Improving Multilingual Neural Machine Translation with Language-Family Adapters

Alexandra Chronopoulou

DG CNECT workshop on large language models

Language-Family Adapters for Multilingual Neural Machine Translation [Chronopoulou A., Stojanovski D., Fraser A., 2022]

Motivation

Machine Translation permits information flow

> 7000 languages in the world

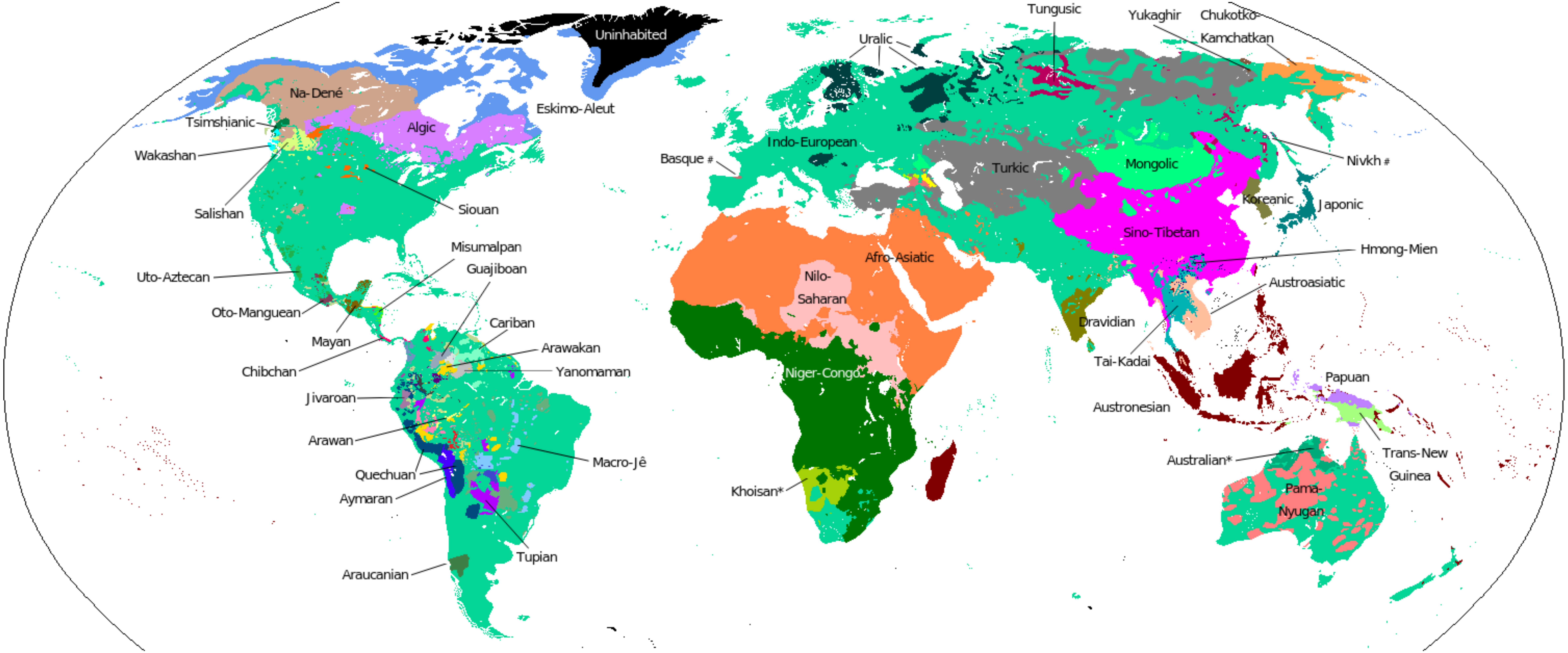
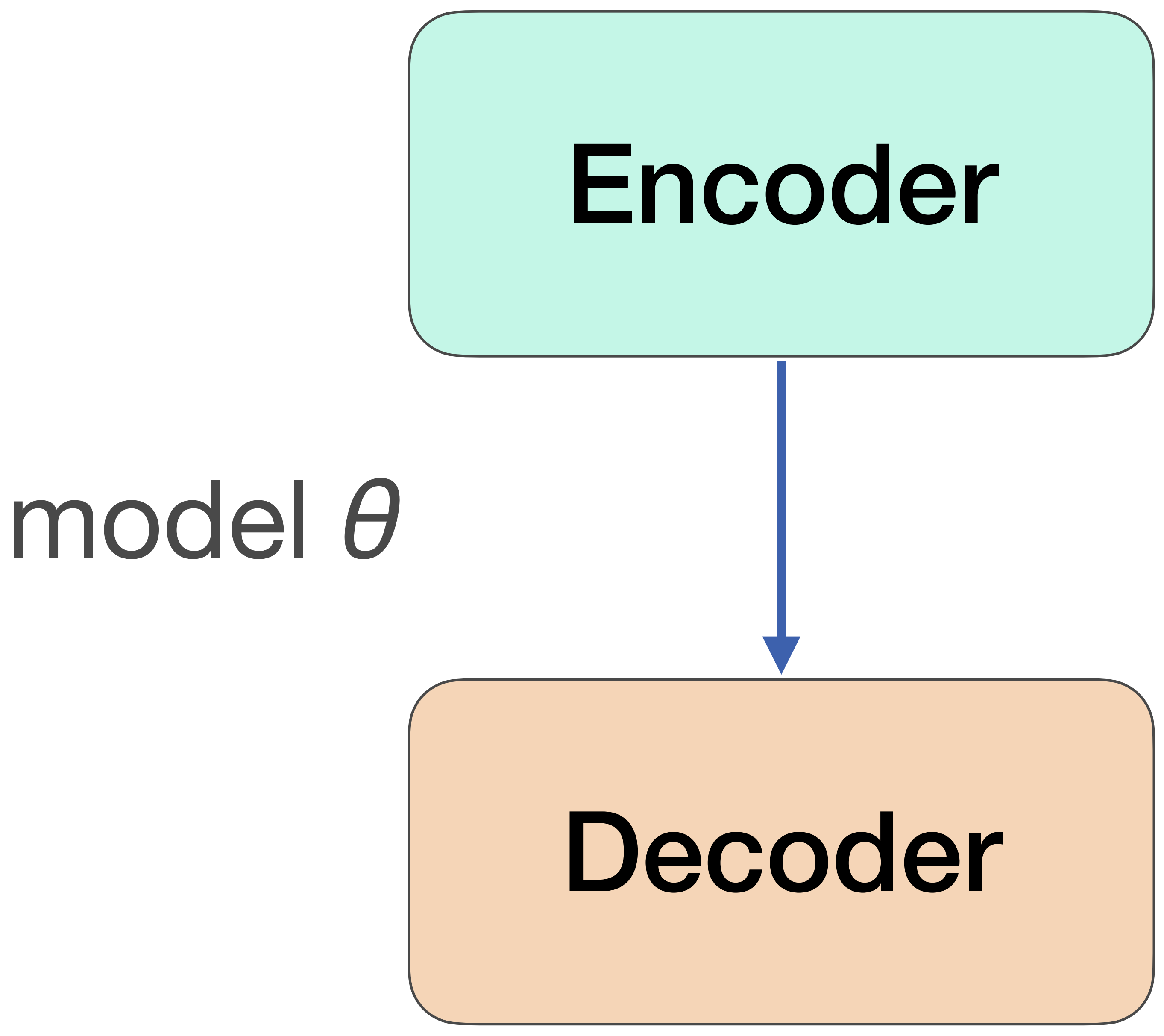


Image credit: Wikipedia

Multilingual Neural Machine Translation



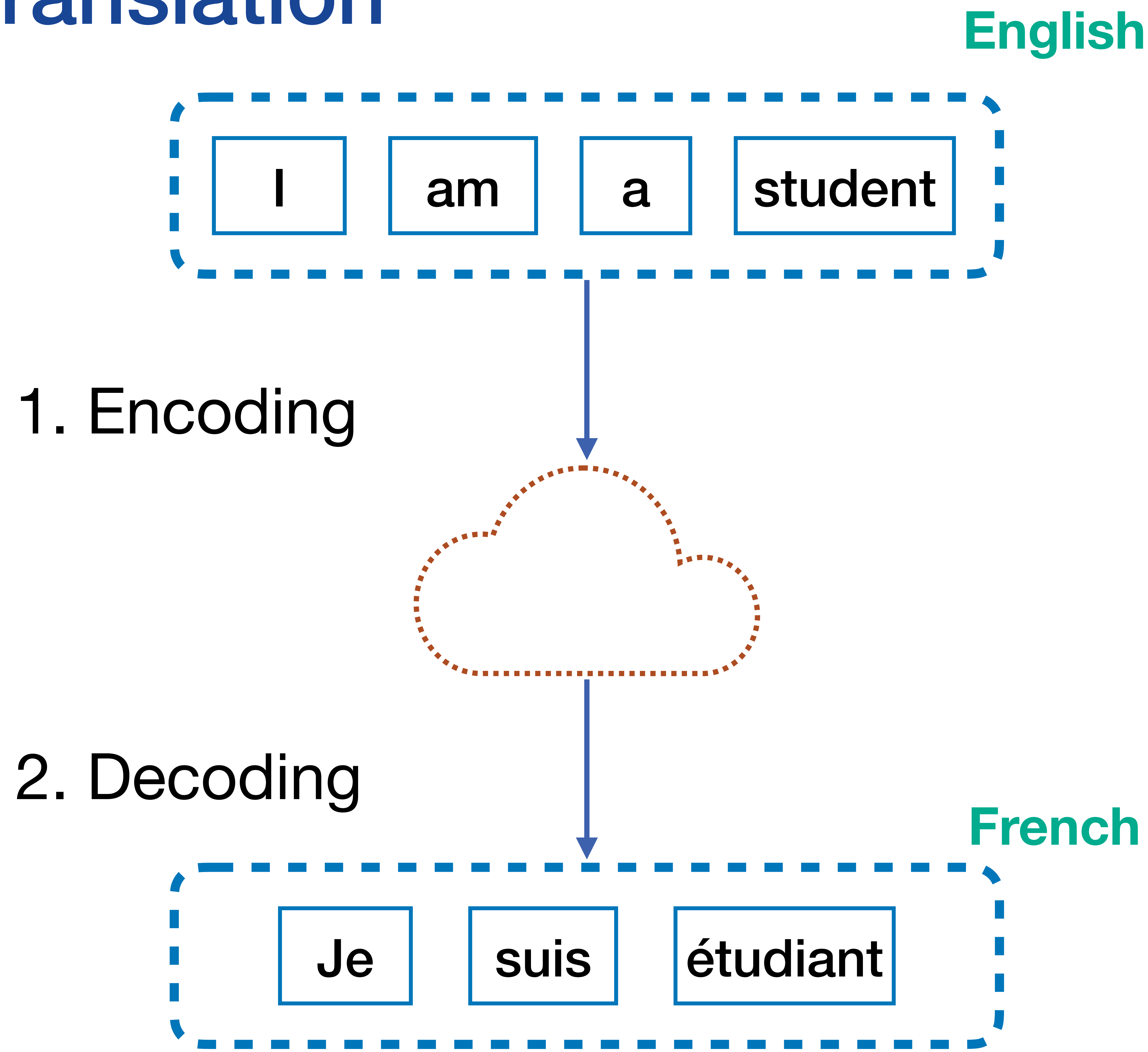
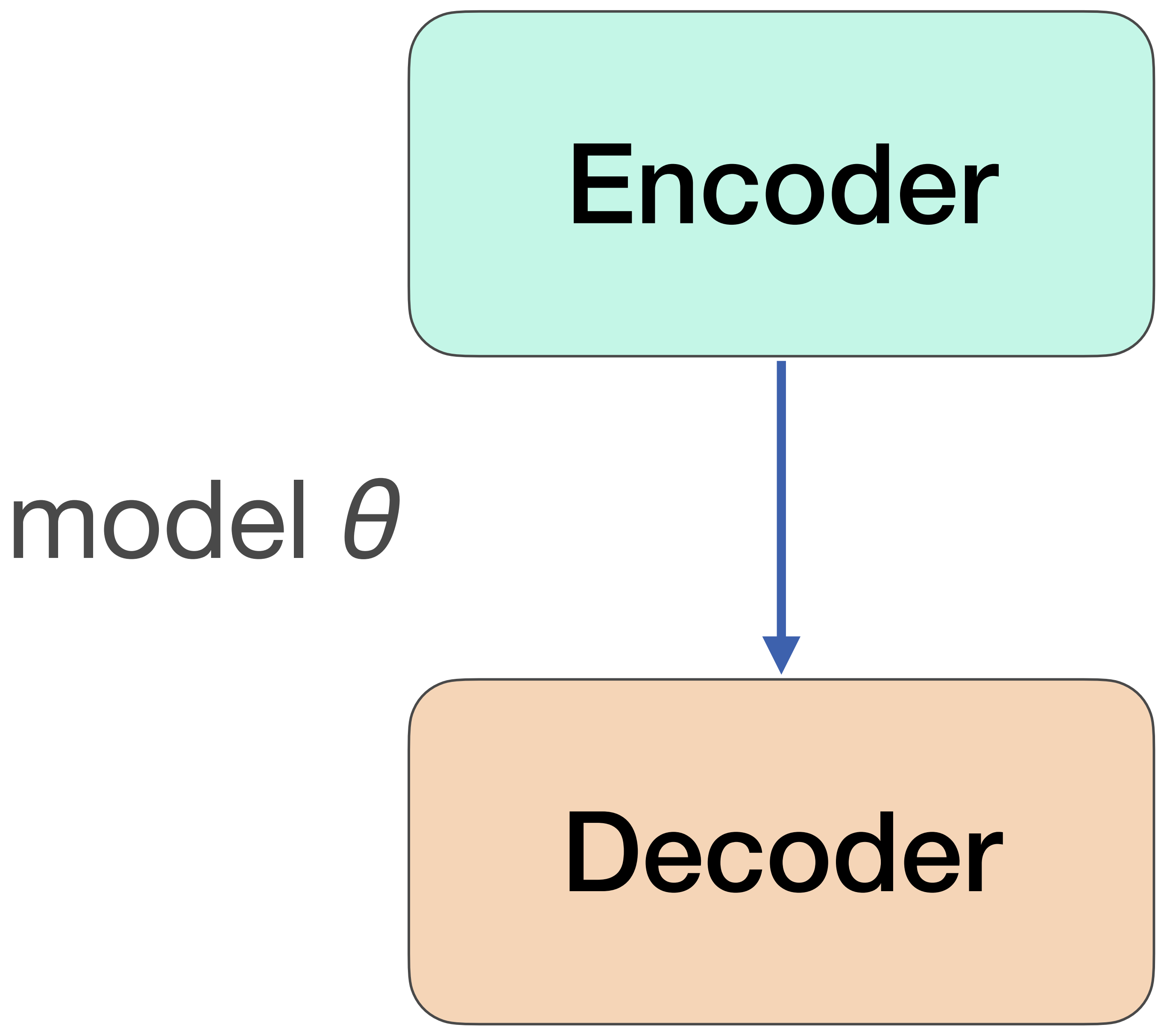


Image credit: https://sites.cs.ucsb.edu/~lilei/TALKS/2021-ACL/pre-training_nmt_ACL_tutorial_2021.pdf

Multilingual Neural Machine Translation



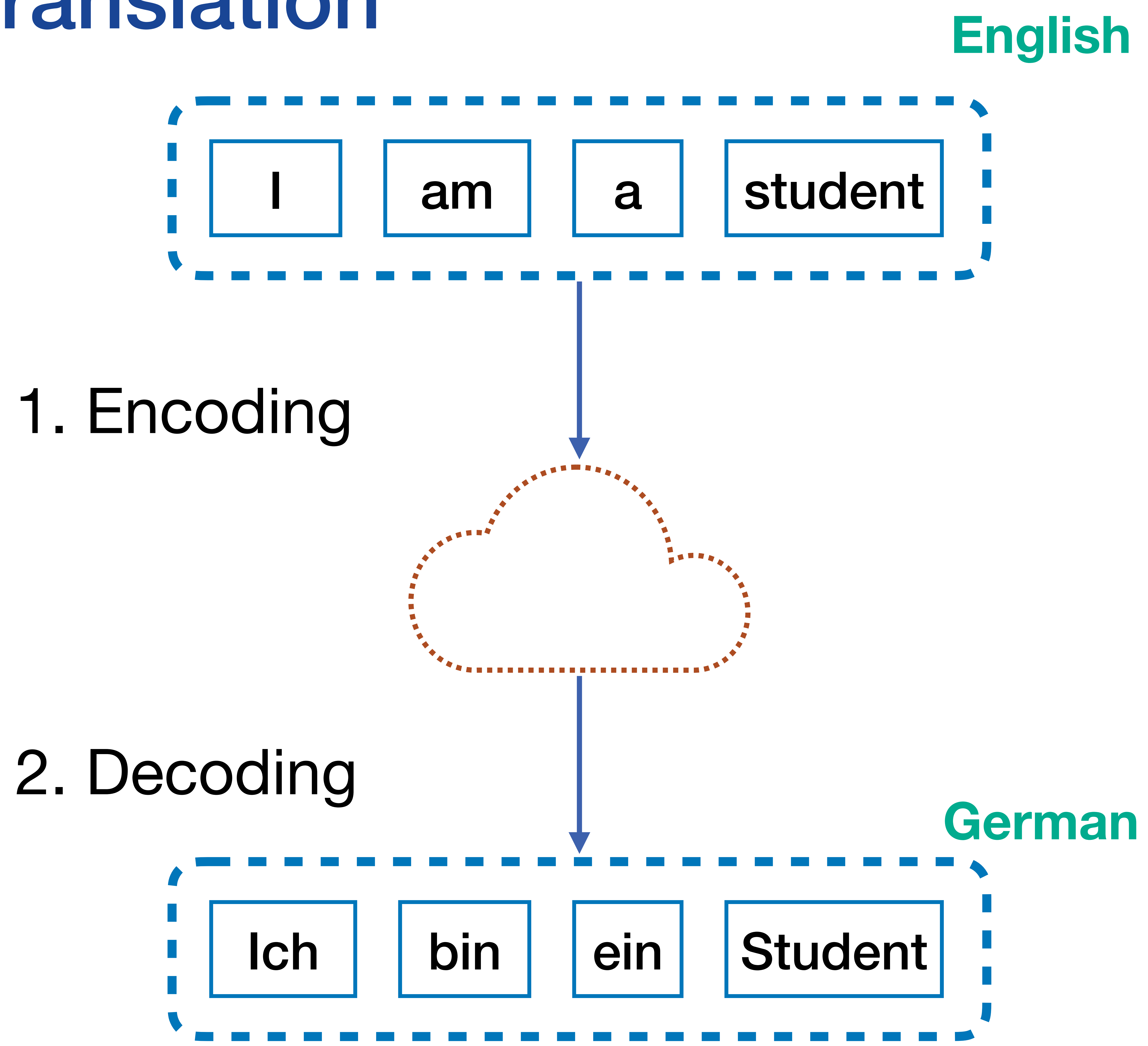
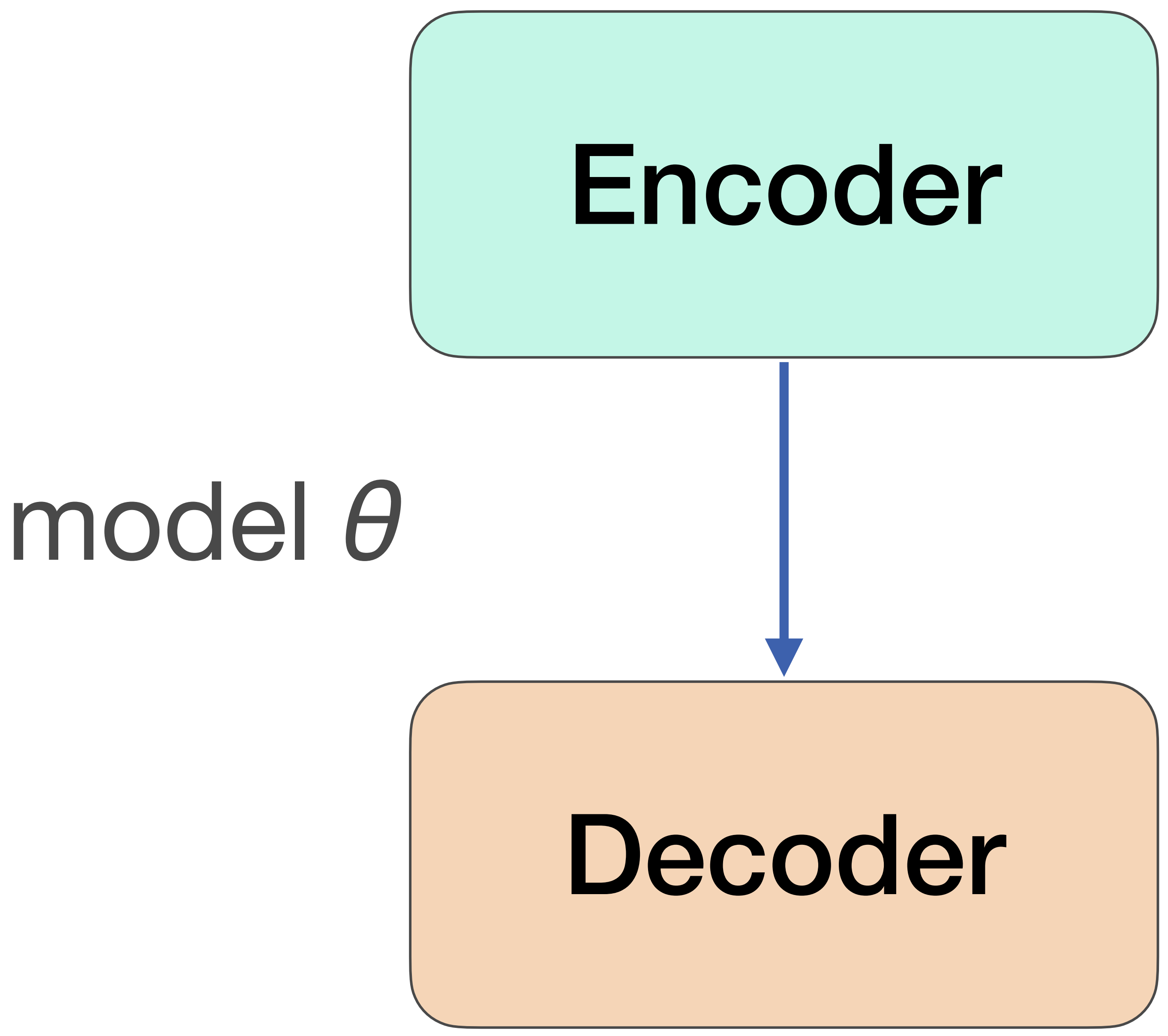


Image credit: https://sites.cs.ucsb.edu/~lilei/TALKS/2021-ACL/pre-training_nmt_ACL_tutorial_2021.pdf

Multilingual Neural Machine Translation



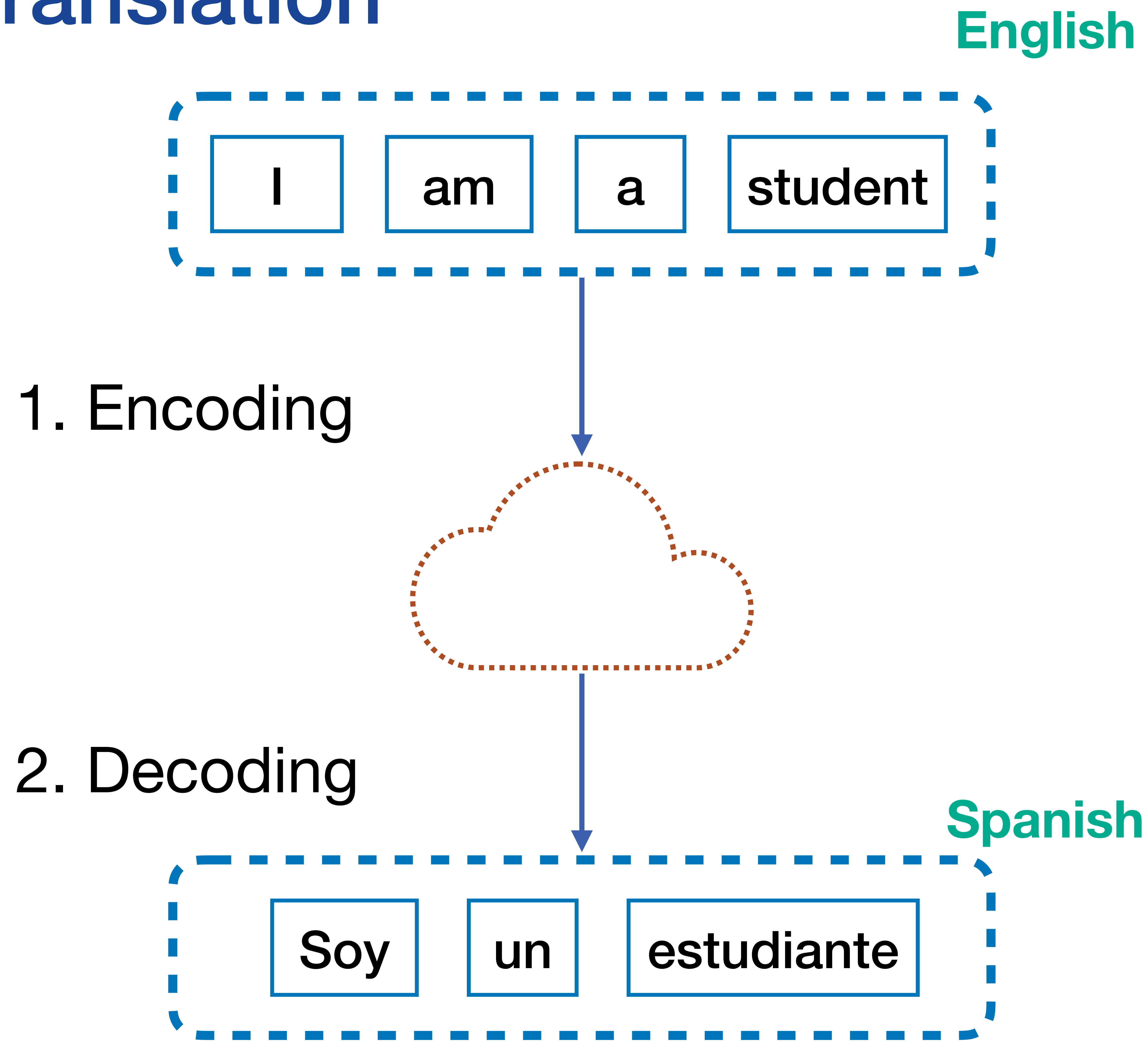
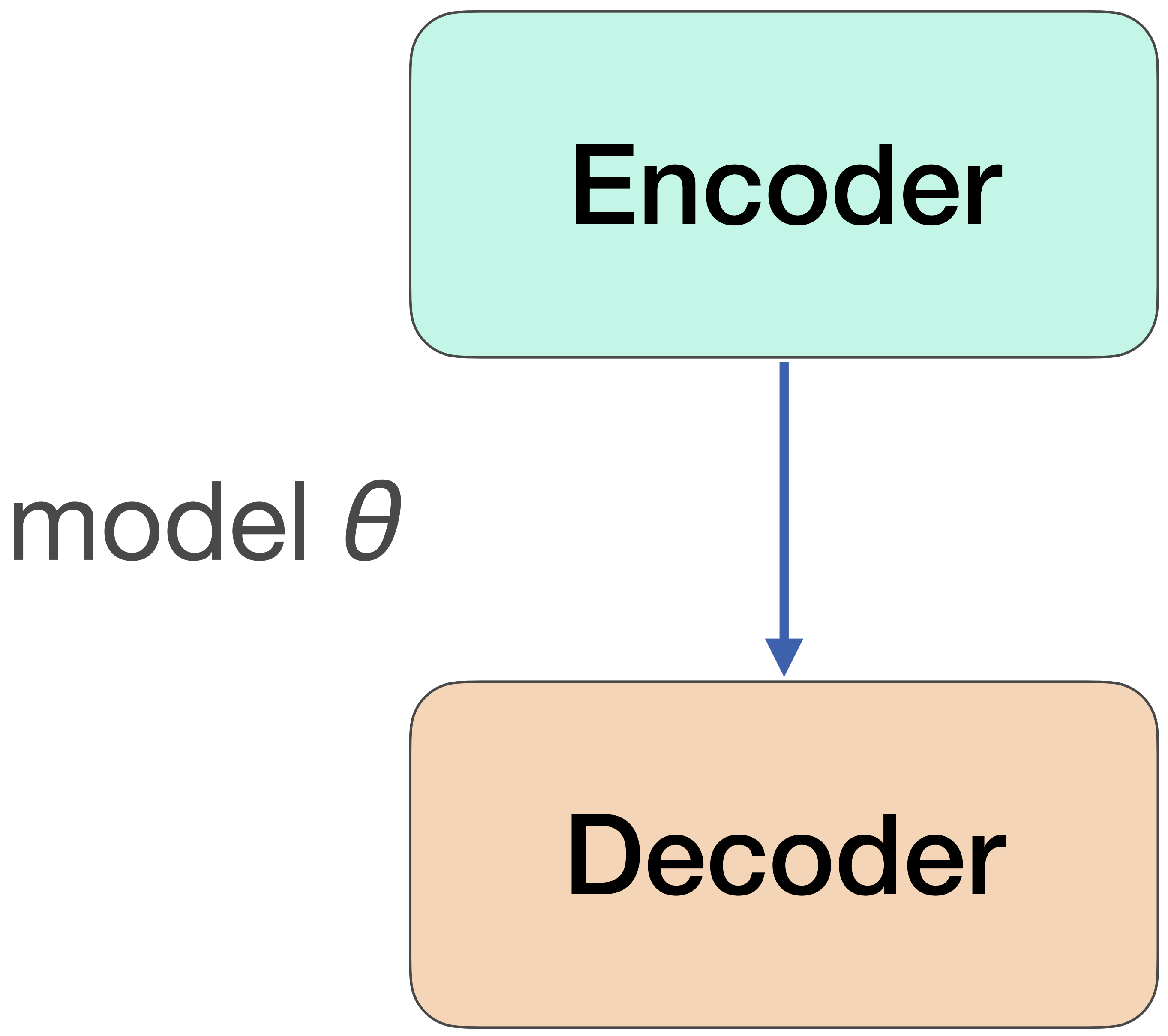


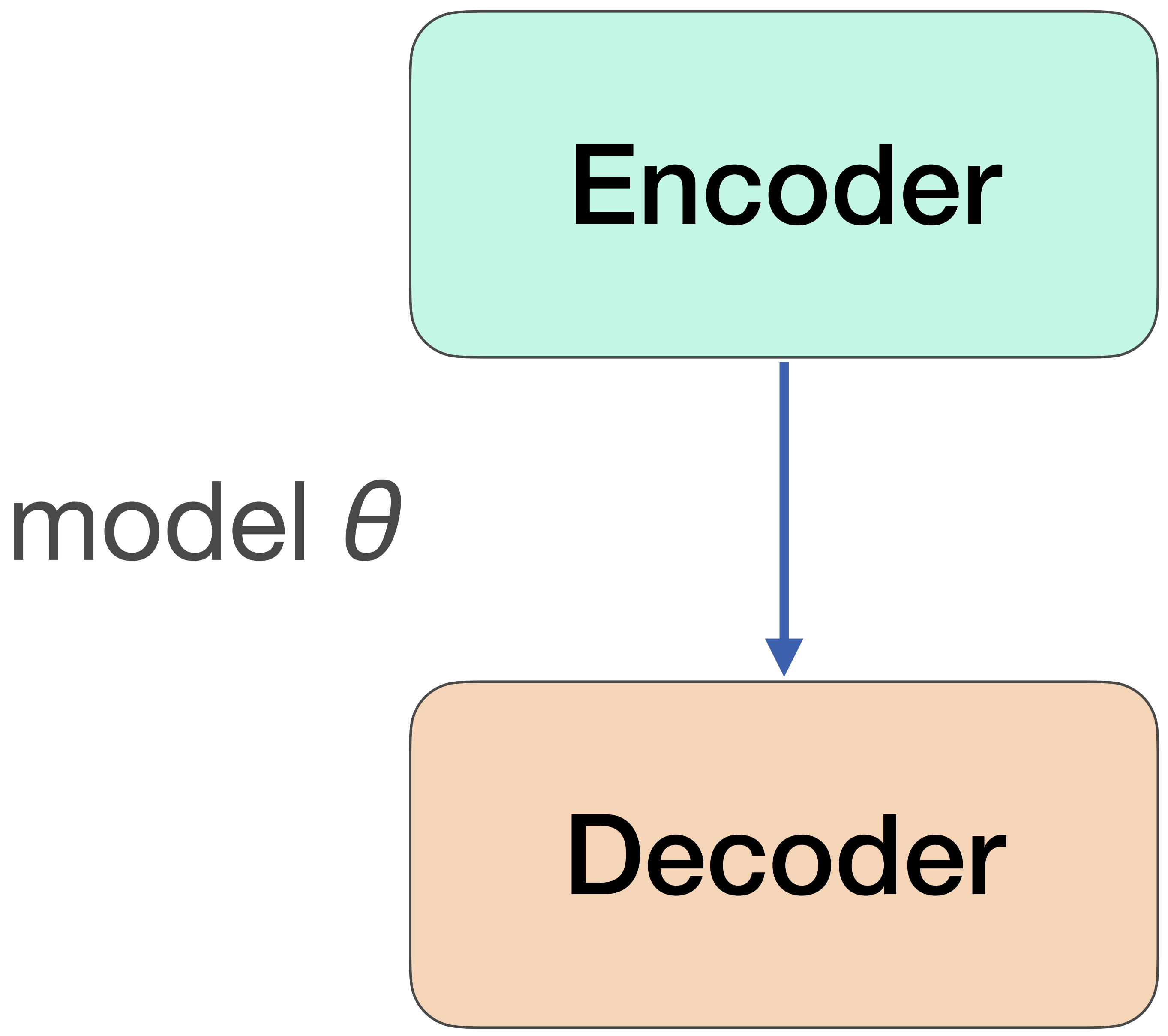
Image credit: https://sites.cs.ucsb.edu/~lilei/TALKS/2021-ACL/pre-training_nmt_ACL_tutorial_2021.pdf

Why choose multilingual NMT?



Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation [Johnson et al., 2016]

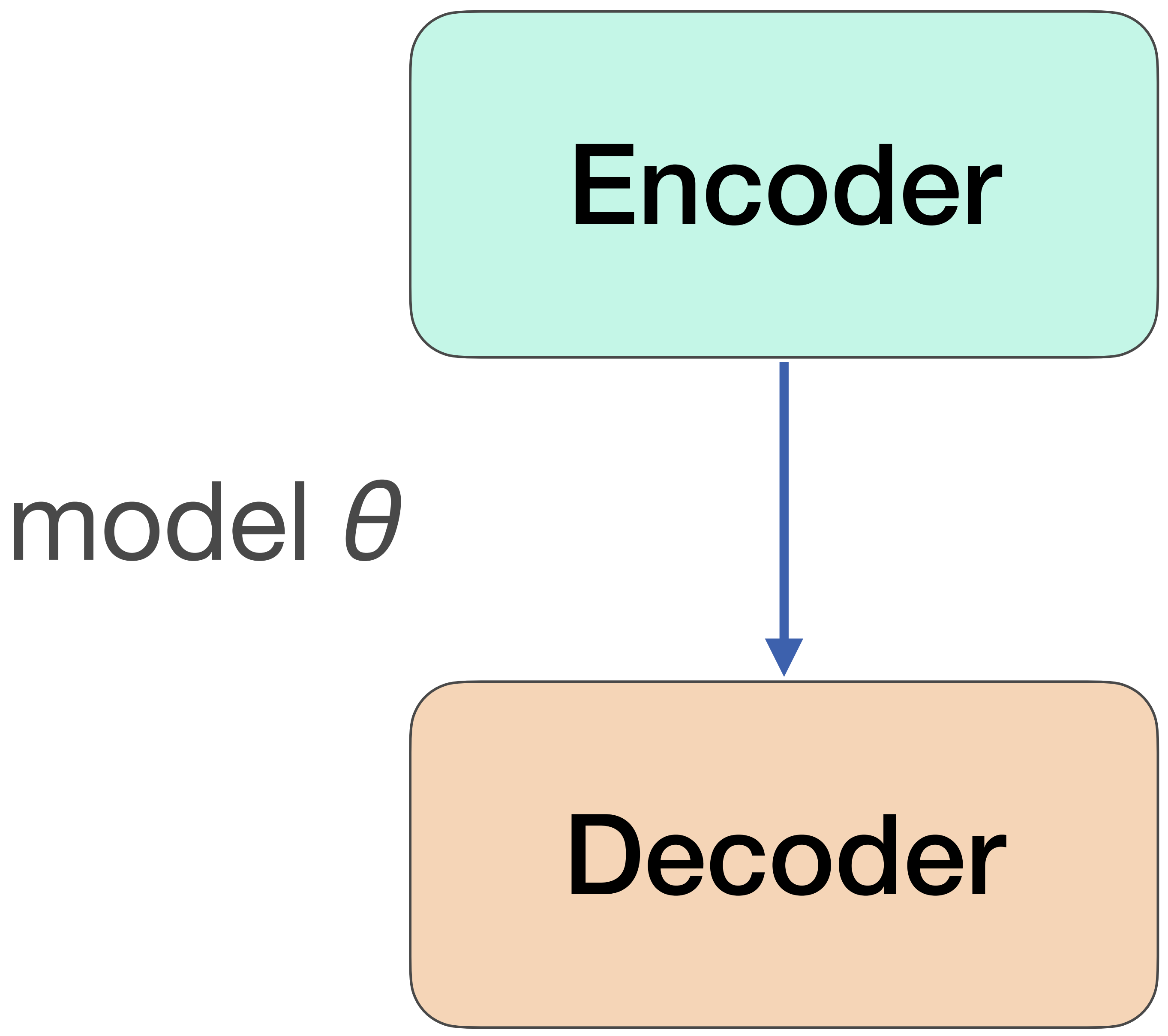
Why choose multilingual NMT?



- Low-resource languages benefit from sharing the same representation space as high-resource languages

Google’s Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation [Johnson et al., 2016]
Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges [Arivazhagan et al., 2019]
Massively Multilingual Neural Machine Translation [Aharoni et al., 2019]

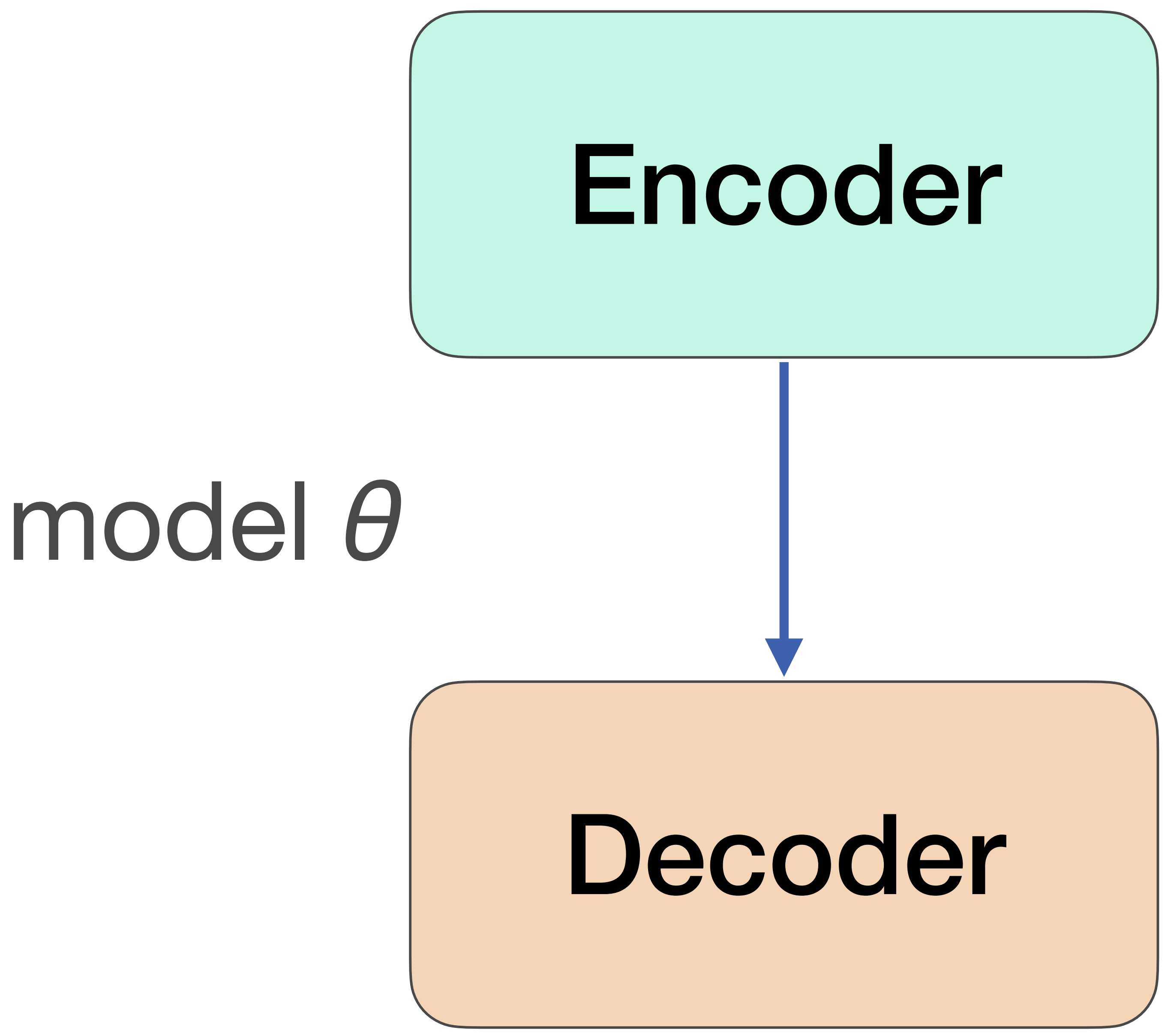
Why choose multilingual NMT?



- Low-resource languages benefit from sharing the same representation space as high-resource languages
- Operational costs are reduced

Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation [Johnson et al., 2016]
Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges [Arivazhagan et al., 2019]
Massively Multilingual Neural Machine Translation [Aharoni et al., 2019]

Why choose multilingual NMT?



- Low-resource languages benefit from sharing the same representation space as high-resource languages
- Operational costs are reduced
- Multilingual NMT scales to a large number of language pairs

Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation [Johnson et al., 2016]
Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges [Arivazhagan et al., 2019]
Massively Multilingual Neural Machine Translation [Aharoni et al., 2019]

Pre-training and fine-tuning

- Training a MNMT model from scratch is computationally expensive
- Parallel data scarcity for low/zero resource languages
- It is easier to find abundant monolingual data
- Pre-training is a way to leverage this data

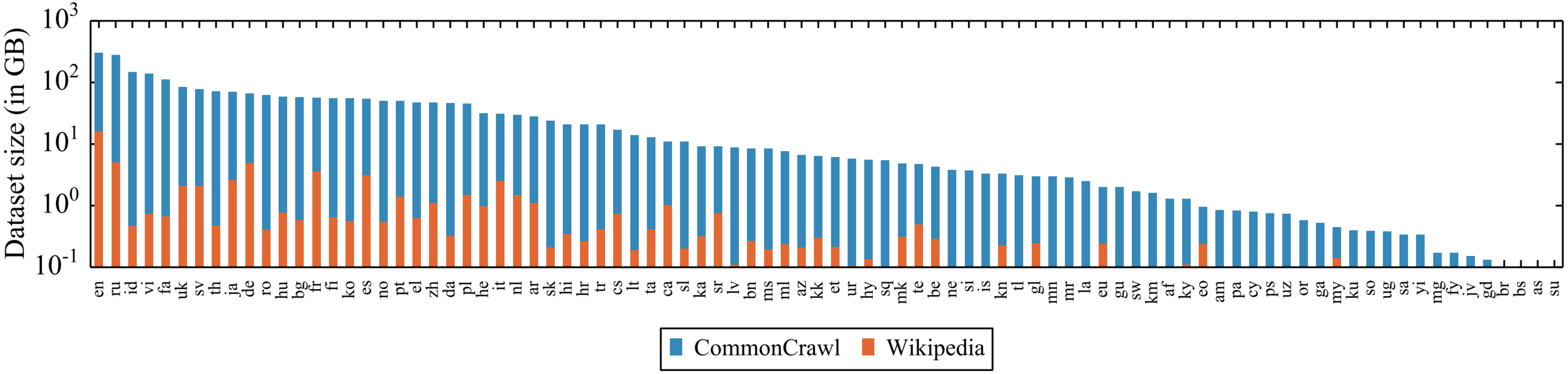


Image credit: Unsupervised Cross-lingual Representation Learning at Scale (XLM-R) [Conneau et al., 2020]

Pre-training and fine-tuning

- Fine-tuning large pretrained models is effective in many NLU tasks
- Perhaps it can also work for machine translation?

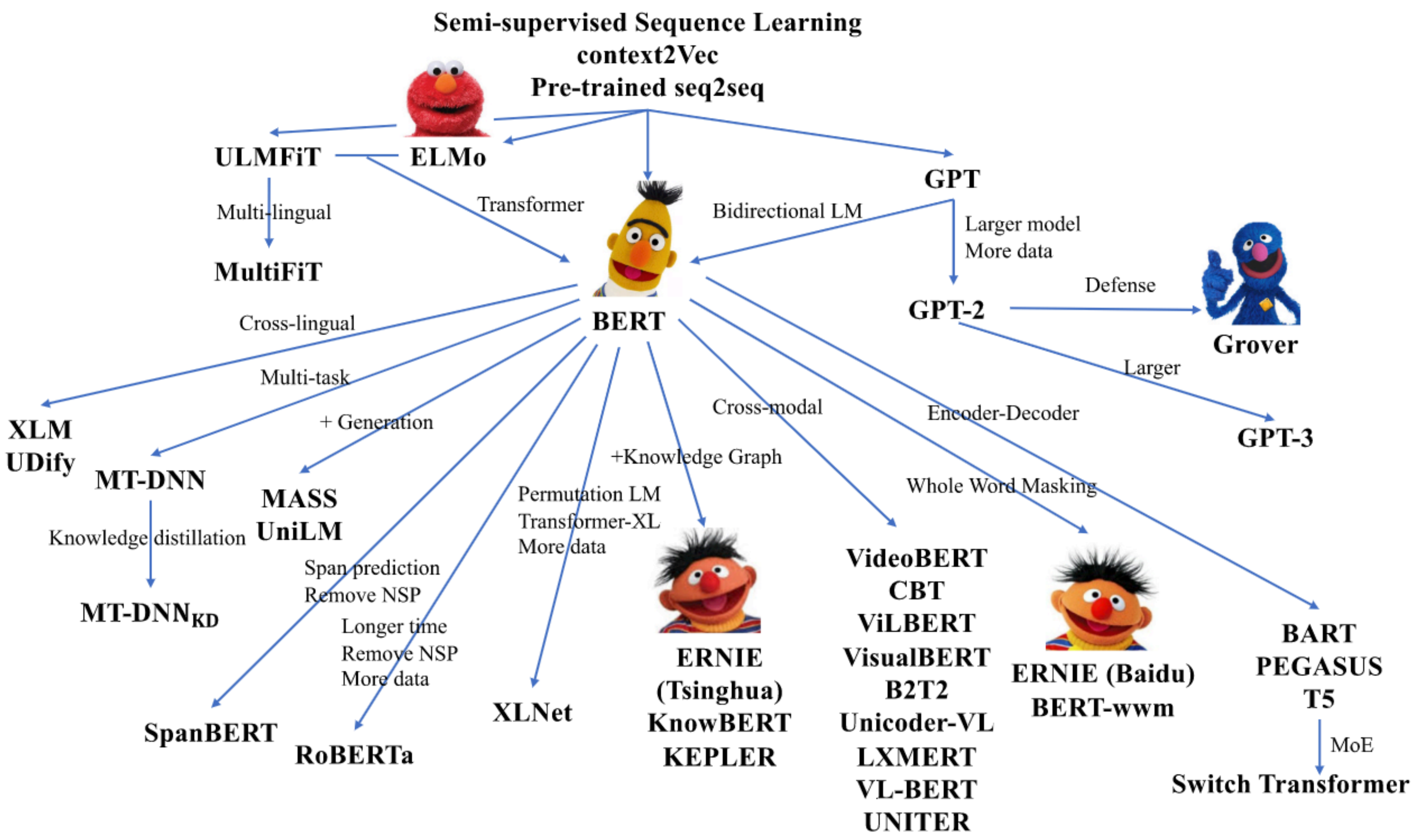
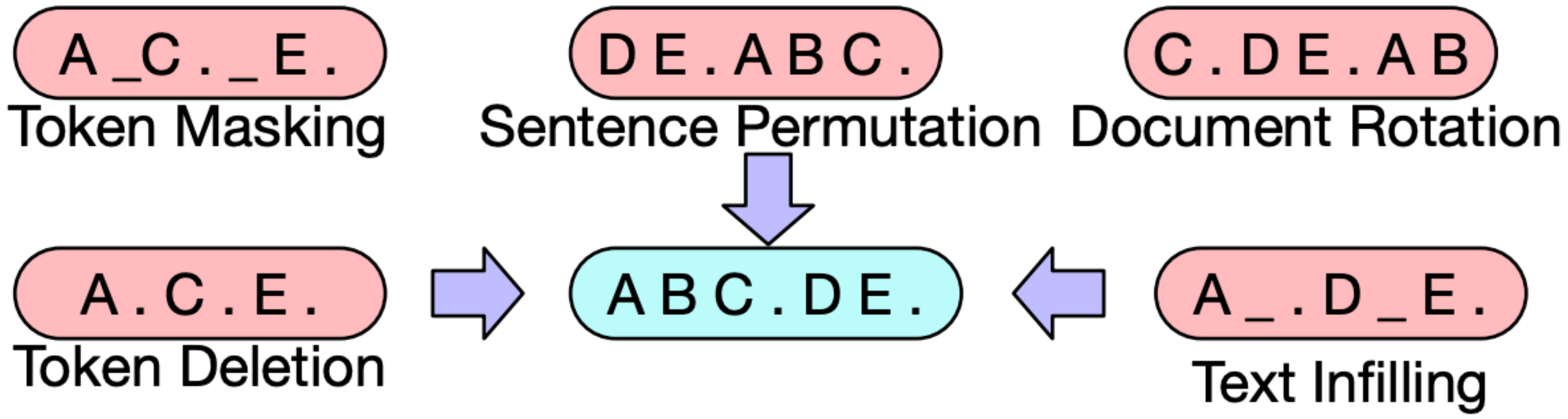


Image credit: Pre-trained models: Past, present and future [Han et al., 2021]

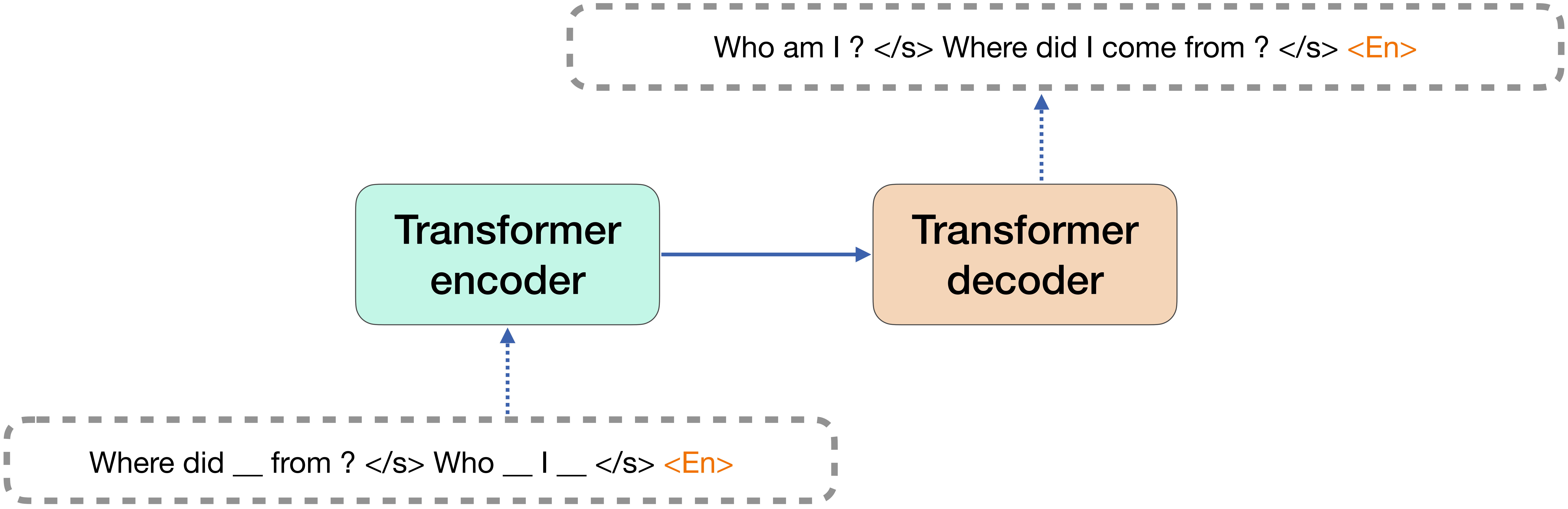
mBART: A multilingual pretraining model for NMT

- Encoder-decoder Transformer
- Denoising autoencoding in multiple languages
- The model uses purely monolingual data during pre-training



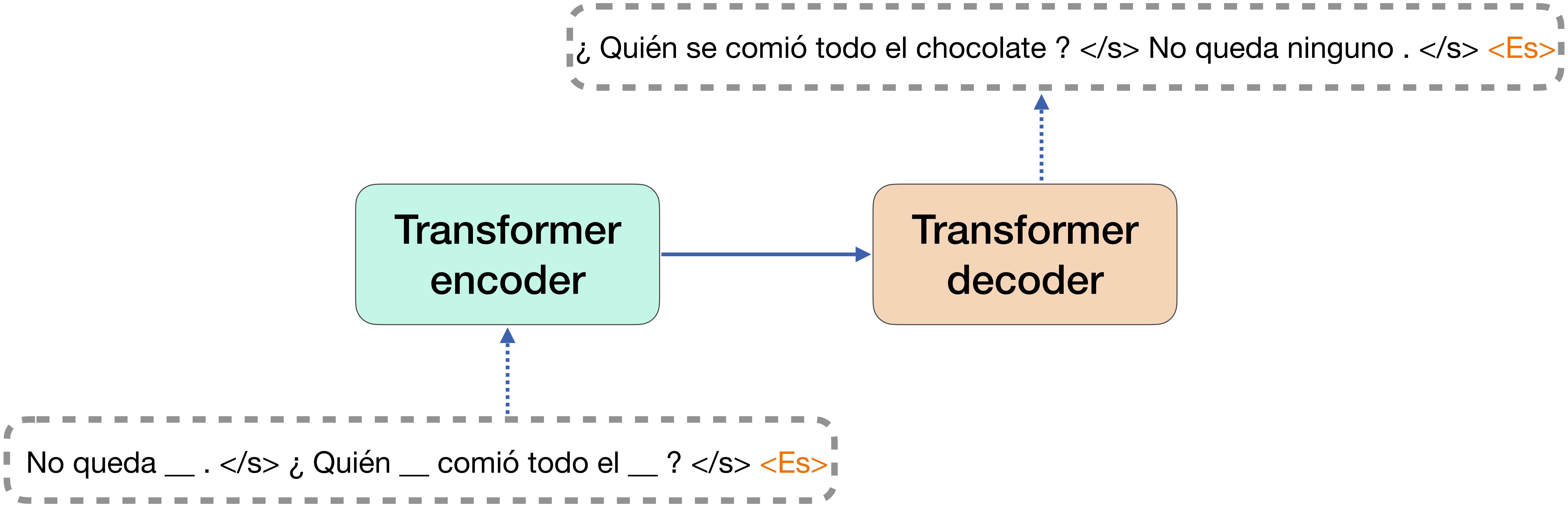
BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension [Lewis et al., 2020]
mBART: Multilingual Denoising Pre-training for Neural Machine Translation [Liu et al., 2020]

mBART: A multilingual pretraining model for NMT



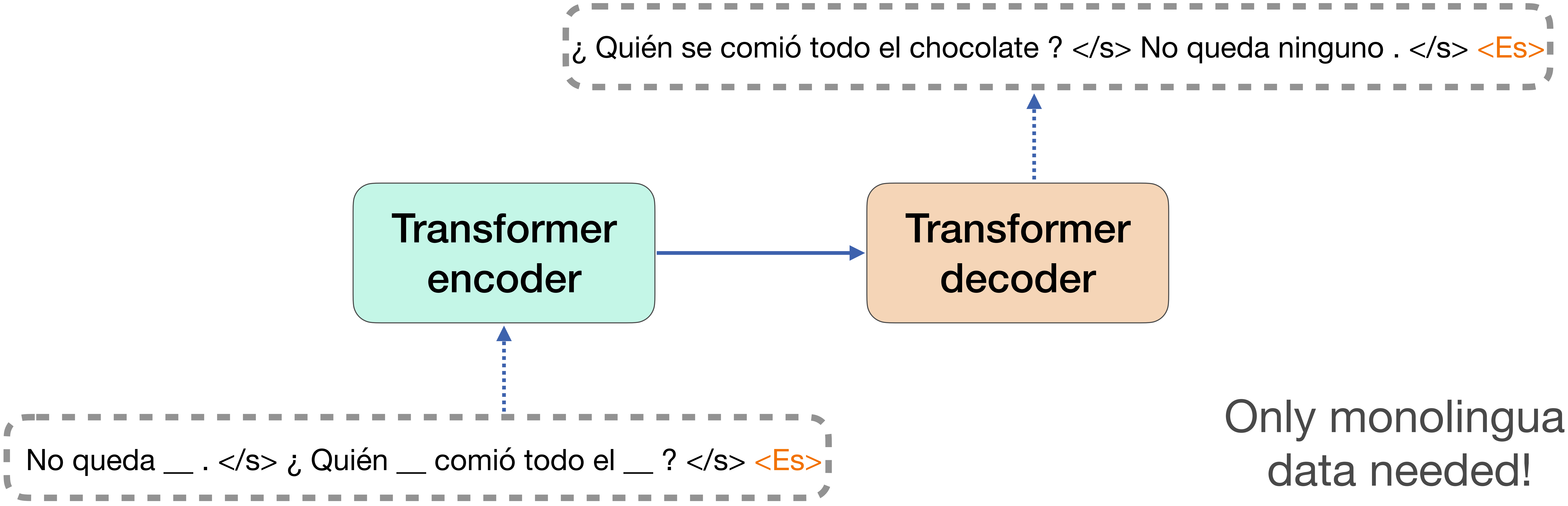
BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension [Lewis et al., 2020]
mBART: Multilingual Denoising Pre-training for Neural Machine Translation [Liu et al., 2020]

mBART: A multilingual pretraining model for NMT



BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension [Lewis et al., 2020]
mBART: Multilingual Denoising Pre-training for Neural Machine Translation [Liu et al., 2020]

mBART: A multilingual pretraining model for NMT

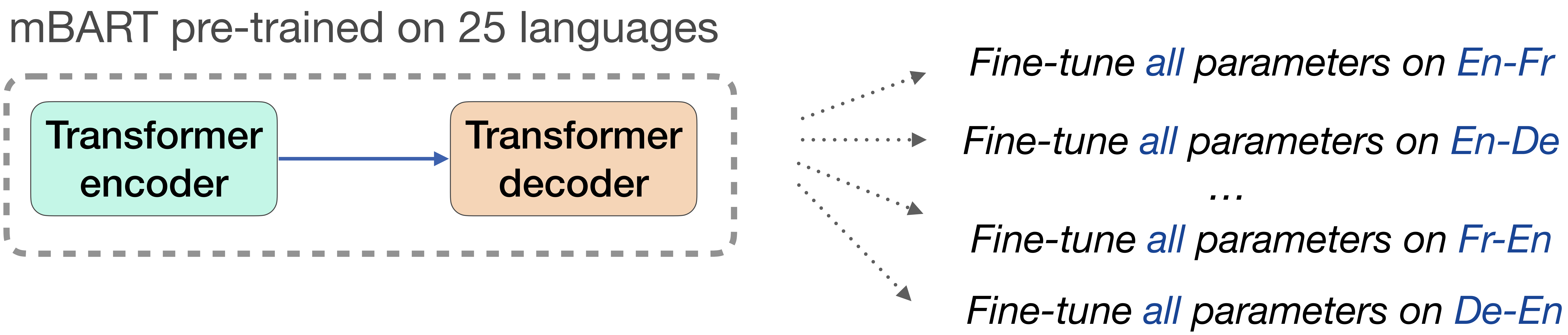


Only monolingual data needed!

BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension [Lewis et al., 2020]
mBART: Multilingual Denoising Pre-training for Neural Machine Translation [Liu et al., 2020]

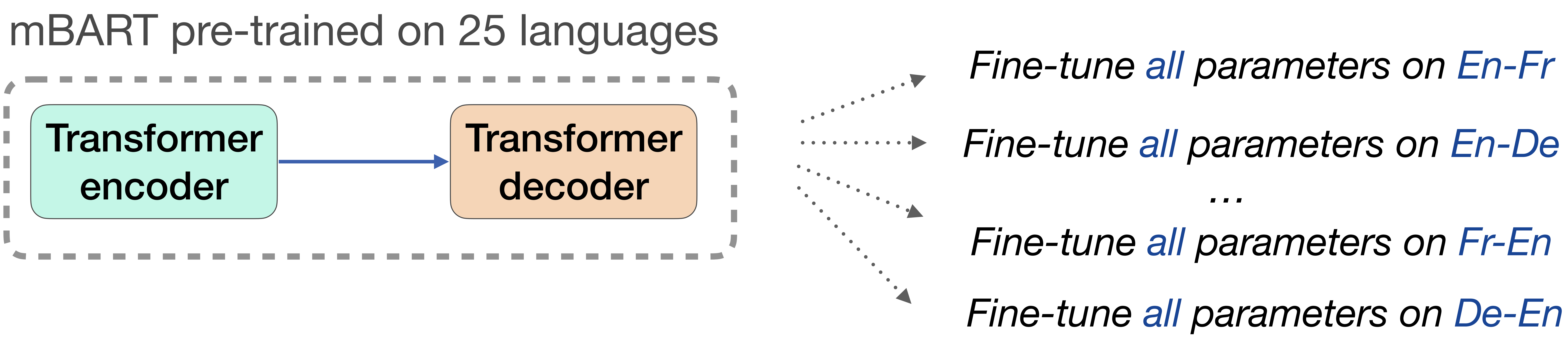
mBART: A multilingual pretraining model for NMT

- After that, the encoder-decoder model is fine-tuned for MT
- This is not efficient



mBART: A multilingual pretraining model for NMT

- After that, the encoder-decoder model is fine-tuned for MT
- This is not efficient

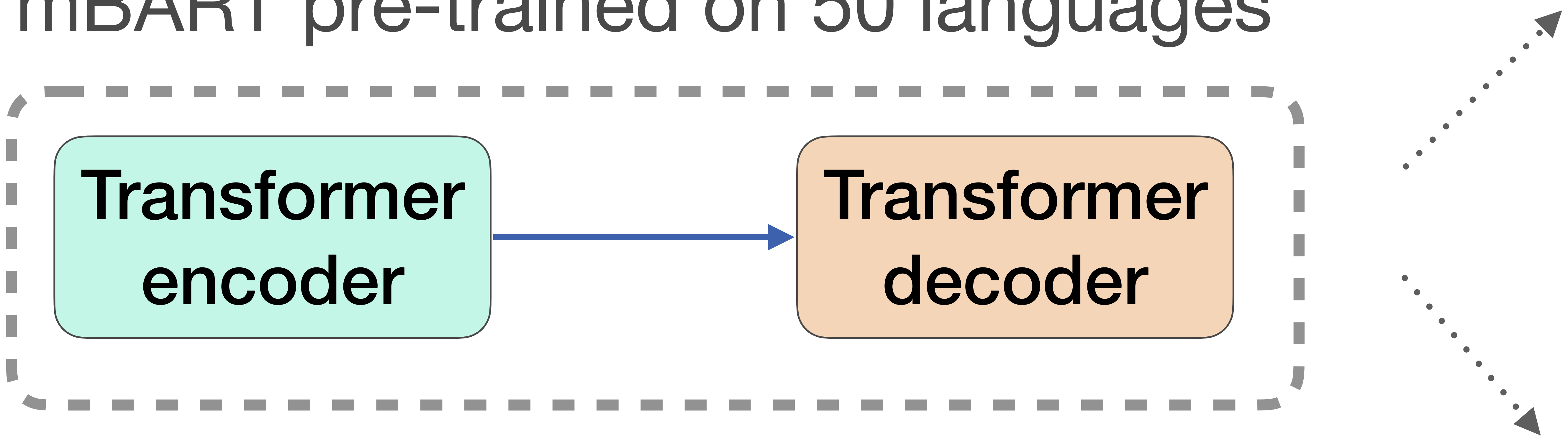


- **Idea**: multilingually fine-tune mBART for machine translation

Multilingual fine-tuning for NMT

- Improvement: fine-tune mBART in a multilingual manner

mBART pre-trained on 50 languages



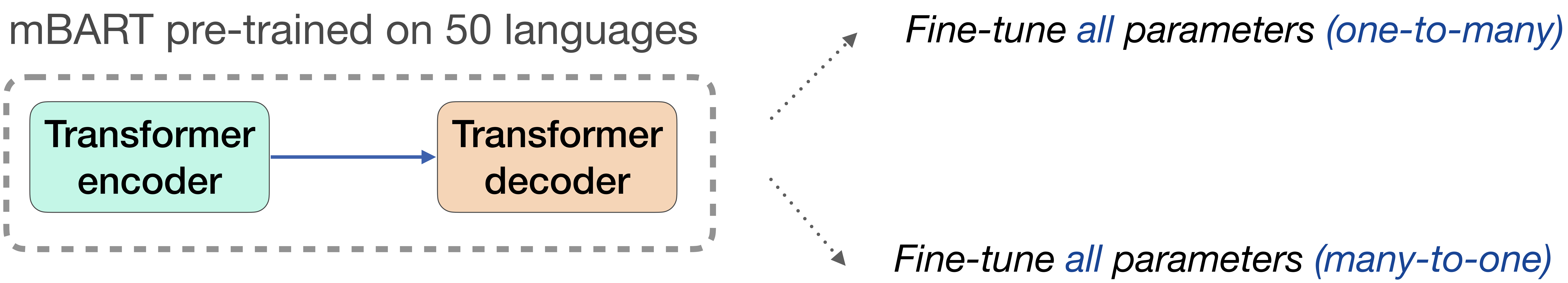
Fine-tune all parameters (one-to-many)

Fine-tune all parameters (many-to-one)

mBART-50: Multilingual Translation with Extensible Multilingual Pretraining and Finetuning [Tang et al., 2020]
Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges [Arivazhagan et al., 2019]

Multilingual fine-tuning for NMT

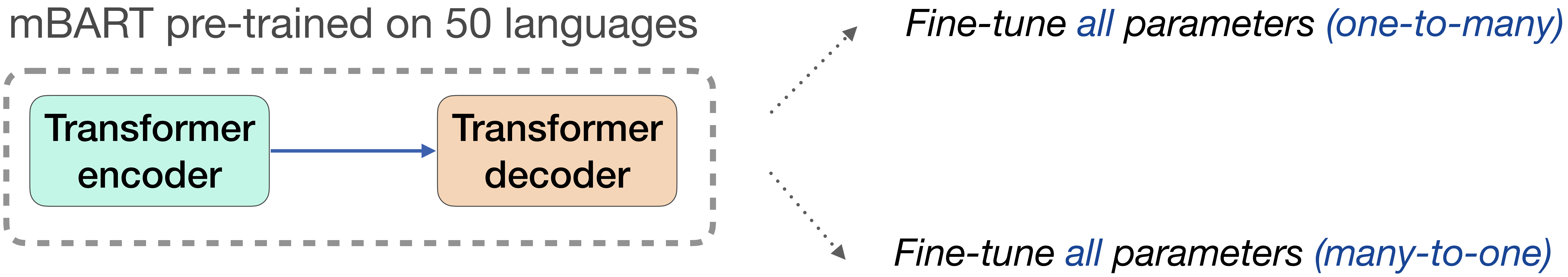
- Improvement: fine-tune mBART in a multilingual manner
- This step fails to model all languages equally well



mBART-50: Multilingual Translation with Extensible Multilingual Pretraining and Finetuning [Tang et al., 2020]
Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges [Arivazhagan et al., 2019]

Multilingual fine-tuning for NMT

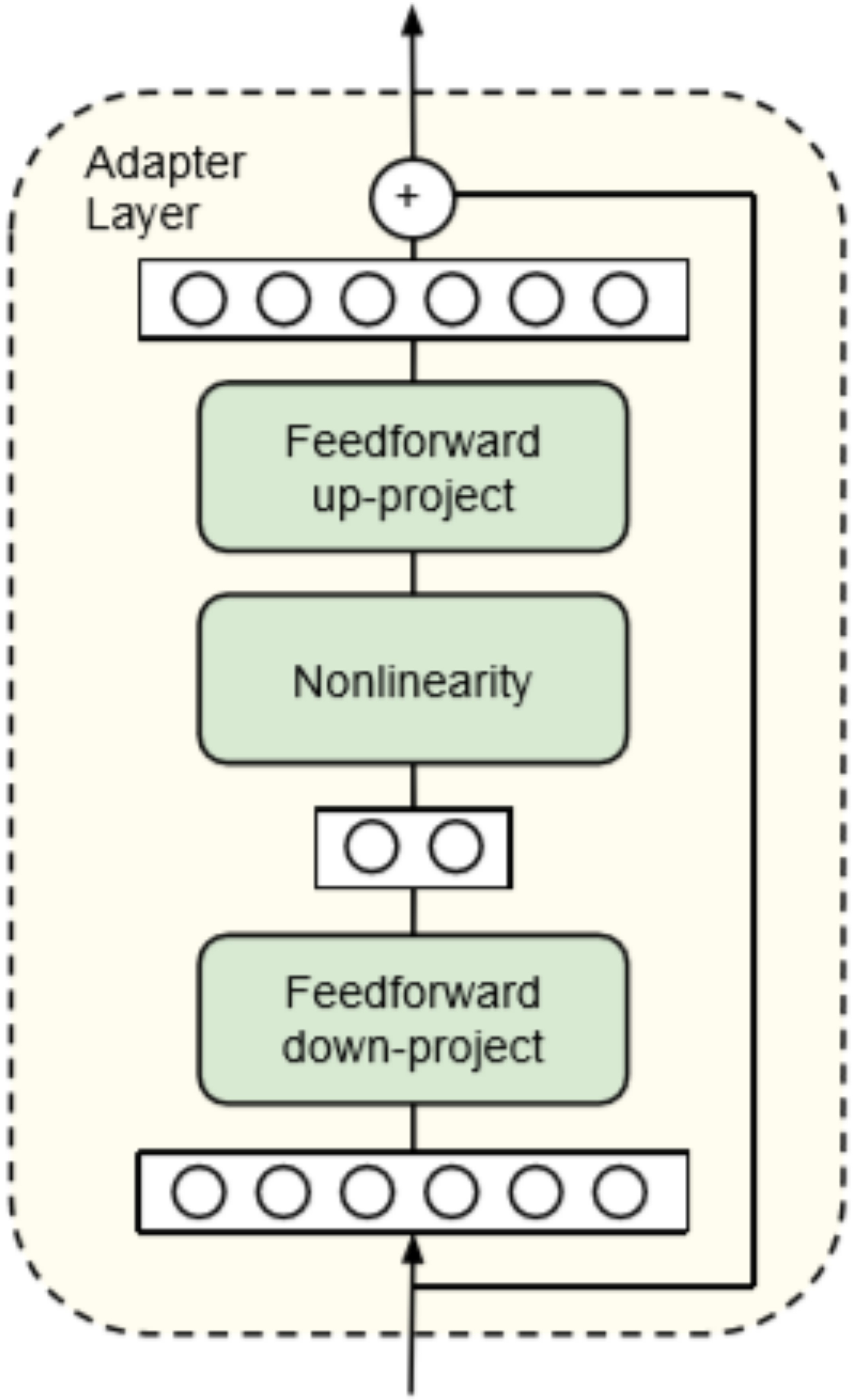
- Improvement: fine-tune mBART in a multilingual manner
- This step fails to model all languages equally well
- It is still required to fine-tune the entire model (680M params)



mBART-50: Multilingual Translation with Extensible Multilingual Pretraining and Finetuning [Tang et al., 2020]
Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges [Arivazhagan et al., 2019]

Efficient fine-tuning for NMT

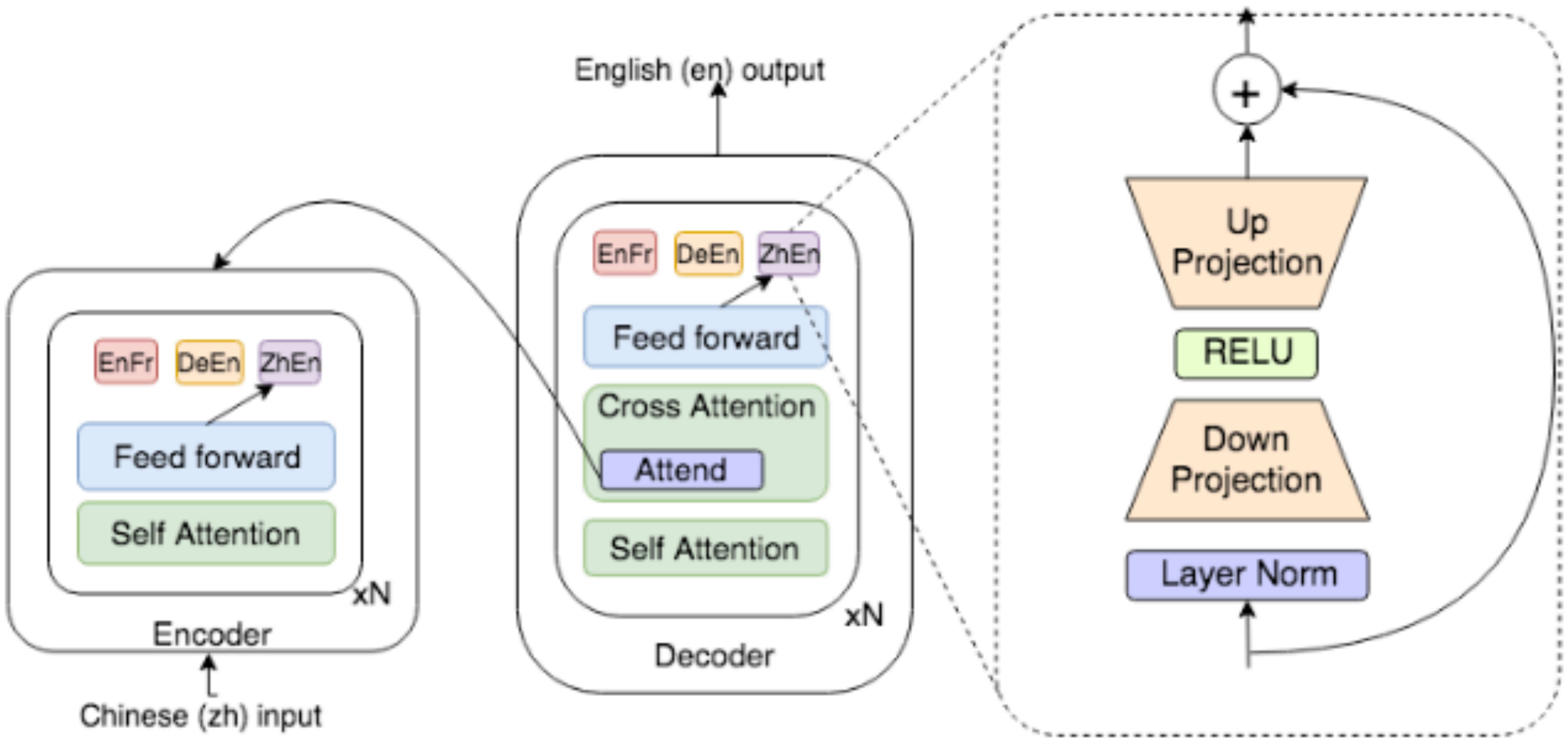
- An efficient alternative: adapters
- Adapters are inserted in each Transformer layer
- The pretrained model remains fixed, only the adapters are fine-tuned for MT



Learning multiple visual domains with residual adapters [Rebuffi et al., 2017]
Parameter-Efficient Transfer Learning for NLP [Houlsby et al., 2019]

Efficient fine-tuning for NMT

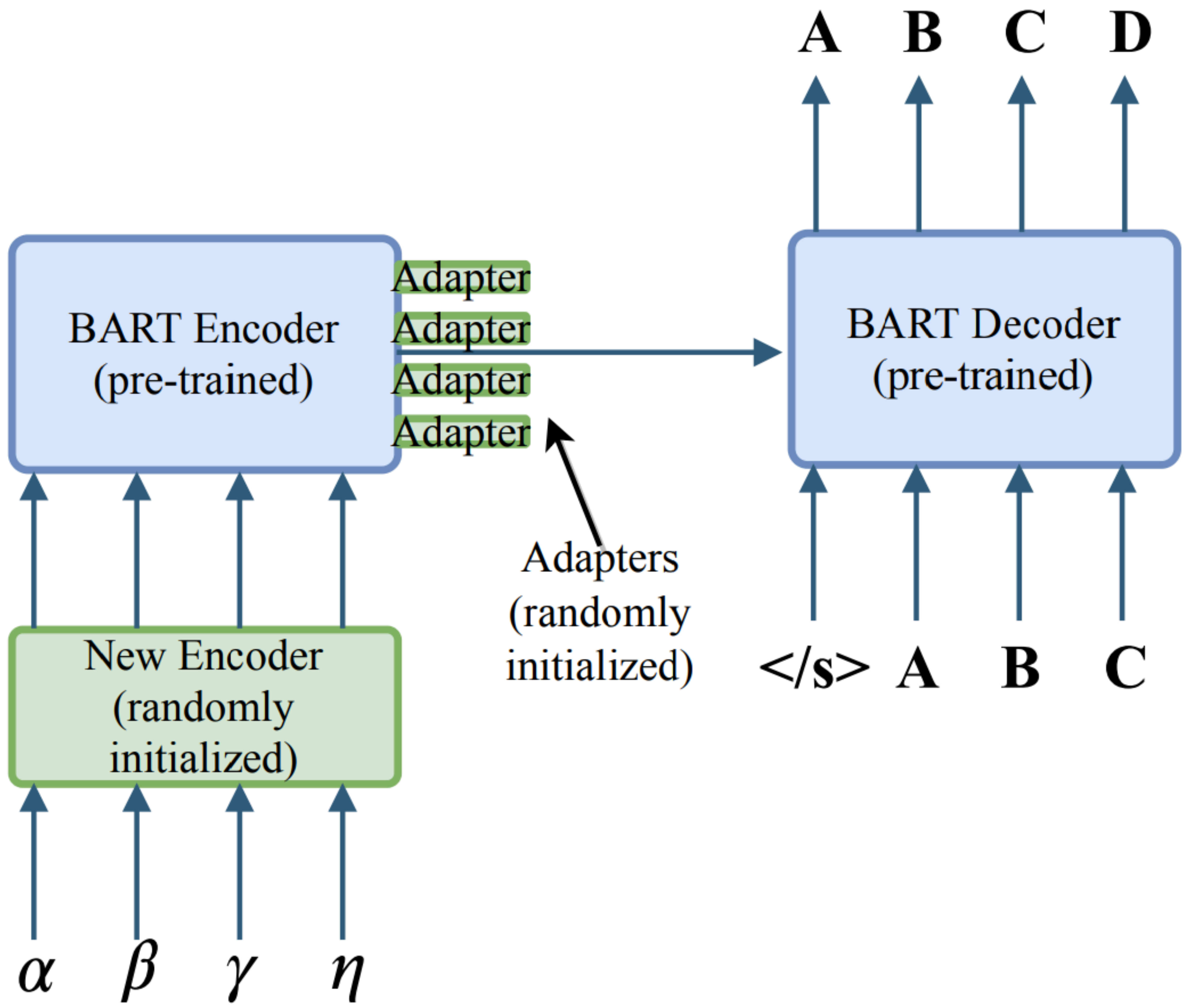
- A new set of adapters can be trained for **each** language pair
- This works well for high-resource languages
- But does not work for low-resource languages, because there is no sharing between related languages



Simple, Scalable Adaptation for Neural Machine Translation [Bapna and Firat, 2019]

Efficient fine-tuning for NMT

- A new set of adapters can be trained for **all** language pairs
- This suffers from negative interference between unrelated languages



Recipes for Adapting Pre-trained Monolingual and Multilingual Models to Machine Translation [Cooper Stickland et al., 2021]

Approach

Language-family adapters for multilingual NMT

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

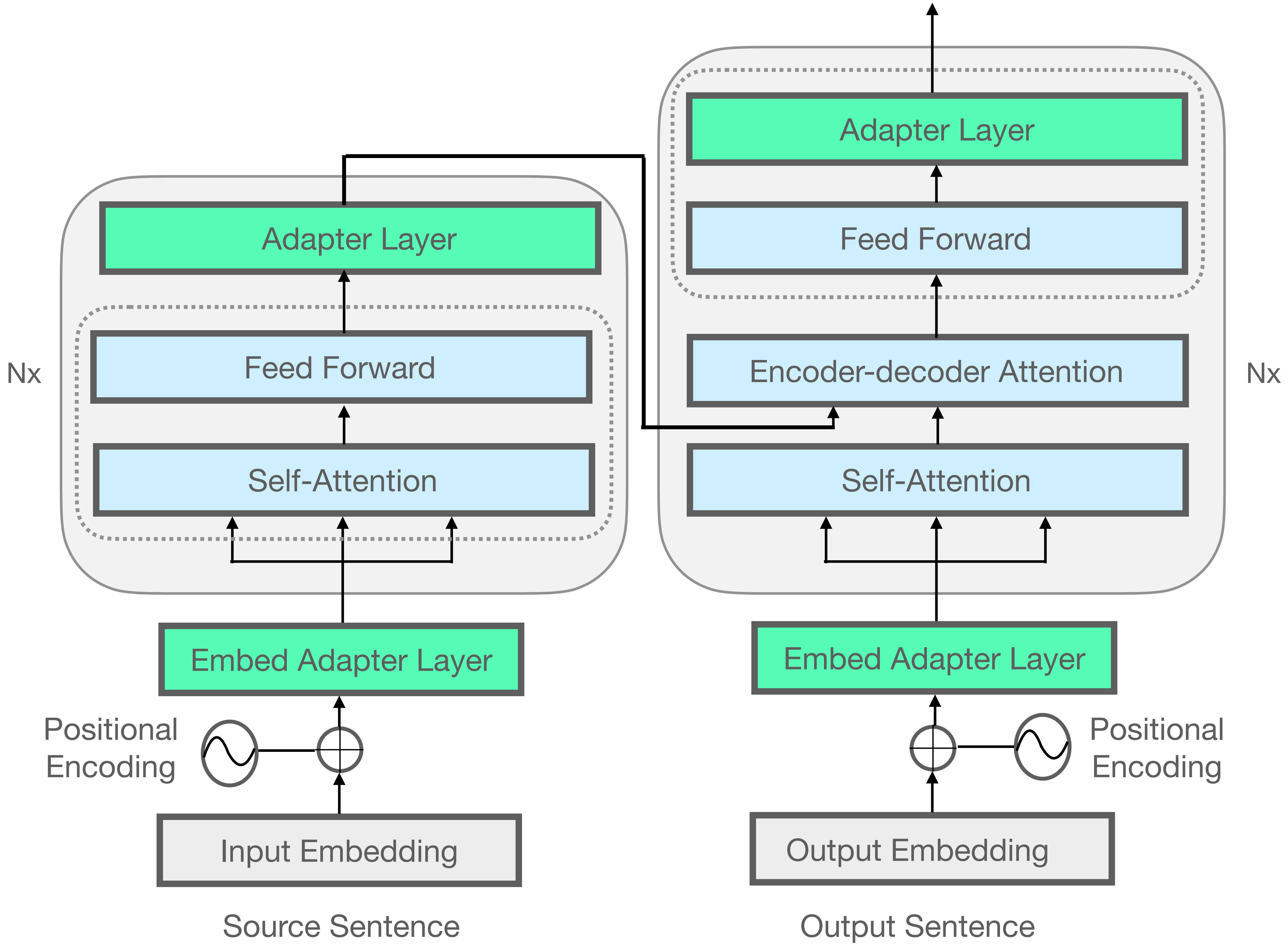
Language-family adapters for multilingual NMT

Our approach:

- Permits positive cross-lingual transfer, avoids negative interference
- Outperforms other adapter baselines
- Improves MT scores for various mid- and low-resource language pairs

Language-family adapters for multilingual NMT

- We add a set of adapters for each language family after the FFN in each Transformer layer of mBART-50
- We add an embedding-layer adapter
- We fine-tune only the adapters for multilingual NMT



Results + Analysis

Data + baselines

- 17 language pairs from 3 language families (based on linguistic knowledge)
- Starred languages have not been used for mBART-50 pre-training

Baselines: lang-pair (one set of adapters per language pair), lang-agnostic (one set of adapters for all language pairs)

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

OPUS-100: Improving Massively Multilingual Neural Machine Translation and Zero-Shot Translation [Zhang et al., 2020]
TED: When and Why are Pre-trained Word Embeddings Useful for Neural Machine Translation? [Qi et al., 2018]

Main findings

Model	Balto-Slavic	Austronesian	Indo-Iranian	Average
OPUS-100				
Lang-pair	**22.8**	25.8	13.7	20.3
Lang-agnostic	20.0	25.2	13.4	18.6
Lang-family	22.5	**28.4**	**16.3**	**21.3**
TED				
Lang-pair	24.2	**23.7**	10.4	20.3
Lang-agnostic	23.3	22.5	9.8	19.2
Lang-family	**24.6**	22.7	**12.7**	**20.7**

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

Language-family adapters outperform the baselines on both parallel datasets

Main findings

Model	Balto-Slavic	Austronesian	Indo-Iranian	Average
OPUS-100				
Lang-pair	**22.8**	25.8	13.7	20.3
Lang-agnostic	20.0	25.2	13.4	18.6
Lang-family	22.5	**28.4**	**16.3**	**21.3**
TED				
Lang-pair	24.2	**23.7**	10.4	20.3
Lang-agnostic	23.3	22.5	9.8	19.2
Lang-family	**24.6**	22.7	**12.7**	**20.7**

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

Language-family adapters outperform the baselines on both parallel datasets

Method	Trainable parameters
mBART-50	100%
Lang-pair	52.2%
Lang-agnostic	8.4%
Lang-family	**11.9%**

Trade-off between performance and efficiency

Is the embedding-layer adapter helpful?

	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 when translating out of English (*en* -> *xx*) on OPUS-100.

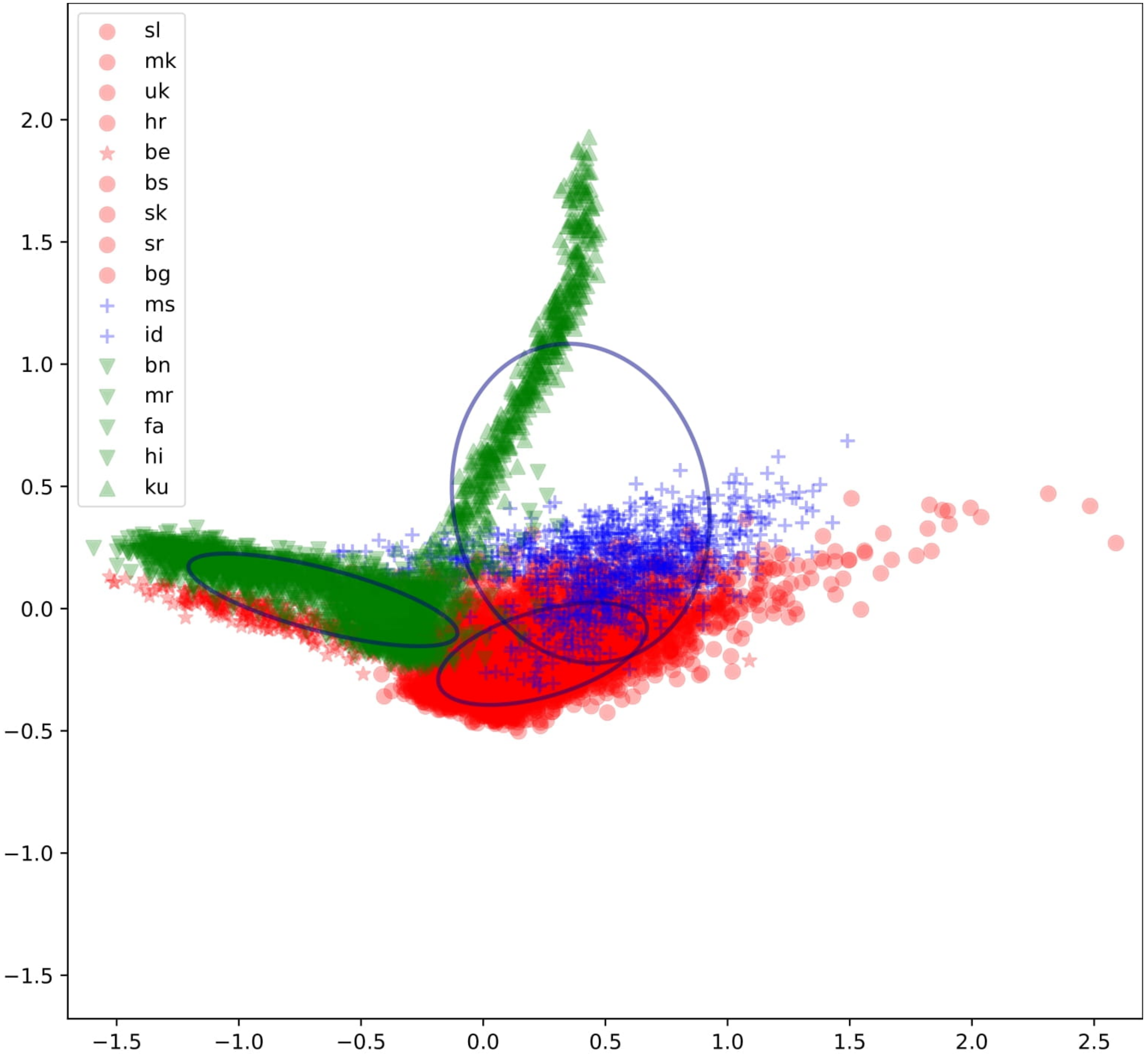
- Embedding-layer adapters consistently improve translation scores while they only add +0.1% of the parameters of mBART-50
- They encode lexical-level information for the languages of interest

How to create language groups

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

Unsupervised clustering to form language groups

- We used XLM-R to get representations of our data and clustered them using a Gaussian Mixture Model (GMM)
- The color shows the language family based on linguistic knowledge
- Clusters are mostly corresponding to the language families (except for *be* and *ku*)



Unsupervised clustering to form language groups

	Bulgarian	Indonesian	Persian	Belarusian	Kurdish	Average
Language family	25.4	31.3	9.8	15.2	15.3	**21.3**
GMM	23.9	29.7	9.2	14.9	14.3	19.4
Random	22.9	27.8	7.0	12.1	15.0	18.4

Evaluation of different methods to form language groups in *en->xx* (BLEU) on OPUS-100.

- Main takeaway: in the presence of linguistic knowledge, it is beneficial to use it create language groups
- If we do not have information on our data (or it is noisy), GMM clustering can be helpful

Conclusion

Conclusion

- We presented an approach that encodes the relations between languages using language-family adapters
- This is an effective and efficient method for translation from English to mid- and low-resource languages
- Clustering languages together with a GMM might be helpful in the absence of linguistic knowledge bases

Code/Paper: will be available soon!

Questions? achron@cis.lmu.de

Thanks!