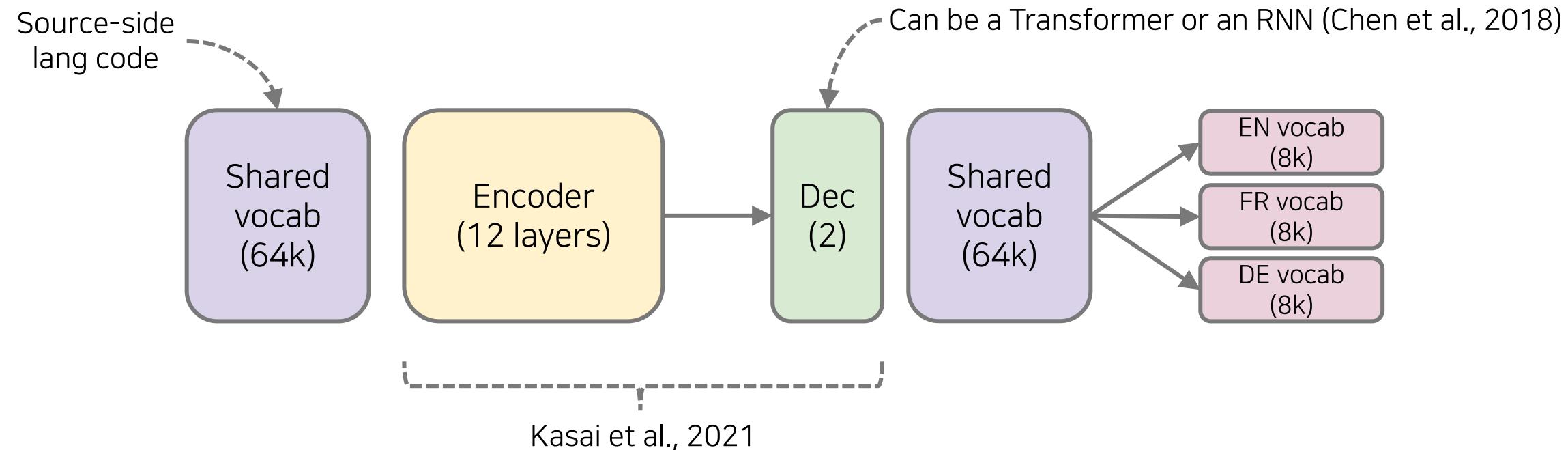








2.2 Techniques: fast MNMT model

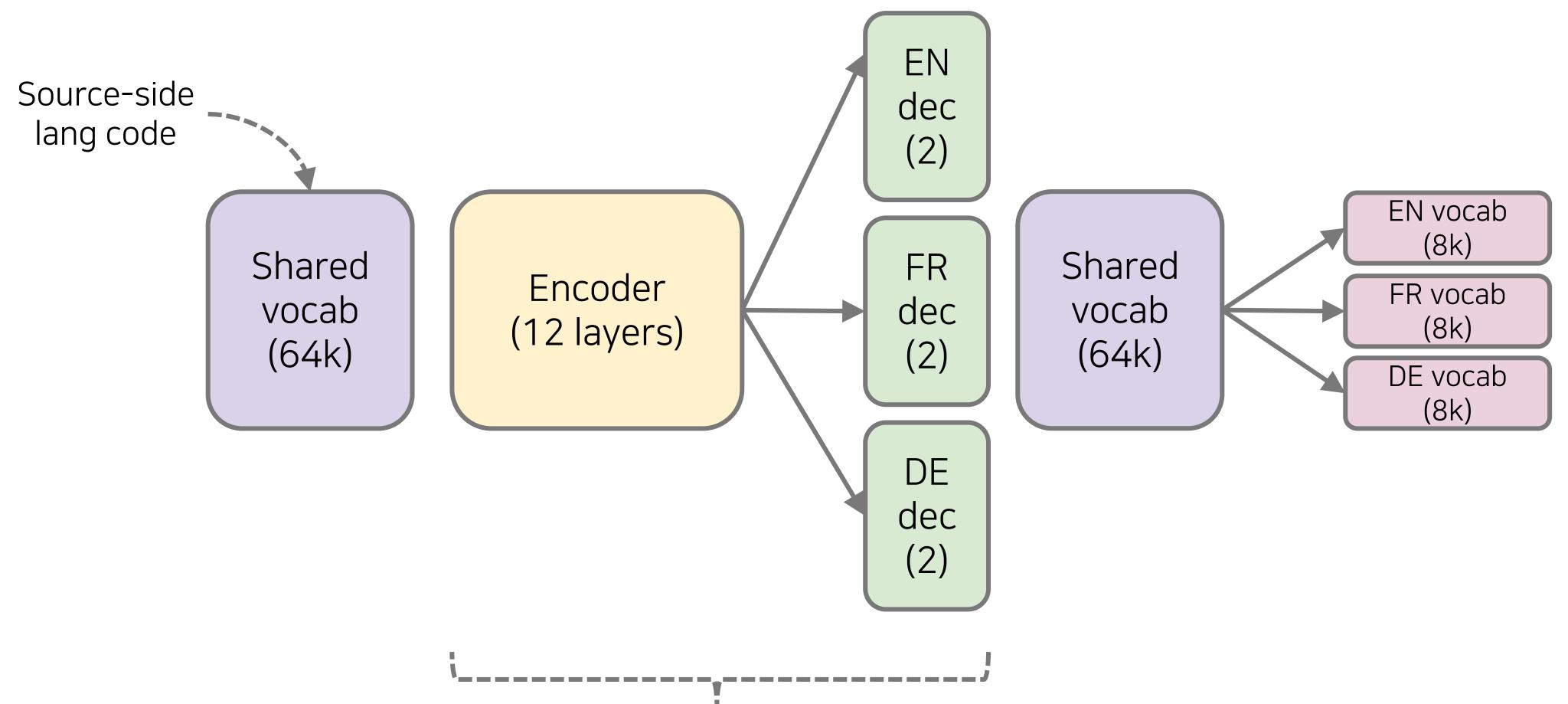








2.2 Techniques: multi-decoder model



Kong et al., 2021





2.2 Techniques: language-specific vocabulary filtering

- 1. Tokenize German training data with shared BPE
- 2. Count token frequencies
- 3. Build German-specific vocab with top 8k tokens (subset of shared vocab)





2.2 Techniques: language-specific vocabulary filtering

- 1. Tokenize German training data with shared BPE
- 2. Count token frequencies
- 3. Build German-specific vocab with top 8k tokens (subset of shared vocab)

Test-time filtering:

Filter target embedding matrix to only keep German tokens —





2.2 Techniques: language-specific vocabulary filtering

- 1. Tokenize German training data with shared BPE
- 2. Count token frequencies
- 3. Build German-specific vocab with top 8k tokens (subset of shared vocab)

Test-time filtering:

Filter target embedding matrix to only keep German tokens _

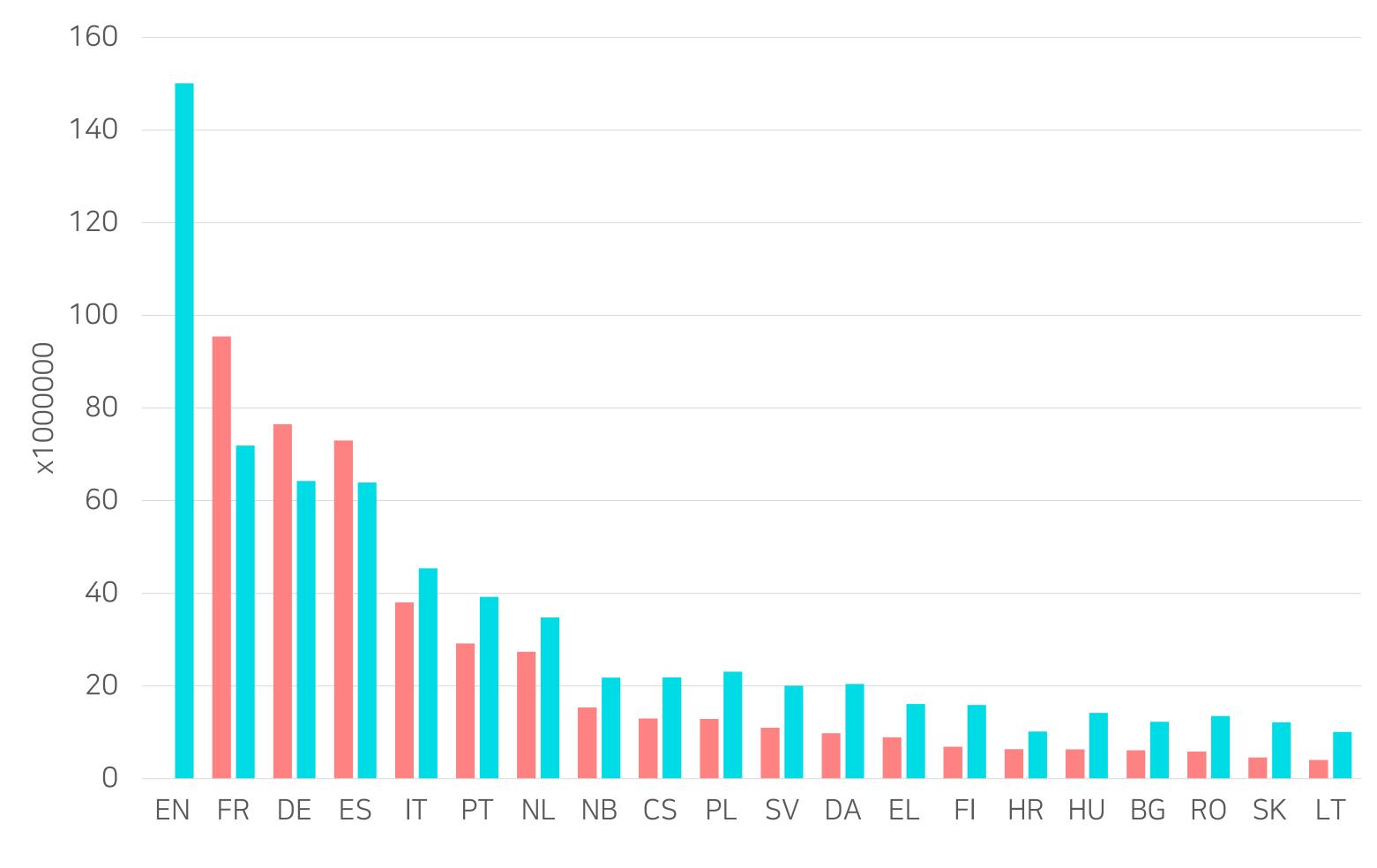
Train-time filtering:

Continue training with shared vocab, but force target tokenization to only _ generate German tokens





2.3 Experiments: ParaCrawl Top 20



6 language families, 3 scripts

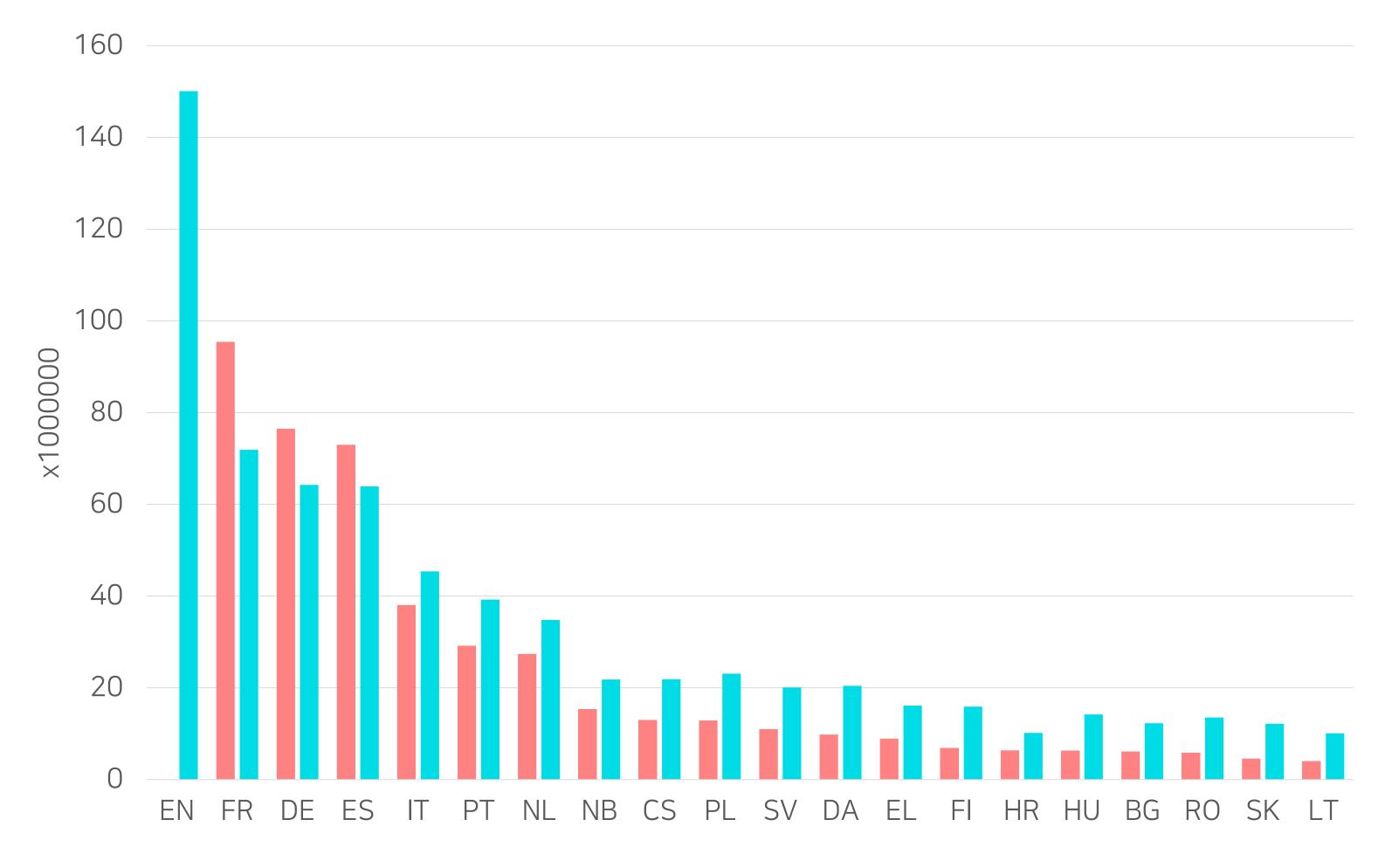


English-centric total: 900M line pairs

Multi-parallel (x3) total: 2B line pairs



2.3 Experiments: ParaCrawl Top 20



6 language families, 3 scripts Test set: FLORES (in all 380 directions)



English-centric total: 900M line pairs

Multi-parallel (x3) total: 2B line pairs



- 2-stage training: English-centric \rightarrow multi-parallel





- 2-stage training: English-centric \rightarrow multi-parallel
- Initialize 12-2 model with 6-6 model (only for TED Talks experiments)





- 2-stage training: English-centric \rightarrow multi-parallel
- Initialize 12-2 model with 6-6 model (only for TED Talks experiments) —
- Initialize Hybrid model with Transformer _





- 2-stage training: English-centric \rightarrow multi-parallel
- Initialize 12-2 model with 6-6 model (only for TED Talks experiments)
- Initialize Hybrid model with Transformer
- Initialize multi-decoder model with single-decoder model



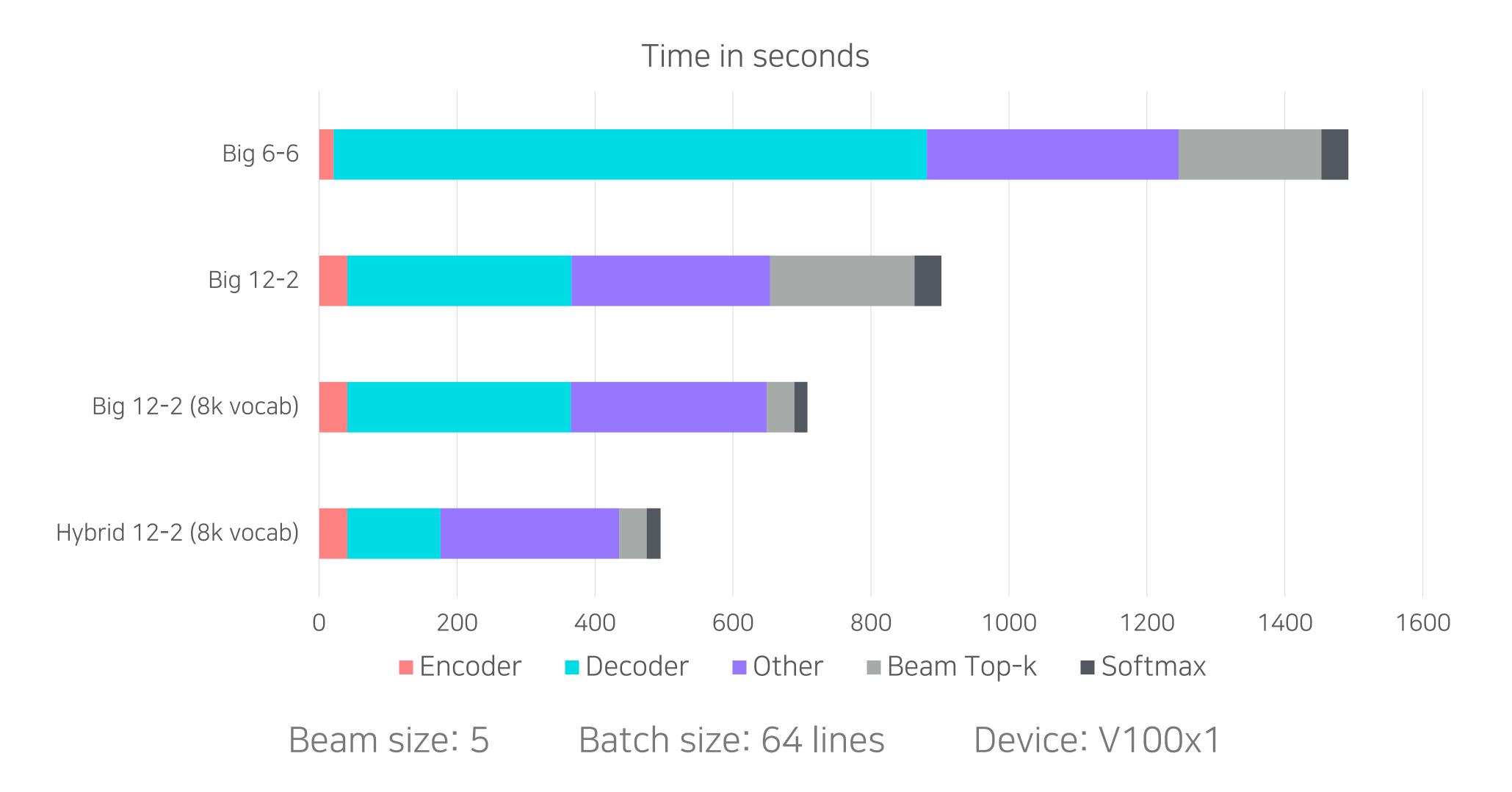


- 2-stage training: English-centric \rightarrow multi-parallel
- Initialize 12-2 model with 6-6 model (only for TED Talks experiments)
- Initialize Hybrid model with Transformer
- Initialize multi-decoder model with single-decoder model
- Language codes must be on the source side
 - For zero-shot translation
 - For the 12-2 architectures





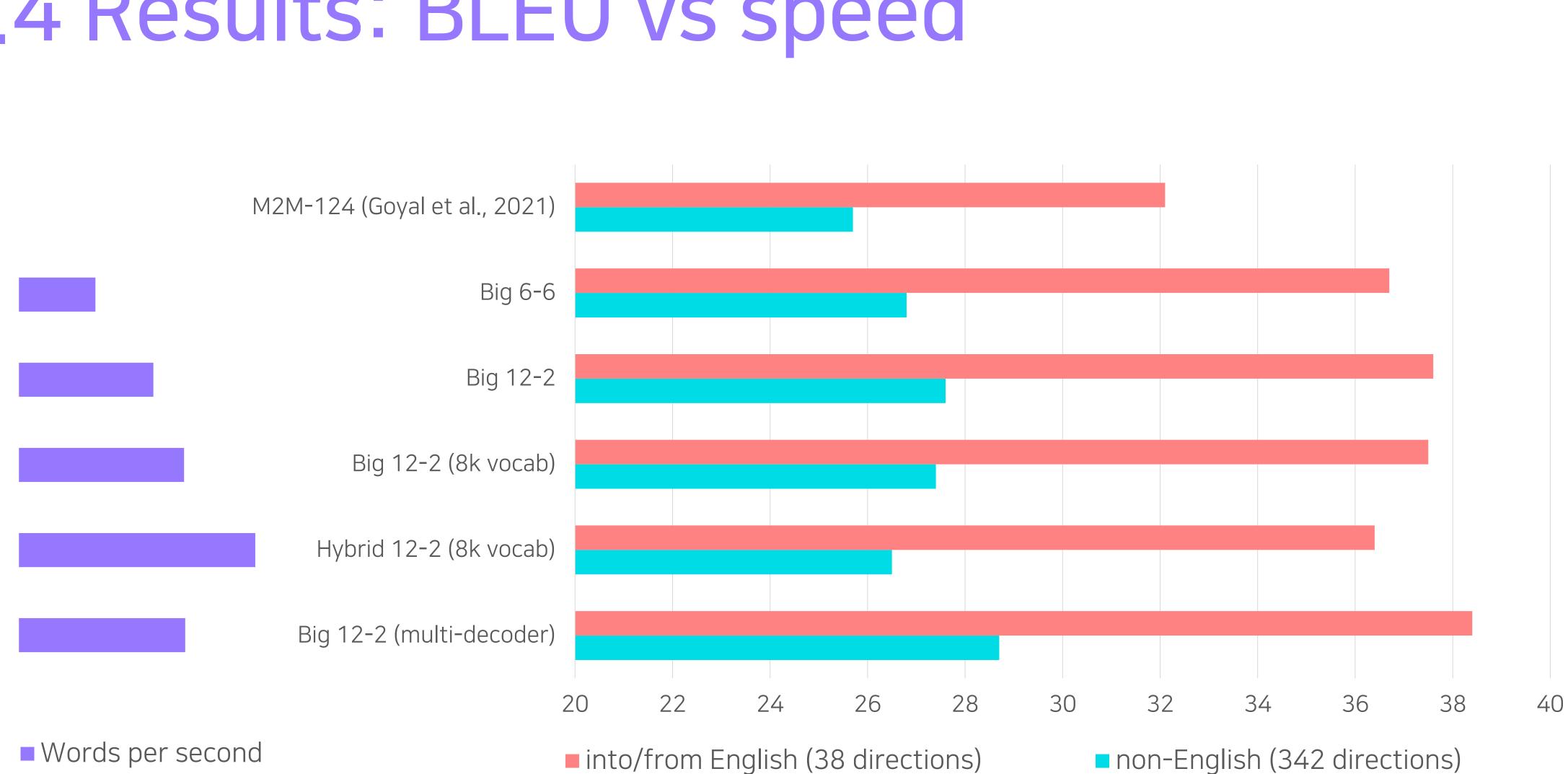
2.4 Results: inference time







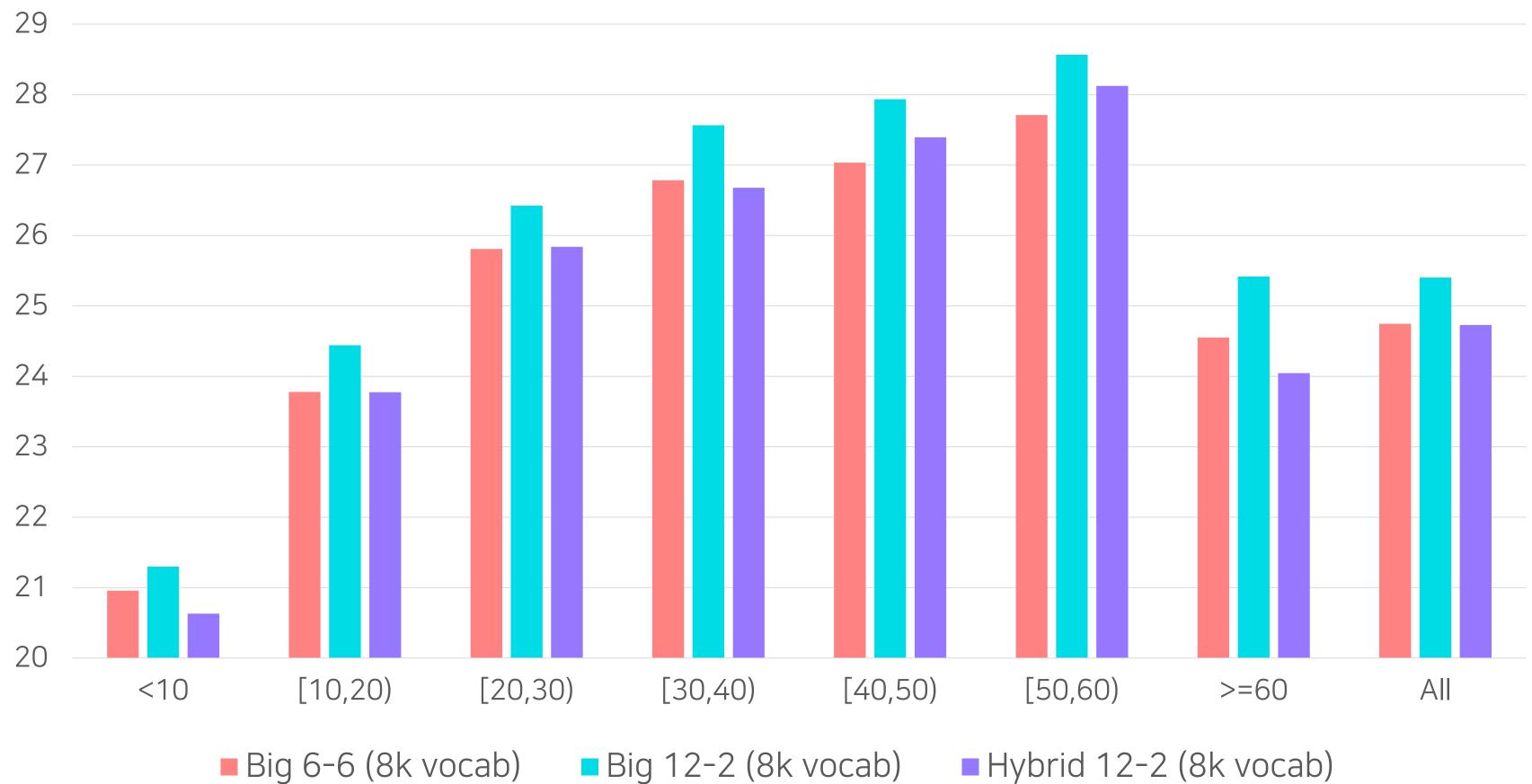
2.4 Results: BLEU vs speed



non-English (342 directions)



2.4 Results: robustness to length



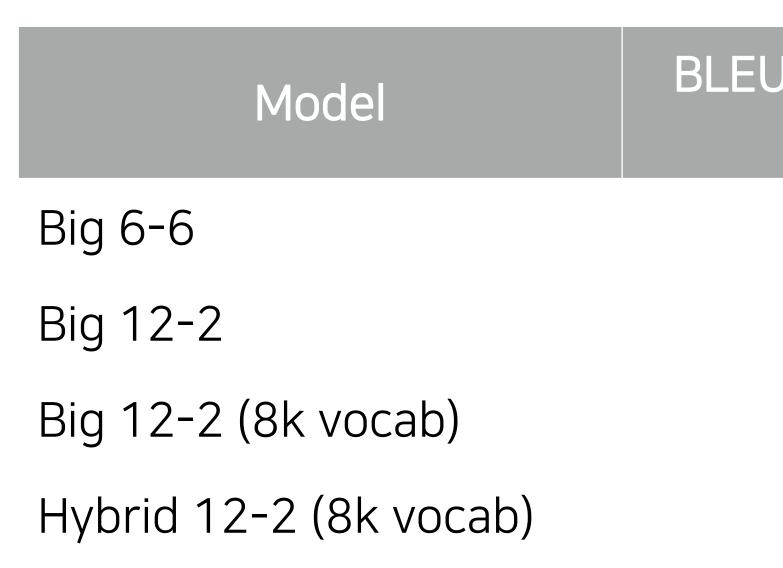


BLEU by length

Hybrid 12-2 (8k vocab)



2.4 Results: robustness to noise



- UNK: unknown symbol inserted at the beginning, middle or end
- Char: 3 random char-level operations (del, ins, sub, swap)



J consistency (UNK)	BLEU consistency (Char)
73.3	54.2
76.4	56.1
73.7	55.5
75.0	55.3

- BLEU consistency: BLEU between translations of clean and noisy inputs



2.5 Conclusion

- 12-2 > 6-6 for multilingual MT (faster and even better quality)
- Lang-specific vocab filtering improves speed
- RNN decoder: very good speed/BLEU tradeoff
- New MNMT setup on ParaCrawl



ster and even better quality) oves speed FU tradeoff



3. Efficient domain adaptation

Multilingual Domain Adaptation for NMT: Decoupling Language and Domain Information with Adapters

> A. Cooper Stickland, A. Berard and V. Nikoulina WMT 2021



A single model for many domains and languages How can we adapt an MNMT model to a new domain in a parameter-efficient way?





A single model for many domains and languages How can we adapt an MNMT model to a new domain in a parameter-efficient way?

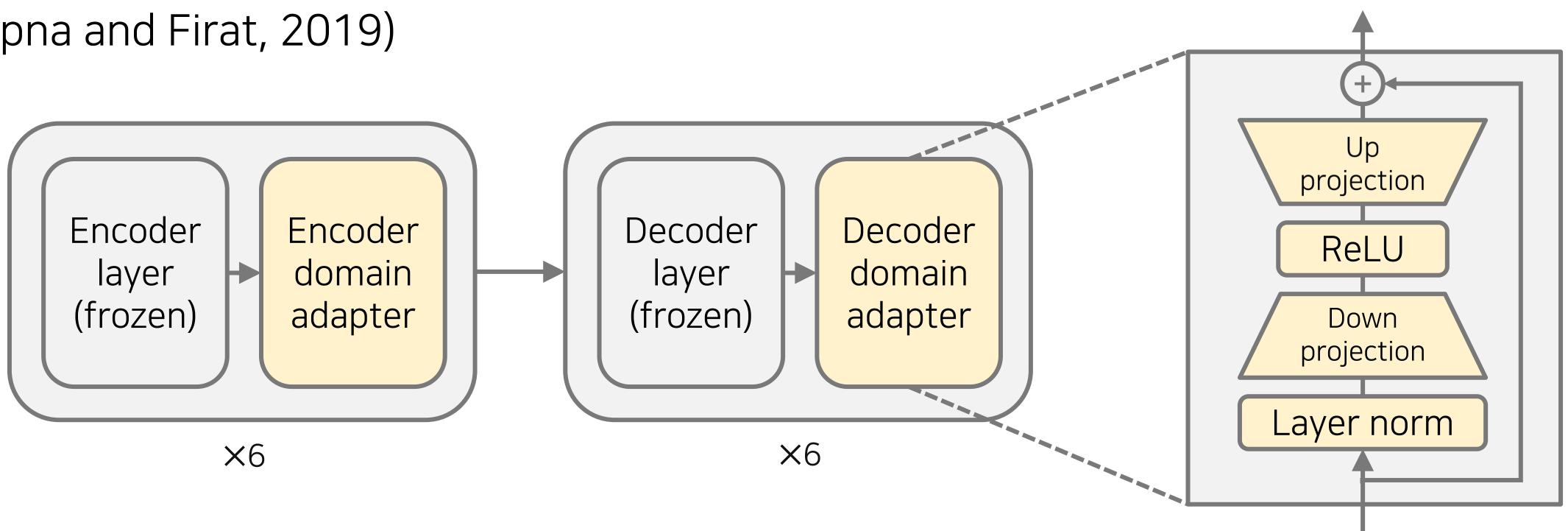
Covering high and low resource languages How can we do multilingual domain adaptation with *incomplete* in-domain data?





3.2 Adapter layers

Adapter layers are lightweight modules inserted in-between layers. They can be trained to specialize to language pairs or domains (Bapna and Firat, 2019)

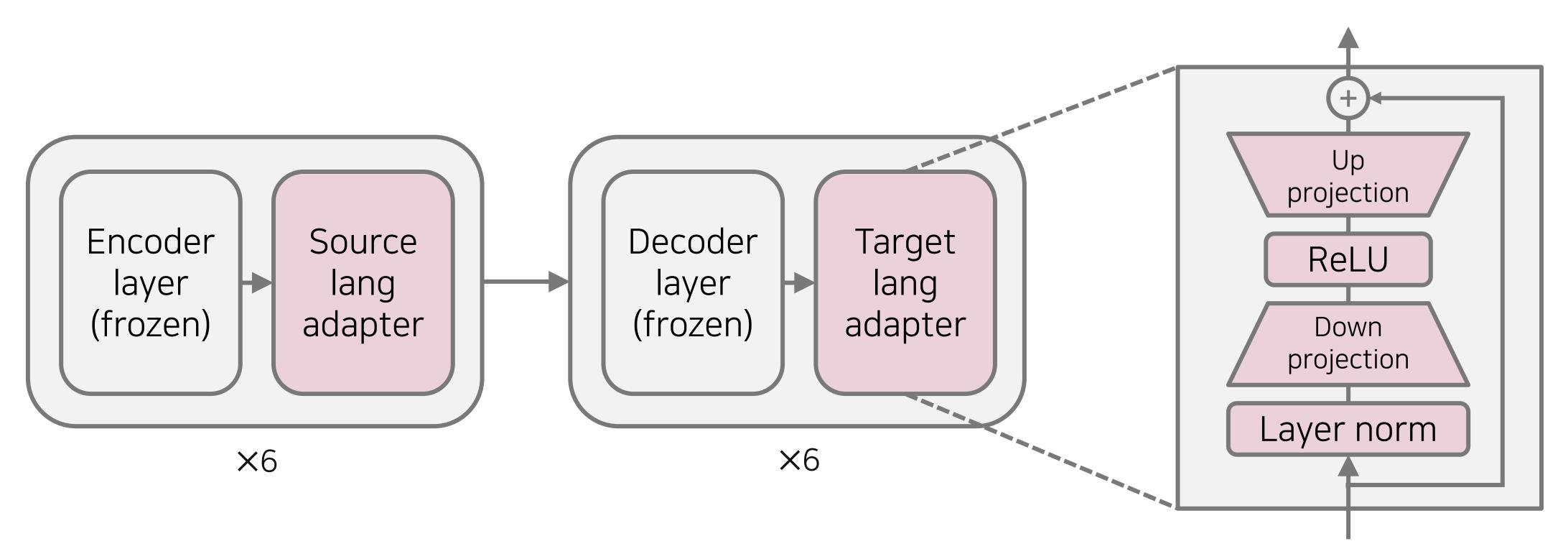






3.2 Adapter layers

zero-shot machine translation (Philip et al., 2020)



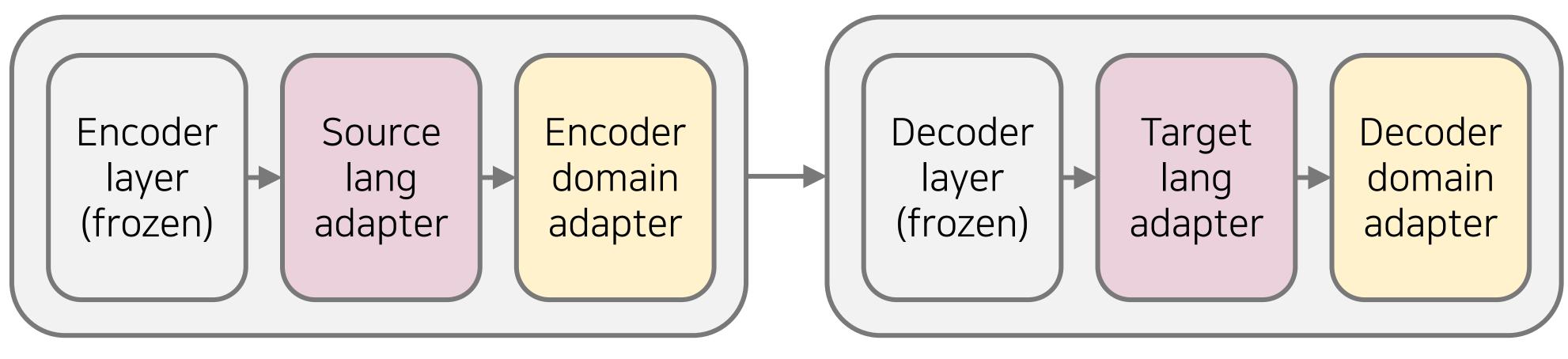


Language adapters are specific to one language and can be composed to perform



3.2 Adapter layers

Can we stack **domain** and **language adapters?** Pfeiffer et al. (2020) propose a similar approach for classification tasks.







3.3 Technique

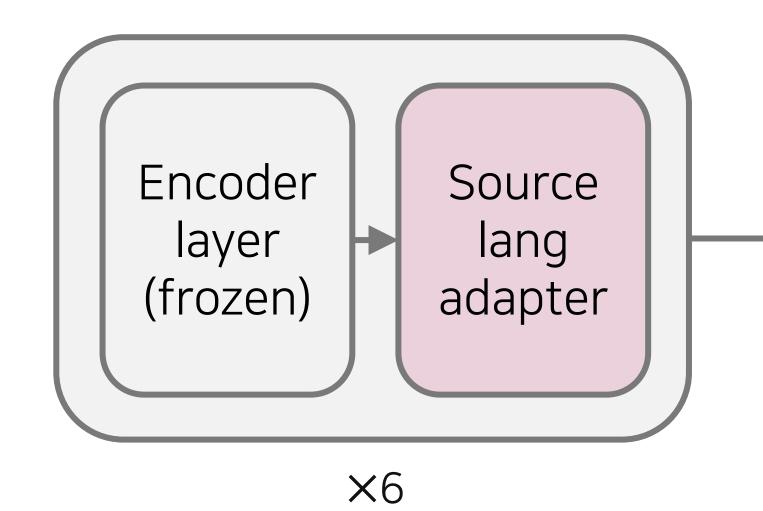


Train a baseline Transformer on English-centric ParaCrawl data



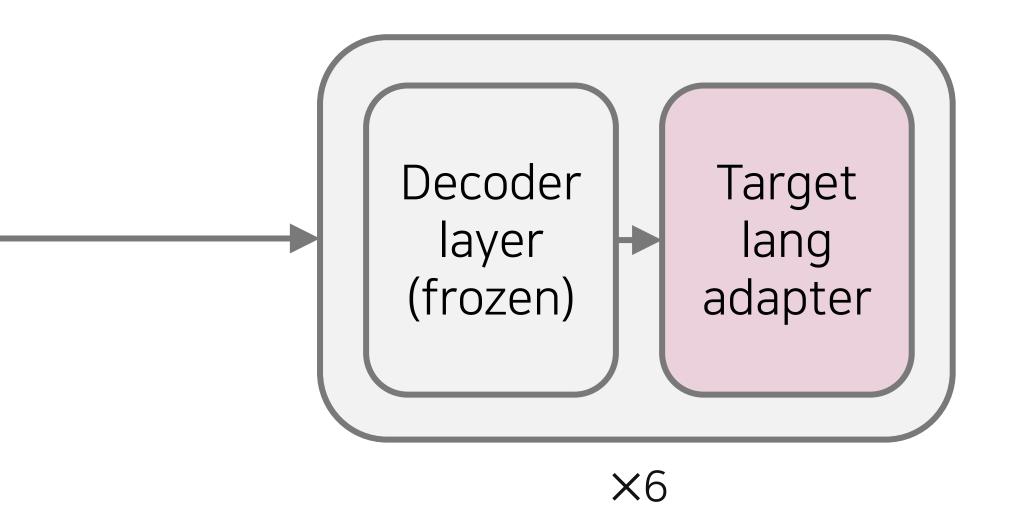


3.3 Technique



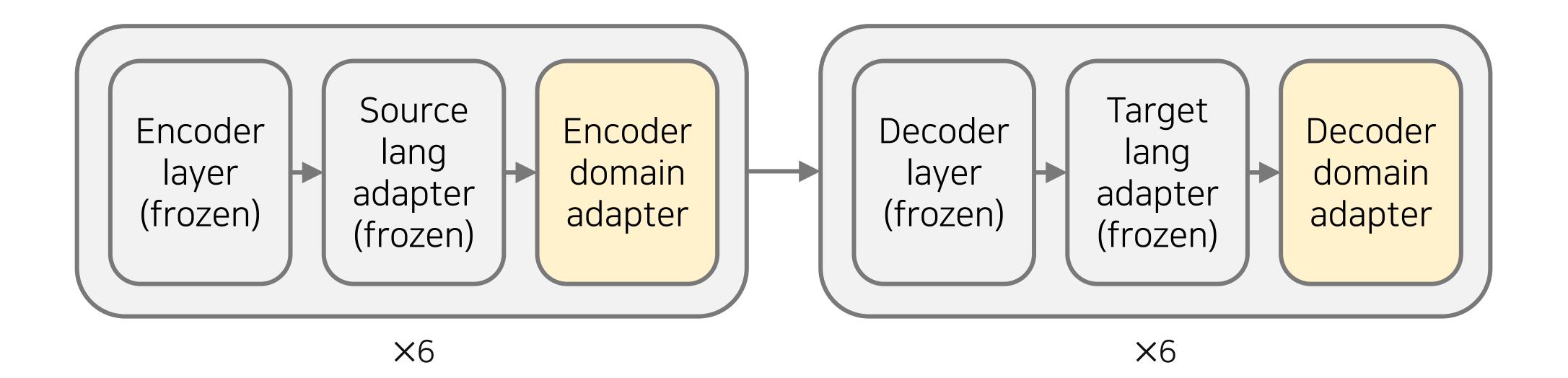
Train language adapters on multi-parallel ParaCrawl data







3.3 Technique





Train domain adapters on in-domain data



Baseline MNMT model

- Trained on ParaCrawl Top 12: {FR, DE, ES, IT, PT, NL, NO, CS, PL, SV, DA} ↔ EN
- With *language adapters* trained on multi-parallel data (12x11 language pairs)





Baseline MNMT model

- Trained on ParaCrawl Top 12: {FR, DE, ES, IT, PT, NL, NO, CS, PL, SV, DA} ↔ EN
- With *language adapters* trained on multi-parallel data (12x11 language pairs)

Domain adaptation

- In-domain data for Medical domain (+ Quran, IT and Ted Talks)
- Fine-tune on a subset of languages (EN, FR, DE, CS)
- Evaluate on all 132 language pairs



, NL, NO, CS, PL, SV, DA} ↔ EN el data (12x11 language pairs)

and Ted Talks) , CS)



Baseline MNMT model

- Trained on ParaCrawl Top 12: {FR, DE, ES, IT, PT, NL, NO, CS, PL, SV, DA} ↔ EN
- With *language adapters* trained on multi-parallel data (12x11 language pairs) _

Domain adaptation

- In-domain data for Medical domain (+ Quran, IT and Ted Talks)
- Fine-tune on a subset of languages (EN, FR, DE, CS) -
- Evaluate on all 132 language pairs

Models

- Stacking domain and language adapters (at all layers, encoder-only or decoder only) —
- (Vanilla fine-tuning and domain tags) —





Baseline MNMT model

- Trained on ParaCrawl Top 12: {FR, DE, ES, IT, PT, NL, NO, CS, PL, SV, DA} ↔ EN
- With *language adapters* trained on multi-parallel data (12x11 language pairs) —

Domain adaptation

- In-domain data for Medical domain (+ Quran, IT and Ted Talks)
- Fine-tune on a subset of languages (EN, FR, DE, CS) -
- Evaluate on all 132 language pairs _

Models

- Stacking domain and language adapters (at all layers, encoder-only or decoder only) —
- (Vanilla fine-tuning and domain tags)

Terminology

- in = languages with in-domain data (EN, FR, DE, CS)
- out = languages without in-domain data (ES, IT, PT, NL, NO, PL, SV, DA)





3.5 Vanilla adapter stacking (all layers)

en ·	0	16	16	18	-8.6	-11	-11	-13	-9.8	-11	-7.3
fr	14	0	15	16	-5	-8.3	-7.2	-11	-12	-15	-5.6
de ·	16	17	0	18	-4.1	-6.8	-4.3	-5.6	-4.4	-6	-2.7
Эœ	20	17	17	0	-10	-5	-11	-7.8	-6.8	-15	-10
🗒 da ·	6.6	6.8	4.8	5.4	0	-8	-13	-9	-7.1	-11	-7.9
⊟ da· ⊞ nl·	- 7.3	8.1	4.3	11	-6.3	0	-9	-10	-7.5	-9.4	-4.7
- so So Lelta	- 7.3	8	6.7	4.9	-11	-7.9	0	-10	-8	-12	-8.9
പ്പ് es ·	- 6.8	7.2	9.3	11	-6	-7.9	-8.5	0	-9.9	-10	-7
it ·	- 5.4	5.5	8.2	8	-7.2	-8.3	-8.5	-12	0	-11	-7.2
pt ·	8.4	6.6	9	5.1	-11	-7.8	-12	-12	-9	0	-9
pl ·	8.2	7.3	8.3	1.6	-7.2	-3.7	-8.8	-6	-5.8	-10	0
	en	fr	de	ćs	da	nl	sv	és	it	pt	pl
	CII			~	aa		54		15	P	P'
en ·	0	0	0	0	0.2	0.2	0.2	0.1	0.1	0.2	0.3
en.∘ ত_fr.∘		0 0	0 0	0 0	0.2 0.2	0.2 0.2	0.2 0.2	0.1 0.2	0.1 0.2	0.2 0.6	0.3 0.3
⊡ fr											
⊡ fr	0	0	0	0	0.2	0.2	0.2	0.2	0.2	0.6	0.3
tgt Ing ep fr	0 0 0	0 0	0 0	0 0	0.2 0.3	0.2 0.2	0.2 0.3	0.2 0.2	0.2 0.2	0.6 0.4	0.3 0.3
tgt Ing ep fr	0 0 0	0 0 0	0 0 0	0 0 0	0.2 0.3 0.6	0.2 0.2 0.2	0.2 0.3 0.3	0.2 0.2 0.2	0.2 0.2 0.2	0.6 <mark>0.4</mark> 0.7	0.3 0.3 0.3
vrong tgt Ing so tgt Ing v so v v sv	0 0 0 0.1	0 0 0 0	0 0 0 0.1	0 0 0 0.1	0.2 0.3 0.6 0	0.2 0.2 0.2 0.2	0.2 0.3 0.3 0.2	0.2 0.2 0.2 0.2	0.2 0.2 0.2 0.2	0.6 0.4 0.7 0.3	0.3 0.3 0.3 0.3
of wrong tgt Ing of wrong tgt Ing so so so so so so so	0 0 0 0.1 0	0 0 0 0	0 0 0 0.1 0	0 0 0.1 0	0.2 0.3 0.6 0 0.2	0.2 0.2 0.2 0.2 0	0.2 0.3 0.3 0.2 0.2	0.2 0.2 0.2 0.2 0.2	0.2 0.2 0.2 0.2 0.2	0.6 0.4 0.7 0.3 0.3	0.3 0.3 0.3 0.3 0.2
of wrong tgt Ing of wrong tgt Ing so so so so so so so	0 0 0.1 0	0 0 0 0 0	0 0 0.1 0	0 0 0.1 0	0.2 0.3 0.6 0.2 0.2	0.2 0.2 0.2 0.2 0.2	0.2 0.3 0.2 0.2 0.2	0.2 0.2 0.2 0.2 0.2 0.2	0.2 0.2 0.2 0.2 0.2 0.2	0.6 0.4 0.7 0.3 0.3 0.3	0.3 0.3 0.3 0.3 0.2 0.3
atio of wrong tgt Ing in Sa su po so ap it	0 0 0.1 0 0	0 0 0 0 0	0 0 0.1 0 0	0 0 0.1 0 0	0.2 0.3 0.6 0.2 0.2 0.2	0.2 0.2 0.2 0.2 0.2 0.2 0.2	0.2 0.3 0.2 0.2 0.2 0.2	0.2 0.2 0.2 0.2 0.2 0.2 0.2	0.2 0.2 0.2 0.2 0.2 0.2 0.2	0.6 0.4 0.7 0.3 0.3 0.3 0.2	0.3 0.3 0.3 0.2 0.3 0.3 0.3
ratio of wrong tgt lng td n s s in b t n s s s o t n s s s in b	0 0 0.1 0 0 0 0	0 0 0 0 0 0	0 0 0.1 0 0 0 0	0 0 0.1 0 0 0	0.2 0.3 0.6 0.2 0.2 0.2 0.2	0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2	0.2 0.3 0.2 0.2 0.2 0.2 0.2	0.2 0.2 0.2 0.2 0.2 0.2 0.2	0.2 0.2 0.2 0.2 0.2 0.2 0.1 0.1	0.6 0.4 0.7 0.3 0.3 0.3 0.2 0.2	0.3 0.3 0.3 0.2 0.3 0.3 0.3 0.3
ratio of wrong tgt lng td i so so n b t i so so n b t i so so n t i t i t i t i t i t i t i t i t i t i	0 0 0.1 0 0 0 0 0	0 0 0 0 0 0 0	0 0 0.1 0 0 0 0 0	0 0 0.1 0 0 0 0 0	0.2 0.3 0.6 0.2 0.2 0.2 0.2 0.2	0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2	0.2 0.3 0.2 0.2 0.2 0.2 0.2 0.2 0.2	0.2 0.2 0.2 0.2 0.2 0.2 0.1 0.1	0.2 0.2 0.2 0.2 0.2 0.2 0.1 0.1 0	0.6 0.4 0.7 0.3 0.3 0.2 0.2 0.2	0.3 0.3 0.3 0.3 0.2 0.3 0.3 0.3 0.3

Freeze LA + enc. & dec. DA



- Big improvements for in-in language pairs



N DEVIEW 2021 3.5 Vanilla adapter stacking (all layers)

	en -	0	16	16	18	-8.6	-11	-11	-13	-9.8	-11	-7.3
	fr -	14	0	15	16	-5	-8.3	-7.2	-11	-12	-15	-5.6
	de -	16	17	0	18	-4.1	-6.8	-4.3	-5.6	-4.4	-6	-2.7
ŝ	cs -	20	17	17	0	-10	-5	-11	-7.8	-6.8	-15	-10
	da -	6.6	6.8	4.8	5.4	0	-8	-13	-9	-7.1	-11	-7.9
Delta(BLE	nl -	7.3	8.1	4.3	11	-6.3	0	-9	-10	-7.5	-9.4	-4.7
lta	sv -	7.3	8	6.7	4.9	-11	-7.9	0	-10	-8	-12	-8.9
മ്	es -	6.8	7.2	9.3	11	-6	-7.9	-8.5	0	-9.9	-10	-7
	it -	5.4	5.5	8.2	8	-7.2	-8.3	-8.5	-12	0	-11	-7.2
	pt -	8.4	6.6	9	5.1	-11	-7.8	-12	-12	-9	0	-9
	pl -	8.2	7.3	8.3	1.6	-7.2	-3.7	-8.8	-6	-5.8	-10	0
		en	fr	de	Ś	da	nl	sv	és	it	pt	pl
											I	1
	en -	0	0	0	0	0.2	0.2	0.2	0.1	0.1	0.2	0.3
ğ	en - fr -		0 0	0 0	0 0	0.2 0.2	0.2 0.2	0.2 0.2	0.1 0.2	0.1 0.2	0.2 0.6	0.3 0.3
lng	fr -											
		0	0	0	0	0.2	0.2	0.2	0.2	0.2	0.6	0.3
tgt	fr - de -	0 0 0	0 0	0 0	0 0	0.2 0.3	0.2 0.2	0.2 0.3	0.2 0.2	0.2 0.2	0.6 0.4	0.3 0.3
tgt	fr - de - cs -	0 0 0.1	0 0 0	0 0 0	0 0 0	0.2 0.3 0.6	0.2 0.2 0.2	0.2 0.3 0.3	0.2 0.2 0.2	0.2 0.2 0.2	0.6 0.4 0.7	0.3 0.3 0.3
tgt	fr - de - cs - da -	0 0 0.1 0	0 0 0 0	0 0 0 0.1	0 0 0 0.1	0.2 0.3 0.6 0	0.2 0.2 0.2 0.2	0.2 0.3 0.3 0.2	0.2 0.2 0.2 0.2	0.2 0.2 0.2 0.2	0.6 0.4 0.7 0.3	0.3 0.3 0.3 0.3
of wrong tgt	fr - de - cs - da - nl -	0 0 0.1 0	0 0 0 0	0 0 0.1 0	0 0 0.1 0	0.2 0.3 0.6 0 0.2	0.2 0.2 0.2 0.2 0	0.2 0.3 0.3 0.2 0.2	0.2 0.2 0.2 0.2 0.2	0.2 0.2 0.2 0.2 0.2	0.6 0.4 0.7 0.3 0.3	0.3 0.3 0.3 0.3 0.2
of wrong tgt	fr - de - cs - da - nl - sv -	0 0 0.1 0 0	0 0 0 0 0	0 0 0.1 0	0 0 0.1 0	0.2 0.3 0.6 0.2 0.2	0.2 0.2 0.2 0.2 0.2	0.2 0.3 0.2 0.2 0.2	0.2 0.2 0.2 0.2 0.2 0.2	0.2 0.2 0.2 0.2 0.2 0.2	0.6 0.4 0.7 0.3 0.3 0.3	0.3 0.3 0.3 0.3 0.2 0.3
tgt	fr- de- cs- da- nl- sv- es- it-	0 0 0.1 0 0 0 0	0 0 0 0 0	0 0 0.1 0 0	0 0 0.1 0 0	0.2 0.3 0.6 0.2 0.2 0.2	0.2 0.2 0.2 0.2 0.2 0.2 0.2	0.2 0.3 0.2 0.2 0.2 0.2	0.2 0.2 0.2 0.2 0.2 0.2 0.2	0.2 0.2 0.2 0.2 0.2 0.2 0.2	0.6 0.4 0.7 0.3 0.3 0.3 0.3	0.3 0.3 0.3 0.2 0.3 0.3
of wrong tgt	fr- de- cs- da- nl- sv- es- it-	0 0 0.1 0 0 0 0	0 0 0 0 0 0	0 0 0.1 0 0 0	0 0 0.1 0 0 0 0	0.2 0.3 0.6 0.2 0.2 0.2 0.2	0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2	0.2 0.3 0.2 0.2 0.2 0.2 0.2 0.2	0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2	0.2 0.2 0.2 0.2 0.2 0.2 0.1 0.1	0.6 0.4 0.7 0.3 0.3 0.2 0.2	0.3 0.3 0.3 0.2 0.3 0.3 0.3 0.3

Freeze LA + enc. & dec. DA

- Big improvements for in-in language pairs
- Some improvements for out-in language pairs —



N DEVIEW 2021 3.5 Vanilla adapter stacking (all layers)

en	- 0	16	16	18	-8.6	-11	-11	-13	-9.8	-11	-7.3
fr	14	0	15	16	-5	-8.3	-7.2	-11	-12	-15	-5.6
de	16	17	0	18	-4.1	-6.8	-4.3	-5.6	-4.4	-6	-2.7
) S	20	17	17	0	-10	-5	-11	-7.8	-6.8	-15	-10
🗒 da 🖞	6.6	6.8	4.8	5.4	0	-8	-13	-9	-7.1	-11	-7.9
🖲 nl -	7.3	8.1	4.3	11	-6.3	0	-9	-10	-7.5	-9.4	-4.7
Delta (BLE s s es	7.3	8	6.7	4.9	-11	-7.9	0	-10	-8	-12	-8.9
പ്പട	6.8	7.2	9.3	11	-6	-7.9	-8.5	0	-9.9	-10	-7
it	5.4	5.5	8.2	8	-7.2	-8.3	-8.5	-12	0	-11	-7.2
pt	8.4	6.6	9	5.1	-11	-7.8	-12	-12	-9	0	-9
pl -	8.2	7.3	8.3	1.6	-7.2	-3.7	-8.8	-6	-5.8	-10	0
	en	fr	de	Ġ	da	nl	sv	es	it	pt	pl
										[I
en	0	0	0	0	0.2	0.2	0.2	0.1	0.1	0.2	0.3
en. ⊡ fr		0 0	0 0	0 0	0.2 0.2	0.2 0.2	0.2 0.2	0.1 0.2	0.1 0.2	0.2 0.6	0.3 0.3
E fr	0										
E fr	0	0	0	0	0.2	0.2	0.2	0.2	0.2	0.6	0.3
tgt Ing so tg	0 0 0	0 0	0 0	0 0	0.2 0.3	0.2 0.2	0.2 0.3	0.2 0.2	0.2 0.2	0.6 0.4	0.3 0.3
tgt Ing so tg	0 0 0	0 0 0	0 0 0	0 0 0	0.2 0.3 0.6	0.2 0.2 0.2	0.2 0.3 0.3	0.2 0.2 0.2	0.2 0.2 0.2	0.6 0.4 0.7	0.3 0.3 0.3
wrong tgt lng so p v	0 0 0 0.1	0 0 0 0	0 0 0 0.1	0 0 0 0.1	0.2 0.3 0.6 0	0.2 0.2 0.2 0.2	0.2 0.3 0.3 0.2	0.2 0.2 0.2 0.2	0.2 0.2 0.2 0.2	0.6 0.4 0.7 0.3	0.3 0.3 0.3 0.3
of wrong tgt lng s v l v s a s s	0 0 0 0.1 0	0 0 0 0	0 0 0 0.1 0	0 0 0.1 0	0.2 0.3 0.6 0 0.2	0.2 0.2 0.2 0.2 0	0.2 0.3 0.3 0.2 0.2	0.2 0.2 0.2 0.2 0.2	0.2 0.2 0.2 0.2 0.2	0.6 0.4 0.7 0.3 0.3	0.3 0.3 0.3 0.3 0.2
of wrong tgt lng s v l v s a s s	0 0 0.1 0	0 0 0 0 0	0 0 0.1 0	0 0 0.1 0	0.2 0.3 0.6 0.2 0.2	0.2 0.2 0.2 0.2 0.2	0.2 0.3 0.2 0.2 0.2	0.2 0.2 0.2 0.2 0.2 0.2	0.2 0.2 0.2 0.2 0.2 0.2	0.6 0.4 0.7 0.3 0.3 0.3	0.3 0.3 0.3 0.3 0.2 0.3
atio of wrong tgt lng i i s s u p s p j	0 0 0.1 0 0 0	0 0 0 0 0	0 0 0.1 0 0	0 0 0.1 0 0	0.2 0.3 0.6 0.2 0.2 0.2	0.2 0.2 0.2 0.2 0.2 0.2 0.2	0.2 0.3 0.2 0.2 0.2 0.2	0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2	0.2 0.2 0.2 0.2 0.2 0.2 0.2	0.6 0.4 0.7 0.3 0.3 0.3 0.2	0.3 0.3 0.3 0.2 0.3 0.3
atio of wrong tgt lng n S s u p S p J	0 0 0.1 0 0 0 0	0 0 0 0 0 0	0 0 0.1 0 0 0 0	0 0 0.1 0 0 0 0	0.2 0.3 0.6 0.2 0.2 0.2 0.2	0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2	0.2 0.3 0.2 0.2 0.2 0.2 0.2	0.2 0.2 0.2 0.2 0.2 0.2 0.2	0.2 0.2 0.2 0.2 0.2 0.2 0.1 0.1	0.6 0.4 0.7 0.3 0.3 0.3 0.2 0.2	0.3 0.3 0.3 0.2 0.3 0.3 0.3 0.3

Freeze LA + enc. & dec. DA

- Big improvements for in-in language pairs
- Some improvements for out-in language pairs ____
- Degradation for in-out and out-out language pairs
 - Partly due to generation in the wrong language



N DEVIEW 2021 3.5 Vanilla adapter stacking (all layers)

en	- 0	16	16	18	-8.6	-11	-11	-13	-9.8	-11	-7.3
fr	14	0	15	16	-5	-8.3	-7.2	-11	-12	-15	-5.6
de	16	17	0	18	-4.1	-6.8	-4.3	-5.6	-4.4	-6	-2.7
) S	20	17	17	0	-10	-5	-11	-7.8	-6.8	-15	-10
🗒 da 🖞	6.6	6.8	4.8	5.4	0	-8	-13	-9	-7.1	-11	-7.9
🖲 nl -	7.3	8.1	4.3	11	-6.3	0	-9	-10	-7.5	-9.4	-4.7
Delta (BLE s s es	7.3	8	6.7	4.9	-11	-7.9	0	-10	-8	-12	-8.9
പ്പട	6.8	7.2	9.3	11	-6	-7.9	-8.5	0	-9.9	-10	-7
it	5.4	5.5	8.2	8	-7.2	-8.3	-8.5	-12	0	-11	-7.2
pt	8.4	6.6	9	5.1	-11	-7.8	-12	-12	-9	0	-9
pl -	8.2	7.3	8.3	1.6	-7.2	-3.7	-8.8	-6	-5.8	-10	0
	en	fr	de	Ġ	da	nl	sv	es	it	pt	pl
										[I
en	0	0	0	0	0.2	0.2	0.2	0.1	0.1	0.2	0.3
en. ⊡ fr		0 0	0 0	0 0	0.2 0.2	0.2 0.2	0.2 0.2	0.1 0.2	0.1 0.2	0.2 0.6	0.3 0.3
E fr	0										
E fr	0	0	0	0	0.2	0.2	0.2	0.2	0.2	0.6	0.3
tgt Ing so tg	0 0 0	0 0	0 0	0 0	0.2 0.3	0.2 0.2	0.2 0.3	0.2 0.2	0.2 0.2	0.6 0.4	0.3 0.3
tgt Ing so tg	0 0 0	0 0 0	0 0 0	0 0 0	0.2 0.3 0.6	0.2 0.2 0.2	0.2 0.3 0.3	0.2 0.2 0.2	0.2 0.2 0.2	0.6 0.4 0.7	0.3 0.3 0.3
wrong tgt lng so p v	0 0 0 0.1	0 0 0 0	0 0 0 0.1	0 0 0 0.1	0.2 0.3 0.6 0	0.2 0.2 0.2 0.2	0.2 0.3 0.3 0.2	0.2 0.2 0.2 0.2	0.2 0.2 0.2 0.2	0.6 0.4 0.7 0.3	0.3 0.3 0.3 0.3
of wrong tgt lng s v l v s a s s	0 0 0 0.1 0	0 0 0 0	0 0 0 0.1 0	0 0 0.1 0	0.2 0.3 0.6 0 0.2	0.2 0.2 0.2 0.2 0	0.2 0.3 0.3 0.2 0.2	0.2 0.2 0.2 0.2 0.2	0.2 0.2 0.2 0.2 0.2	0.6 0.4 0.7 0.3 0.3	0.3 0.3 0.3 0.3 0.2
of wrong tgt lng s v l v s a s s	0 0 0.1 0	0 0 0 0 0	0 0 0.1 0	0 0 0.1 0	0.2 0.3 0.6 0.2 0.2	0.2 0.2 0.2 0.2 0.2	0.2 0.3 0.2 0.2 0.2	0.2 0.2 0.2 0.2 0.2 0.2	0.2 0.2 0.2 0.2 0.2 0.2	0.6 0.4 0.7 0.3 0.3 0.3	0.3 0.3 0.3 0.3 0.2 0.3
atio of wrong tgt lng i i s s u p s p j	0 0 0.1 0 0 0	0 0 0 0 0	0 0 0.1 0 0	0 0 0.1 0 0	0.2 0.3 0.6 0.2 0.2 0.2	0.2 0.2 0.2 0.2 0.2 0.2 0.2	0.2 0.3 0.2 0.2 0.2 0.2	0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2	0.2 0.2 0.2 0.2 0.2 0.2 0.2	0.6 0.4 0.7 0.3 0.3 0.3 0.2	0.3 0.3 0.3 0.2 0.3 0.3
atio of wrong tgt lng n S s u p S p J	0 0 0.1 0 0 0 0	0 0 0 0 0 0	0 0 0.1 0 0 0	0 0 0.1 0 0 0 0	0.2 0.3 0.6 0.2 0.2 0.2 0.2	0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2	0.2 0.3 0.2 0.2 0.2 0.2 0.2	0.2 0.2 0.2 0.2 0.2 0.2 0.2	0.2 0.2 0.2 0.2 0.2 0.2 0.1 0.1	0.6 0.4 0.7 0.3 0.3 0.3 0.2 0.2	0.3 0.3 0.3 0.2 0.3 0.3 0.3 0.3

Freeze LA + enc. & dec. DA

- Big improvements for in-in language pairs
- Some improvements for out-in language pairs
- Degradation for in-out and out-out language pairs
 - Partly due to generation in the wrong language
- Domain adapters seem to "erase out" language knowledge from the model
- Hard to decouple language knowledge from domain knowledge



3.6 Encoder-only or decoder-only domain adapters

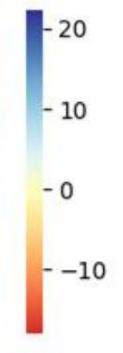
			Fr	reeze	LA +	enc	. & d	ec. D	A							Fre	eze l	-A +	enc.	DA							F	ree	ze L	A +	dec.	DA			
en	1 - 0	16	16	18	-8.6	-11	-11	-13	-9.8	-11	-7.3	en ·	0	14	14	16	-1.7	-4.3	-2.6	-5.2	-5.1	-3.9	-1.8	en -	0	14	13	16	-6.1	-8.4	-8.3 -	8.9 -	7.7 -5	.8 -6.	1
fr	- 14	0	15	16	-5	-8.3	-7.2	-11	-12	-15	-5.6	fr -	12	0	13	15	1.4	-0.6*	-0.5*	-3.3	-3.7	-3.9	-0.8	fr -	11	0	12	14	-5	-7.2	-6.8 -	8.5 -	7.8 -6	.5 -5.	5
de	- 16	17	0	18	-4.1	-6.8	-4.3	-5.6	-4.4	-6	-2.7	de ·	15	15	0	17	2.3	-1	2.2	1 *	0.1*	1.7	1.4*	de -	14	14	0	16	-2.8	-7	-2.7	4.3 -	3.6 -2	.3 -2	
5 9	- 20	17	17	0	-10	-5	-11	-7.8	-6.8	-15	-10	CS ·	17	15	15	0	-5.9	1.8	-3.4	-0.6*	-1.7	-6.9	-4.8	cs -	16	13	14	0	-7.2	-5.5	-8.5	-7 -	6.6 -8	1.2 -9	
🗒 da	- 6.6	6.8	4.8	5.4	0	-8	-13	-9	-7.1	-11	-7.9	da -	3.7	4.2	3	2.7	0	-3.9	-7.1	-4.7	-3.8	-6.9	-4	da -	11	11	9.7	9.6	0	-6.1	-11 -	6.2 -	5.5 -6	.5 -6.	2
🖲 n	- 7.3	8.1	4.3	11	-6.3	0	-9	-10	-7.5	-9.4	-4.7	nl -	3.8	5.6	2.1	8	-2.1	0	-4.3	-5.5	-3.8	-4	-0.6*	nl -	12	13	9.5	15	-4.7	0	-6.3	-7.4 -	-5.5 -4	.8 -3.	9
ve Ita	- 7.3	8	6.7	4.9	-11	-7.9	0	-10	-8	-12	-8.9	SV -	4.2	4.6	4.1	2.1	-5.7	-3.5	0	-5.9	-4.7	-7.5	-4.9	SV -	12	13	11	9.2	-11	-6.5	0	-7 -	6.2 -	8 -7.	2
å es	- 6.8	7.2	9.3	11	-6	-7.9	-8.5	0	-9.9	-10	-7	es -	4.2	4.7	6.5	7.4	-2	-2.6	-2.9	0	-6	-3.9	-2.5	es -	9.4	10	11	13	-6.1	-7.8	-7.5	0 -	9.6 -8	.5 -6.	1
it	- 5.4	5.5	8.2	8	-7.2	-8.3	-8.5	-12	0	-11	-7.2	it -	2.9	2.3	5	5.4	-3.2	-3.9	-3.7	-6.7	0	-4.7	-2.5	it -	11	11	13	14	-4.7	-6.6	-6.4	-10	0 -7	.3 -5.	5
pt	- 8.4	6.6	9	5.1	-11	-7.8	-12	-12	-9	0	-9	pt -	4.7	3.7	6.7	1.9	-6.3	-3.4	-6.8	-5.5	-5.1	0	-5	pt -	12	12	13	9.7	-7.7	-6.3	-10 -	8.8 -	7.4	0 -7.	7
p	- 8.2	7.3	8.3	1.6	-7.2	-3.7	-8.8	-6	-5.8	-10	0	pl -	4.5	4.1	5.5	0.1*	-3.6	-0.8	-5	-2.5	-3.3	-5.5	0	pl -	12	11	12	6.7	-6	-4.3	-7.8 -	5.9 -	6.2 -7	.2 0	
	en	fr	de	cs	da	nl	sv	es	it	pt	pl	62/2	én	fr	de	cs	da	nl	sv	es	it	pt	pl	2000	en	fr	de	cs	da	nl	sv	es	it r	ot pl	

Both: slight decrease in in-in translations (due to lower DA capacity)

Franza I A L and DA

Franza I A I das DA





3.6 Encoder-only or decoder-only domain adapters

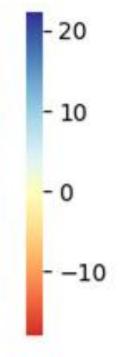
		F	reeze	ELA -	+ enc	. & d	ec. D	A							Fre	eze l	-A +	enc.	DA							F	ree	ze L	A +	dec	. DA			
en - 0	16	16	18	-8.6	-11	-11	-13	-9.8	-11	-7.3	en ·	0	14	14	16	-1.7	-4.3	-2.6	-5.2	-5.1	-3.9	-1.8	en -	0	14	13	16	-6.1	-8.4	-8.3	-8.9	-7.7 -	5.8 -	6.1
fr - 14	0	15	16	-5	-8.3	-7.2	-11	-12	-15	-5.6	fr	12	0	13	15	1.4	-0.6*	-0.5*	-3.3	-3.7	-3.9	-0.8	fr -	11	0	12	14	-5	-7.2	-6.8	-8.5 -	-7.8 -(6.5 -	5.5
de - 16	17	0	18	-4.1	-6.8	-4.3	-5.6	-4.4	-6	-2.7	de ·	15	15	0	17	2.3	-1	2.2	1 *	0.1*	1.7	1.4*	de -	14	14	0	16	-2.8	-7	-2.7	-4.3 -	-3.6 -2	2.3	-2
S CS - 20	17	17	0	-10	-5	-11	-7.8	-6.8	-15	-10	CS ·	17	15	15	0	-5.9	1.8	-3.4	-0.6*	-1.7	-6.9	-4.8	CS ·	16	13	14	0	-7.2	-5.5	-8.5	-7 .	-6.6 -8	8.2	-9
🗒 da - 6.0	6.8	4.8	5.4	0	-8	-13	-9	-7.1	-11	-7.9	da ·	3.7	4.2	3	2.7	0	-3.9	-7.1	-4.7	-3.8	-6.9	-4	da -	11	11	9.7	9.6	0	-6.1	-11	-6.2 .	-5.5 -(6.5 -	6.2
🖲 nl - 7.3	8.1	4.3	11	-6.3	0	-9	-10	-7.5	-9.4	-4.7	nl ·	3.8	5.6	2.1	8	-2.1	0	-4.3	-5.5	-3.8	-4	-0.6*	nl -	12	13	9.5	15	-4.7	0	-6.3	-7.4 -	-5.5 -4	4.8 -	3.9
sv - 7.3	8	6.7	4.9	-11	-7.9	0	-10	-8	-12	-8.9	SV ·	4.2	4.6	4.1	2.1	-5.7	-3.5	0	-5.9	-4.7	-7.5	-4.9	SV -	12	13	11	9.2	-11	-6.5	0	-7 -	-6.2	-8 -	7.2
ළ es - 6.8	3 7.2	9.3	11	-6	-7.9	-8.5	0	-9.9	-10	-7	es ·	4.2	4.7	6.5	7.4	-2	-2.6	-2.9	0	-6	-3.9	-2.5	es -	9.4	10	11	13	-6.1	-7.8	-7.5	0	-9.6 -8	8.5 -	6.1
it - 5.4	5.5	8.2	8	-7.2	-8.3	-8.5	-12	0	-11	-7.2	it	2.9	2.3	5	5.4	-3.2	-3.9	-3.7	-6.7	0	-4.7	-2.5	it -	11	11	13	14	-4.7	-6.6	-6.4	-10	0 -	7.3 -	5.5
pt - 8.4	6.6	9	5.1	-11	-7.8	-12	-12	-9	0	-9	pt ·	4.7	3.7	6.7	1.9	-6.3	-3.4	-6.8	-5.5	-5.1	0	-5	pt -	12	12	13	9.7	-7.7	-6.3	-10	-8.8 -	-7.4	0 -	7.7
pl - 8.2	2 7.3	8.3	1.6	-7.2	-3.7	-8.8	-6	-5.8	-10	0	pl -	4.5	4.1	5.5	0.1*	-3.6	-0.8	-5	-2.5	-3.3	-5.5	0	pl -	12	11	12	6.7	-6	-4.3	-7.8	-5.9	-6.2 -7	7.2	0
en	fr	de	(S	da	nl	sv	es	it	nt	n	2200	en	fr	de	CS.	da	nl	sv	es	it	nt	n		en	fr	de	's	da	nl	SV	es	it	nt	n
CI		uc	6	uu		24	0.5	i.	pr	Pi		CII		uc	0	uu		54	00	i.	pr	Pi		CII		uc	0	uu		34	00	i.	pe	P

- Both: slight decrease in in-in translations (due to lower DA capacity) -----
- Encoder only: less off-target translation, but lower out-in performance -



Franza I A I das DA

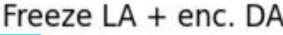




3.6 Encoder-only or decoder-only domain adapters

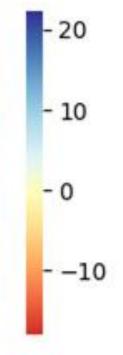
			Fr	reeze	LA +	+ enc	. & d	ec. D	A							Fre	eze L	A +	enc.	DA							F	ree	ze L/	A + 1	dec.	. DA			
en -	0	16	16	18	-8.6	-11	-11	-13	-9.8	-11	-7.3	en -	0	14	14	16	-1.7	-4.3	-2.6	-5.2	-5.1	-3.9	-1.8	en -	0	14	13	16	-6.1	-8.4	-8.3	-8.9	-7.7 -	5.8 -	6.1
fr -	14	0	15	16	-5	-8.3	-7.2	-11	-12	-15	-5.6	fr -	12	0	13	15	1.4	-0.6*	-0.5*	-3.3	-3.7	-3.9	-0.8	fr -	11	0	12	14	-5	-7.2 -	6.8	-8.5 -	-7.8 -	6.5 -	5.5
de -	16	17	0	18	-4.1	-6.8	-4.3	-5.6	-4.4	-6	-2.7	de ·	15	15	0	17	2.3	-1	2.2	1 *	0.1*	1.7	1.4*	de -	14	14	0	16	-2.8	-7 -	2.7	-4.3	-3.6 -	2.3	-2
<u> 3</u> -	20	17	17	0	-10	-5	-11	-7.8	-6.8	-15	-10	CS ·	17	15	15	0	-5.9	1.8	-3.4	-0.6*	-1.7	-6.9	-4.8	cs -	16	13	14	0	-7.2	-5.5	-8.5	-7 -	-6.6 -	8.2	-9
띸 da -	6.6	6.8	4.8	5.4	0	-8	-13	-9	-7.1	-11	-7.9	da ·	3.7	4.2	3	2.7	0	-3.9	-7.1	-4.7	-3.8	-6.9	-4	da -	11	11	9.7	9.6	0	-6.1	-11 -	-6.2 -	-5.5 -	6.5 -	6.2
🖲 nl -	7.3	8.1	4.3	11	-6.3	0	-9	-10	-7.5	-9.4	-4.7	nl ·	3.8	5.6	2.1	8	-2.1	0	-4.3	-5.5	-3.8	-4	-0.6*	nl -	12	13	9.5	15	-4.7	0	6.3	-7.4 -	-5.5 -	4.8 -	3.9
- va <u>ta</u>	7.3	8	6.7	4.9	-11	-7.9	0	-10	-8	-12	-8.9	SV ·	4.2	4.6	4.1	2.1	-5.7	-3.5	0	-5.9	-4.7	-7.5	-4.9	SV -	12	13	11	9.2	-11	-6.5	0	-7 -	-6.2	-8 -	7.2
d es -	6.8	7.2	9.3	11	-6	-7.9	-8.5	0	-9.9	-10	-7	es ·	4.2	4.7	6.5	7.4	-2	-2.6	-2.9	0	-6	-3.9	-2.5	es -	9.4	10	11	13	-6.1	-7.8 -	-7.5	0 -	-9.6 -	8.5 -	6.1
it -	5.4	5.5	8.2	8	-7.2	-8.3	-8.5	-12	0	-11	-7.2	it ·	2.9	2.3	5	5.4	-3.2	-3.9	-3.7	-6.7	0	-4.7	-2.5	it -	11	11	13	14	-4.7	6.6	6.4	-10	0 -	7.3 -	5.5
pt -	8.4	6.6	9	5.1	-11	-7.8	-12	-12	-9	0	-9	pt ·	4.7	3.7	6.7	1.9	-6.3	-3.4	-6.8	-5.5	-5.1	0	-5	pt -	12	12	13	9.7	-7.7	-6.3	-10 -	-8.8 -	-7.4	0 -	7.7
pl -	8.2	7.3	8.3	1.6	-7.2	-3.7	-8.8	-6	-5.8	-10	0	pl -	4.5	4.1	5.5	0.1*	-3.6	-0.8	-5	-2.5	-3.3	-5.5	0	pl -	12	11	12	6.7	-6	-4.3 ·	7.8	-5.9 -	-6.2 -	7.2	0
6-10 A	-	5	de		da	-		-	4	nt	n l	6.00	-	5	de	-	da	-	-		4	nt	, ,	2000	-	4	do	-	da	n l	-	-		nt	
	en	If	de	cs	da	ni	sv	es	it	pt	р		en	ir	de	cs	ua	ni	SV	es	IC	pt	pi		en	IL	de	CS	ua	ni	SV	es	IC	pt	рі

- Both: slight decrease in in-in translations (due to lower DA capacity) ____
- Encoder only: less off-target translation, but lower out-in performance —
- Decoder only: better out-in performance, but worse out-out and in-out performance _



Franza I A I das DA





3.7 Regularization and data augmentation

DADrop: domain adapter drop

- Randomly drop adapters during training
- Motivation: reduce "language overfitting" effect



aining 'fitting" effect



3.7 Regularization and data augmentation

DADrop: domain adapter drop

- Randomly drop adapters during training
- Motivation: reduce "language overfitting" effect

Back-translation Use baseline model to back-translate in-domain data from and into English for out languages



aining 'fitting" effect



3.7 Regularization and data augmentation

Freeze LA + enc. & dec. DA

c -	0	16	16	18	-8.6	-11	-11	-13	-9.8	-11	-7.3	- en	0
fr en	14	0	15	16	-5	-8.3	-7.2	-11	-12	-15	-5.6	- fr e	13
- de -	16	17	0	18	-4.1	-6.8	-4.3	-5.6	-4.4	-6	-2.7	- de	14
Н с С Ц С С С	20	17	17	0	-10	-5	-11	-7.8	-6.8	-15	-10	- CS	18
– <u>a</u> –	6.6	6.8	4.8	5.4	0	-8	-13	-9	-7.1	-11	-7.9	- da	1.5
- nd	7.3	8.1	4.3	11	-6.3	0	-9	-10	-7.5	-9.4	-4.7	- <u> </u>	1.2
- ^s	7.3	8	6.7	4.9	-11	-7.9	0	-10	-8	-12	-8.9	- S	1.2
es a	6.8	7.2	9.3	11	-6	-7.9	-8.5	0	-9.9	-10	-7	- es	0.1 *
_ <u>ب</u> ک	5.4	5.5	8.2	8	-7.2	-8.3	-8.5	-12	0	-11	-7.2	- <u>ب</u>	1.1
- Ħ	8.4	6.6	9	5.1	-11	-7.8	-12	-12	-9	0	-9	ъ-	0.9
ਕ -	8.2	7.3	8.3	1.6	-7.2	-3.7	-8.8	-6	-5.8	-10	0		0.9
	en	fr	de	ı CS	da	nl	sv	es	it	pt	pl		en

Freeze LA + enc. & dec.DA + BT 4 16 3 4.4 4.1 2.6 9 4.7 6.6 11 2.8 8.6 12 8.5 3.8 5.1 1.6 5.5 5.9 -4.2 -2 -2 -1.1 1.2 -3.6 -2.1 4.5 6.9 2 0 -0.4 * -3.9 -4.2 1 -5.7 0 6.2 -2.4 -1.5 0 0.7 * -2.8 0.2 3.6 0 -2.9 -6.1 -4.6 -7.4 -6.5 -3.6 0 -2.3 -3 -2.1 -1.5 -0.1* 2.2 7.5 0 4.1 -1.7 da CS

ب -	13	0	14
- de	14	15	0
cs de -	18	16	16
	1.5	8.7	7.
nl da -	1.2	11	7.
- S	1.2	10	8.
- es	0.1 *	8.2	9.
.≓ -	1.1	7.5	9.
- b	0.9	8.9	1
<u>d</u> -	0.9	9.6	9.
	ı en	fr	de

Freeze | A + enc & dec DA + DADrop

		110	EEZE		enc	. a u	ec. D	AT	DADI	υþ						enc		uec	DA	TL	JAD	ιυρ	ты
- en	0	14	14	17	-5.7	-8.8	-8.3	-10	-7.3	-7.6	-5.1	- eu	0	15	14	16	8.7	6.5	8.4	7.7	9.9	16	11
- ب	12	0	13	15	-3.3	-5.9	-5.1	-7.2	-8	-9.5	-4.1	<u> </u>	12	0	13	15	5.3	3.8	5	3.4	5.5	9.6	9.3
- de -	14	15	0	17					-2.8		-1.1*	- de	13	14	0	16	8.7	3.4	9.1	5.6	6.7	12	11
LEU a cs c	17	15	15	0	-5.3	-3.1	-7.2	-6.6	-4	-10	-8.6	- C	16	15	15	0	3.1	3.9	3.1	4.8	6.1	7.2	5.8
- da 📙	7.4	6.9	5.1	5.2	0	-7.1			-5.6									-2.7					
- n d	7.4	8.2	4.8	11	-5.2	0	-7.5	-8.7	-5.6	-7.2	-3												
- sv	8	8	7.1	4.4	-9.1	-6.5	0	-9.2	-5.9	-10	-7.5	- S	1	10	8.9	7.3	0.8	-1.9	0	-1.7	-0.7*	1.8	2.3
es -	6.5	7.2	8.8	10	-5	-6.8	-6.4	0	-8.4	-7.7	-5.2	es -	0.5	9.1	10	12	2.1	-0.7*	0.5*	0	1.6	6.8	6.9
_{- א} ר	6.8	5.9	8.1	8.4	-5.5	-7	-7.4	-9.7	0	-7.8	-5.1												
- pt	9	6.7	9.4	5.6	-8.6	-6.5	-10	-9.7	-7	0	-8.2	_											
<u>a</u> -	7.7	6.9	7.8	1.9	-5.7	-3.4	-8.3	-5.5	-4.3	-8	0	<u>a</u> -	1.2	9	10	4.2	-0.9	-0.5*	-2.1	-0.4*	0.3*	1.5	0
	en	fr	de	cs	da	nl	sv	es	it	pt	pl		en	fr	de	cs	da	'nl	sv	es	it	pt	pl



Freeze LA + enc. & dec. DA + DADrop + BT

DADrop

- Improves out-out translation —
- But off-target translations persist





3.7 Regularization and data augmentation

Freeze LA + enc. & dec. DA

- en	0	16	16	18	-8.6	-11	-11	-13	-9.8	-11	-7.3	en
- fr	14	0	15	16	-5	-8.3	-7.2	-11	-12	-15	-5.6	fr
	16	17	0	18	-4.1	-6.8	-4.3	-5.6	-4.4	-6	-2.7	de
EU CS -	20	17	17	0	-10	-5	-11	-7.8	-6.8	-15	-10	S
BLE da	6.6	6.8	4.8	5.4	0	-8	-13	-9	-7.1	-11	-7.9	da
<u> </u>	7.3	8.1	4.3	11	-6.3	0	-9	-10	-7.5	-9.4	-4.7	ᄃ
lta sv	7.3	8	6.7	4.9	-11	-7.9	0	-10	-8	-12	-8.9	SV
Del ces	6.8	7.2	9.3	11	-6	-7.9	-8.5	0	-9.9	-10	-7	es
ш _{.±} -	F 4	5.5	8.2	8	-7.2	-8.3	-8.5	-12	0	-11	-7.2	<u>ب</u> :
- pt	8.4	6.6	9	5.1	-11	-7.8	-12	-12	-9	0	-9	pt
[d -	8.2	7.3	8.3	1.6	-7.2	-3.7	-8.8	-6	-5.8	-10	0	d
	en	fr	de	cs	da	nl	sv	es	it	pt	pl	

	Freeze LA <u>+ enc. & dec.DA + BT</u>														
- en	0	15	15	18	9.2	6.6	9.3	7.5	10	16	11				
- fr	13	0	14	16	4	3	4.4	2.6	4.1	9	9				
- de	14	15	0	17	8.5	2.8	8.6	4.7	6.6	12	11				
- CS	18	16	16	0	2	1	1.6	3.8	5.1	5.9	5.5				
- da	1.5	8.7	7.4	7.7	0	-4	-2	-4.2	-2	-1.1	1.2				
- <u>-</u>	1.2	11	7.4	13	3.1	0	2	-3.6	-2.1	4.5	6.9				
- S	1.2	10	8.1	6.2	-0.4*	-3.9	0	-5.7	-4.2	-0.9	1				
- es	0.1 *	8.2	9.7	11	0.7 *	-2.8	-2.4	0	-1.5	0.2 *	6.2				
.± -	1.1	7.5	9.2	10	-1.5	-4.7	-2.9	-6.1	0	-1 *	3.6				
- Ħ	0.9	8.9	11	7.5	-4.1	-4.6	-7.4	-6.5	-3.6	0	2.2				
<u>a</u> -	0.9	9.6	9.9	4.1	-1.7	-2.3	-3	-2.1	-1.5	-0.1*	0				
	en	fr	de	CS	da	nl	sv	es	it	pt	pl				

Freeze | A + enc & dec DA + DADron

	Treeze LA + enc. & dec. DA + DADTOP											T = 2 = 1 + 2 = 10										ты	
- en	0	14	14	17	-5.7	-8.8	-8.3	-10	-7.3	-7.6	-5.1	- en	0	15	14	16	8.7	6.5	8.4	7.7	9.9	16	11
- Ť	12	0	13	15	-3.3	-5.9	-5.1	-7.2	-8	-9.5	-4.1	- L	12	0			5.3	3.8	5	3.4	5.5	9.6	9.3
- de -	14	15	0	17	-2.5	-4.8	-2.5	-4	-2.8	-2	-1.1*	- de	13	14	0	16	8.7	3.4	9.1	5.6	6.7	12	11
(BLEU nl da cs d	17	15	15	0	-5.3	-3.1	-7.2	-6.6	-4	-10	-8.6	- CS	16	15	15	0	3.1	3.9	3.1	4.8	6.1	7.2	5.8
	7.4	6.9	5.1	5.2	0	-7.1	-11	-8.8	-5.6	-9.5	-7.3	- qa	1.7	8.9	7.4	7.4	0	-2.7	-0.3*	-1.5	0.1*	1.5	2.2
	7.4	8.2	4.8	11	-5.2	0	-7.5	-8.7	-5.6	-7.2	-3							0					
- ^s	8	8	7.1	4.4	-9.1	-6.5	0	-9.2	-5.9	-10	-7.5	- S	1	10	8.9	7.3	0.8	-1.9	0	-1.7	-0.7*	1.8	2.3
- es	6.5	7.2	8.8	10	-5	-6.8	-6.4	0	-8.4	-7.7	-5.2	- es	0.5	9.1	10	12	2.1	-0.7*	0.5*	0	1.6	6.8	6.9
_ <u>بن</u> _	6.8	5.9	8.1	8.4	-5.5	-7	-7.4	-9.7	0	-7.8	-5.1							-2.2					
- pt	9	6.7	9.4	5.6	-8.6	-6.5	-10	-9.7	-7	0	-8.2							-2.2					
<u>a</u> -	7.7	6.9	7.8	1.9	-5.7	-3.4	-8.3	-5.5	-4.3	-8	0	<u>a</u> -	1.2	9	10	4.2	-0.9	-0.5*	-2.1	-0.4*	0.3*	1.5	0
	en	fr	de	cs	da	nl	sv	es	it	pt	pl	, ,	en	fr	de	cs	da	'nl	sv	es	it	pt	pl



Freeze LA + enc. & dec. DA + DADrop + BT

DADrop

- Improves out-out translation -
- But off-target translations persist

Back-translation (BT)

- Solves off-target translation —
- Improves in-out translation







3.7 Regularization and data augmentation

Freeze LA + enc. & dec. DA

- en	0	16	16	18	-8.6	-11	-11	-13	-9.8	-11	-7.3	en
- fr	14	0	15	16	-5	-8.3	-7.2	-11	-12	-15	-5.6	fr
	16	17	0	18	-4.1	-6.8	-4.3	-5.6	-4.4	-6	-2.7	de
EU S -	20	17	17	0	-10	-5	-11	-7.8	-6.8	-15	-10	S
BLE da	6.6	6.8	4.8	5.4	0	-8	-13	-9	-7.1	-11	-7.9	da
<u> </u>	7.3	8.1	4.3	11	-6.3	0	-9	-10	-7.5	-9.4	-4.7	ᄃ
lta sv	7.3	8	6.7	4.9	-11	-7.9	0	-10	-8	-12	-8.9	SV
Del ces	6.8	7.2	9.3	11	-6	-7.9	-8.5	0	-9.9	-10	-7	es
ш _{.±} -	F 4	5.5	8.2	8	-7.2	-8.3	-8.5	-12	0	-11	-7.2	<u>ب</u> :
- pt	8.4	6.6	9	5.1	-11	-7.8	-12	-12	-9	0	-9	pt
[d -	8.2	7.3	8.3	1.6	-7.2	-3.7	-8.8	-6	-5.8	-10	0	d
	en	fr	de	cs	da	nl	sv	es	it	pt	pl	

	Freeze LA <u>+ enc. & dec.DA + BT</u>														
- en	0	15	15	18	9.2	6.6	9.3	7.5	10	16	11				
- fr	13	0	14	16	4	3	4.4	2.6	4.1	9	9				
- de	14	15	0	17	8.5	2.8	8.6	4.7	6.6	12	11				
- CS	18	16	16	0	2	1	1.6	3.8	5.1	5.9	5.5				
- da	1.5	8.7	7.4	7.7	0	-4	-2	-4.2	-2	-1.1	1.2				
- <u>-</u>	1.2	11	7.4	13	3.1	0	2	-3.6	-2.1	4.5	6.9				
- S	1.2	10	8.1	6.2	-0.4*	-3.9	0	-5.7	-4.2	-0.9	1				
- es	0.1 *	8.2	9.7	11	0.7 *	-2.8	-2.4	0	-1.5	0.2 *	6.2				
.≓ -	1.1	7.5	9.2	10	-1.5	-4.7	-2.9	-6.1	0	-1 *	3.6				
- Þ	0.9	8.9	11	7.5	-4.1	-4.6	-7.4	-6.5	-3.6	0	2.2				
<u>a</u> -	0.9	9.6	9.9	4.1	-1.7	-2.3	-3	-2.1	-1.5	-0.1*	0				
	en	fr	de	cs	da	nl	sv	es	it	pt	pl				

Freeze | A + enc & dec DA + DADron

	THEEZE LA T ETC. & GEC. DA T DADTOP												1667		АТ	enc		uec.	DA		JAD	ιop	ты
- G	0	14	14	17	-5.7	-8.8	-8.3	-10	-7.3	-7.6	-5.1	- eu	0	15	14	16	8.7	6.5	8.4	7.7	9.9	16	11
- fr	12	0	13	15	-3.3	-5.9	-5.1	-7.2	-8	-9.5	-4.1	ل	12	0	13	15	5.3	3.8	5	3.4	5.5	9.6	9.3
- de _	14	15	0	17	-2.5	-4.8	-2.5	-4	-2.8	-2	-1.1*	77	13	14	0	16	8.7	3.4	9.1	5.6	6.7	12	11
⊡ õ-	17	15	15	0	-5.3	-3.1	-7.2	-6.6	-4	-10	-8.6	- CS	16	15	15	0	3.1	3.9	3.1	4.8	6.1	7.2	5.8
– a –	7.4	6.9	5.1	5.2	0	-7.1	-11	-8.8	-5.6	-9.5	-7.3	- da	1.7	8.9	7.4	7.4	0	-2.7	-0.3 [*]	-1.5	0.1*	1.5	2.2
- n d	7.4	8.2	4.8	11	-5.2	0	-7.5	-8.7	-5.6	-7.2	-3							0					
- ^{sv}	8	8	7.1	4.4	-9.1	-6.5	0	-9.2	-5.9	-10	-7.5							-1.9					
- es -	6.5	7.2	8.8	10	-5	-6.8	-6.4	0	-8.4	-7.7	-5.2	es -	0.5	9.1	10	12	2.1	-0.7*	0.5*	0	1.6	6.8	6.9
_ <u>ب</u>	6.8	5.9	8.1	8.4	-5.5	-7	-7.4	-9.7	0	-7.8	-5.1	.± -	1.2	8.7	11	11	-0.2*	-2.2	-1	-1.5	0	4	5.8
- pt	9	6.7	9.4	5.6	-8.6	-6.5	-10	-9.7	-7	0	-8.2	- pt	1.1	9.1	11	8.1	-3.2	-2.2	-3.9	-2.9	-0.3*	0	3.7
<u>a</u> -	7.7	6.9	7.8	1.9	-5.7	-3.4	-8.3	-5.5	-4.3	-8	0	d -	1.2	9	10	4.2	-0.9	-0.5*	-2.1	-0.4*	0.3*	1.5	0
	en	fr	de	CS	da	nl	sv	es	it	pt	pl		en	fr	de	cs	da	'nl	sv	es	it	pt	pl



Freeze LA + enc. & dec. DA + DADrop + BT

DADrop

- Improves out-out translation
- But off-target translations persist

Back-translation (BT)

- Solves off-target translation —
- Improves in-out translation
- Effect of DA for out-out is small







3.8 Conclusion

knowledge

- Contrary to Pfeiffer et al. (2020), who use encoder-only classification models
- Generation tasks require good language-specific representations
- Encoder-only or decoder-only adapters have useful properties



It is hard to properly decouple language knowledge from domain



3.8 Conclusion

knowledge

- Contrary to Pfeiffer et al. (2020), who use encoder-only classification models
- Generation tasks require good language-specific representations
- Encoder-only or decoder-only adapters have useful properties

Regularization and data augmentation techniques can help



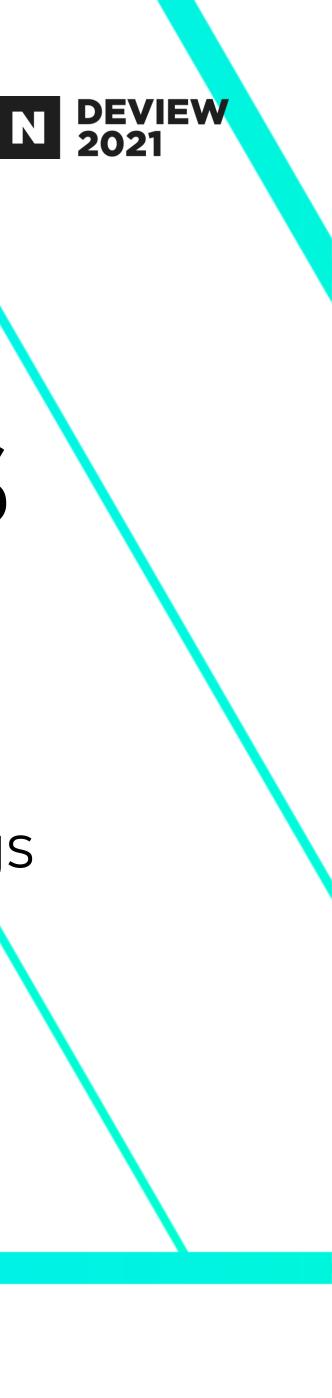
It is hard to properly decouple language knowledge from domain



4. Learning new languages efficiently

Continual Learning in Multilingual NMT via Language-Specific Embeddings

andre Berar



Given an existing MNMT model $\{FR, DE, EN\} \rightarrow \{FR, DE, EN\}$





Given an existing MNMT model $\{FR, DE, EN\} \rightarrow \{FR, DE, EN\}$



How can we efficiently add a new source language? $\{FR, DE, EN, EL\} \rightarrow \{FR, DE, EN\}$





Given an existing MNMT model $\{FR, DE, EN\} \rightarrow \{FR, DE, EN\}$

How can we efficiently add a new source language? $\{FR, DE, EN, EL\} \rightarrow \{FR, DE, EN\}$

or a new target language? $\{FR, DE, EN\} \rightarrow \{FR, DE, EN, EL\}$



Our (self-imposed) constraints:

- No re-training on the initial language pairs





Our (self-imposed) constraints:

- No re-training on the initial language pairs
- No performance drop on the initial language pairs





Our (self-imposed) constraints:

- No re-training on the initial language pairs
- No performance drop on the initial language pairs
- Good zero-shot performance (train on $EL \rightarrow EN$, evaluate on $EL \rightarrow FR$)





4.1 Introduction

Our (self-imposed) constraints:

- No re-training on the initial language pairs
- No performance drop on the initial language pairs
- Good zero-shot performance (train on $EL \rightarrow EN$, evaluate on $EL \rightarrow FR$)
- No significant increase in model size





4.1 Introduction

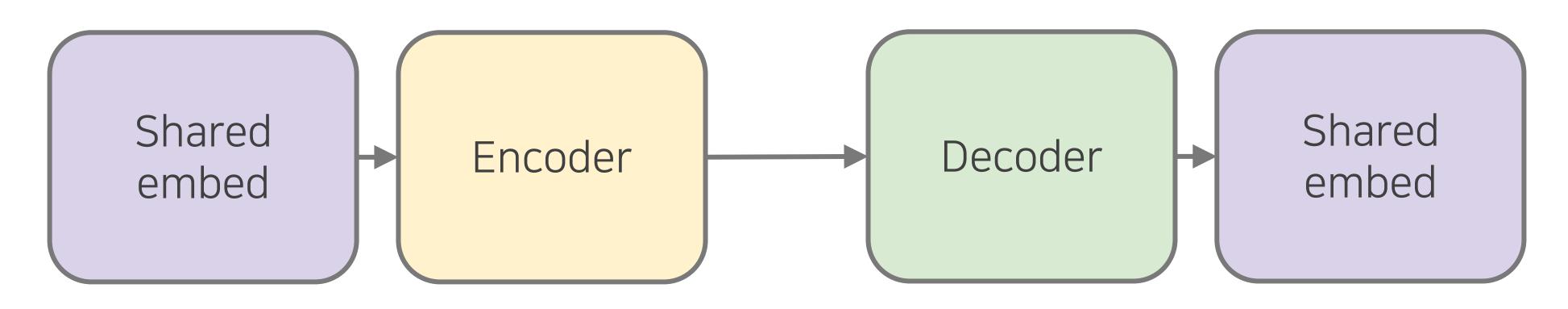
Our (self-imposed) constraints:

- No re-training on the initial language pairs
- No performance drop on the initial language pairs
- Good zero-shot performance (train on $EL \rightarrow EN$, evaluate on $EL \rightarrow FR$)
- No significant increase in model size
- Fast training and inference





4.2 Technique: initial MNMT model

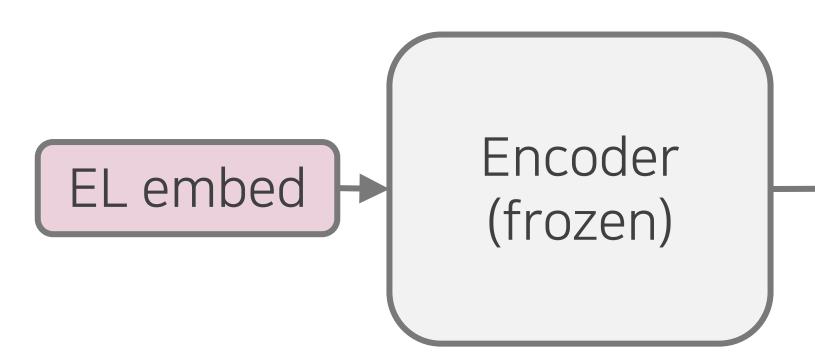




Train on many-to-many data (English-centric or multi-parallel)



4.2 Technique: add a new source language



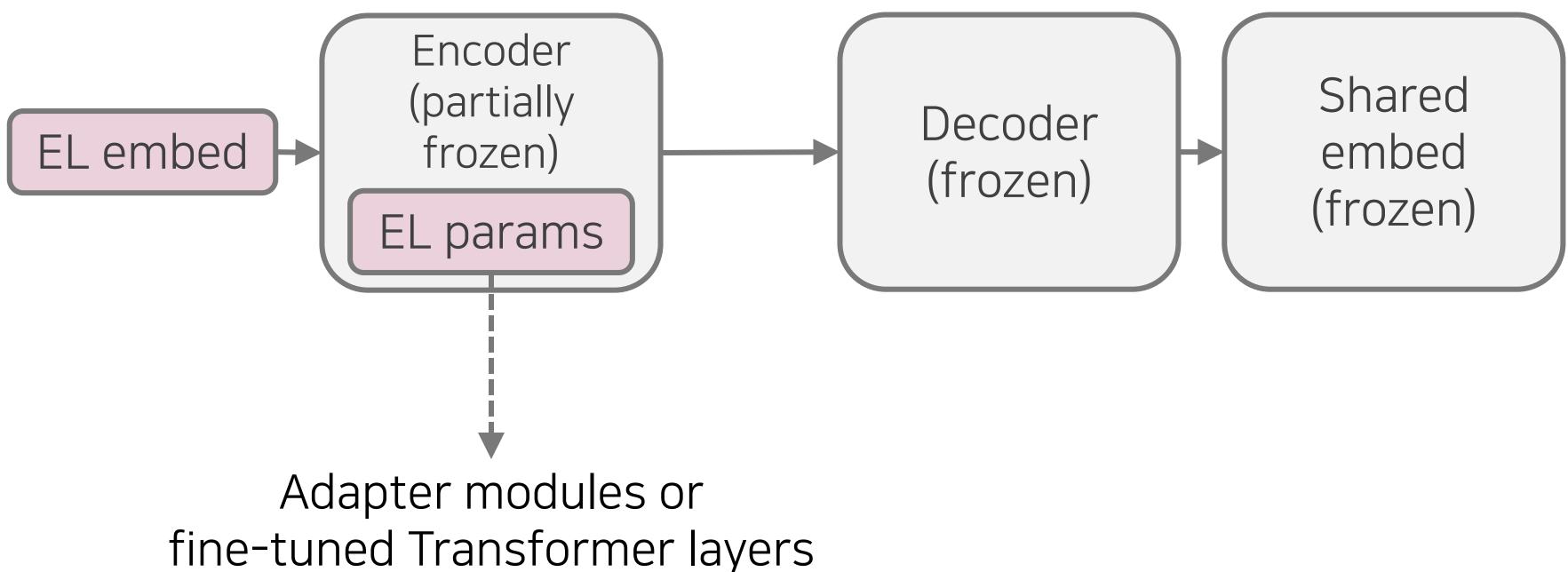


Train on Greek-English data

Shared Decoder embed (frozen) (frozen)



4.2 Technique: add a new source language

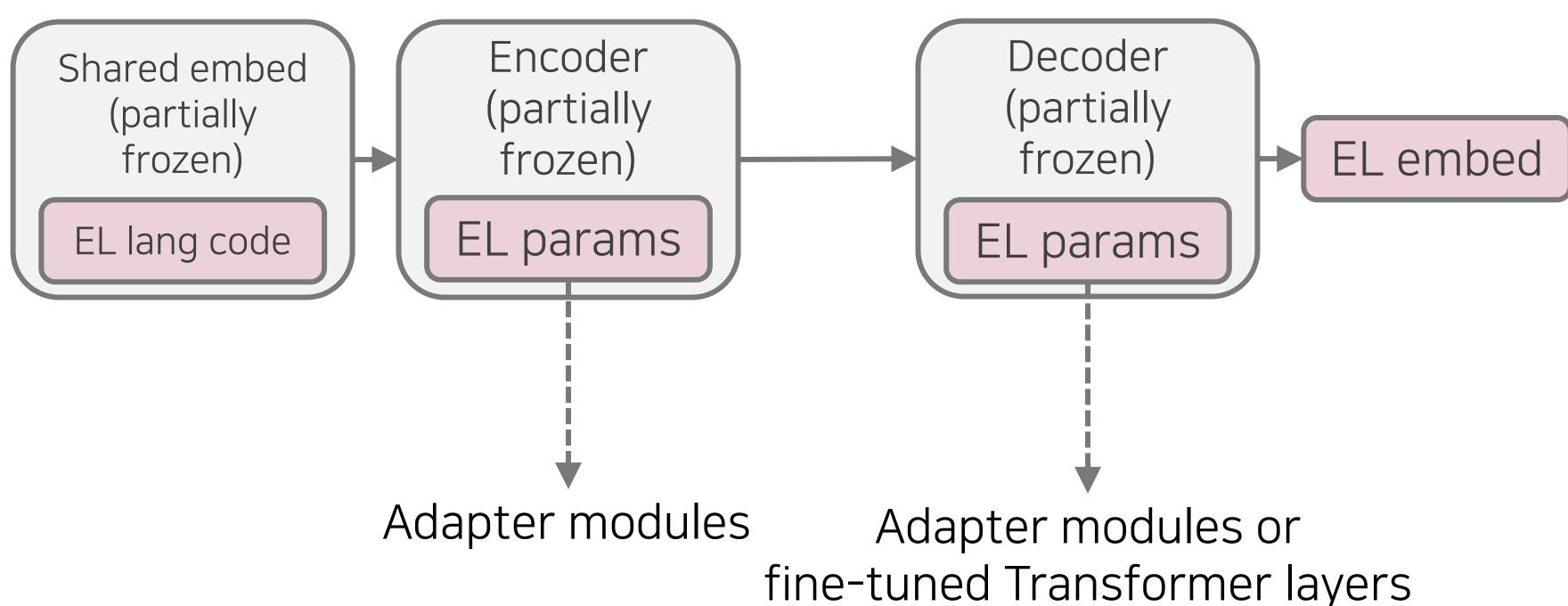




Train on Greek-English data



4.2 Technique: add a new target language





Train on English-Greek data



4.3 Experiments: TED Talks Top 20

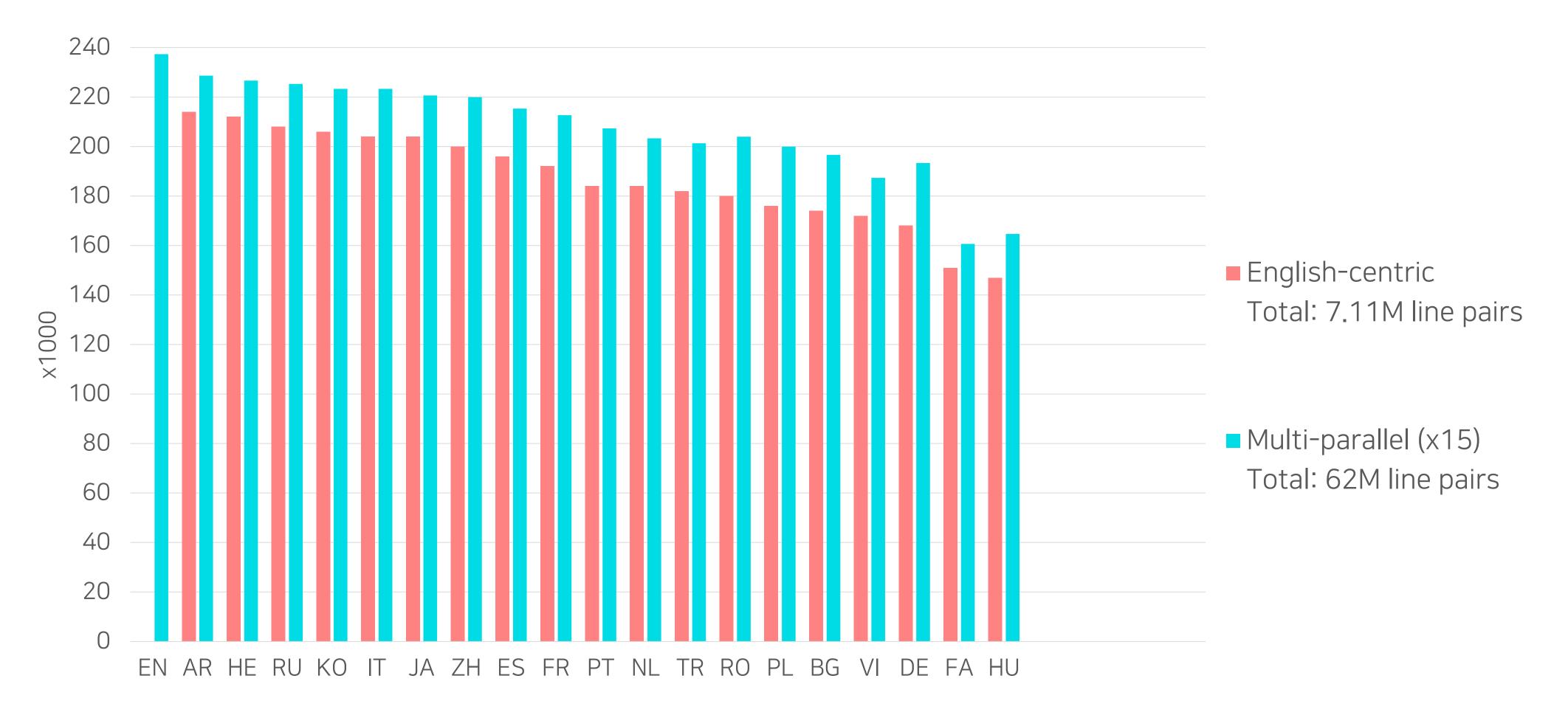


English-centric Total: 7.11M line pairs

Initial model: Transformer Base



4.3 Experiments: TED Talks Top 20



Initial model: Transformer Base



4.3 Experiments: TED Talks Top 20

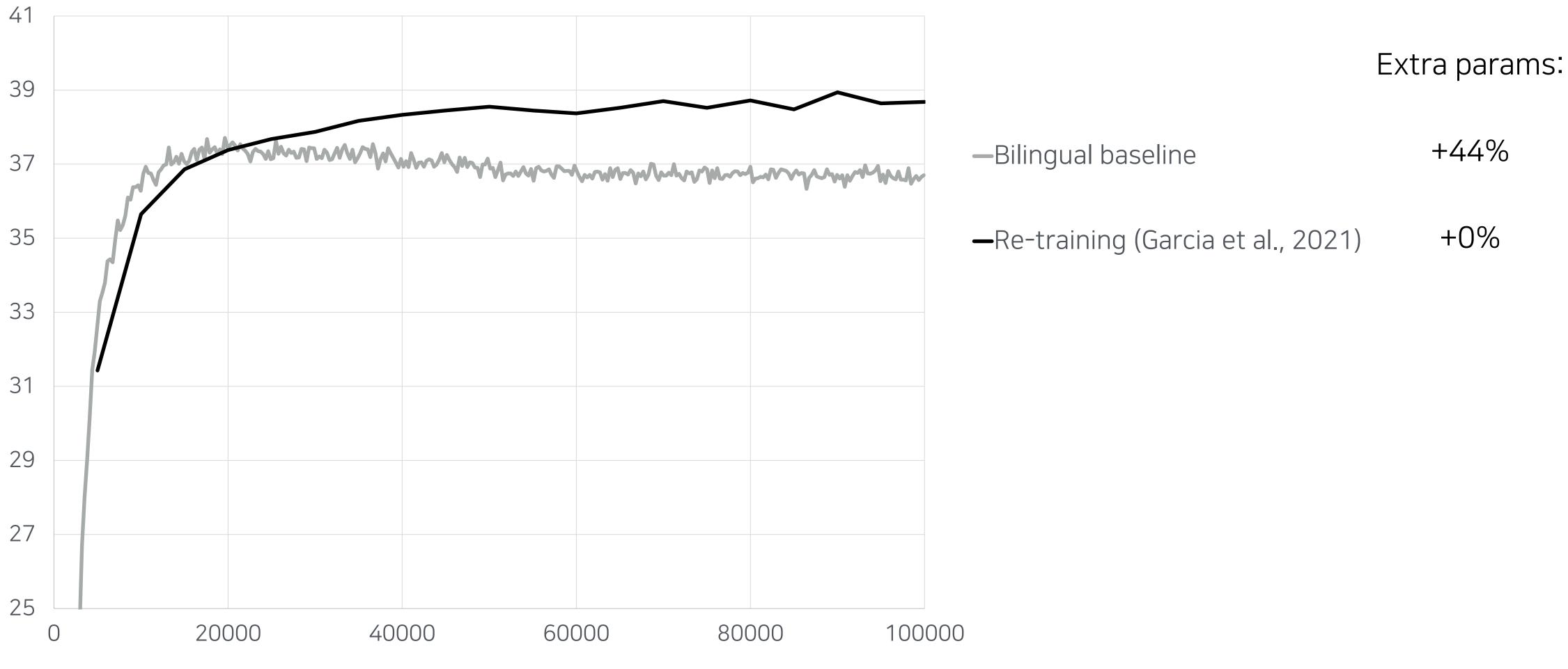


Initial model: Transformer Base



4.4 Results: new source language

Greek-English BLEU by training step

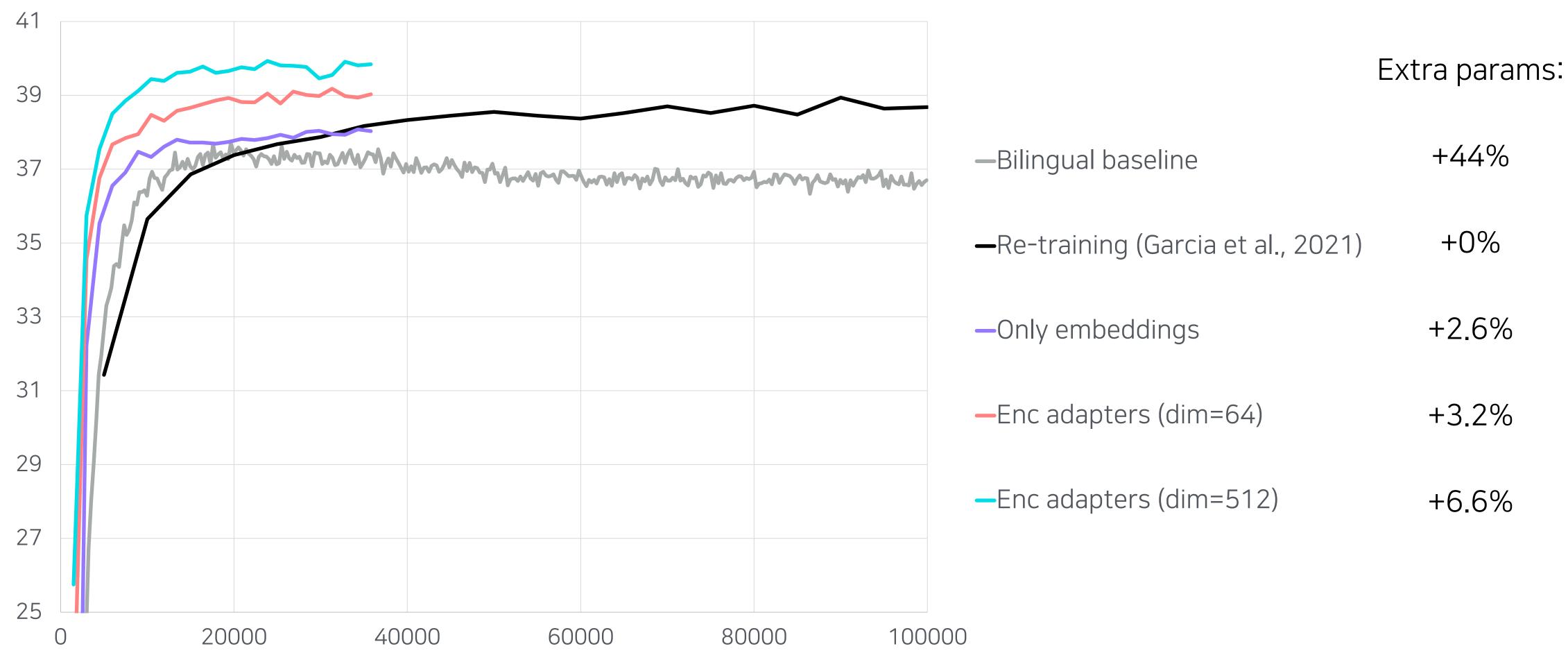






4.4 Results: new source language

Greek-English BLEU by training step

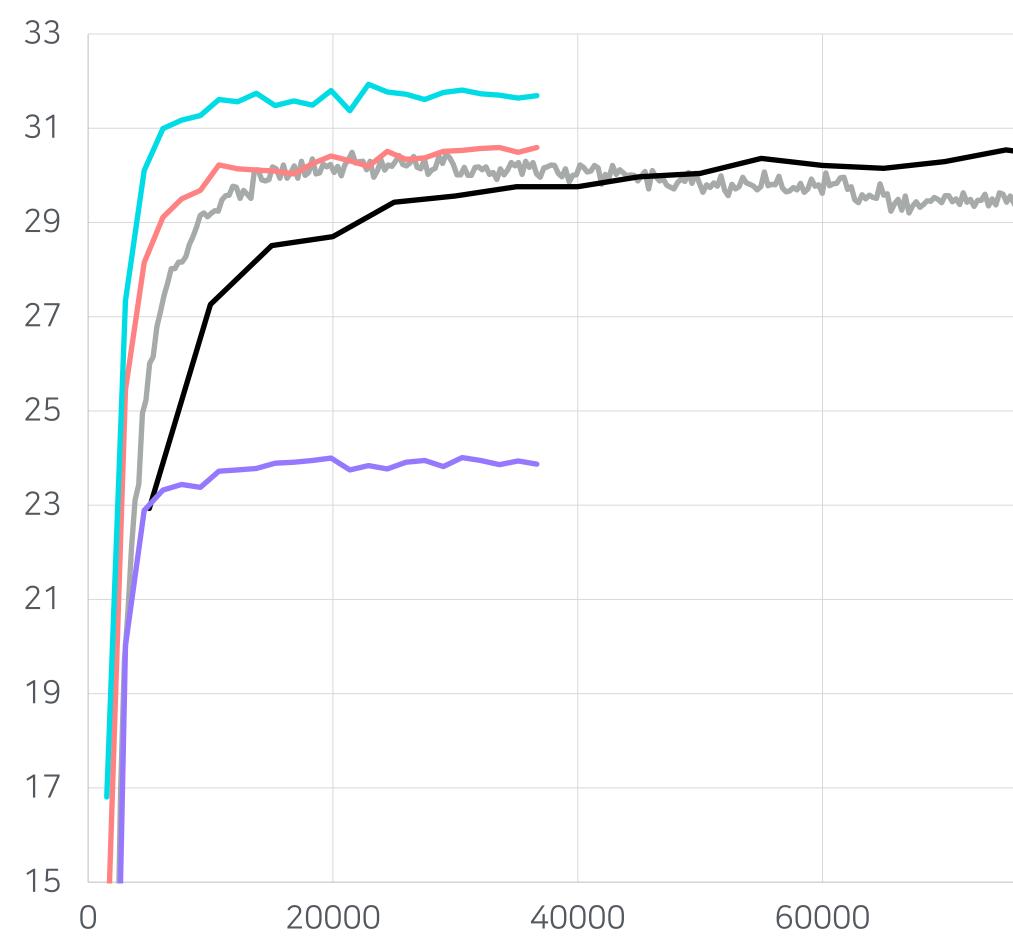






4.4 Results: new target language

English-Greek BLEU by training step





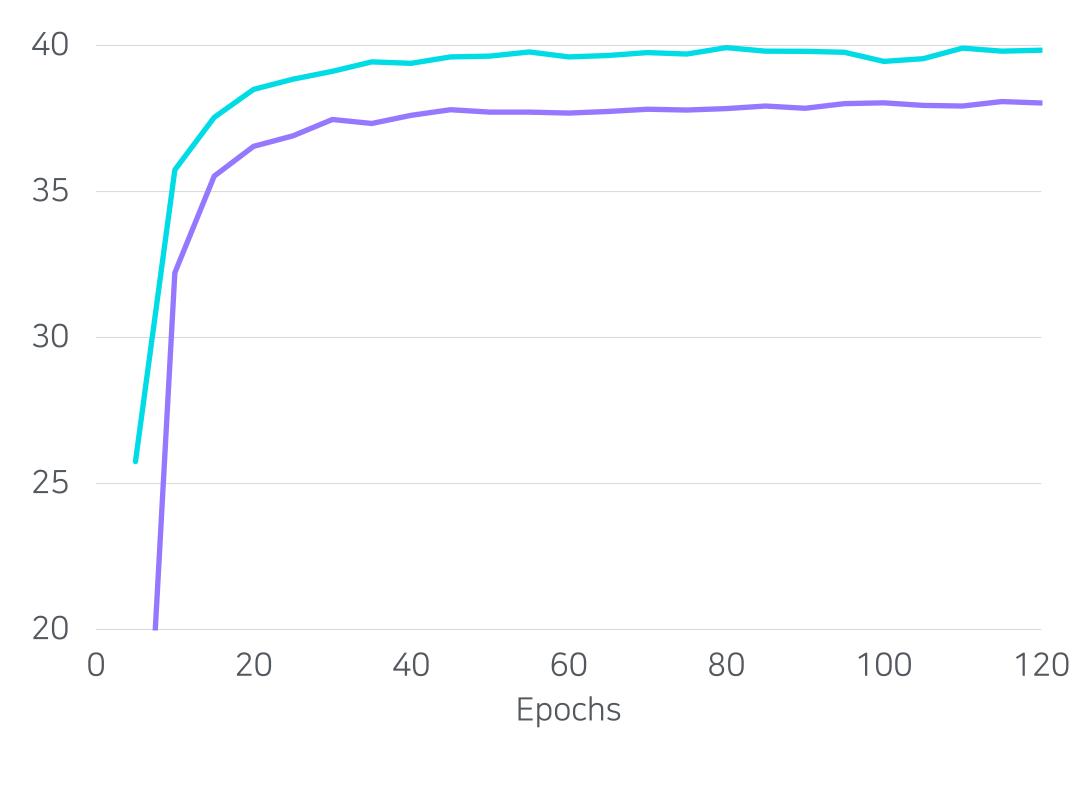
		Extra param
Mul Mul	—Bilingual baseline	+44%
	-Re-training (Garcia et al., 2021)	+0%
	—Only embeddings	+2.6%
	-Adapters (dim=64)	+3.7%
	—Dec adapters (690) + enc adapters last (1024)	+9.2%
80000 100	000	



ns:

4.4 Results: zero-shot translation

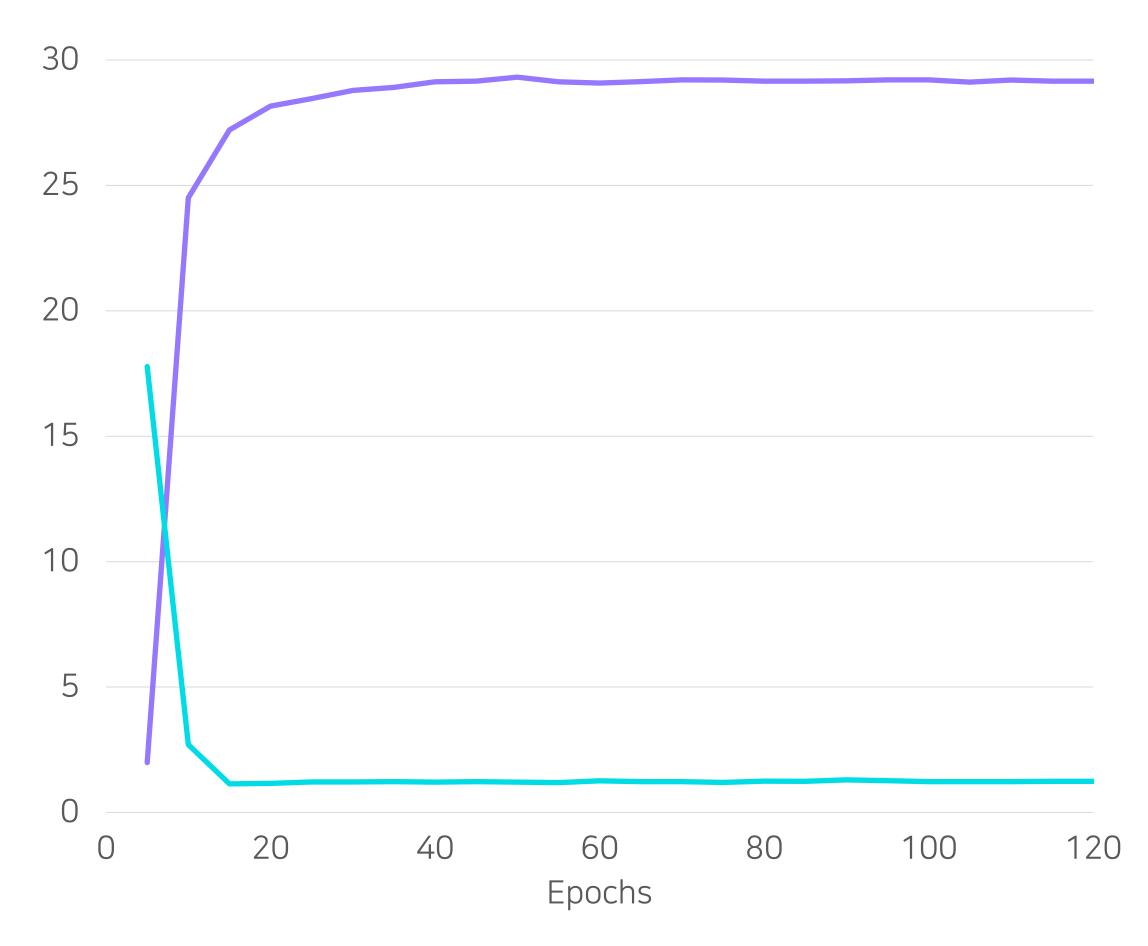
Greek-English BLEU



-Only embeddings Enc adapters (dim=512)



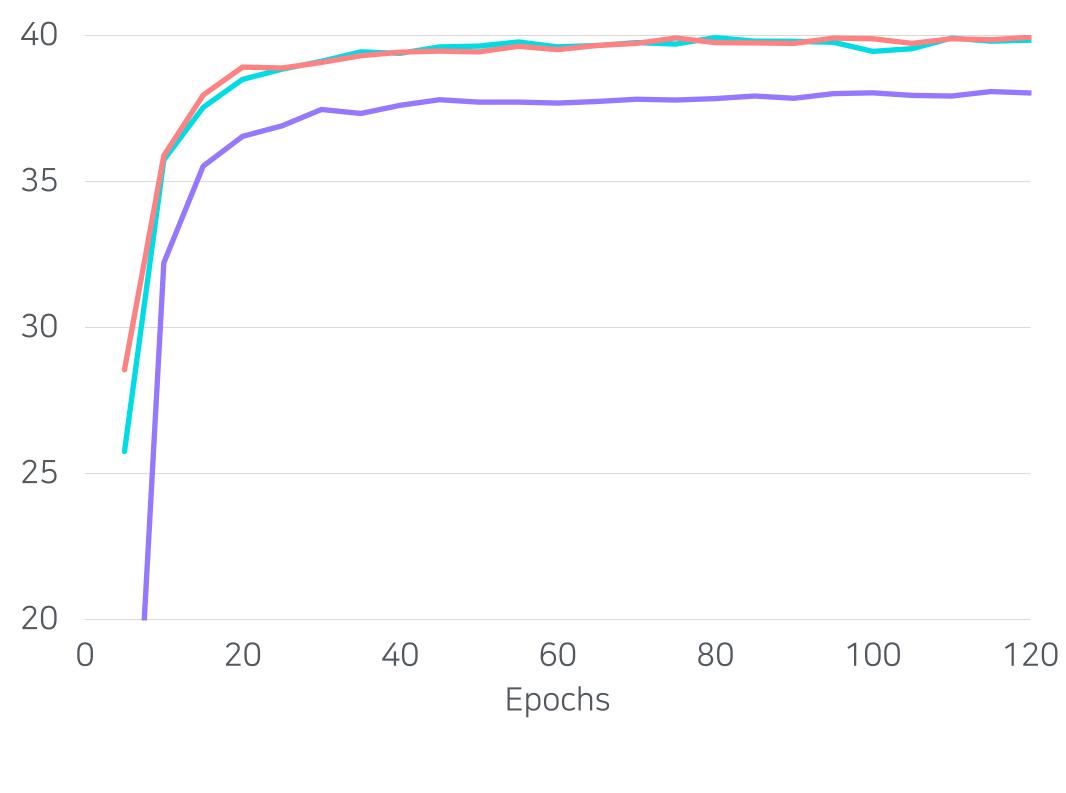
Greek-French BLEU (zero-shot)





4.4 Results: zero-shot translation

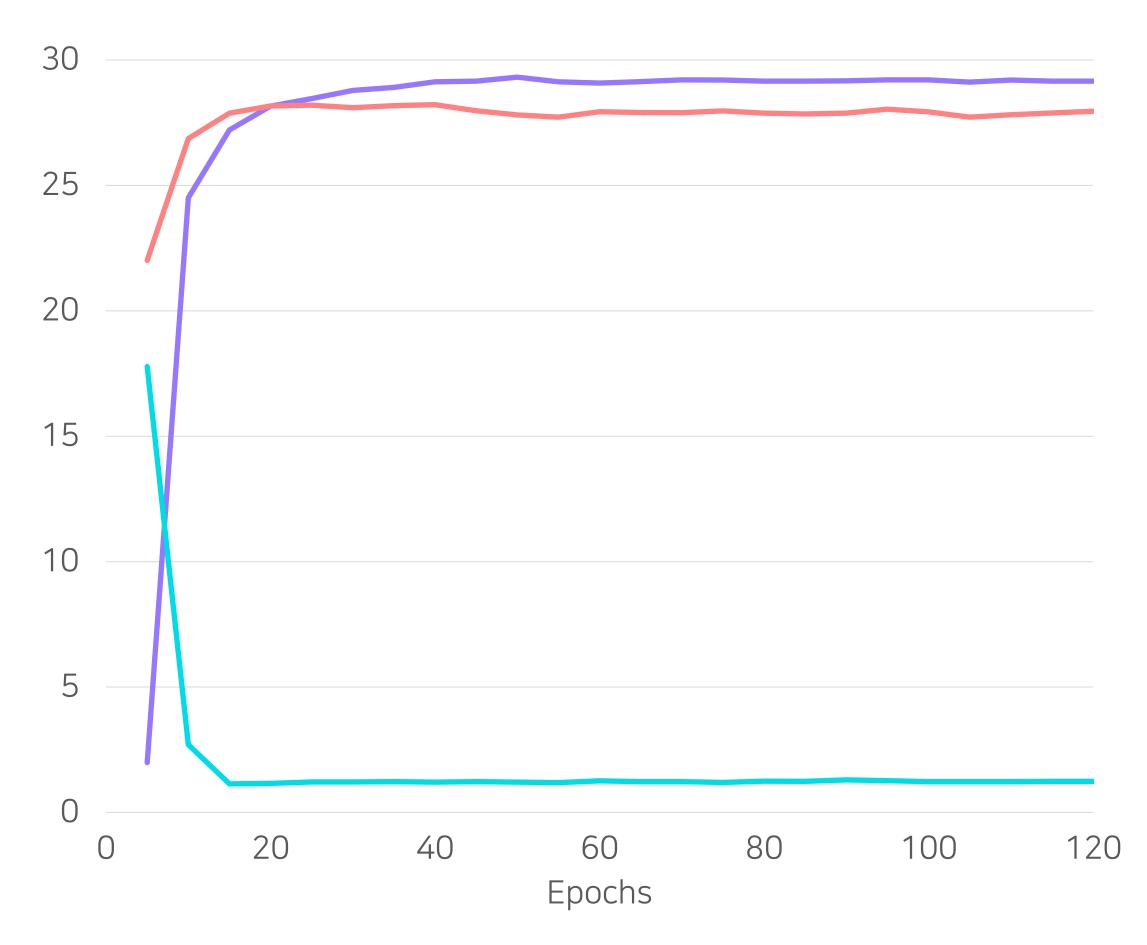
Greek-English BLEU



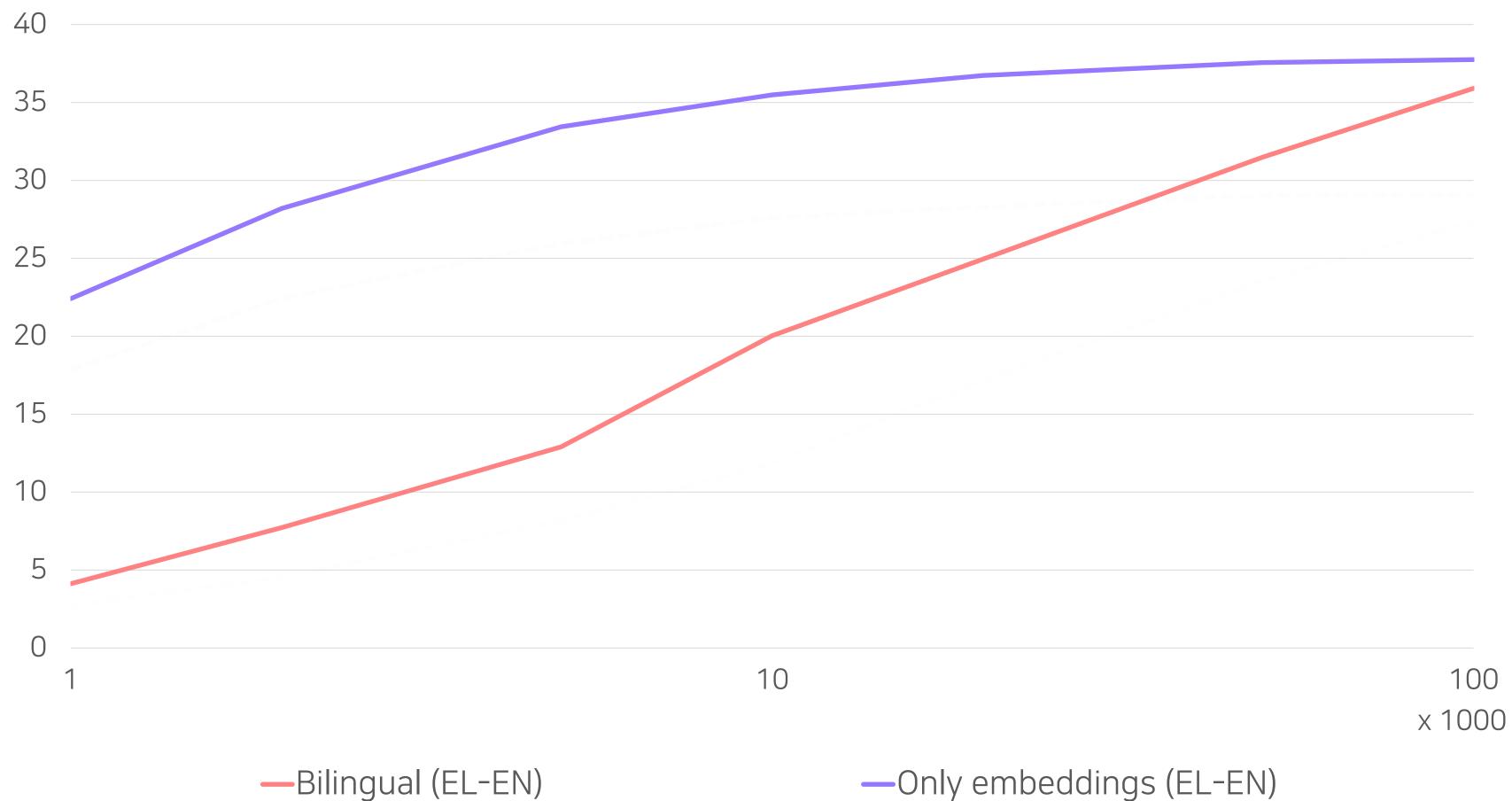
-Only embeddings Enc adapters (dim=512) -Enc adapters (dim=512) + 1k lines per lang (BT)



Greek-French BLEU (zero-shot)





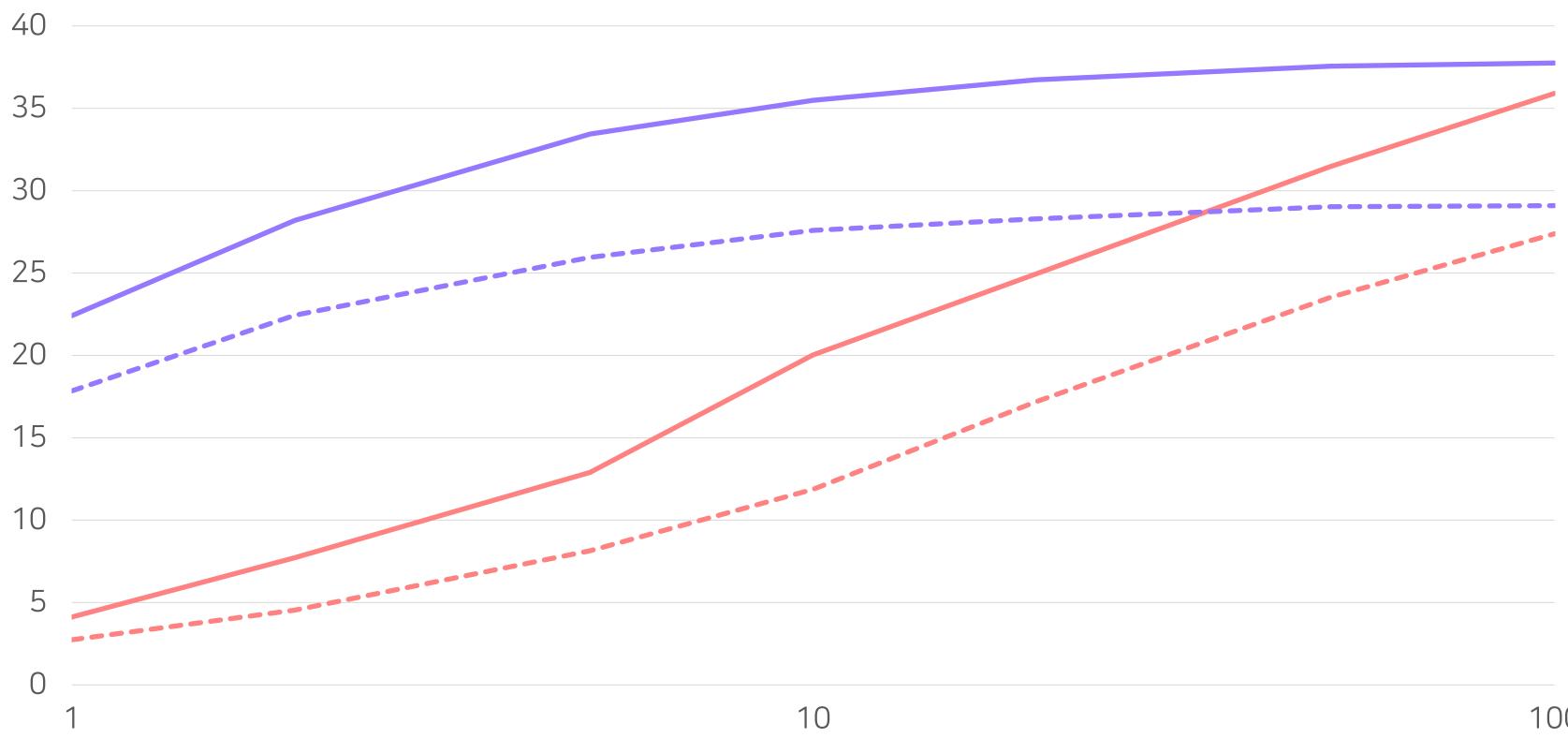


-Bilingual (EL-EN)



BLEU by training data size





—Bilingual (EL-EN) --Bilingual (EL-FR)



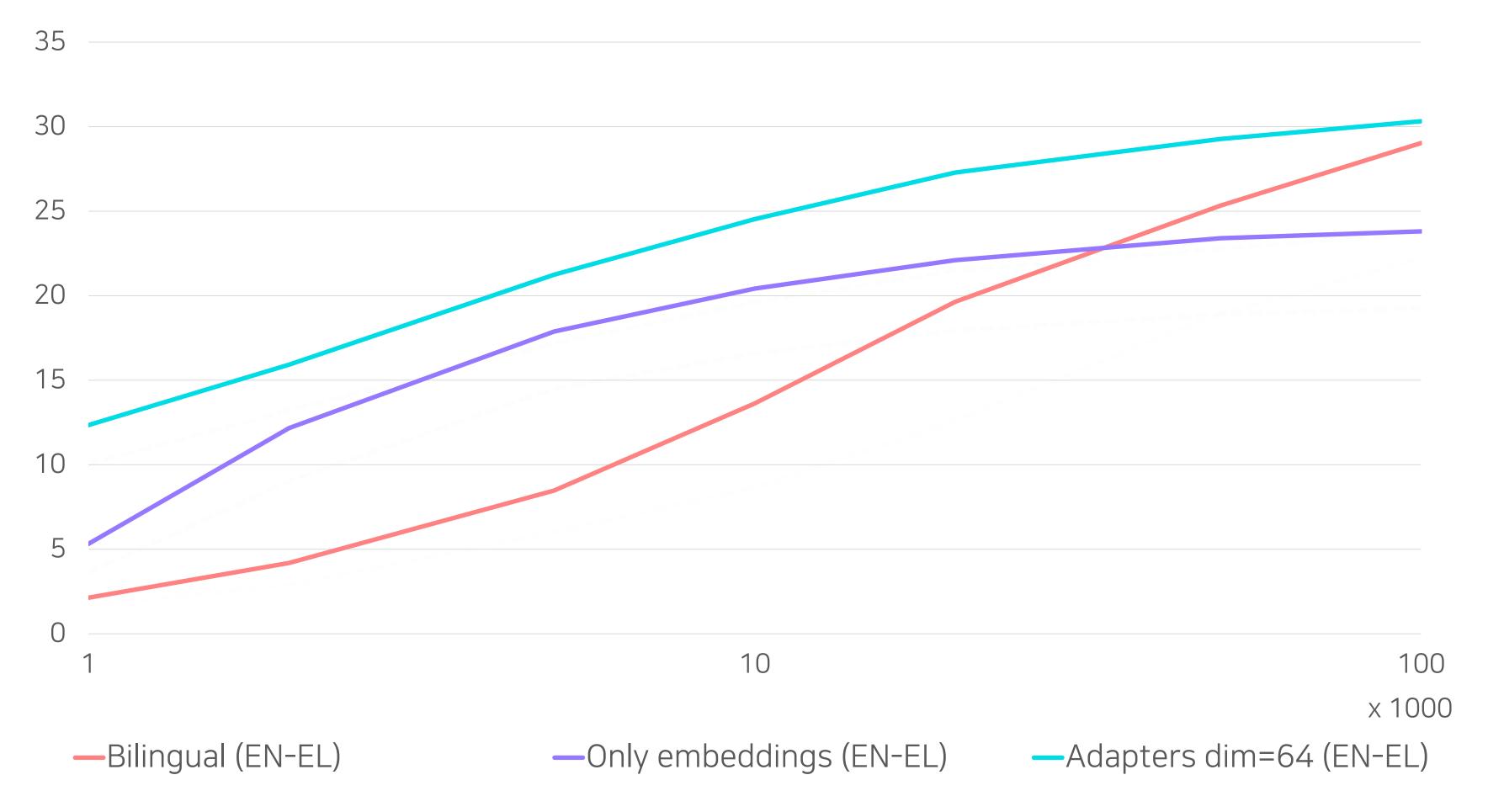
BLEU by training data size

100 x 1000

—Only embeddings (EL-EN)

--Only embeddings (EL-FR)

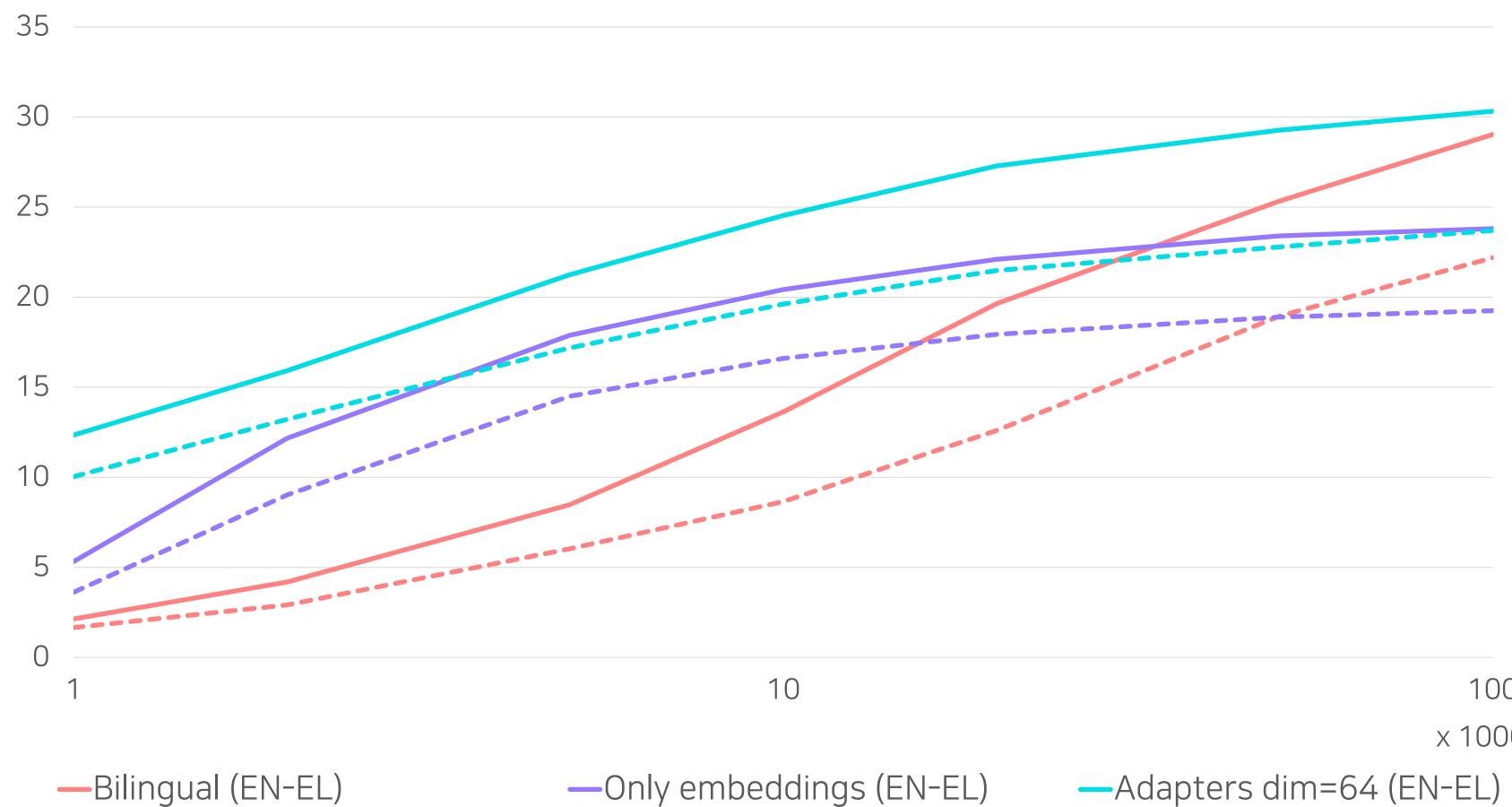






BLEU by training data size





--Bilingual (FR-EL)

- --Only en



BLEU by training data size

10	100
	x 1000
mbeddings (EN-EL)	—Adapters dim=64 (EN-EL)
mbeddings (FR-EL)	Adapters dim=64 (FR-EL)



4.4 Results: new source and target language

Model

- **Bilingual base**
- Re-training + {EL,



	BLEU
elines	14.9
UK, SV, ID}	22.0



4.4 Results: new source and target language

Model

- **Bilingual base**
- Re-training + {EL,

Source model

Only embeddings

Only embeddings

Enc adapters + BT

Test-time combination of source and target params



	BLEU
selines	14.9
, UK, SV, ID}	22.0
Target model	BLEU
Target model Only embeddings	BLEU 19.0



4.5 Conclusion

- How to learn a new source or target language:
 - Create a new vocabulary for that language
- Replace the source (resp. target) shared embeddings by lang-specific ones
- Train the new embeddings plus some adapter modules





4.5 Conclusion

- How to learn a new source or target language:
 - Create a new vocabulary for that language
 - Replace the source (resp. target) shared embeddings by lang-specific ones
- Train the new embeddings plus some adapter modules
- Zero-shot translation issue solved with tiny amounts of back-translation





4.5 Conclusion

- How to learn a new source or target language:
 - Create a new vocabulary for that language
 - ones
- Train the new embeddings plus some adapter modules
- Zero-shot translation issue solved with tiny amounts of back-translation
- Translation between 2 new languages by combining their lang-specific params

- Replace the source (resp. target) shared embeddings by lang-specific





5. Learning new languages without parallel data

Multilingual Unsupervised Neural Machine Translation with Denoising Adapters



Istun, A. Berard, L. Besacier and M. Galle **EMNLP 2021**

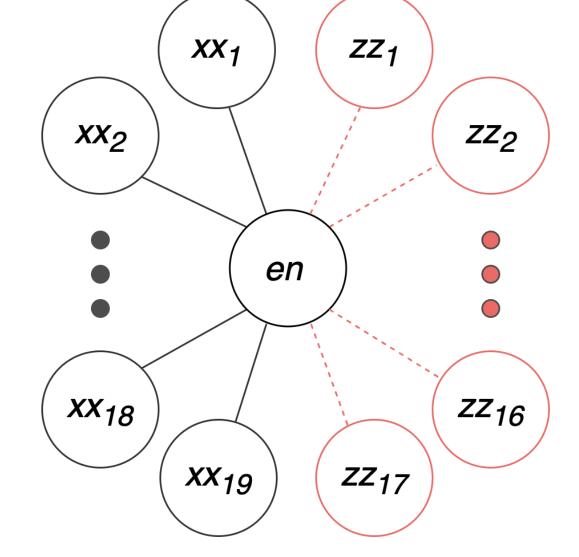


5.1 Introduction

Unsupervised MNMT

- A single multilingual NMT model that can translate from/into multiple languages with *incomplete* parallel data
- Learn from both parallel data ($EN \leftrightarrow XX_n$) and monolingual data (ZZ_n)
- Add a new language (ZZ_n) to an existing MNMT model
 - Without retraining the full model
 - Using only monolingual data





Overview of our multilingual UNMT setup. Figure adapted from Garcia et al. (2020)



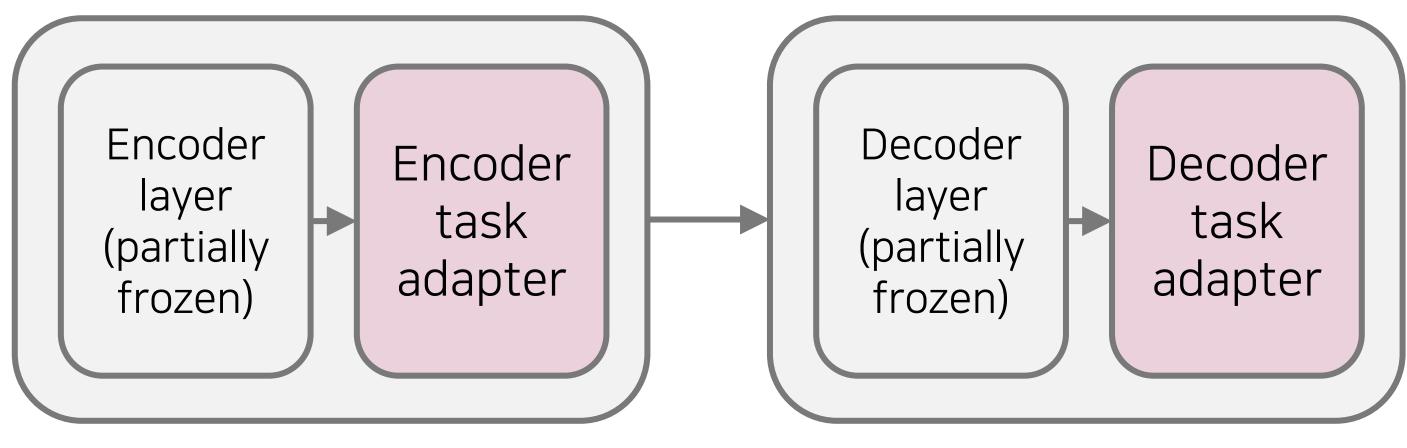


5.2 mBART

Starting point: mBART50 (Tang et al., 2020) A 50-language sequence-to-sequence model trained with a denoising objective. Cannot do MT but can improve final performance when used as initialization.

Can be adapted to MT with parallel data, by:

- Full fine-tuning
- Partial fine-tuning (e.g., cross-attention) + task adapters (Stickland et al., 2021) —

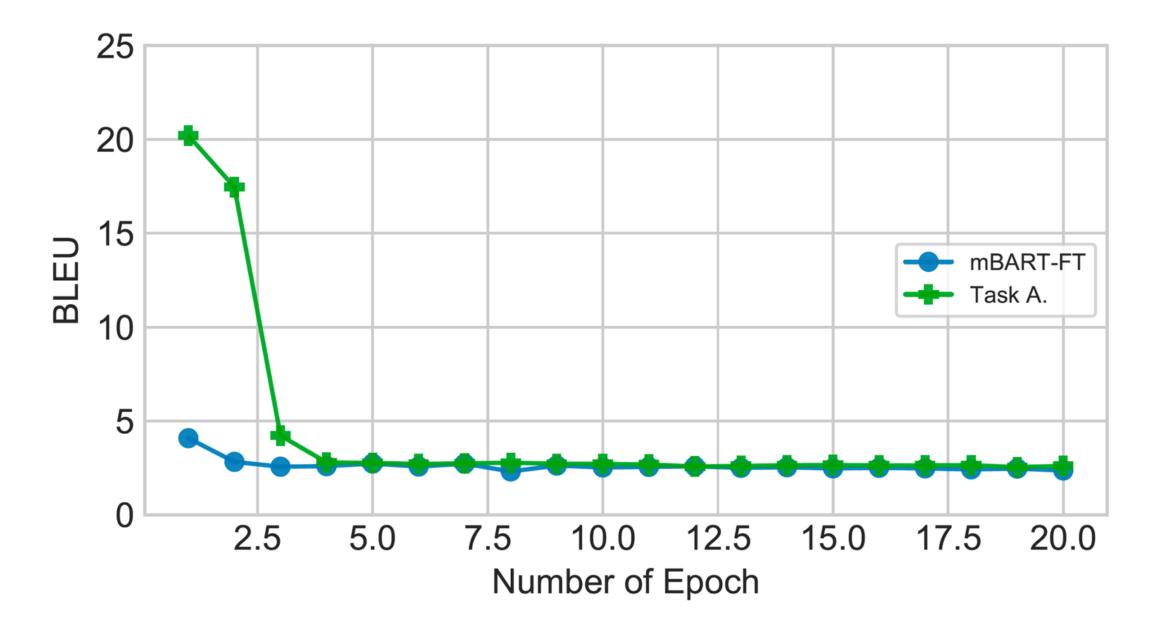






5.2 mBART: unsupervised MNMT

1. Fine-tune mBART with parallel data in a subset of N languages 2. Use it translate into/from the other (50 – N) languages





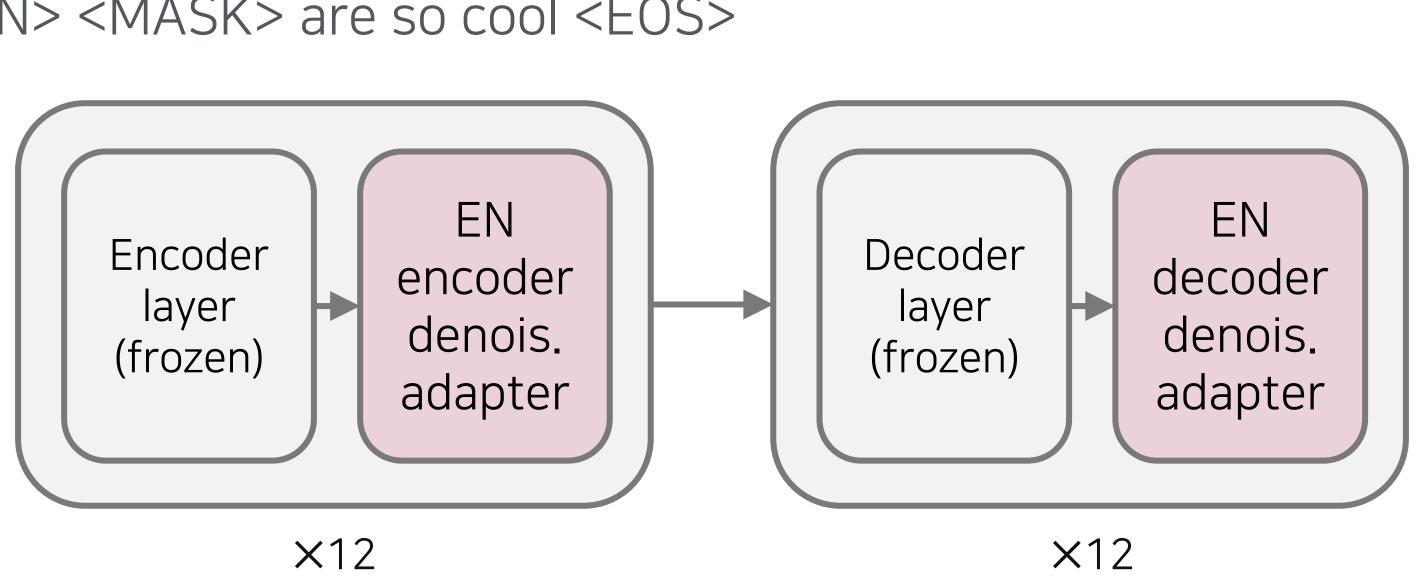
- **Issue:** when fine-tuned (even partially), mBART quickly forgets about the other languages

Unsupervised EN \rightarrow NL performance when fine-tuning mBART on 19 language pairs



5.3 Technique

<EN> <MASK> are so cool <EOS>



1. Train denoising adapters for all languages with monolingual data

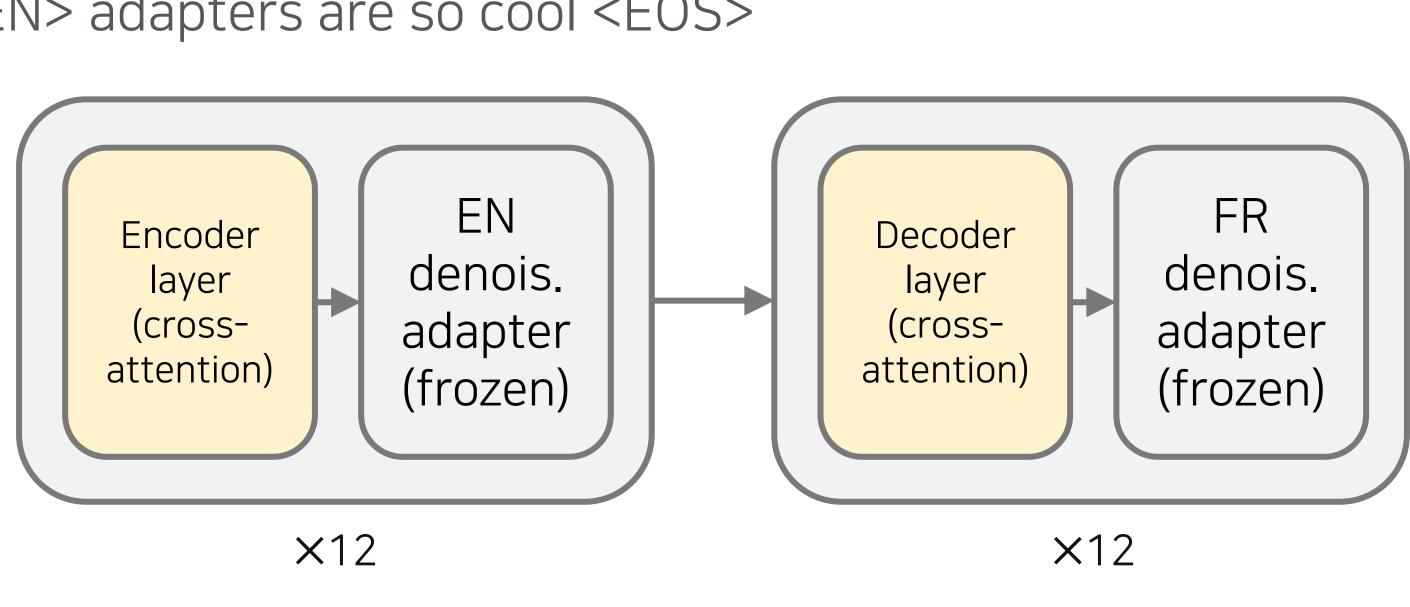


<EN> adapters are so cool <EOS>



5.3 Technique

<EN> adapters are so cool <EOS>



2. Plug-in denoising adapters and fine-tune cross-attention on the available parallel data

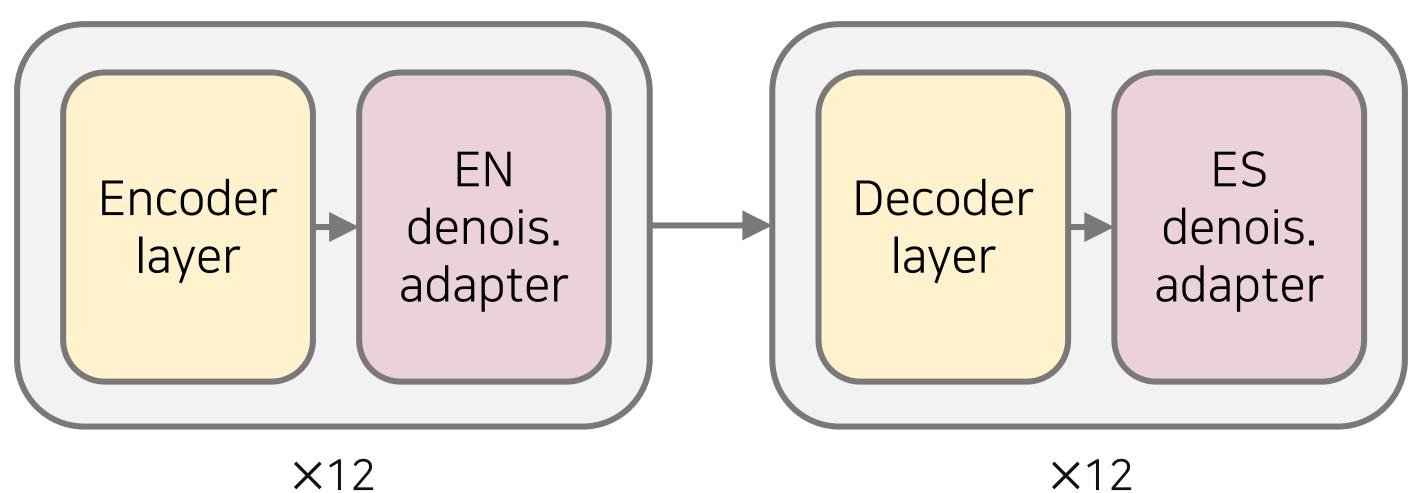


<FR> les adapteurs sont tellement cool <EOS>



5.3 Technique

<EN> adapters are so cool <EOS>



3. Plug-in the denoising adapters of an unsupervised language (ES) and translate



<ES> las adaptadoras son tan geniales <EOS>



5.4 Experiments



20 languages with both English-centric parallel data (TED Talks) and monolingual data (Wikipedia and News)



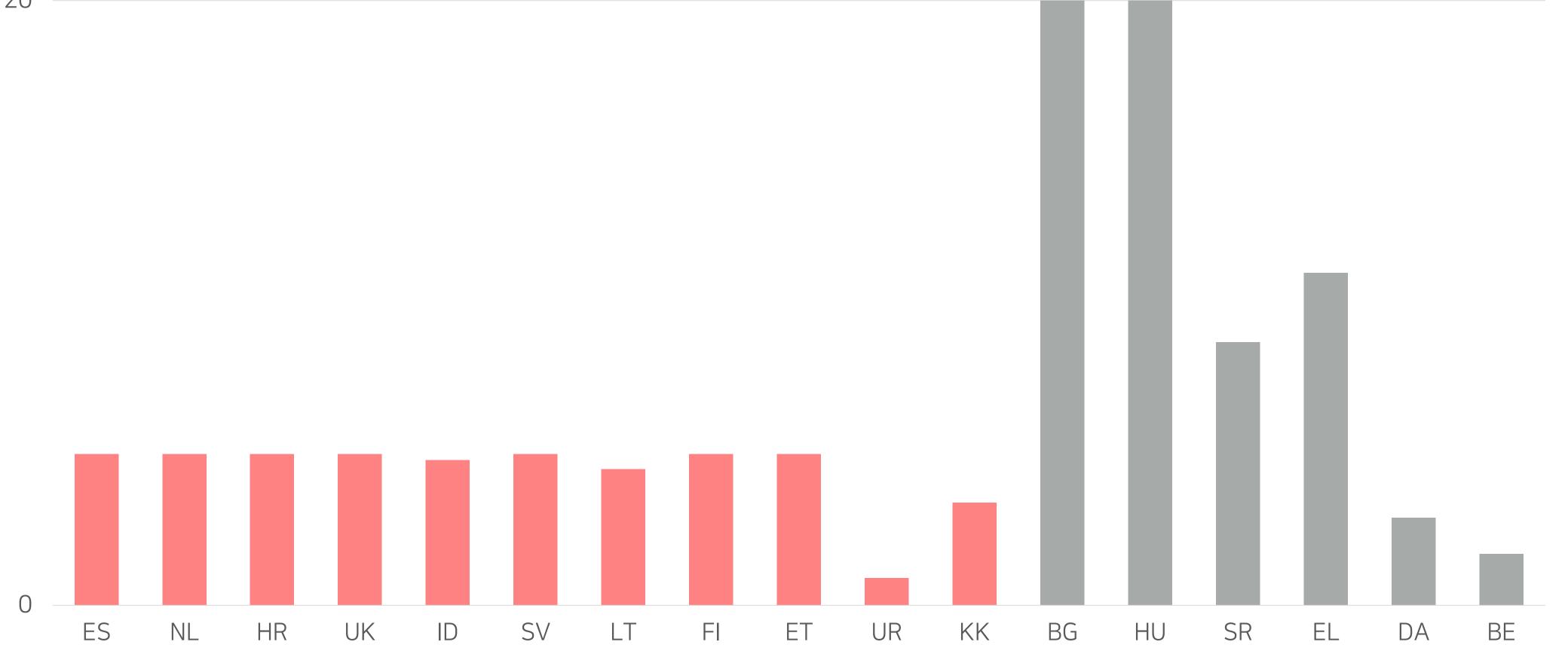




5.4 Experiments

20

×1 000 000



17 languages with only monolingual data 6 languages unknown by mBART





5.4 Experiments

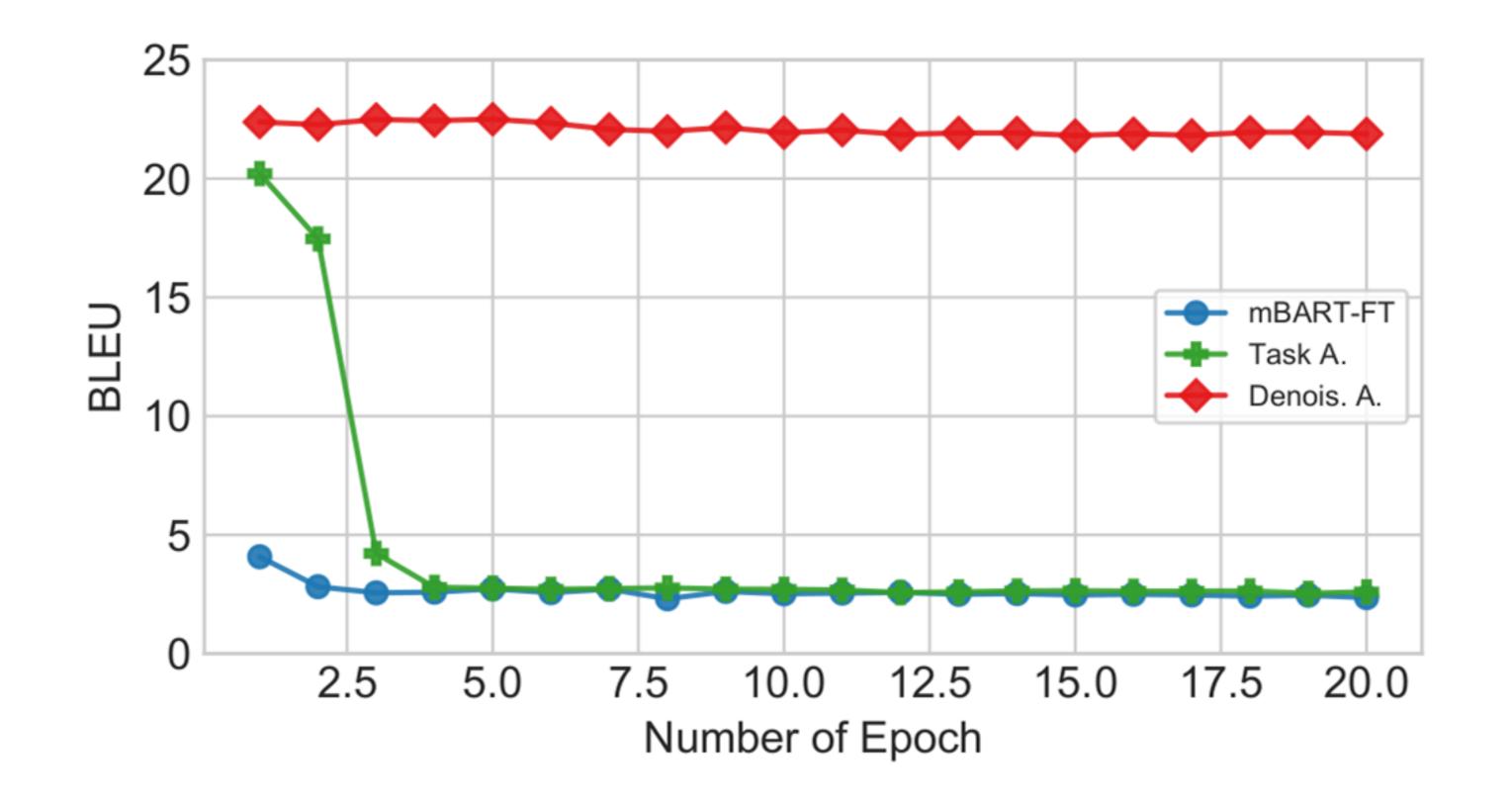
Baselines

- Bilingual supervised models (Bilingual)
- mBART full fine-tuning (mBART-FT)
- Fine-tune cross-attention + task adapters (Task Adapters)
- Same models with back-translation (+ BT)



gual) -T) adapters (Task Adapters) on (+ BT)

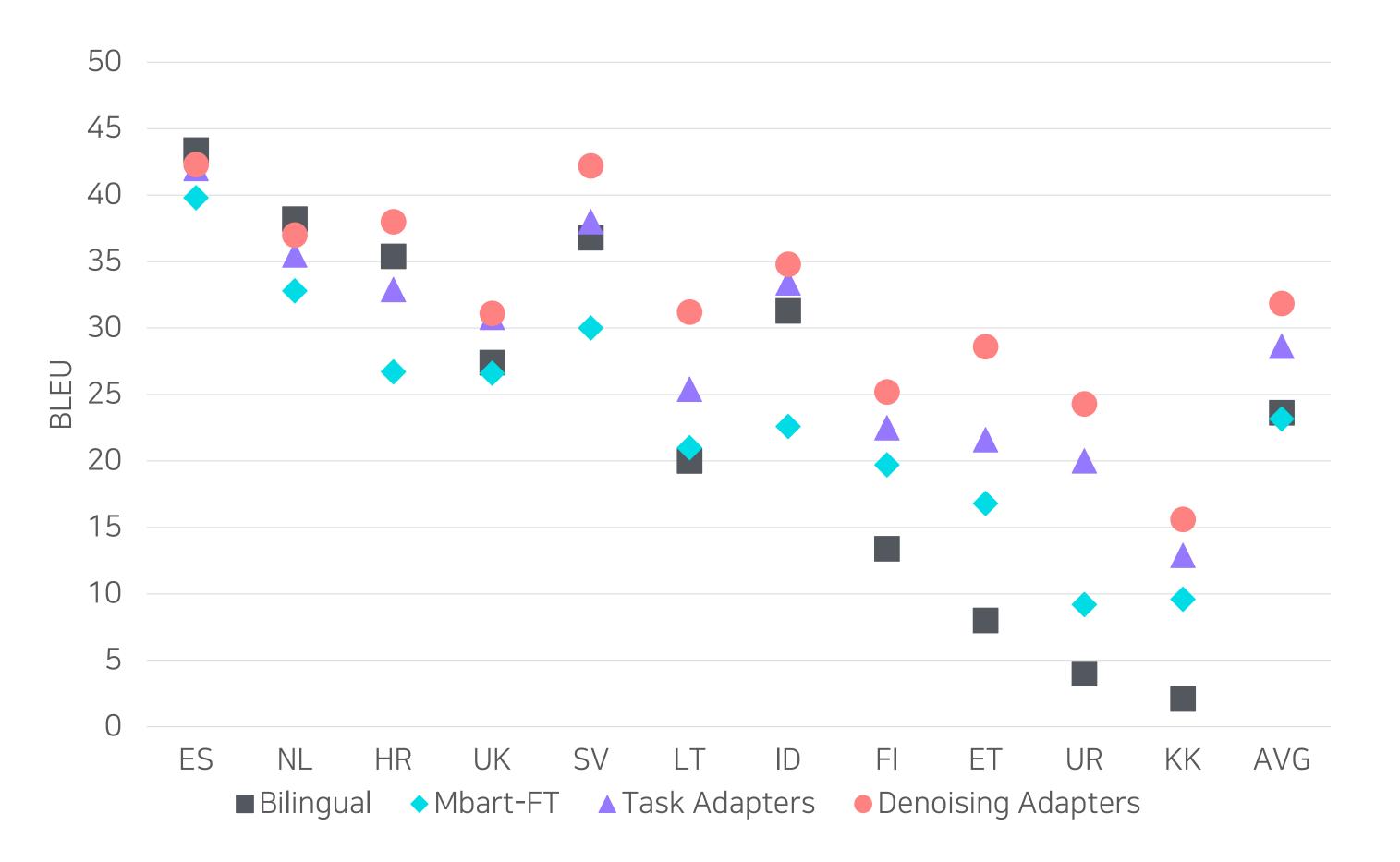




Unsupervised EN \rightarrow NL performance when fine-tuning mBART on 19 language pairs



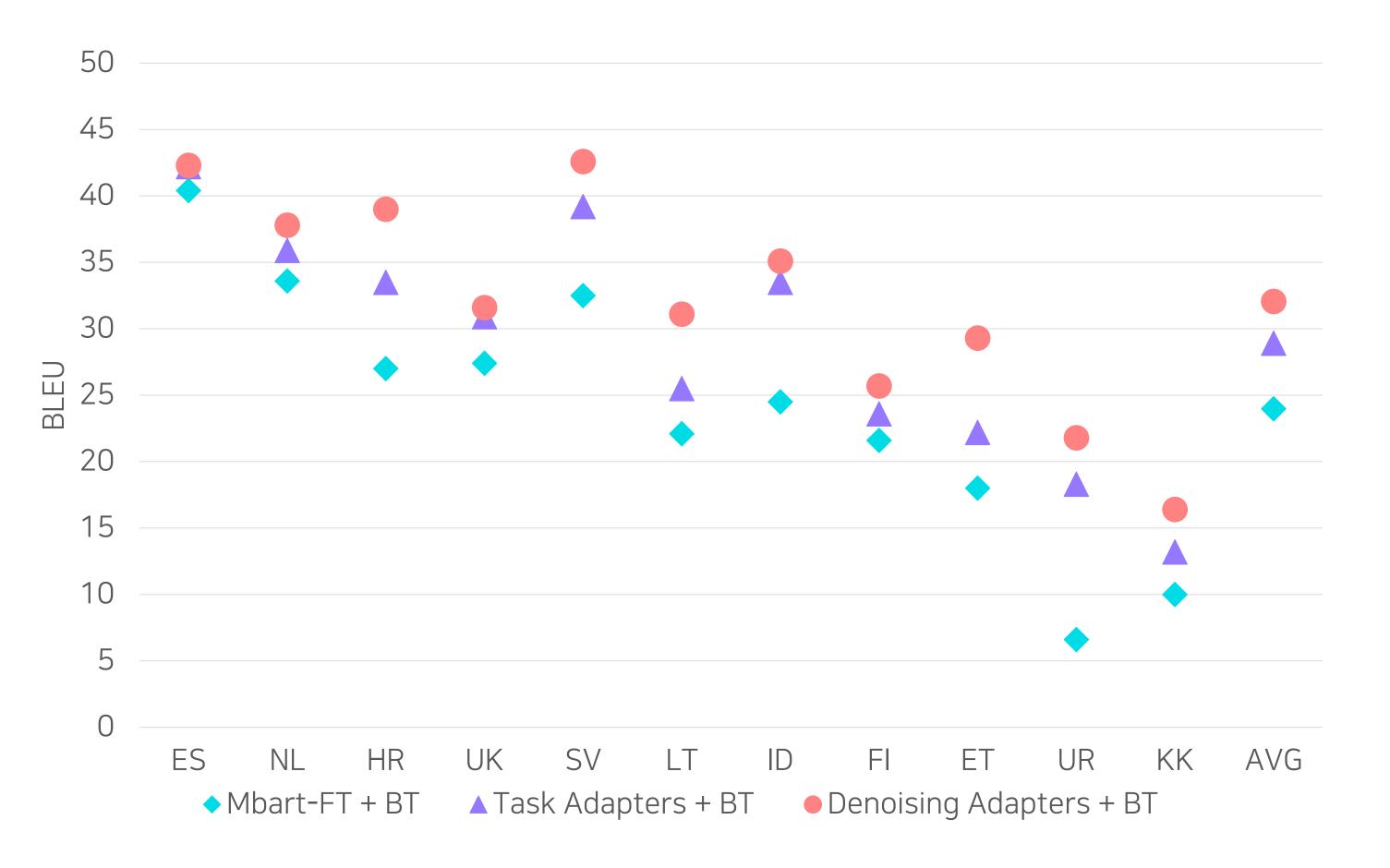






Unsupervised translation into English

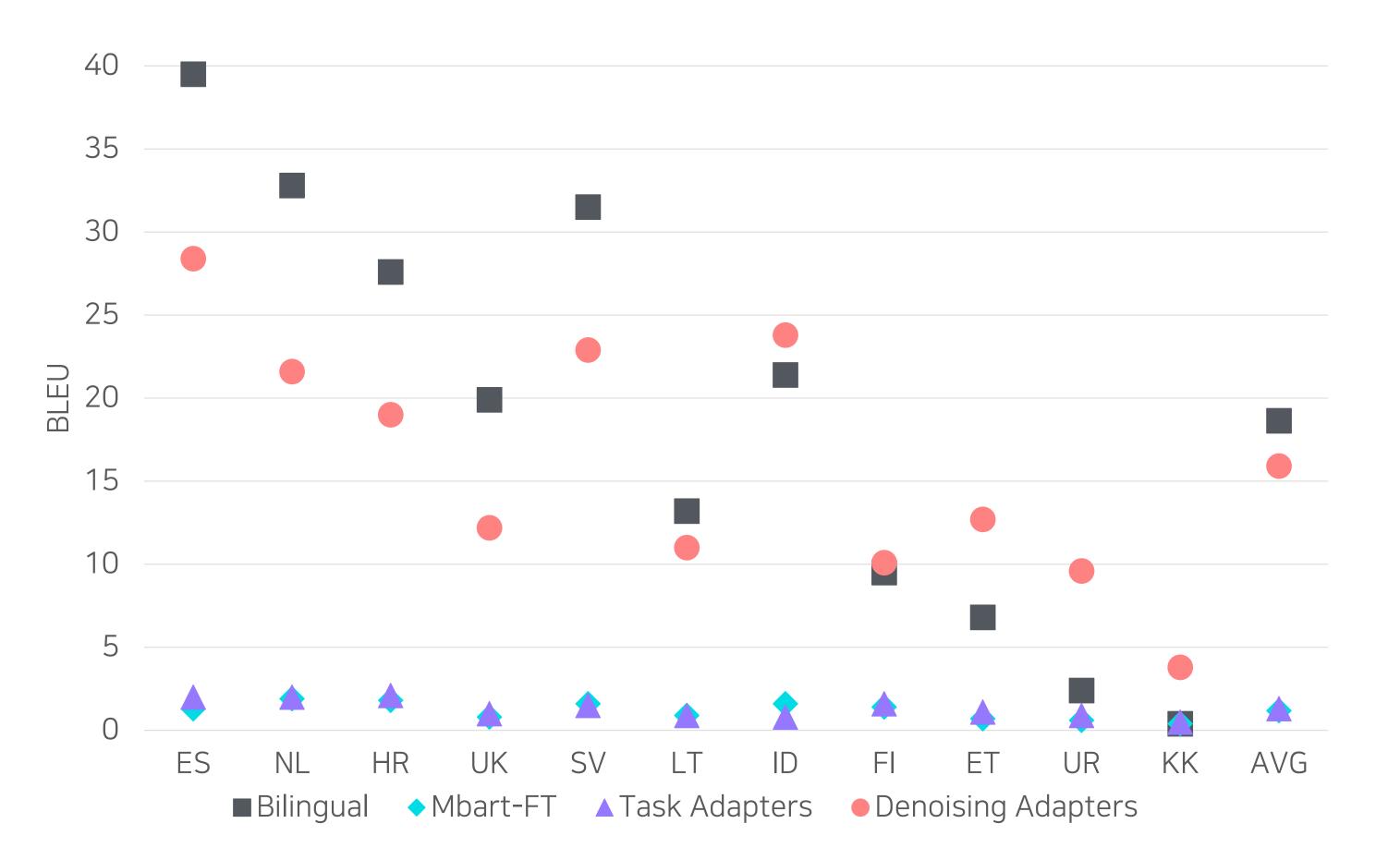






Unsupervised translation into English with back-translation

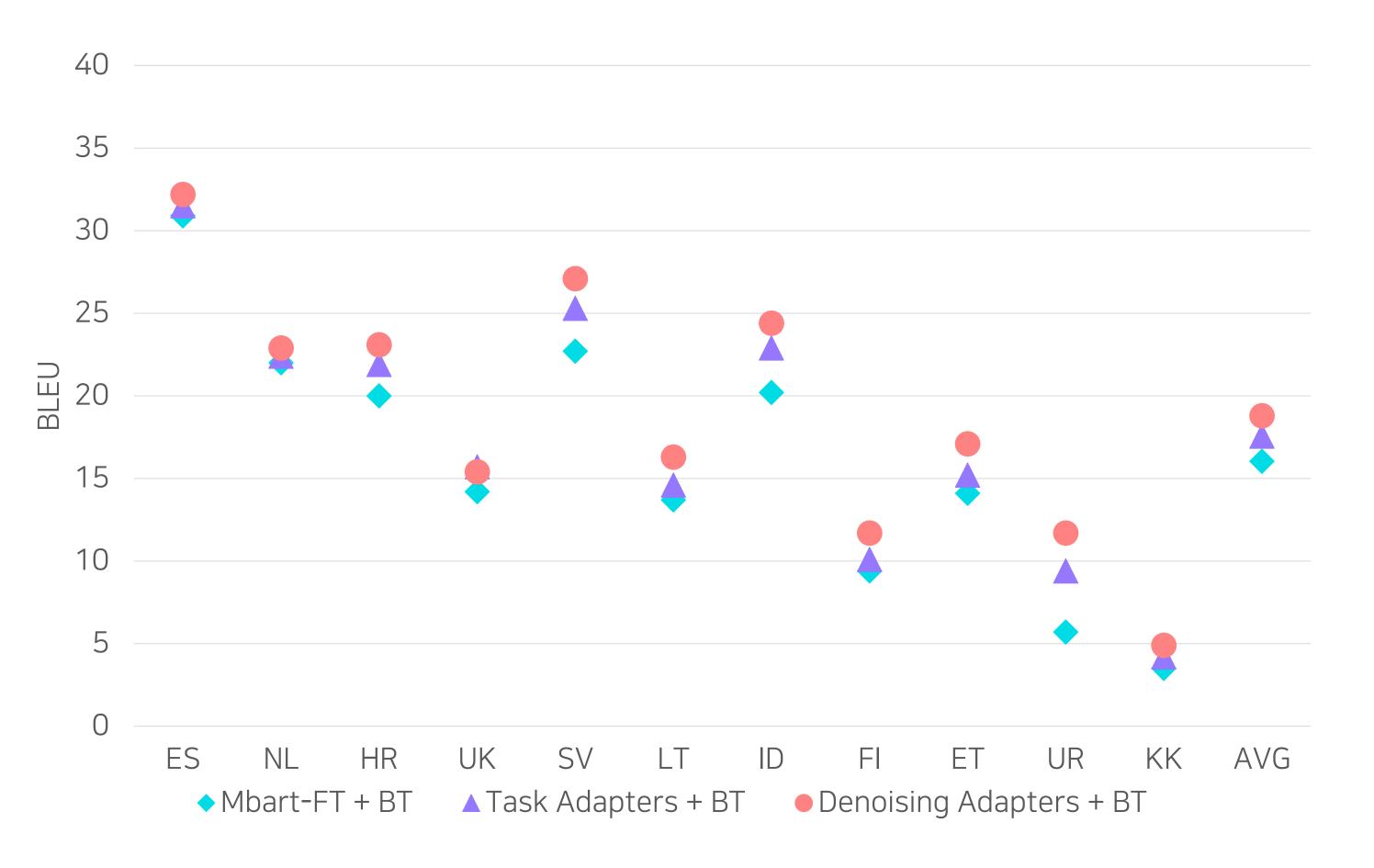




Unsupervised translation from English



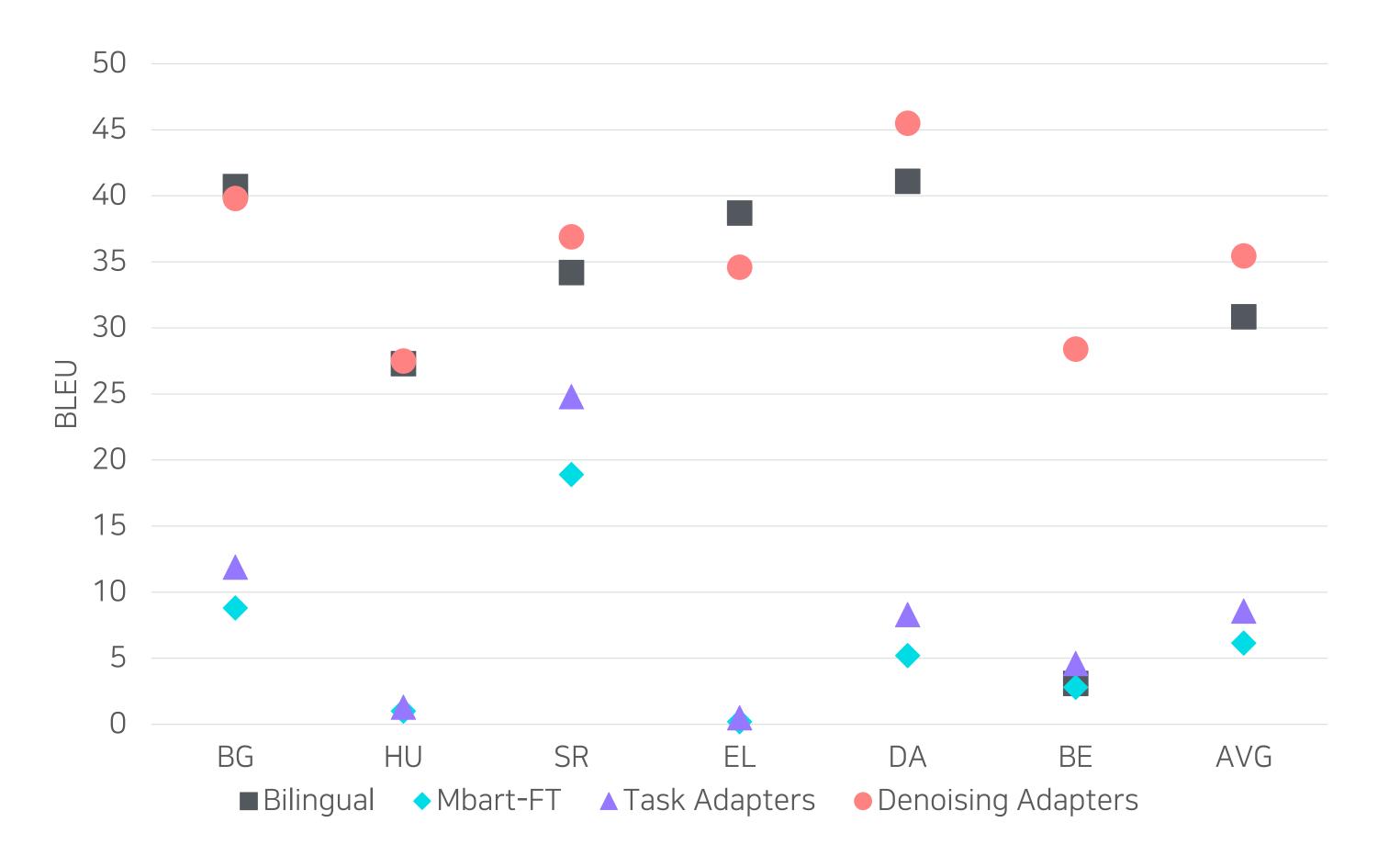






Unsupervised translation from English with back-translation

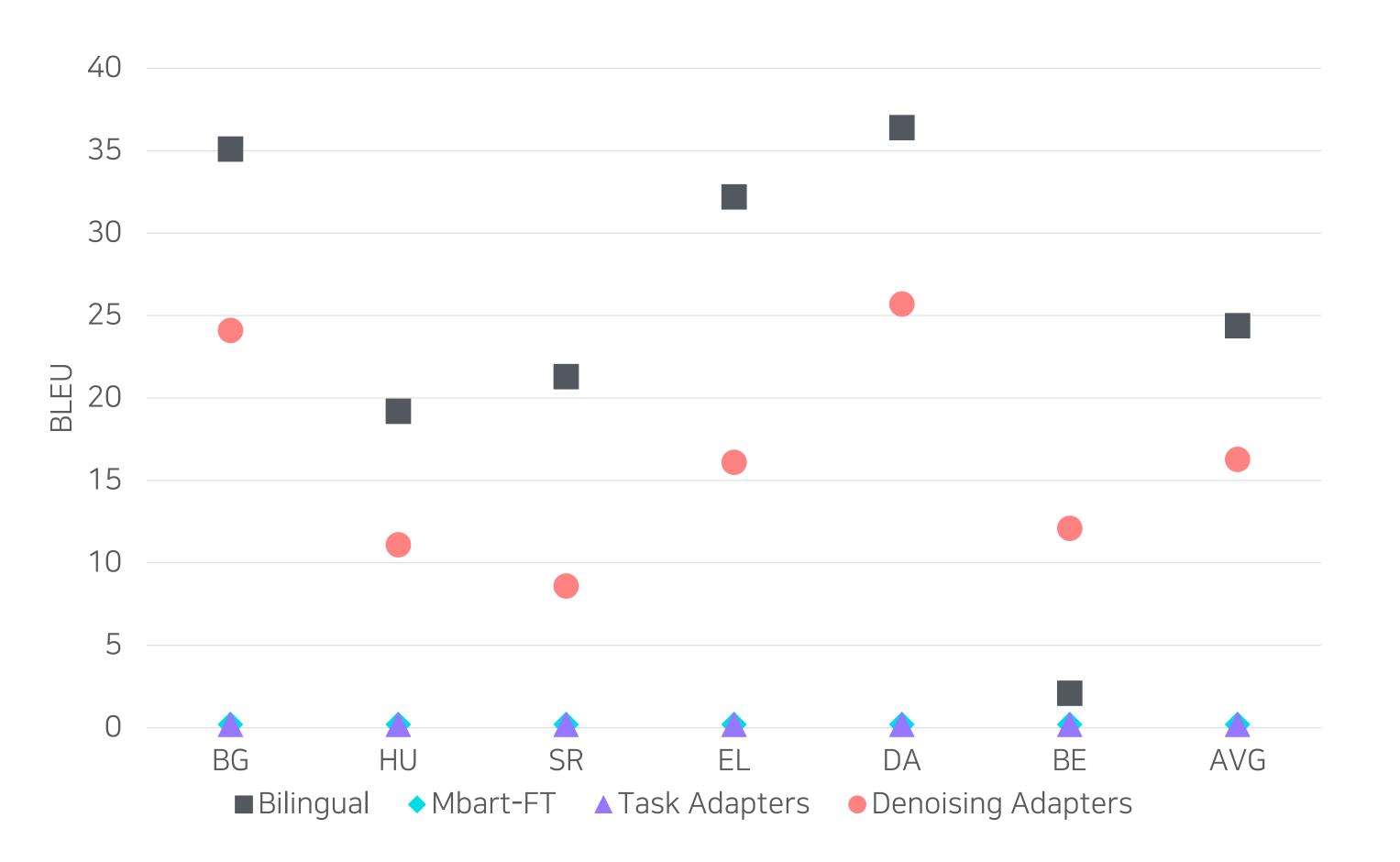




Unsupervised translation into English from languages unseen by mBART







Unsupervised translation from English into languages unseen by mBART





5.6 Conclusion

- Adapting mBART50 to multilingual NMT comes with challenges
 - Multilingual parallel data is needed
 - Poor performance for languages NOT covered by parallel data —
- We propose denoising adapters, monolingually-trained adapter layers to leverage monolingual data for unsupervised MT
- Our experiments on a large set of languages show the effectiveness of denoising adapters with and without BT
- We also show that denoising adapters can be used to add languages unknown by mBART





Takeaways

challenges

- Larger models are slower at inference
- Need for parameter-efficient domain and language adaptation

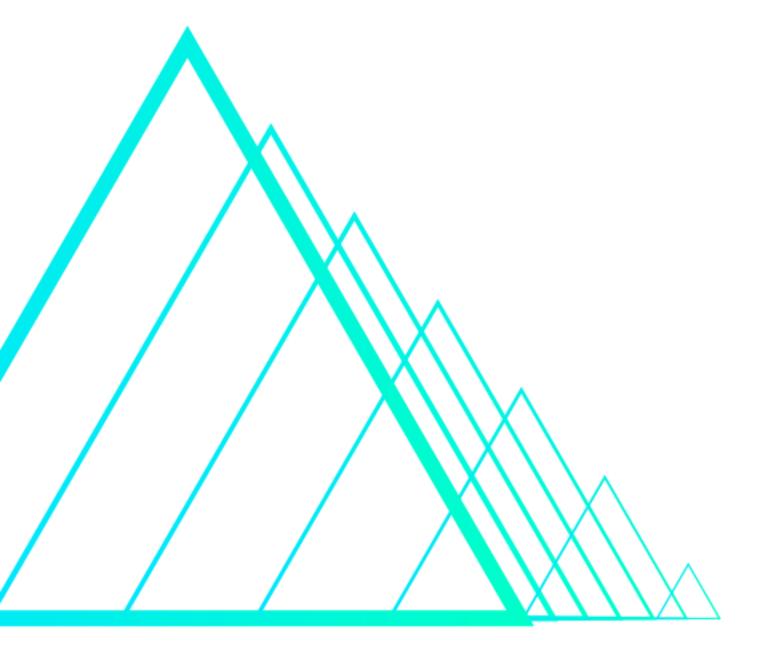
Towards continual learning in Multilingual NMT

- Learn new languages efficiently

- Learn new languages without parallel data

Multilingual NMT is appealing in production but it comes with





Thank You

