Language-Family Adapters for Low-Resource Multilingual Neural Machine Translation

Alexandra Chronopoulou, Dario Stojanovski, Alexander Fraser

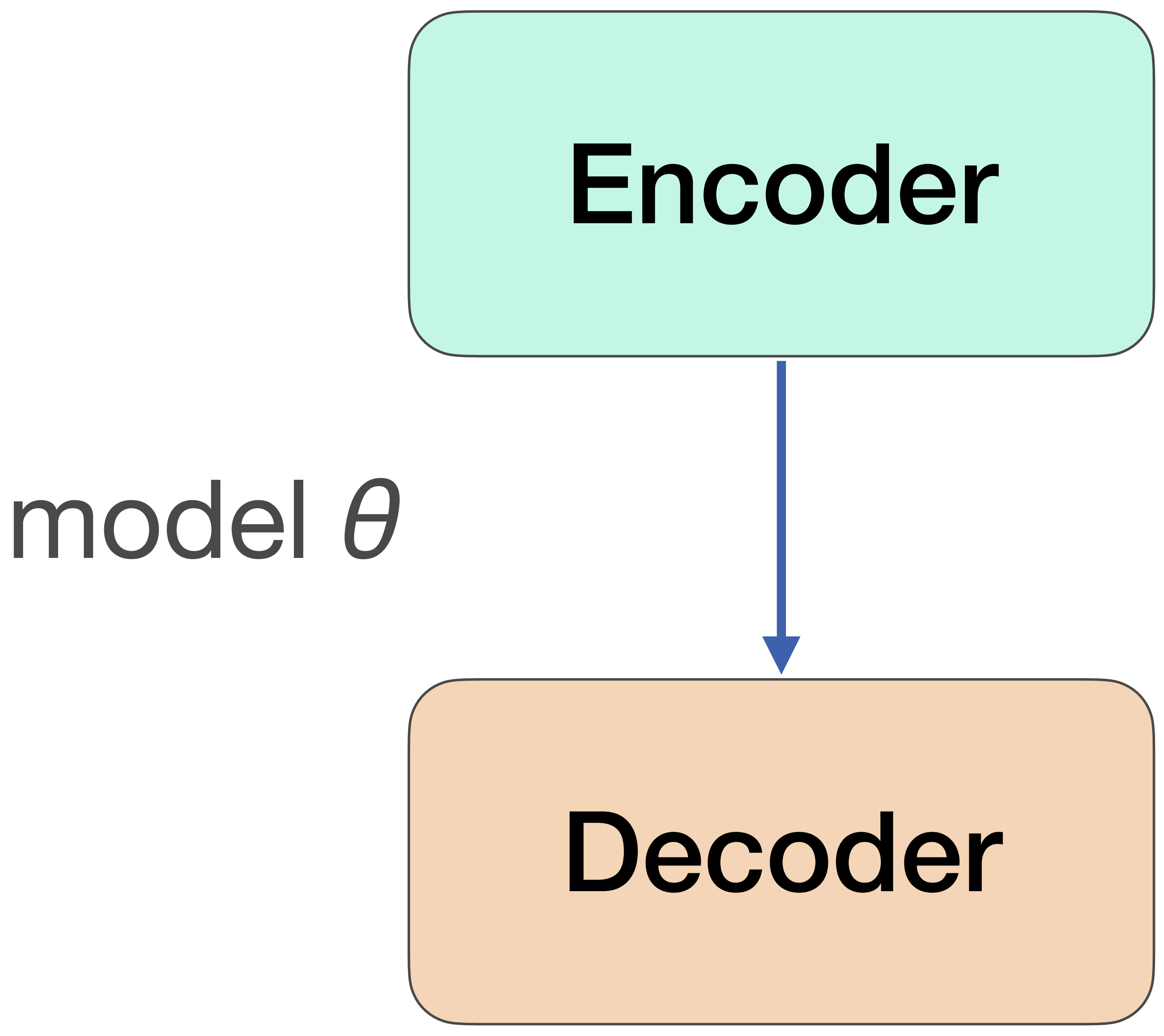
LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Presentation outline

- Motivation
- Proposed Approach
- Experiments
- Conclusion

- **Motivation**
- Proposed Approach
- Experiments
- Conclusion

Overview: Multilingual NMT

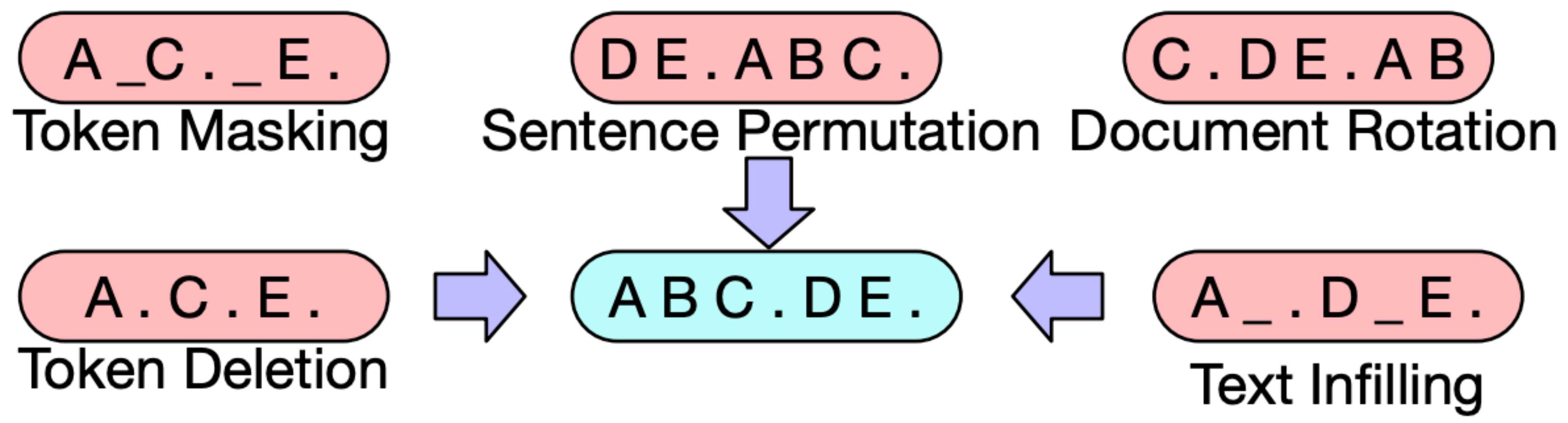


- Low-resource languages benefit from sharing the same representation space as high-resource languages *(Firat et al., 2016; Zoph et al., 2016; Johnson et al., 2017)*

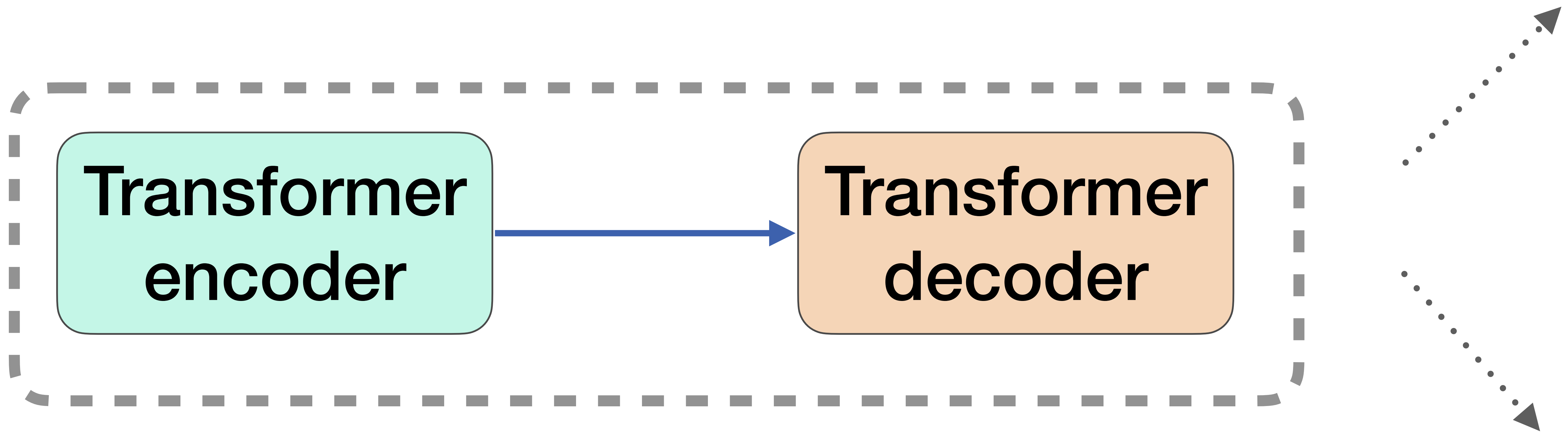
- Operational costs are reduced and models scale to a large number of language pairs *(Arivazhagan et al., 2019; Aharoni et al., 2019)*

mBART-50: A multilingual pretraining model (Tang et al., 2020)

- Encoder-decoder Transformer
- Denoising autoencoding in multiple languages *(Lewis et al., 2020, Liu et al., 2020)*
- Monolingual data of 50 languages during pre-training
- Has not been trained for MT



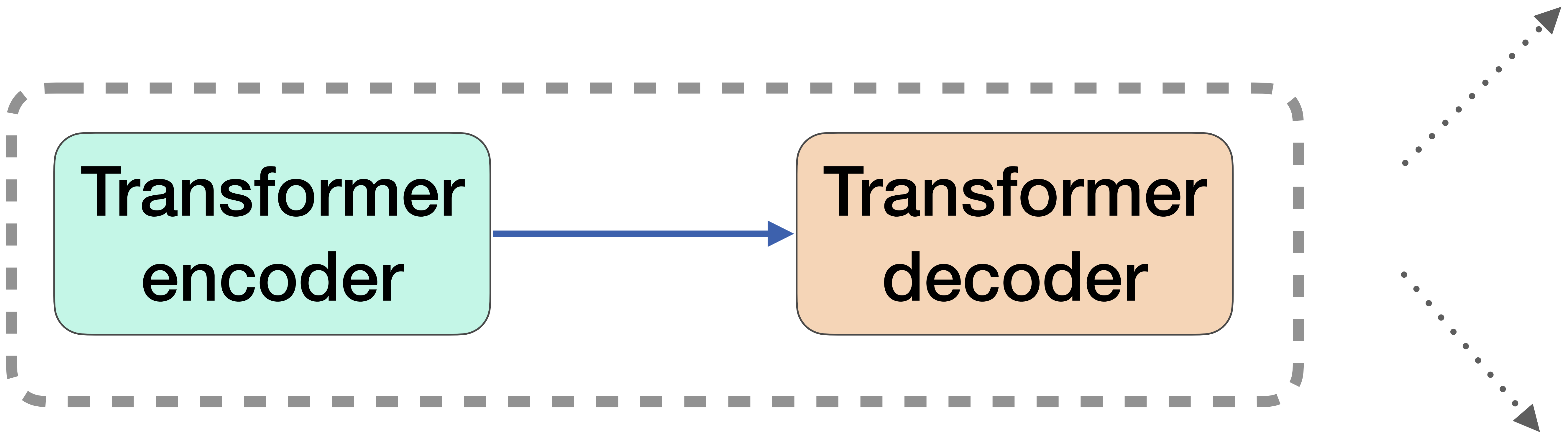
mBART-50 for NMT



Fine-tune all parameters for NMT (English-to-many)

Fine-tune all parameters for NMT (many-to-English)

mBART-50 for NMT

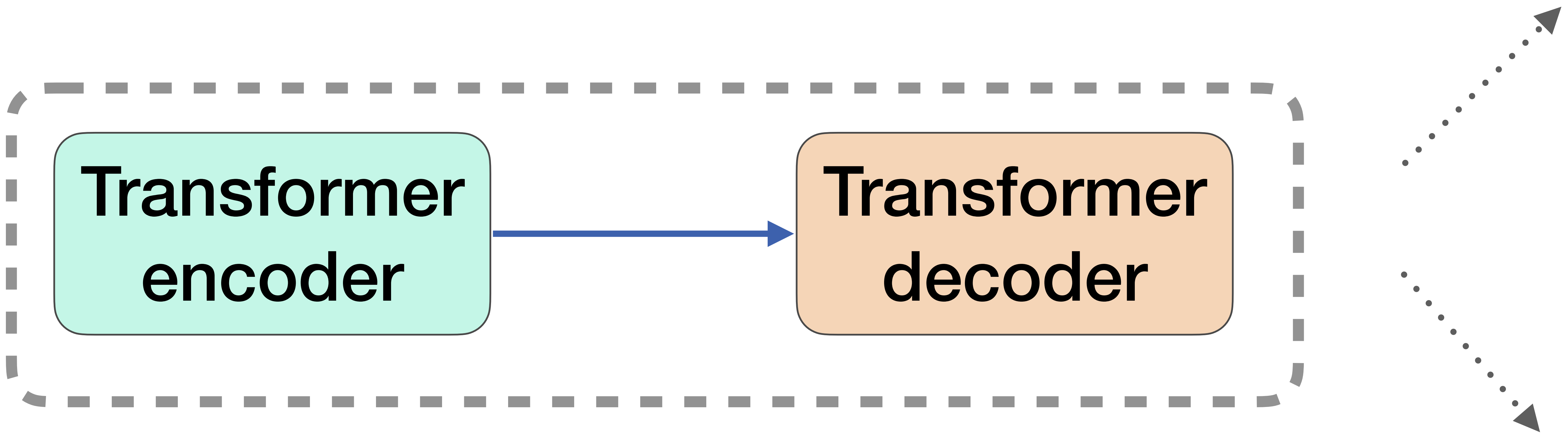


Fine-tune all parameters for NMT (English-to-many)

Fine-tune all parameters for NMT (many-to-English)

- Not all languages are modeled equally well

mBART-50 for NMT



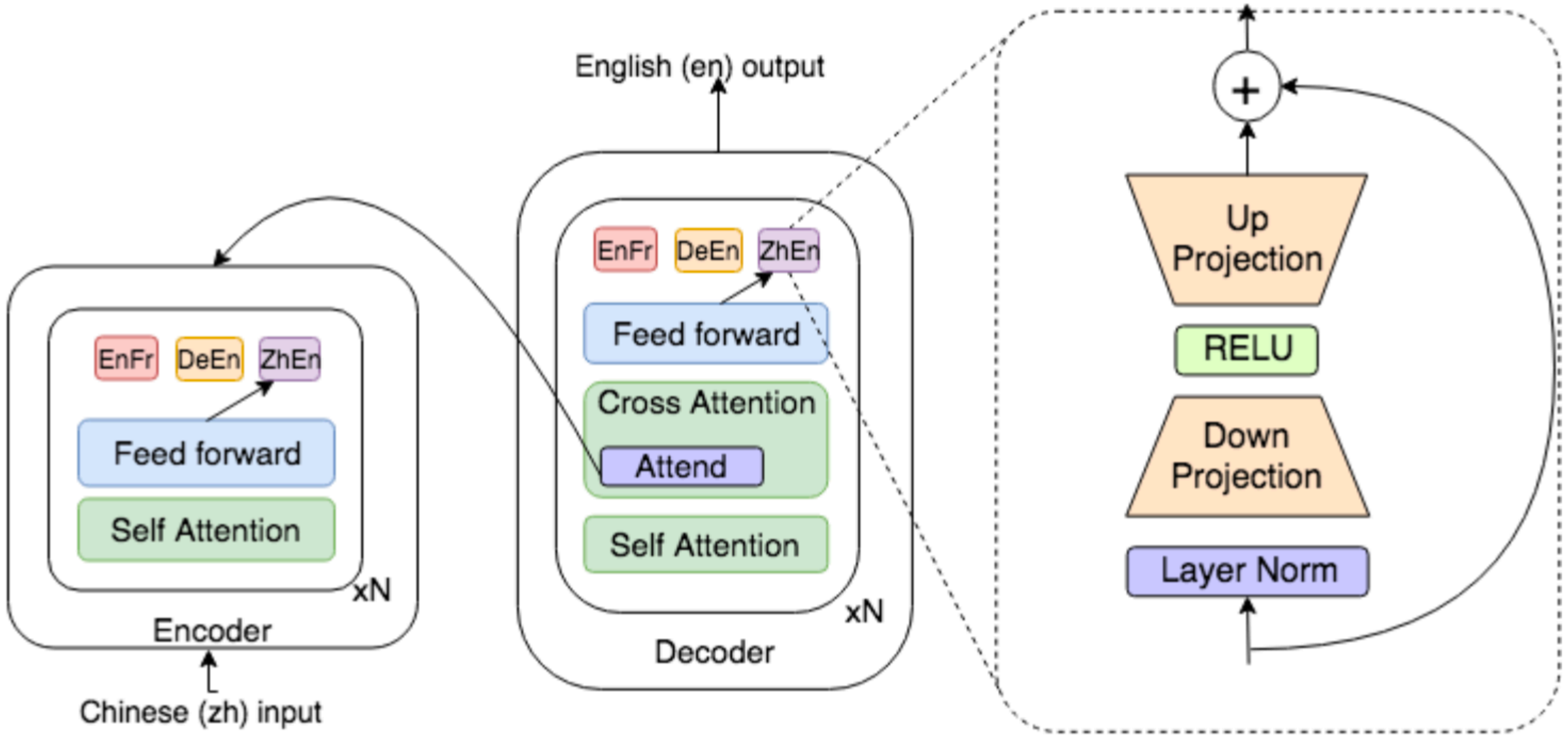
Fine-tune all parameters for NMT (English-to-many)

Fine-tune all parameters for NMT (many-to-English)

- Not all languages are modeled equally well
- The entire model needs to be updated

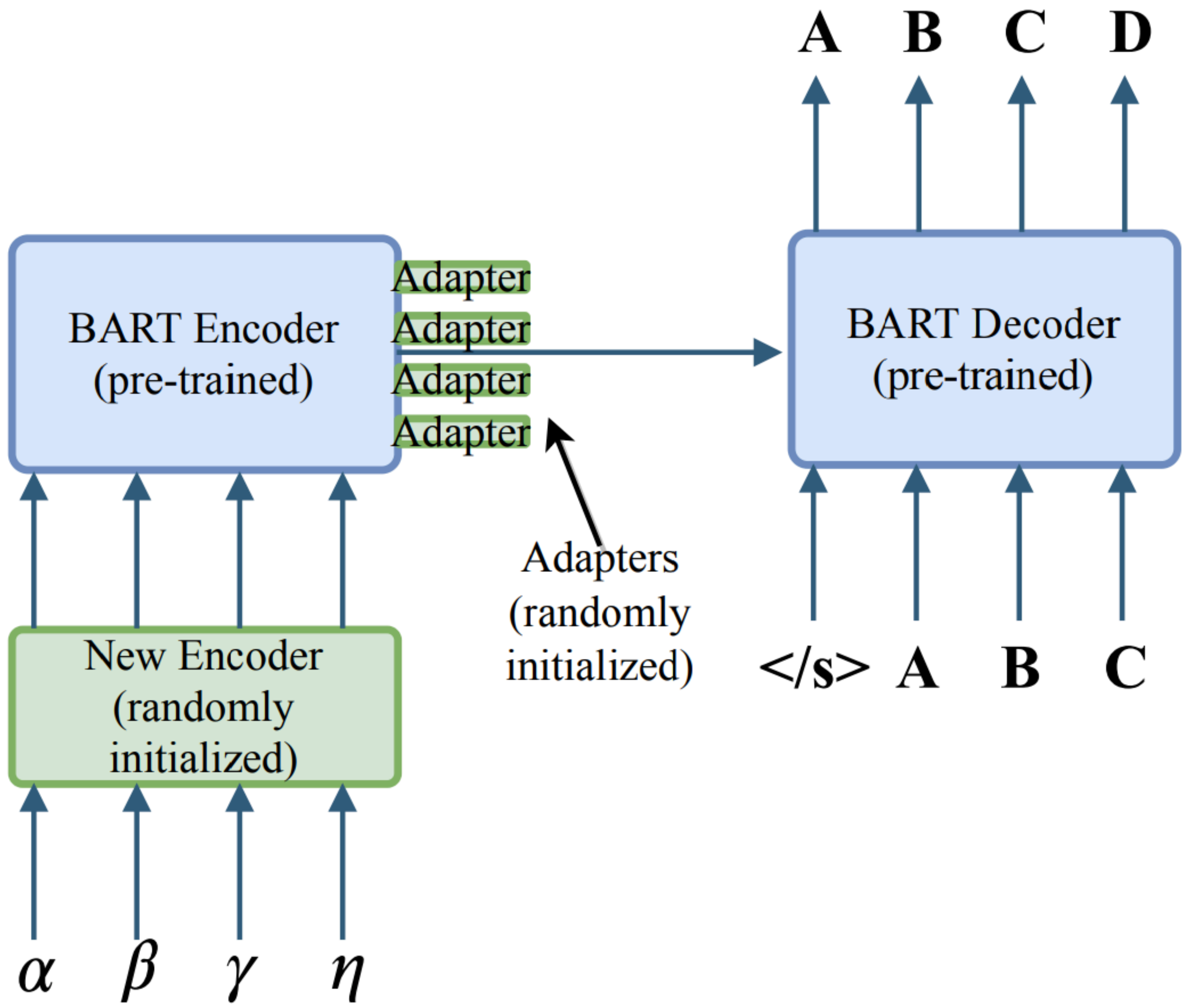
Efficient fine-tuning for NMT: Language-pair adapters

- A new set of adapters can be trained for **each** language pair
- This works well for high-resource languages *(Bapna and Firat, 2019)*
- But does not work for low-resource languages, because there is no sharing between related languages



Efficient fine-tuning for NMT: Language-agnostic adapters

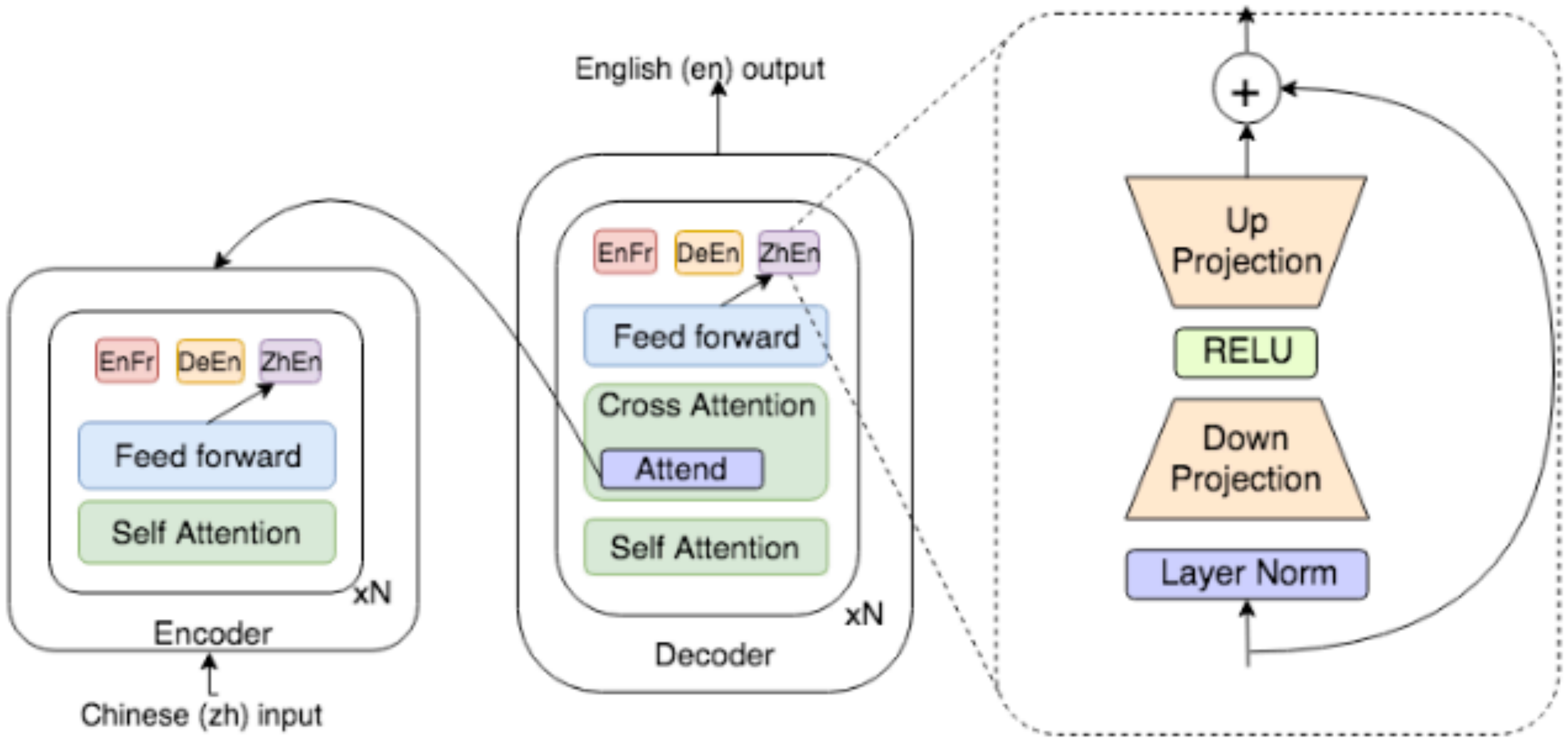
- A new set of adapters can be trained for **all** language pairs *(Stickland et al., 2021)*
- This suffers from negative interference between unrelated languages



Efficient fine-tuning for NMT

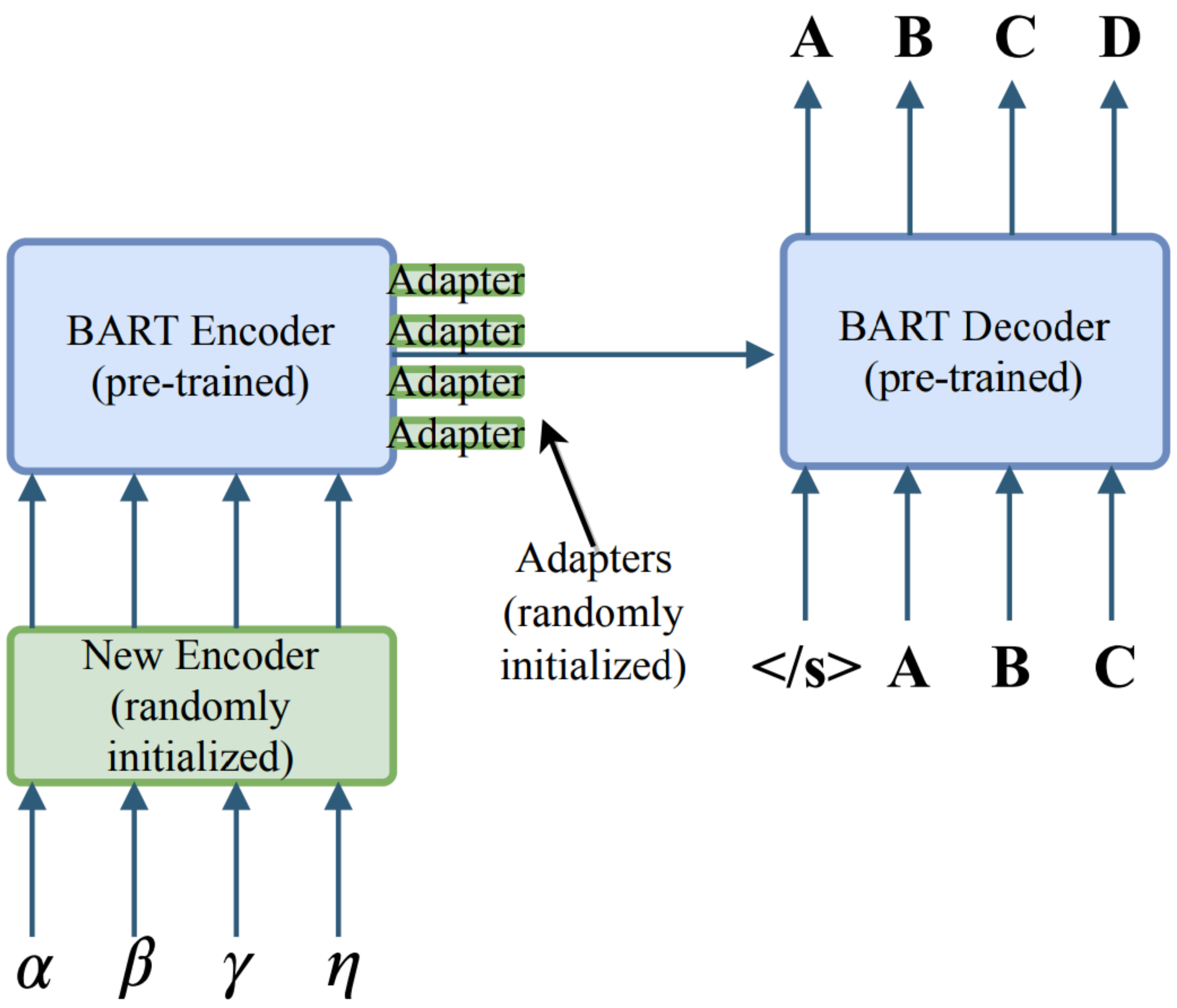
Language-pair adapters

(Bapna and Firat, 2019)



Language-agnostic adapters

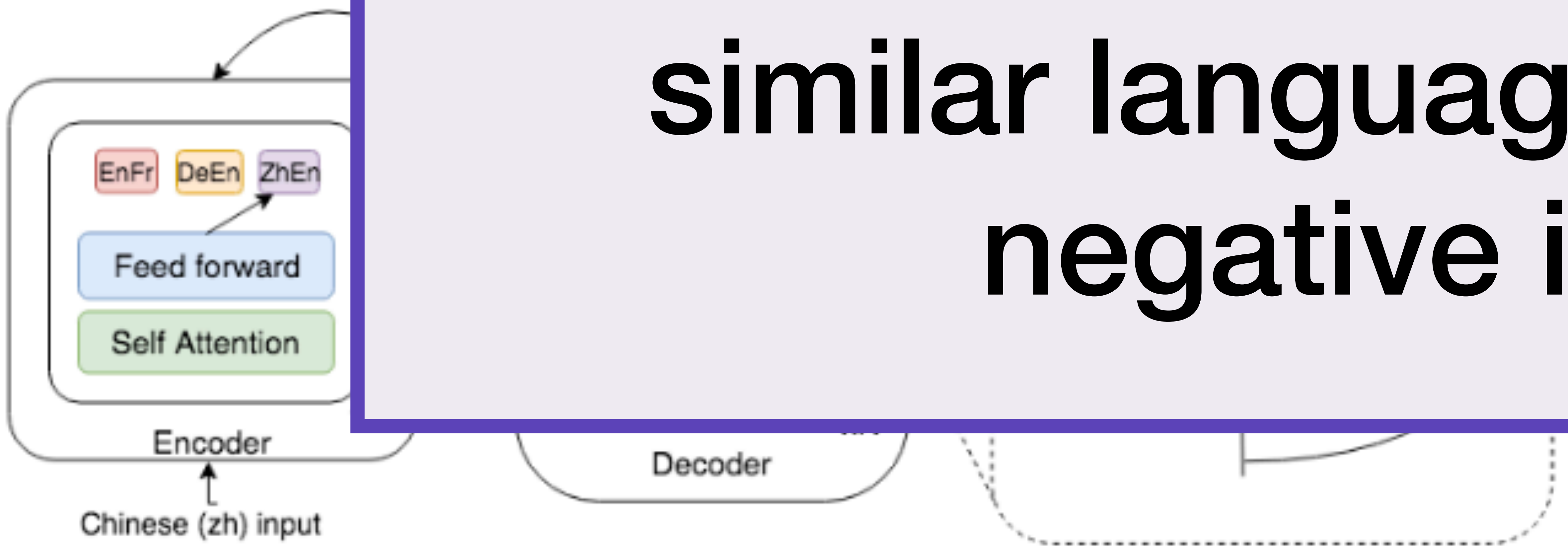
(Stickland et al., 2021)



Efficient fine-tuning for NMT

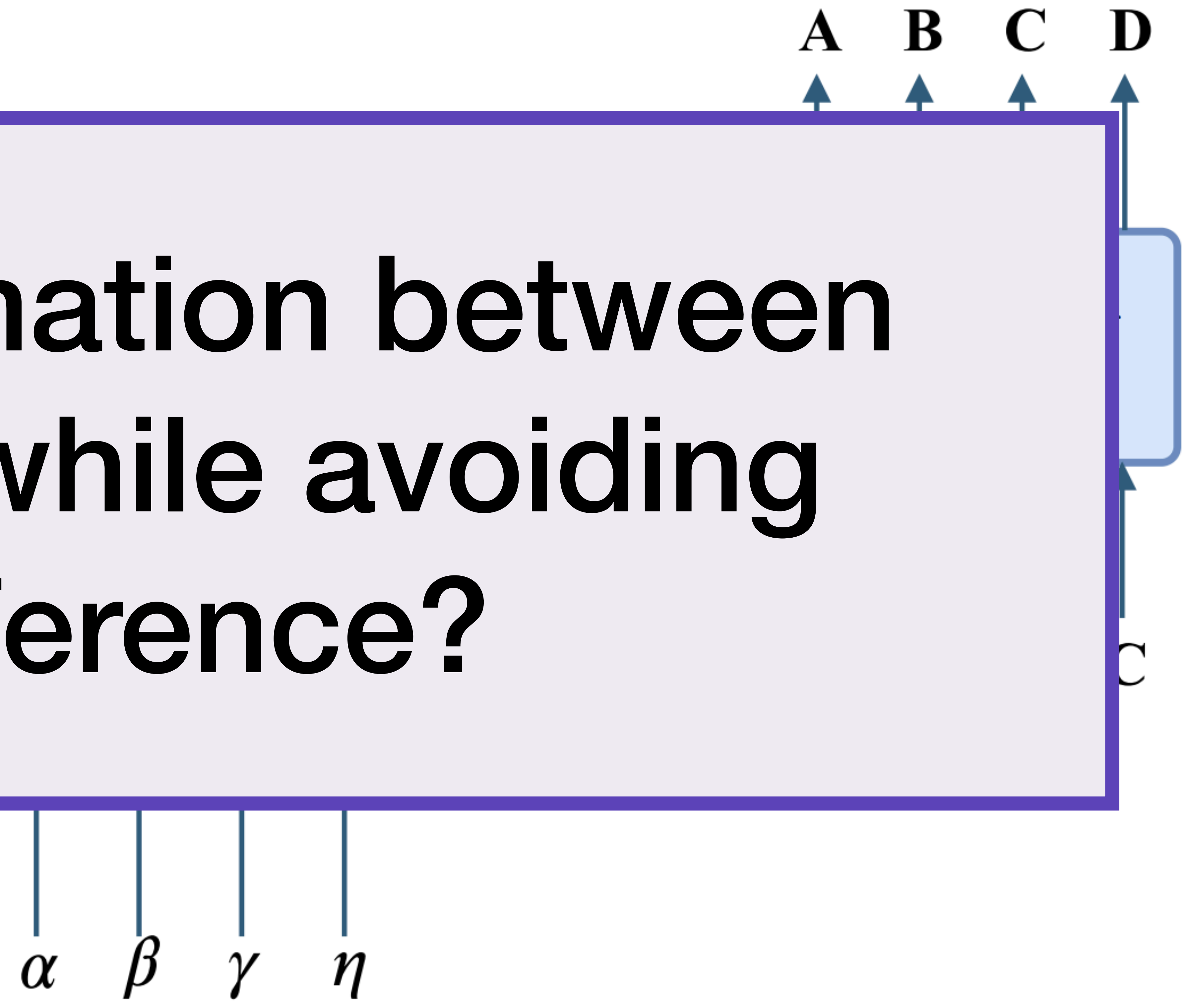
Language-pair adapters

(Bapna and Firat, 2019)



Language-agnostic adapters

(Stickland et al., 2021)



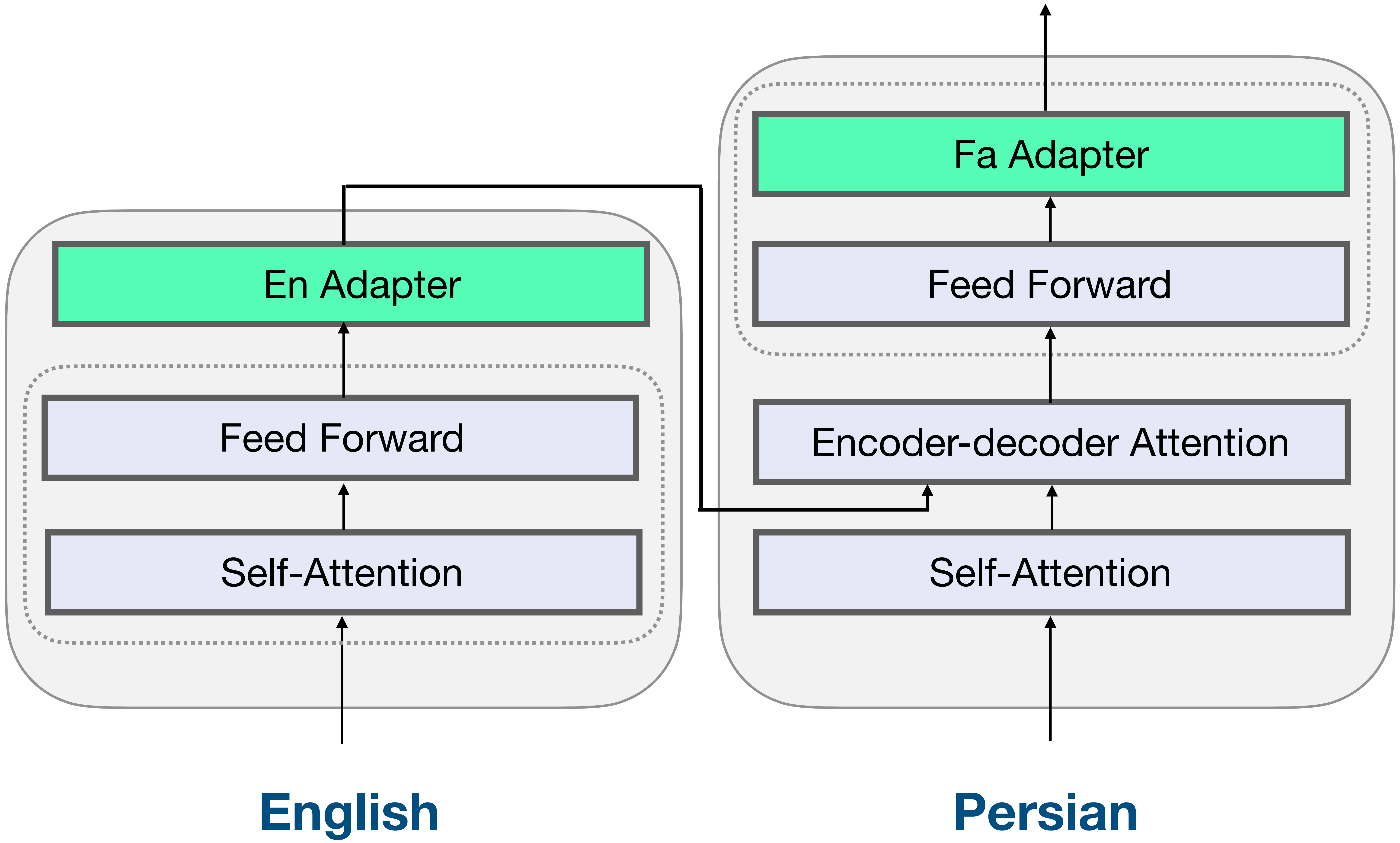
Can we share information between similar languages while avoiding negative inteference?

- Motivation
- **Proposed Approach**
- Experiments
- Conclusion

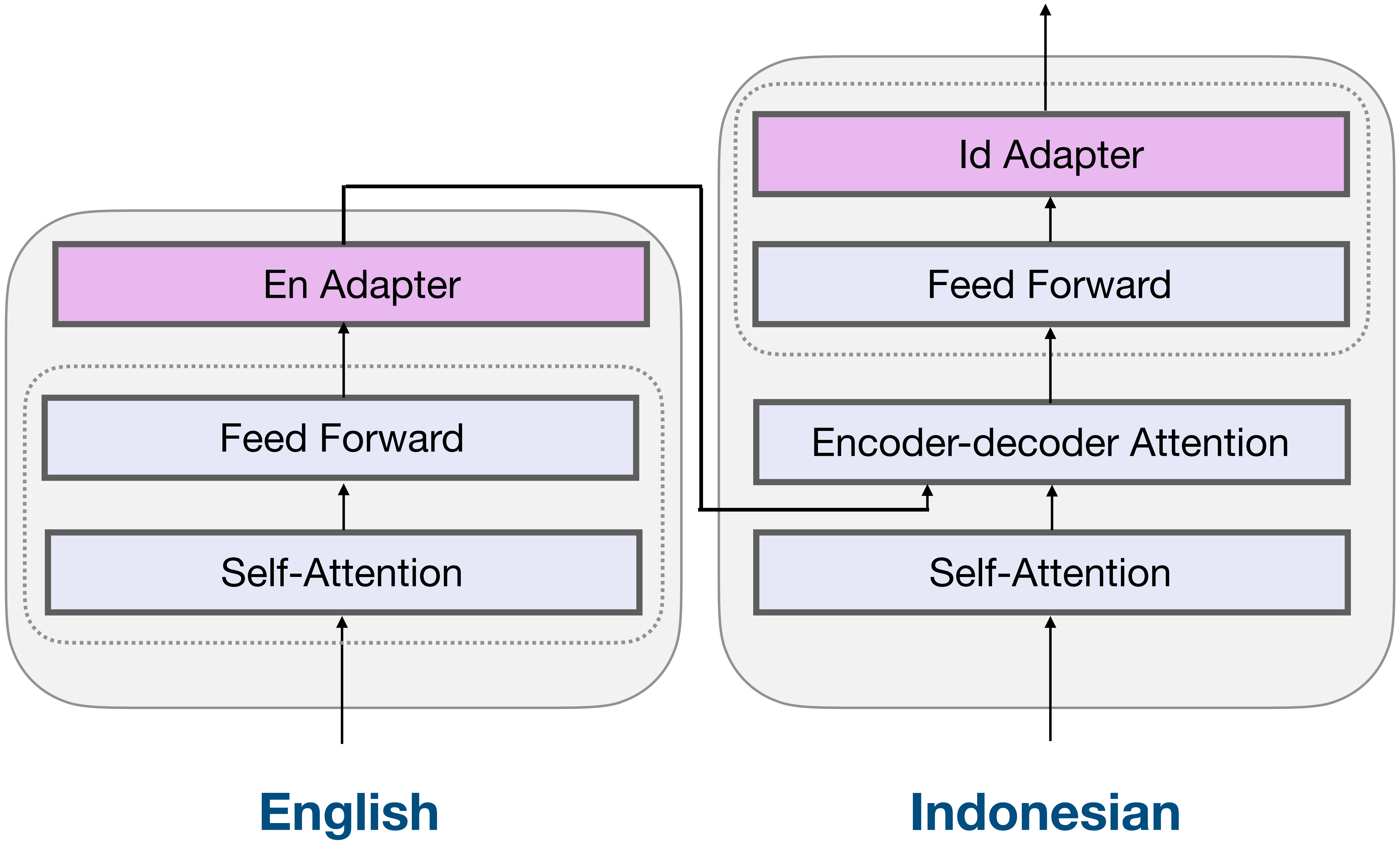
Language-family adapters for low-resource multilingual NMT

Idea: We encode the similarities between related languages with adapters trained on each language family.

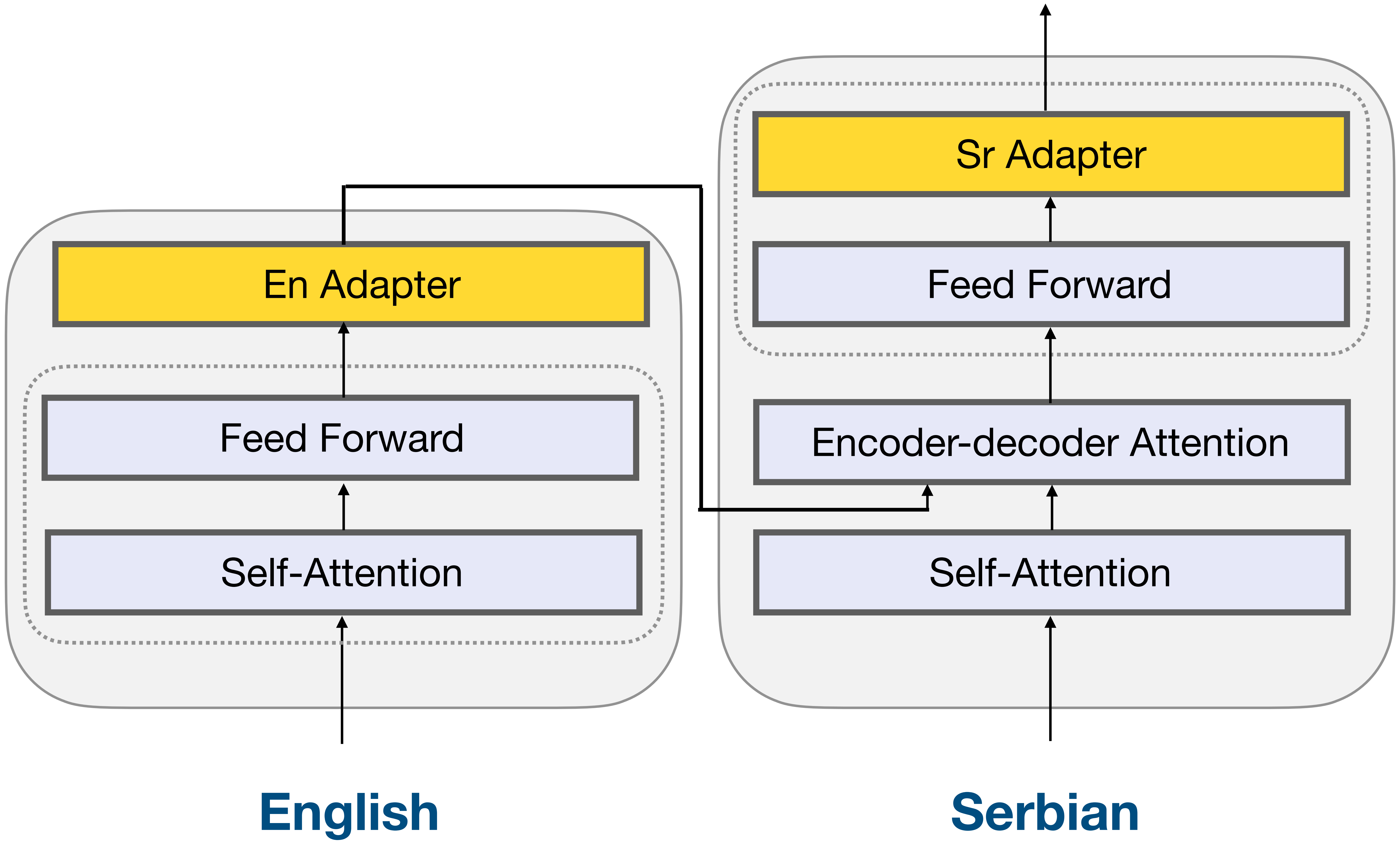
Adding adapters to mBART-50



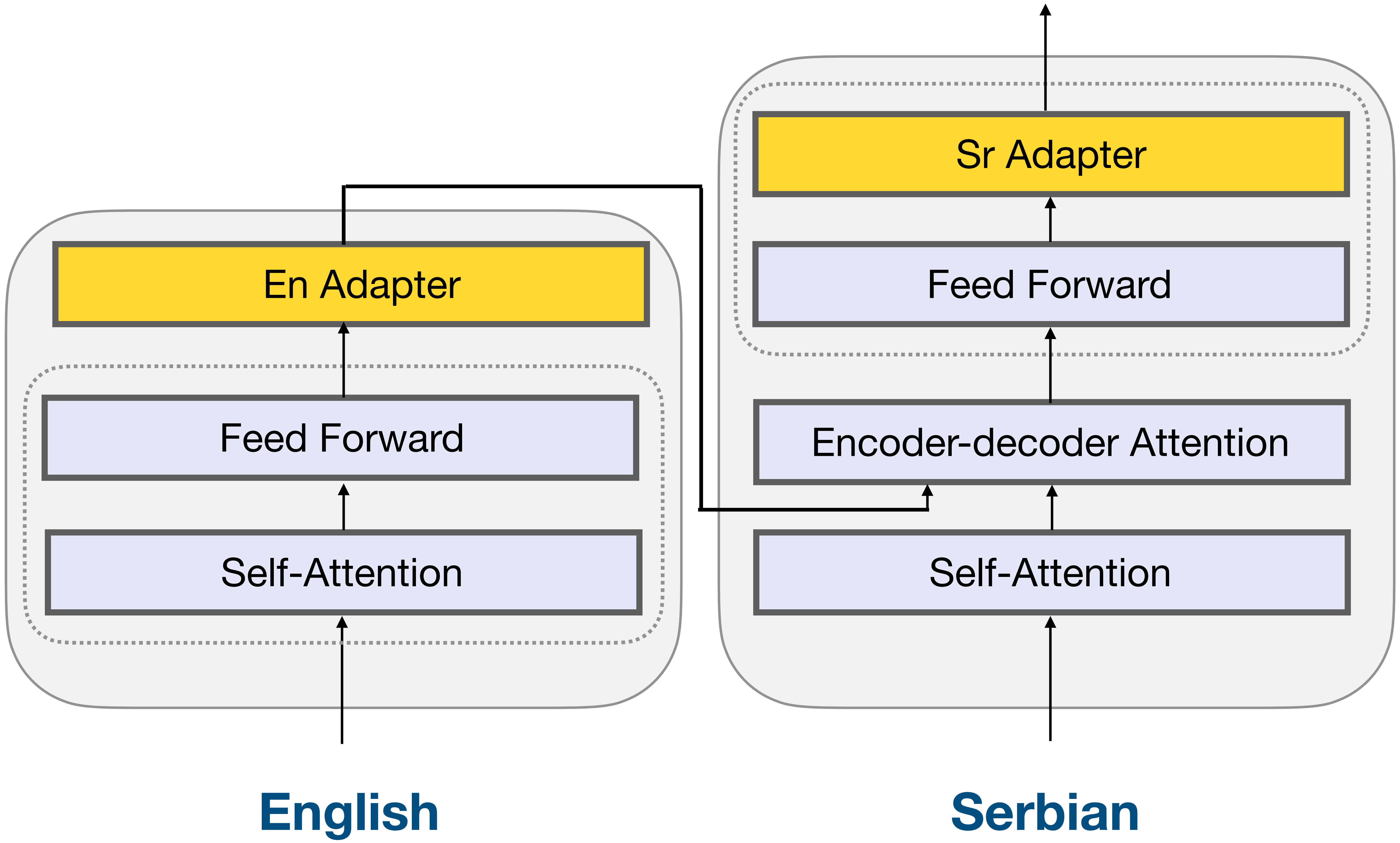
Adding adapters to mBART-50



Adding adapters to mBART-50

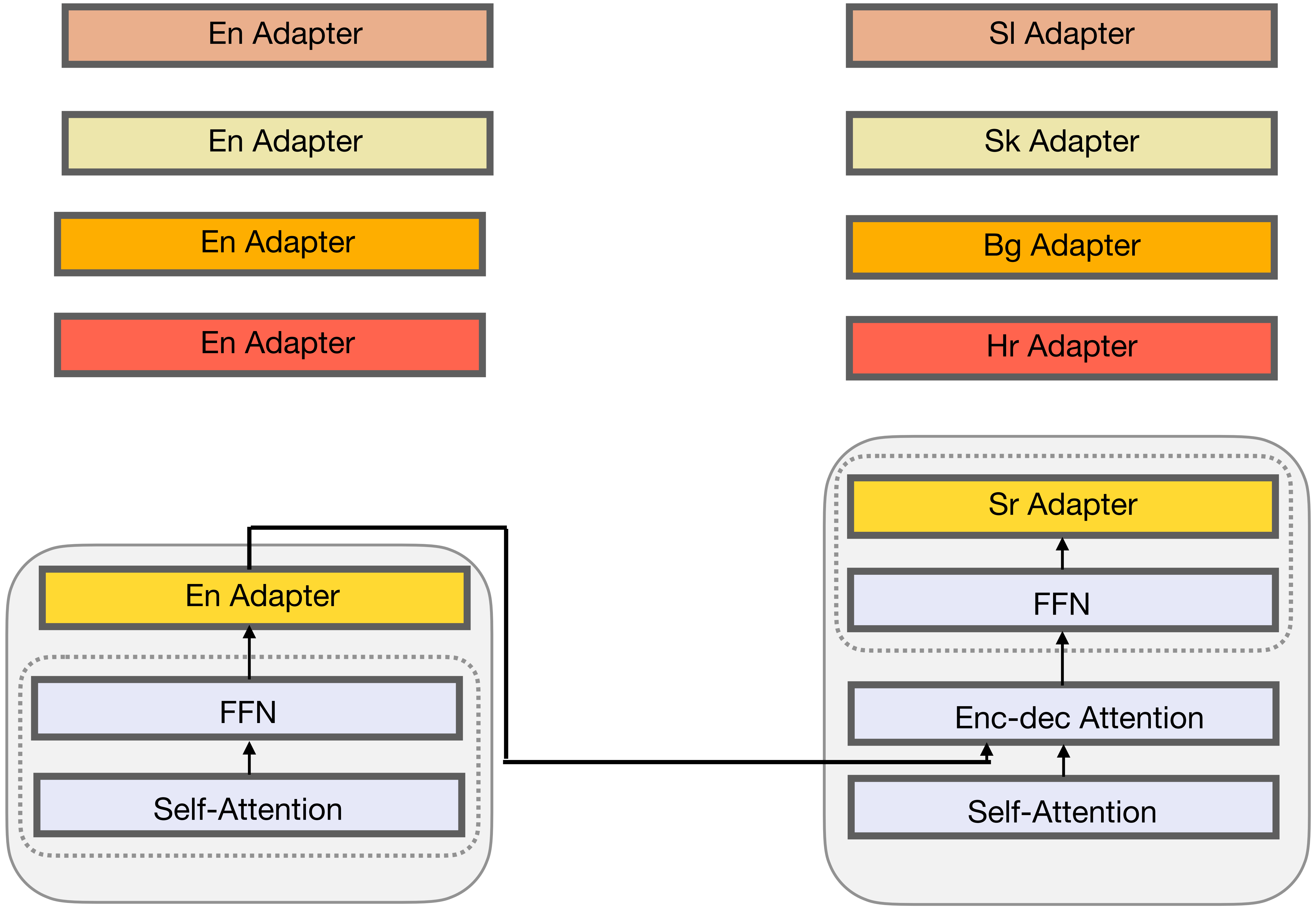


Adding adapters to mBART-50



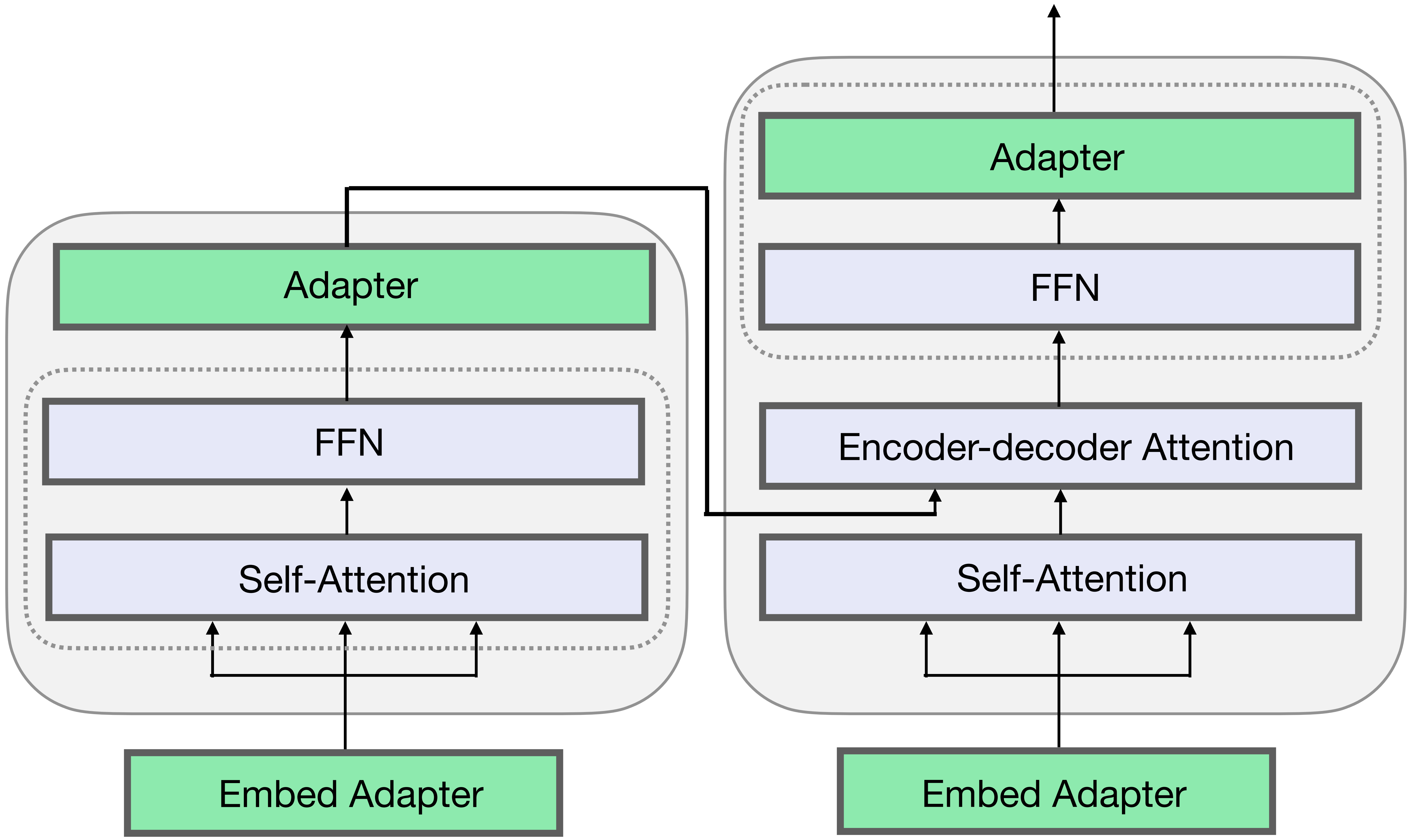
Independently-trained adapters for various language pairs

Adding adapters to mBART-50

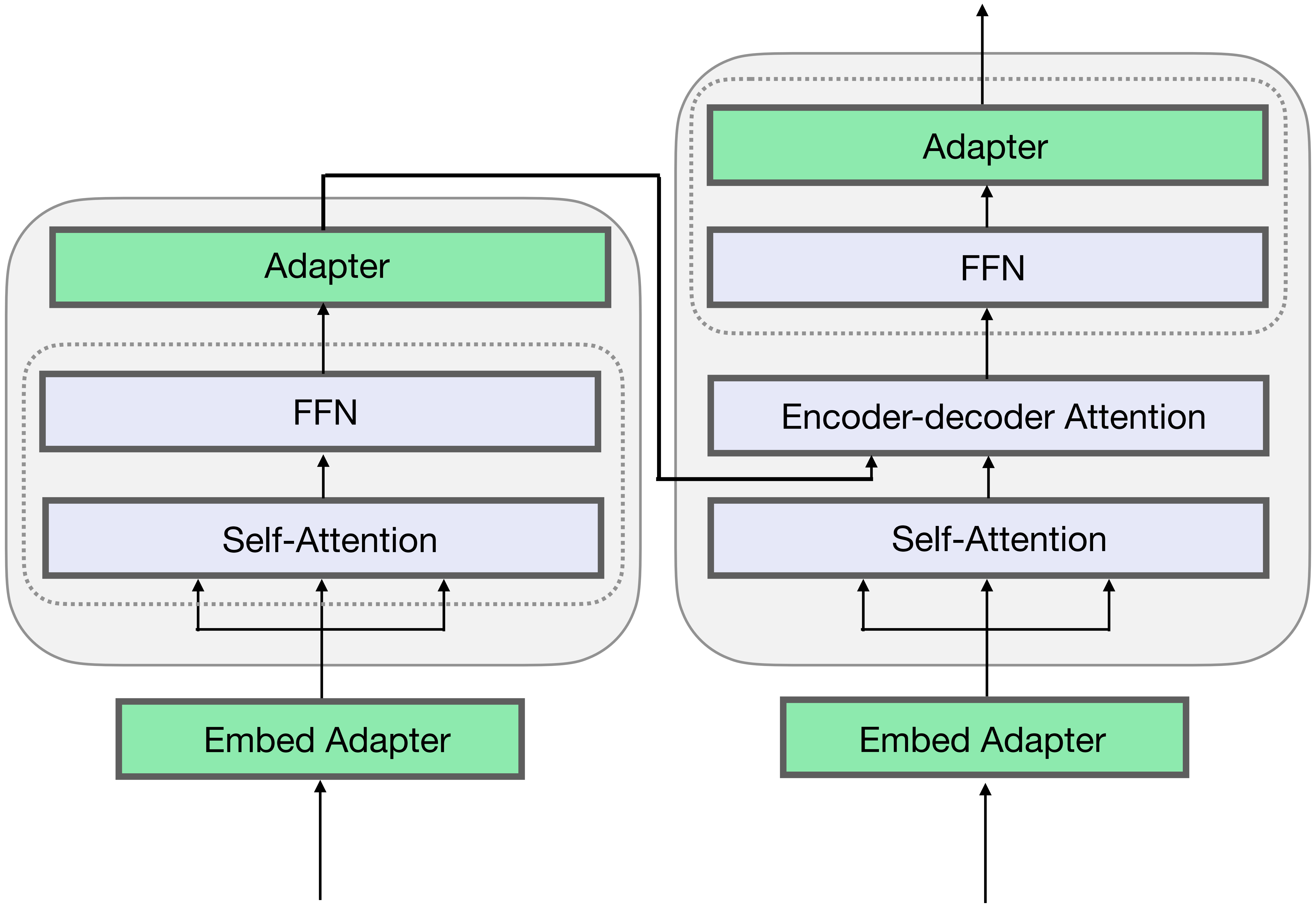


When in same family
-> cluster together?

Our model: Language-family adapters

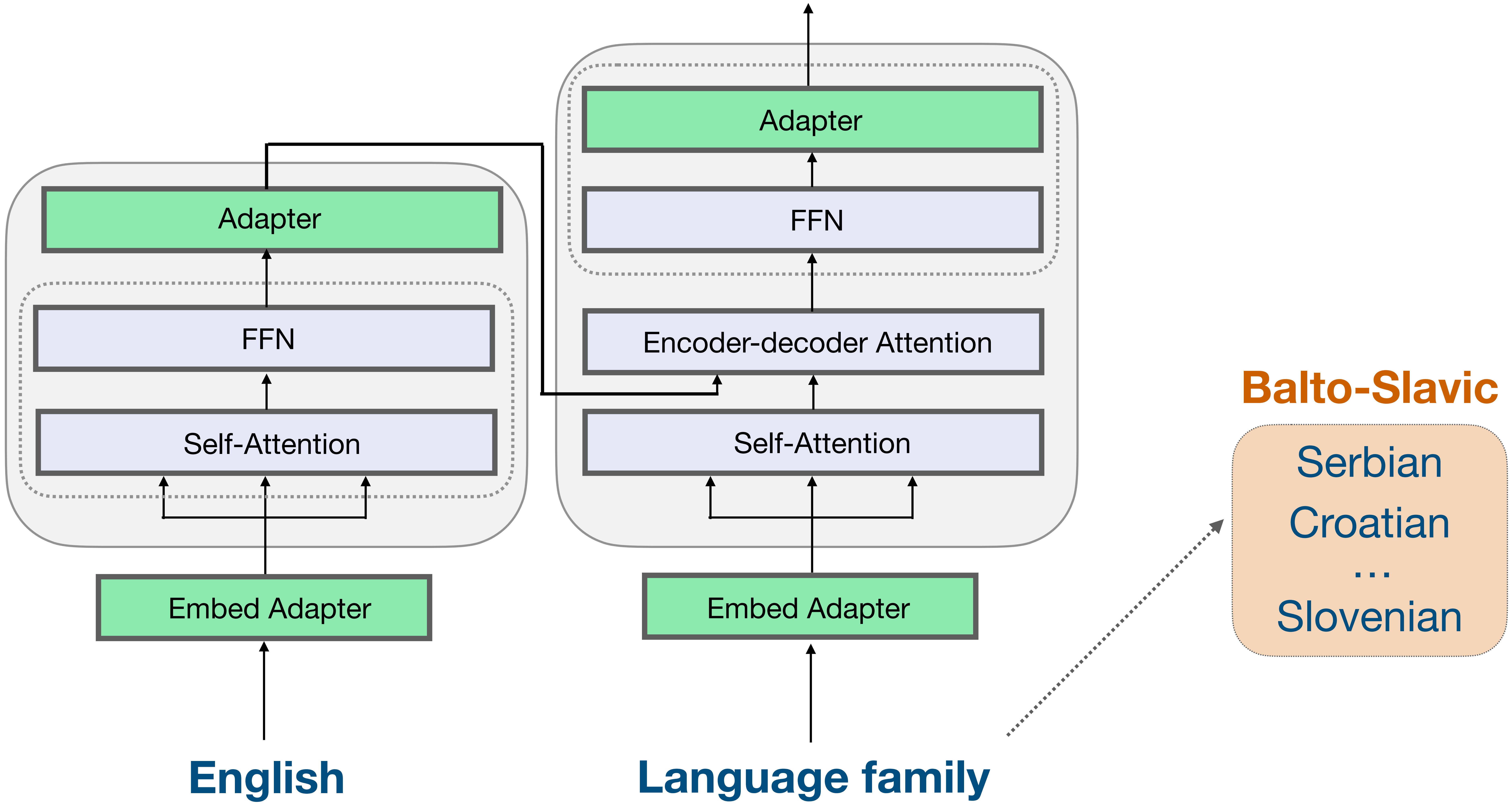


Training for MT

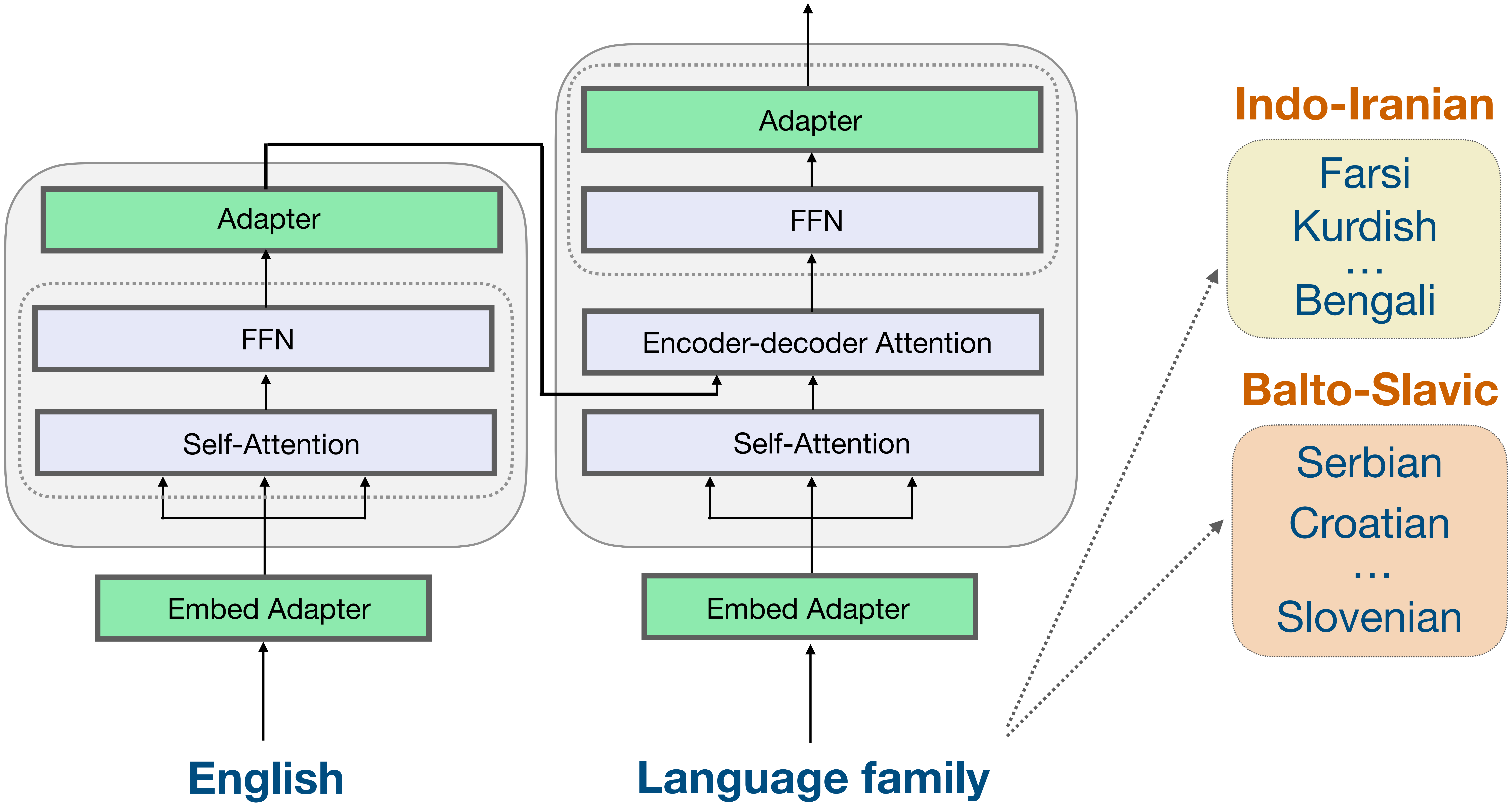


English **Language family**

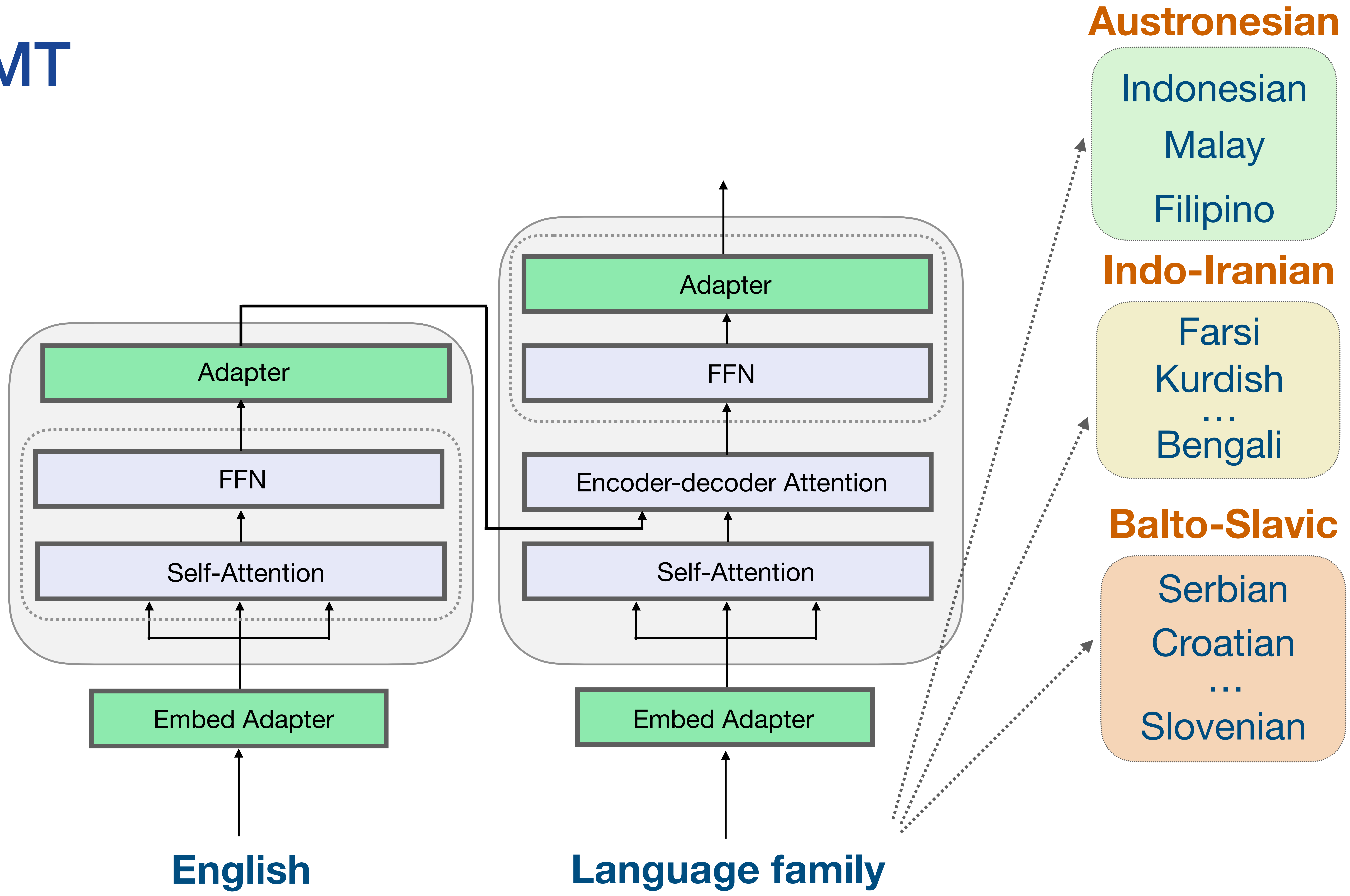
Training for MT



Training for MT



Training for MT



- Motivation
- Proposed Approach
- **Experiments**
- Conclusion

Experimental Setup

- PLM: mBART-50 (trained on monolingual data, ~680M params)
- Adapters: bottleneck size 512
- Translation: En->XX
- Baselines:
    - Language-pair adapters
    - Language-agnostic adapters

Datasets

- TED talks *(Qi et al., 2018)* and OPUS-100 *(Zhang et al., 2020)* for 17 low-resource languages (and English)
- Language families: Indo-Iranian (I), Balto-Slavic (BS), Austronesian (A)
- Starred languages do not appear in mBART-50 pretraining corpus

Language (code)	**Family**	**Train Set**	
		TED	**OPUS-100**
⋆Bulgarian (bg)	BS	174k	1M
Persian (fa)	I	151k	1M
⋆Serbian (sr)	BS	137k	1M
Croatian (hr)	BS	122k	1M
Ukrainian (uk)	BS	108k	1M
Indonesian (id)	A	87k	1M
⋆Slovak (sk)	BS	61k	1M
Macedonian (mk)	BS	25k	1M
Slovenian (sl)	BS	20k	1M
Hindi (hi)	I	19k	534k
Marathi (mr)	I	10k	27k
⋆Kurdish (ku)	I	10k	45k
⋆Bosnian (bs)	BS	6k	1M
⋆Malay (ms)	A	5k	1M
Bengali (bn)	I	5k	1M
⋆Belarusian (be)	BS	5k	67k
⋆Filipino (fil)	A	3k	-

Main results

Model	Balto-Slavic									Austro-Nesian			Indo-Iranian					
	bg*	sr*	hr	uk	sk*	mk	sl	bs*	be*	id	ms*	fil*	fa	hi	mr	ku*	bn	AVG
OPUS-100																		
Lang-pair	**27.8**	17.5	23.7	**17.7**	25.0	**35.0**	**24.1**	**21.0**	10.1	28.0	24.5	-	**10.5**	15.6	17.0	14.1	13.0	20.3
Lang-agnostic	21.6	19.7	21.4	13.8	24.1	28.9	19.6	19.5	11.3	28.6	21.8	-	8.1	16.9	17.8	12.8	11.2	18.6
Lang-family	25.4	**20.9**	23.7	15.1	**27.7**	31.9	22.6	20.3	**15.2**	**31.3**	**25.4**	-	9.8	**18.7**	**25.0**	**15.3**	12.9	**21.3**
TED																		
Lang-pair	**35.7**	21.1	30.5	21.1	**24.2**	27.0	21.4	**28.6**	**12.5**	**35.4**	**23.4**	**12.2**	14.0	14.1	10.0	**4.9**	9.0	20.3
Lang-agnostic	31.7	24.0	29.7	21.9	20.6	26.5	20.2	27.8	7.7	33.8	22.1	11.6	17.0	15.5	7.0	3.3	6.0	19.2
Lang-family	33.8	**25.1**	30.5	22.2	22.8	**28.0**	21.5	27.8	9.5	34.7	22.0	11.5	**17.5**	**19.8**	10.3	4.1	**11.6**	**20.7**

Test set BLEU (↑) scores when translating out of English (*en* -> *xx*).

Main results

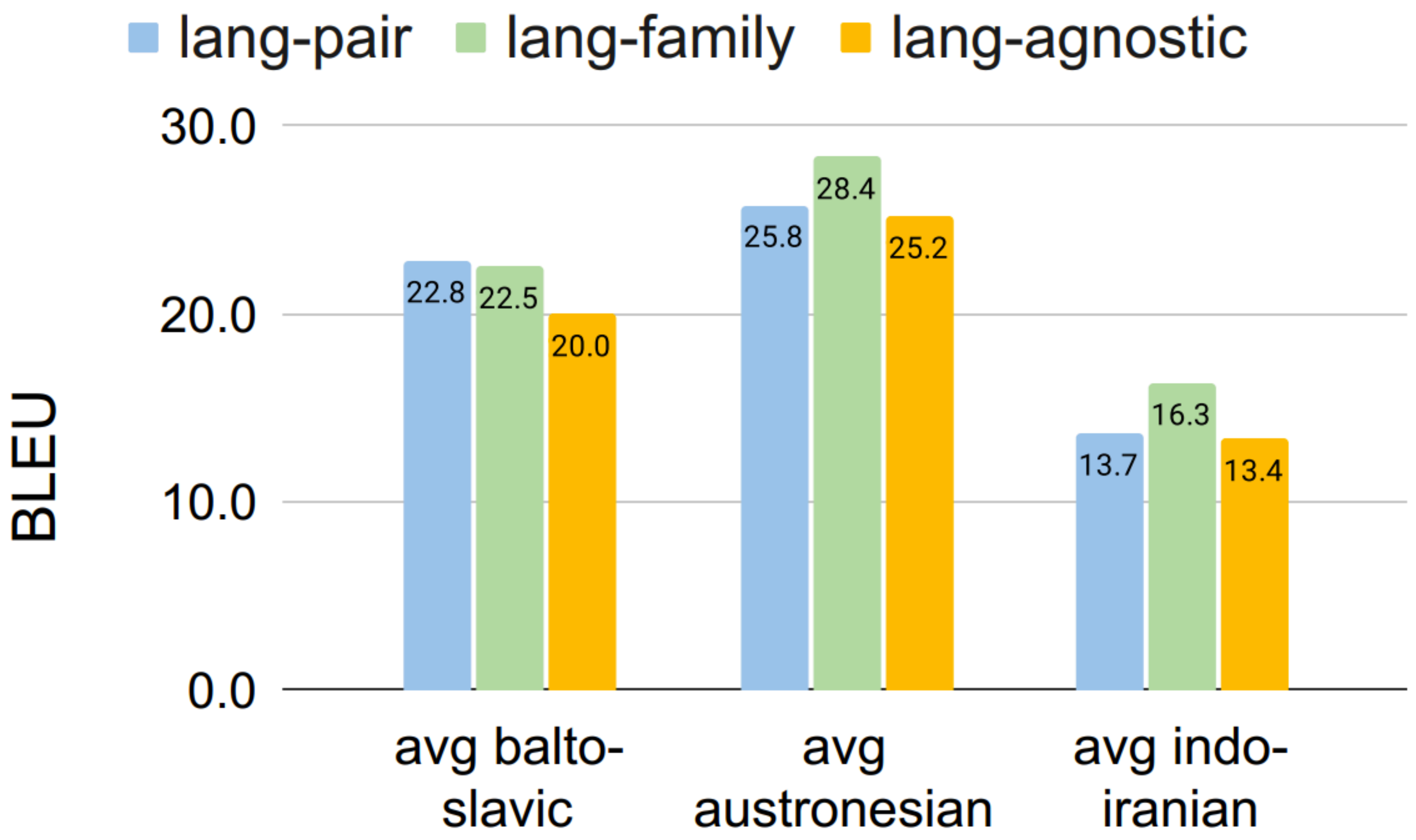
Model	Balto-Slavic									Austro-Nesian			Indo-Iranian					
	bg*	sr*	hr	uk	sk*	mk	sl	bs*	be*	id	ms*	fil*	fa	hi	mr	ku*	bn	AVG
OPUS-100																		
Lang-pair	**27.8**	17.5	23.7	**17.7**	25.0	**35.0**	**24.1**	**21.0**	10.1	28.0	24.5	-	**10.5**	15.6	17.0	14.1	13.0	20.3
Lang-agnostic	21.6	19.7	21.4	13.8	24.1	28.9	19.6	19.5	11.3	28.6	21.8	-	8.1	16.9	17.8	12.8	11.2	18.6
Lang-family	25.4	**20.9**	23.7	15.1	**27.7**	31.9	22.6	20.3	**15.2**	**31.3**	**25.4**	-	9.8	**18.7**	**25.0**	**15.3**	12.9	**21.3**
TED																		
Lang-pair	**35.7**	21.1	30.5	21.1	**24.2**	27.0	21.4	**28.6**	**12.5**	**35.4**	**23.4**	**12.2**	14.0	14.1	10.0	**4.9**	9.0	20.3
Lang-agnostic	31.7	24.0	29.7	21.9	20.6	26.5	20.2	27.8	7.7	33.8	22.1	11.6	17.0	15.5	7.0	3.3	6.0	19.2
Lang-family	33.8	**25.1**	30.5	22.2	22.8	**28.0**	21.5	27.8	9.5	34.7	22.0	11.5	**17.5**	**19.8**	10.3	4.1	**11.6**	**20.7**

Test set BLEU (↑) scores when translating out of English (*en* -> *xx*).

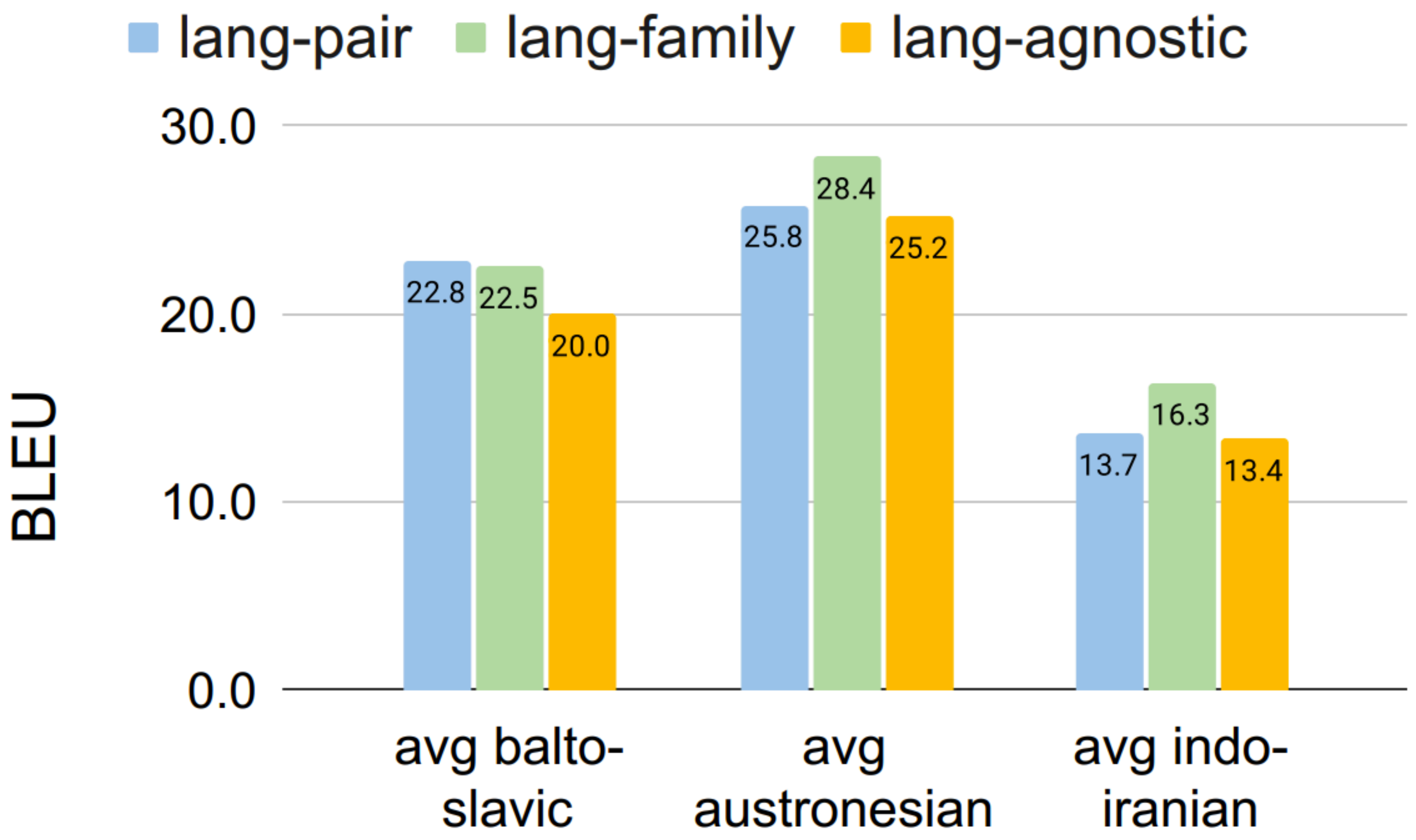
Language-family adapters consistently outperform the baselines on both parallel datasets

How does performance vary per language family?

How does performance vary per language family?



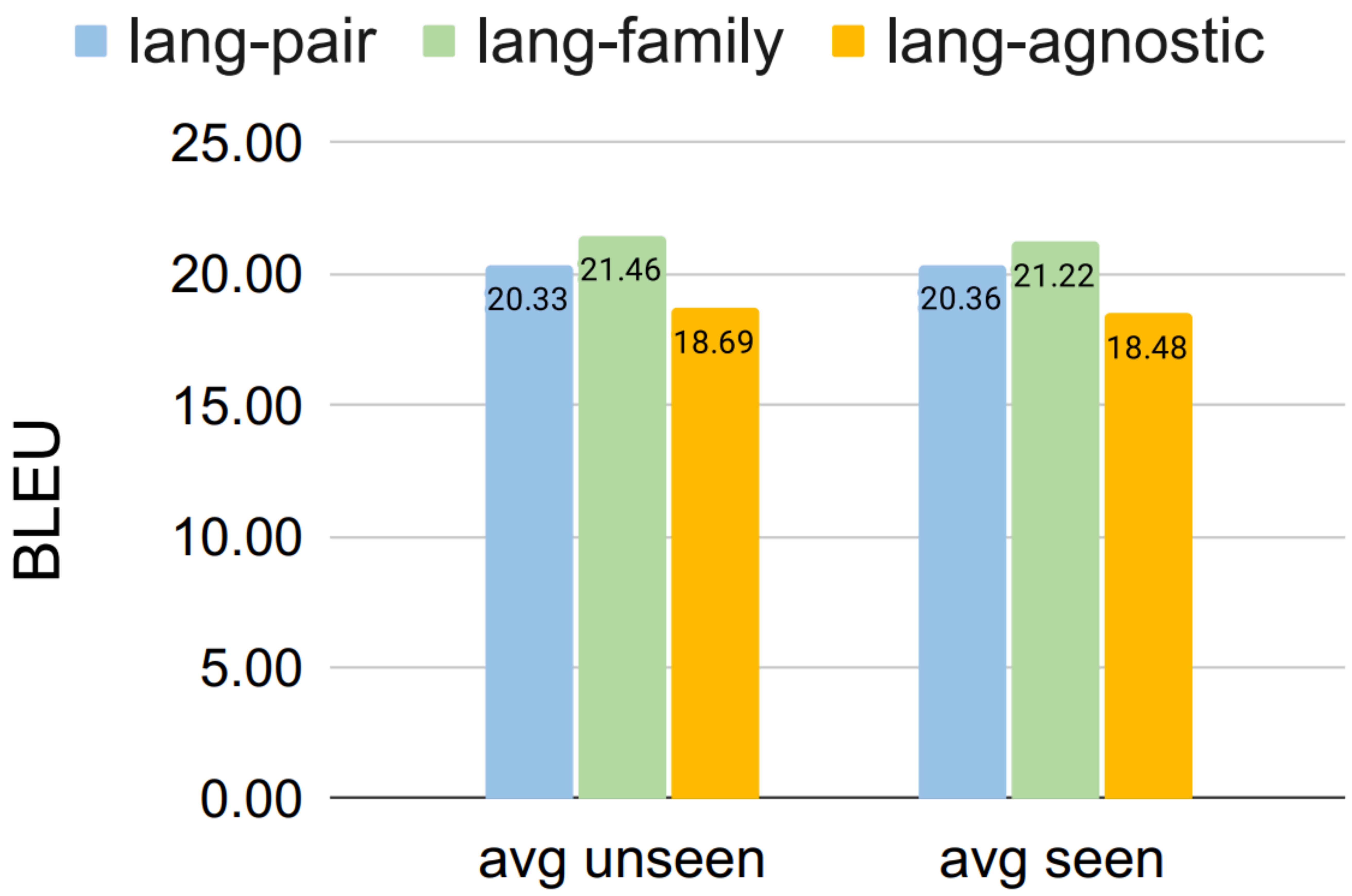
How does performance vary per language family?



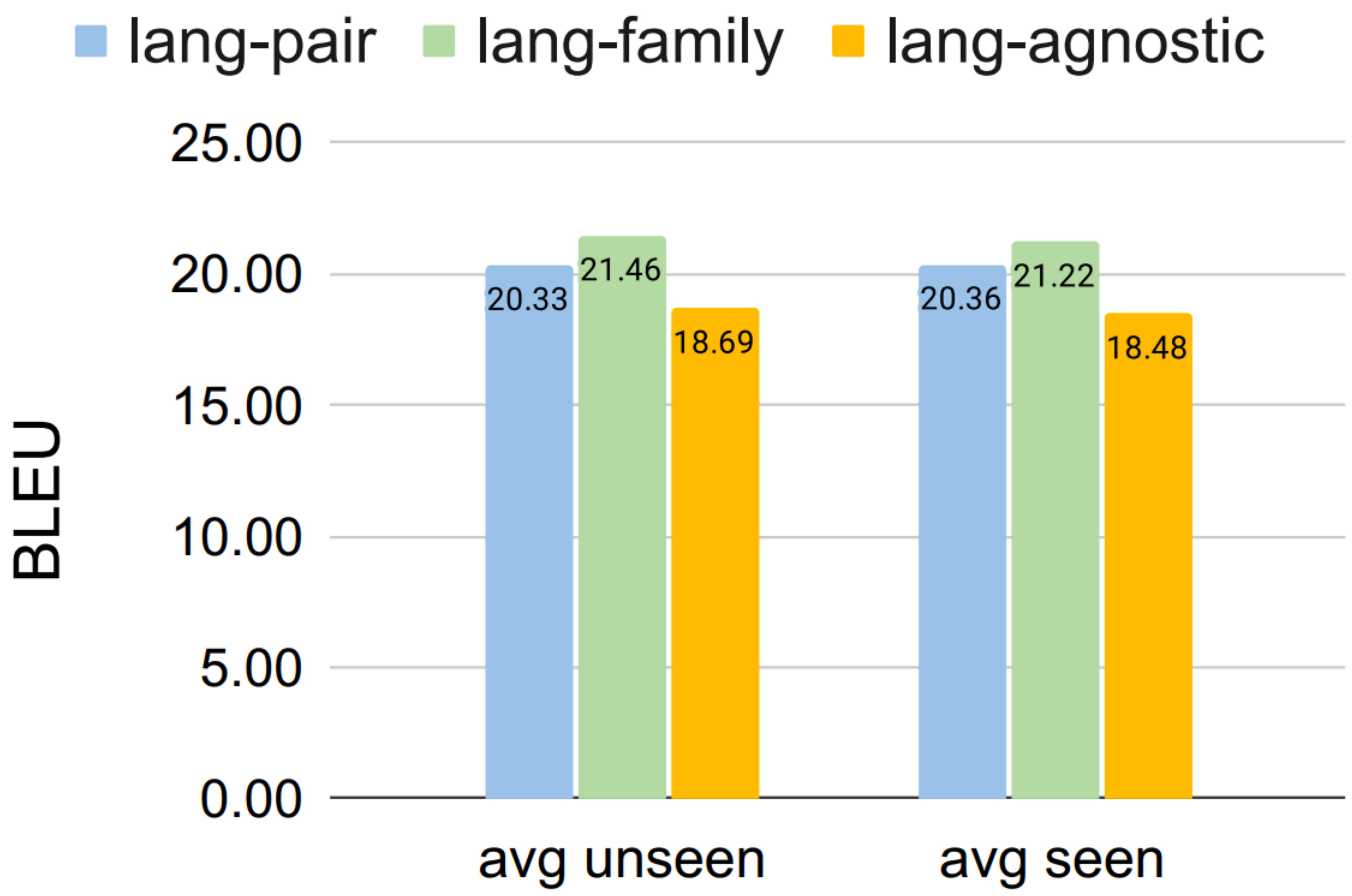
- Compared to **lang-agnostic**, our approach performs better, possibly because of avoiding negative interference

- Compared to **lang-pair**, in BS results equivalent, as many of these languages similar to languages in pretraining corpus

Does the approach perform better in seen or unseen languages?



Does the approach perform better in seen or unseen languages?



- Larger performance improvement for unseen languages
- Caveat: all languages are covered by mBART's vocabulary

Does the embedding layer help?

	Balto-Slavic				Austro-nesian		Indo-Iranian			
	bg	hr	mk	be	id	ms	fa	ku	bn	AVG-16
Lang-Agnostic w/o emb adapter	21.3	21.5	28.3	10.5	28.7	21.5	7.6	12.4	10.9	18.1
Lang-Agnostic with emb adapter (baseline)	21.6	21.4	28.9	11.3	28.6	21.8	8.1	12.8	11.2	**18.6**
Lang-Family w/o emb adapter	24.3	22.6	31.2	13.4	31.4	25.2	9.0	13.7	12.2	20.6
Lang-Family with emb adapter (ours)	25.4	23.7	31.9	15.2	31.3	25.4	9.8	15.3	12.9	**21.3**

Test set BLEU scores (*en -> xx*) on OPUS-100.

Does the embedding layer help?

	BALTO-SLAVIC				AUSTRO-NESIAN		INDO-IRANIAN			
	bg	hr	mk	be	id	ms	fa	ku	bn	AVG-16
LANG-AGNOSTIC w/o emb adapter	21.3	21.5	28.3	10.5	28.7	21.5	7.6	12.4	10.9	18.1
LANG-AGNOSTIC with emb adapter (BASELINE)	21.6	21.4	28.9	11.3	28.6	21.8	8.1	12.8	11.2	**18.6**
LANG-FAMILY w/o emb adapter	24.3	22.6	31.2	13.4	31.4	25.2	9.0	13.7	12.2	20.6
LANG-FAMILY with emb adapter (OURS)	25.4	23.7	31.9	15.2	31.3	25.4	9.8	15.3	12.9	**21.3**

Test set BLEU scores (*en -> xx*) on OPUS-100.

- On average improve translation scores, only add +0.1% of the parameters of mBART-50
- They encode lexical-level information for the languages of interest

Analysis

Should we group languages based on linguistic knowledge or use an unsupervised, data-driven method?

Automatic clustering of languages

	Language Groups			id	fa	ku	AVG
ling. family (ours)	<be, bg, sr, hr, uk, sk, mk, sl, bs>	<id, ms>	<ku, fa, hi, mr, bn>	31.3	9.8	15.3	21.3
GMM	<bg, sr, hr, uk, sk, mk, sl, bs>	<**ku**, id, ms>	<**be**, fa, hi, mr, bn>	29.7	9.2	14.3	19.4
random	<bg, hr, mk, bs, be, ms, hi, mr, ku>	<sl, id>	<sr, uk, sk, fa, bn>	27.8	7.0	15.0	18.4

Test set BLEU scores (*en -> xx*) on OPUS-100.

Automatic clustering of languages

	Language Groups			id	fa	ku	AVG
ling. family (ours)	<be, bg, sr, hr, uk, sk, mk, sl, bs>	<id, ms>	<ku, fa, hi, mr, bn>	31.3	9.8	15.3	21.3
GMM	<bg, sr, hr, uk, sk, mk, sl, bs>	<**ku**, id, ms>	<**be**, fa, hi, mr, bn>	29.7	9.2	14.3	19.4
random	<bg, hr, mk, bs, be, ms, hi, mr, ku>	<sl, id>	<sr, uk, sk, fa, bn>	27.8	7.0	15.0	18.4

Test set BLEU scores (*en* -> *xx*) on OPUS-100.

- Clusters are mostly corresponding to the language families (except for *be* and *ku*)
- Performance is better using linguistic families

- Motivation
- Proposed Approach
- Experiments
- **Conclusion**

Key Takeaways

- We presented an approach that encodes the relations between languages using language-family adapters

- This is an effective and efficient method for MT from English to low-resource languages

- Clustering languages together with a GMM might be helpful in the absence of linguistic knowledge bases

Limitations

- Exploration of non English-centric models
- Covering languages for which the vocabulary is unseen
- More fine-grained grouping of languages

Thanks!

paper: arxiv.org/pdf/2209.15236.pdf

@alexandraxron

achron@cis.lmu.de