Multilingual NLP

CS685 Spring 2023

Advanced Natural Language Processing

Mohit Iyyer

College of Information and Computer Sciences

University of Massachusetts Amherst

Some slides from Simeng Sun

Multilingual transfer

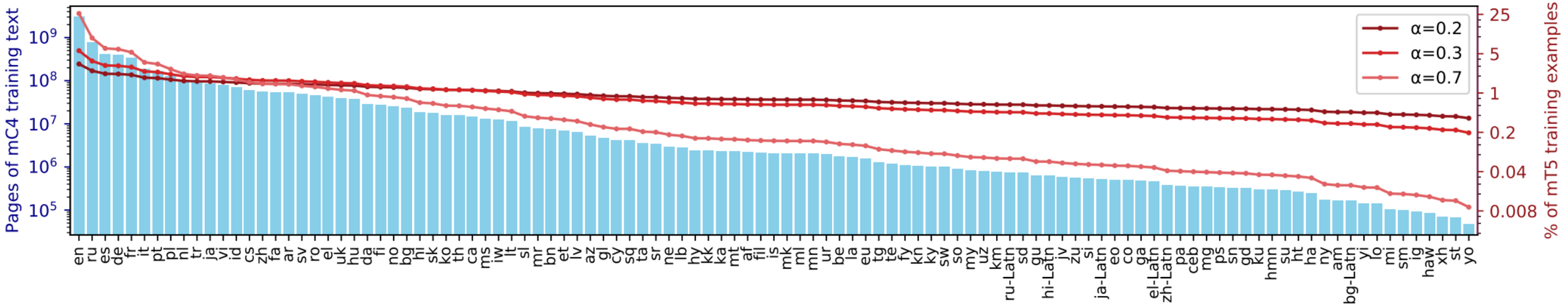
- So far, we've mainly talked about pretraining and fine-tuning models on English text.

- One approach: pretrain BERT-like models on *monolingual* data from a different language
  - "BERTje > Dutch, "FlauBERT" > French, "PhoBERT" > Vietnamese, etc.

- Another approach: pretrain models on a large mixture of many languages
  - mBERT, mBART, XLM-R, mT5, byT5, etc.
  - Allows for transfer learning *across* languages

mC4 dataset

- 107 languages, lower-resource languages upsampled based on their frequency in the dataset



Model	Architecture	Parameters	# languages	Data source
mBERT (Devlin, 2018)	Encoder-only	180M	104	Wikipedia
XLM (Conneau and Lample, 2019)	Encoder-only	570M	100	Wikipedia
XLM-R (Conneau et al., 2020)	Encoder-only	270M – 550M	100	Common Crawl (CCNet)
mBART (Lewis et al., 2020b)	Encoder-decoder	680M	25	Common Crawl (CC25)
MARGE (Lewis et al., 2020a)	Encoder-decoder	960M	26	Wikipedia or CC-News
mT5 (ours)	Encoder-decoder	300M – 13B	101	Common Crawl (mC4)

mT5 paper, Sue et al., ACL 2021

Cross-lingual zero-shot learning

- We are given labeled training data for task X only in language A. Can we build a model that can make predictions for task X in a different language B?

- **Idea**: leverage information from high-resource languages to help improve performance on low-resource languages.

- **Zero-shot** learning: no labeled data is available for the target task X in language B, although unlabeled data in language B might be available for pretraining

XNLI benchmark

Language	Premise / Hypothesis	Genre	Label
English	You don't have to stay there. You can leave.	Face-To-Face	Entailment
French	La figure 4 montre la courbe d'offre des services de partage de travaux. Les services de partage de travaux ont une offre variable.	Government	Entailment
Spanish	Y se estremeció con el recuerdo. El pensamiento sobre el acontecimiento hizo su estremecimiento.	Fiction	Entailment
German	Während der Depression war es die ärmste Gegend, kurz vor dem Hungertod. Die Weltwirtschaftskrise dauerte mehr als zehn Jahre an.	Travel	Neutral
Swahili	Ni silaha ya plastiki ya moja kwa moja inayopiga risasi. Inadumu zaidi kuliko silaha ya chuma.	Telephone	Neutral
Russian	И мы занимаемся этим уже на протяжении 85 лет. Мы только начали этим заниматься.	Letters	Contradiction
Chinese	让我告诉你，美国人最终如何看待你作为独立顾问的表现。 美国人完全不知道您是独立律师。	Slate	Contradiction

XNLI dataset, Conneau et al., ACL 2018

XNLI given only English training data

Model	Sentence pair	
	XNLI	PAWS-X
Metrics	Acc.	Acc.
Cross-lingual zero-shot transfer (models fine-tu		
mBERT	65.4	81.9
XLM	69.1	80.9
InfoXLM	81.4	-
X-STILTs	80.4	87.7
XLM-R	79.2	86.4
VECO	79.9	88.7
RemBERT	80.8	87.5
mT5-Small	67.5	82.4
mT5-Base	75.4	86.4
mT5-Large	81.1	88.9
mT5-XL	82.9	89.6
mT5-XXL	**85.0**	**90.0**

What if we use a machine translation system to get more labeled data (e.g., translate all the labeled English text to other languages)?

Adding translations doesn't improve that much over the zero-shot setting!

Model	Sentence pair	
	XNLI	PAWS-X
Metrics	Acc.	Acc.
Cross-lingual zero-shot transfer (models fine-tu		
mBERT	65.4	81.9
XLM	69.1	80.9
InfoXLM	81.4	-
X-STILTs	80.4	87.7
XLM-R	79.2	86.4
VECO	79.9	88.7
RemBERT	80.8	87.5
mT5-Small	67.5	82.4
mT5-Base	75.4	86.4
mT5-Large	81.1	88.9
mT5-XL	82.9	89.6
mT5-XXL	**85.0**	**90.0**

Translate-train (models fine-tuned on English		
XLM-R	82.6	90.4
FILTER + Self-Teaching	83.9	91.4
VECO	83.0	91.1
mT5-Small	64.7	79.9
mT5-Base	75.9	89.3
mT5-Large	81.8	91.2
mT5-XL	84.8	91.0
mT5-XXL	**87.8**	**91.5**

TyDiQA benchmark

1. **Passage Selection Task:** Given a list of the passages in the article, return either (a) the index of the passage that answers the question or (b) NULL if no such passage exists.

2. **Minimal Answer Span Task:** Given the full text of an article, return one of (a) the start and end byte indices of the minimal span that completely answers the question; (b) YES or NO if the question requires a yes/no answer and we can draw a conclusion from the passage; (c) NULL if it is not possible to produce a minimal answer for this question.

Language	Latin script[a]	White space tokens	Sentence boundaries	Word formation[b]	Gender[c]	Prodrop
English	+	+	+	+	+[d]	—
Arabic	—	+	+	++	+	+
Bengali	—	+	+	+	+	+
Finnish	+	+	+	+++	—	—
Indonesian	+	+	+	+	—	+
Japanese	—	—	+	+	—	+
Kiswahili	+	+	+	+++	—[e]	+
Korean	—	+[f]	+	+++	—	+
Russian	+	+	+	++	+	+
Telugu	—	+	+	+++	+	+
Thai	—	—	—	+	+	+

[a] '—' indicates **Latin script** is not the conventional writing system. Intermixing of Latin script should still be expected.
[b] We include inflectional and derivation phenomena in our notion of **word formation**.
[c] We limit the **gender** feature to sex-based gender systems associated with coreferential gendered personal pronouns.
[d] English has grammatical gender only in third person personal and possessive pronouns.
[e] Kiswahili has morphological noun classes (Corbett, 1991), but here we note sex-based gender systems.
[f] In Korean, tokens are often separated by whitespace, but prescriptive spacing conventions are commonly flouted.

Table 1: Typological features of the 11 languages in TyDi QA. We use + to indicate that this phenomena occurs, ++ to indicate that it occurs frequently, and + + + to indicate very frequently.

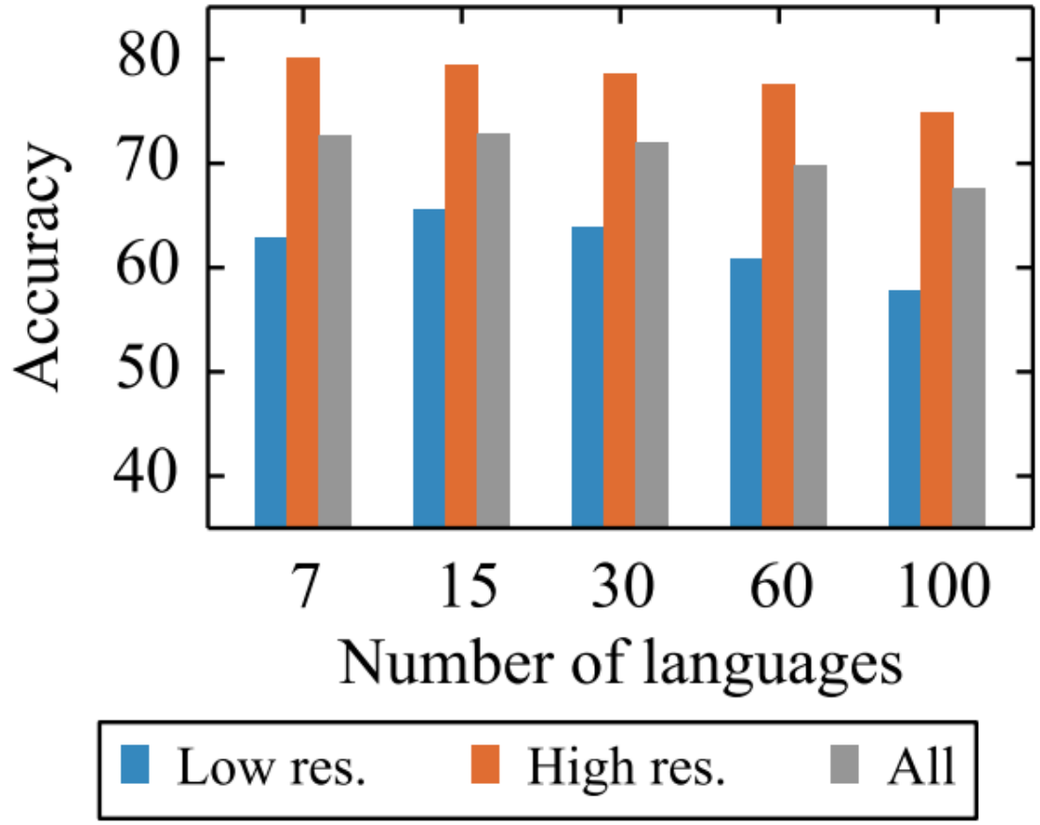
https://ai.google.com/research/tydiqa

Larger multilingual model = better QA performance

Model	TyDiQA-GoldP
Metrics	F1 / EM
Cross-lingu	
mBERT	59.7 / 43.9
XLM	43.6 / 29.1
InfoXLM	- / -
X-STILTs	76.0 / 59.5
XLM-R	65.1 / 45.0
VECO	67.6 / 49.1
RemBERT	77.0 / 63.0
mT5-Small	35.2 / 23.2
mT5-Base	57.2 / 41.2
mT5-Large	69.9 / 52.2
mT5-XL	75.9 / 59.4
mT5-XXL	**80.8 / 65.9**

What if a language is unseen or poorly represented during *pretraining*?

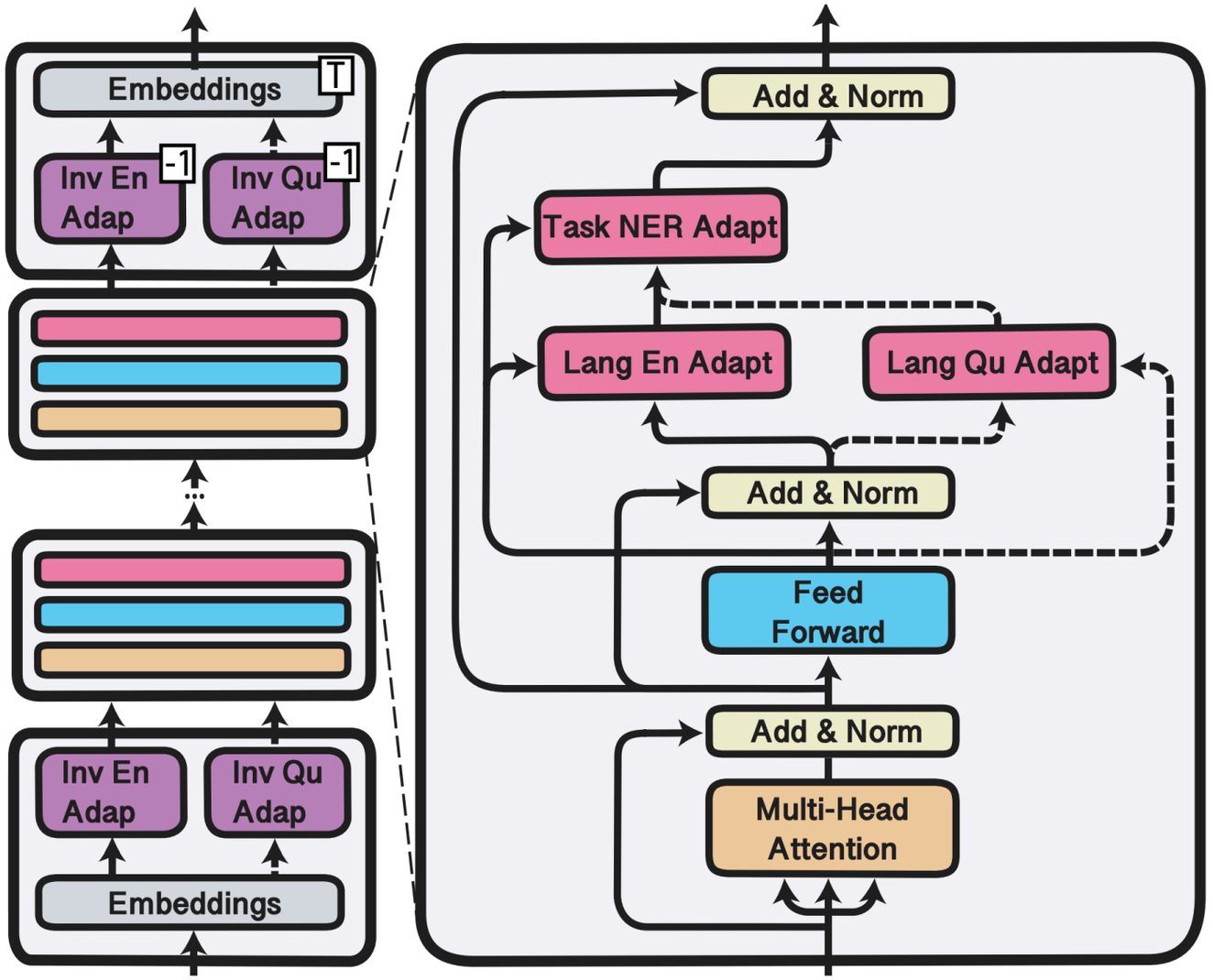
- The “curse of multilinguality” (Conneau et al., 2020): *For a fixed-size model, the per-language capacity decreases as we increase the number of languages…*



Target language adaptation

- If you only care about transferring to a specific target language B, then after normal pretraining on many languages, you can perform a second phase of fine-tuning on only unlabeled data from language B
- However, doing this might result in *catastrophic forgetting* of multilingual knowledge learned during the first stage of pretraining.

One solution: just train a small number of parameters in the second phase!



MAD-X paper, Pfeiffer et al., EMNLP 2020

This research is still in early stages, but it’s very exciting! Let’s move on to machine translation

Do we have enough parallel data?

Parallel Corpus	Sentences	Parallel Corpus	Sentences
Romanian-English	399,375	Greek-English	1,235,976
Bulgarian-English	406,934	Swedish-English	1,862,234
Slovene-English	623,490	Italian-English	1,909,115
Hungarian-English	624,934	German-English	1,920,209
Polish-English	632,565	Finnish-English	1,924,942
Lithuanian-English	635,146	Portuguese-English	1,960,407
Latvian-English	637,599	Spanish-English	1,965,734
Slovak-English	640,715	Danish-English	1,968,800
Czech-English	646,605	Dutch-English	1,997,775
Estonian-English	651,746	French-English	2,007,723

Europarl parallel data: http://www.statmt.org/europarl/

What if we don’t have parallel data?

https://arxiv.org/pdf/1804.07755.pdf

Phrase-Based & Neural Unsupervised Machine Translation

Guillaume Lample[†]
Facebook AI Research
Sorbonne Universités
glample@fb.com

Myle Ott
Facebook AI Research
myleott@fb.com

Alexis Conneau
Facebook AI Research
Université Le Mans
aconneau@fb.com

Ludovic Denoyer[†]
Sorbonne Universités
ludovic.denoyer@lip6.fr

Marc’Aurelio Ranzato
Facebook AI Research
ranzato@fb.com

https://arxiv.org/pdf/1711.00043.pdf

UNSUPERVISED MACHINE TRANSLATION
USING MONOLINGUAL CORPORA ONLY

Guillaume Lample † ‡ , **Alexis Conneau** † , **Ludovic Denoyer** ‡ , **Marc’Aurelio Ranzato** †
† Facebook AI Research,
‡ Sorbonne Universités, UPMC Univ Paris 06, LIP6 UMR 7606, CNRS
{gl,aconneau,ranzato}@fb.com,ludovic.denoyer@lip6.fr

https://arxiv.org/pdf/1901.07291.pdf

Cross-lingual Language Model Pretraining

Guillaume Lample[*]
Facebook AI Research
Sorbonne Universités
glample@fb.com

Alexis Conneau[*]
Facebook AI Research
Université Le Mans
aconneau@fb.com

https://arxiv.org/pdf/1710.11041.pdf

UNSUPERVISED NEURAL MACHINE TRANSLATION

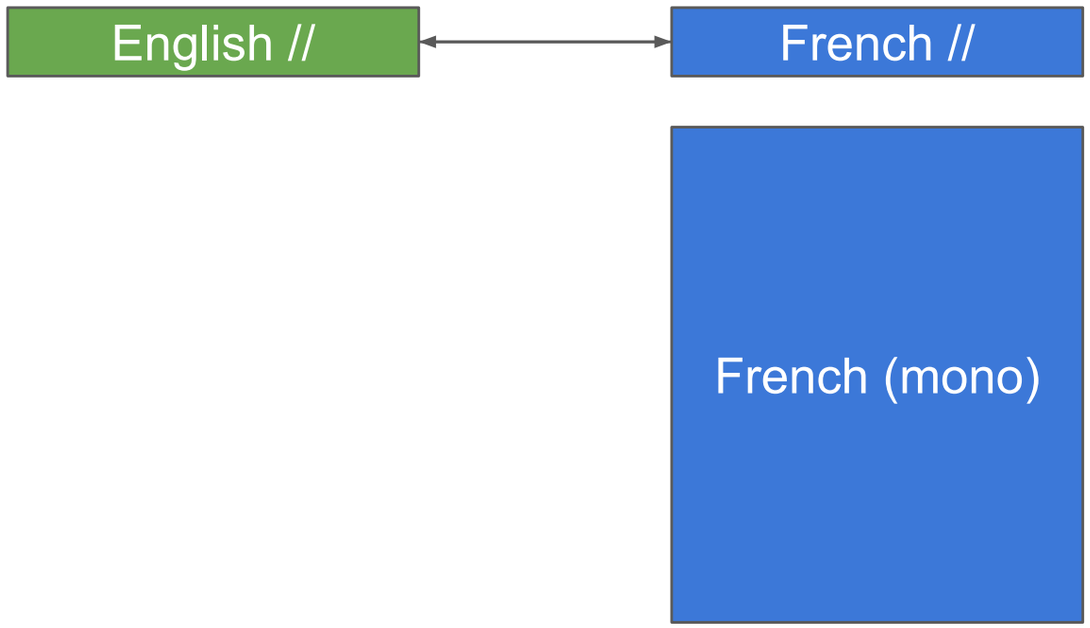
Mikel Artetxe, Gorka Labaka & Eneko Agirre
IXA NLP Group
University of the Basque Country (UPV/EHU)
{mikel.artetxe,gorka.labaka,e.agirre}@ehu.eus

Kyunghyun Cho
New York University
CIFAR Azrieli Global Scholar
kyunghyun.cho@nyu.edu

Back-translation (Sennrich et al. 2016)

Improving Neural Machine Translation Models with Monolingual Data

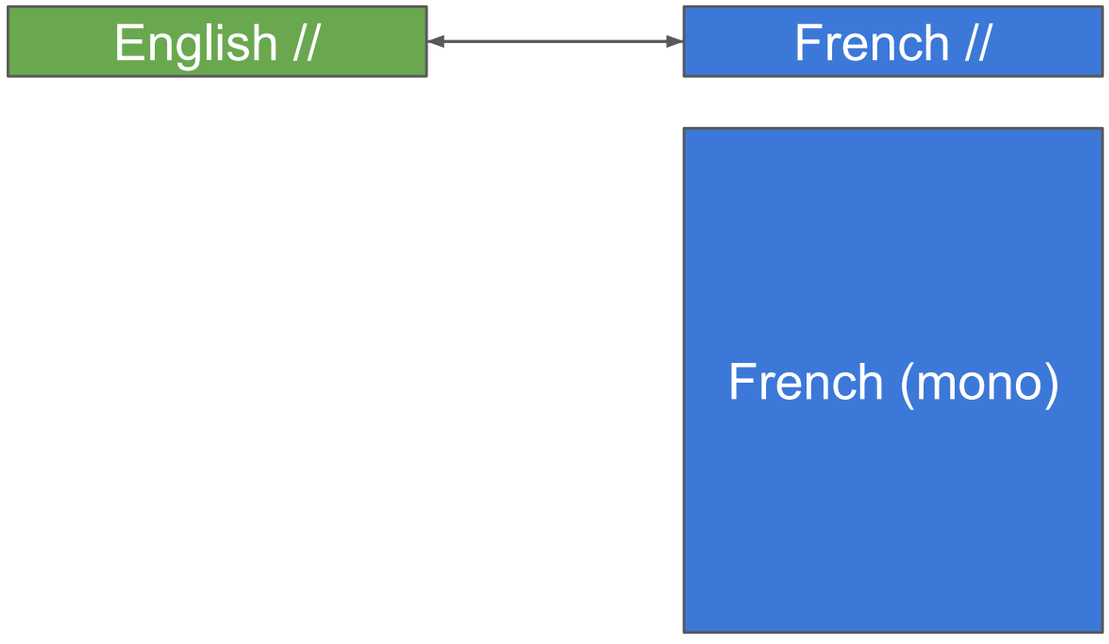
- Small parallel dataset
- Huge monolingual corpus in target language



Back-translation (Sennrich et al. 2016)

Improving Neural Machine Translation Models with Monolingual Data

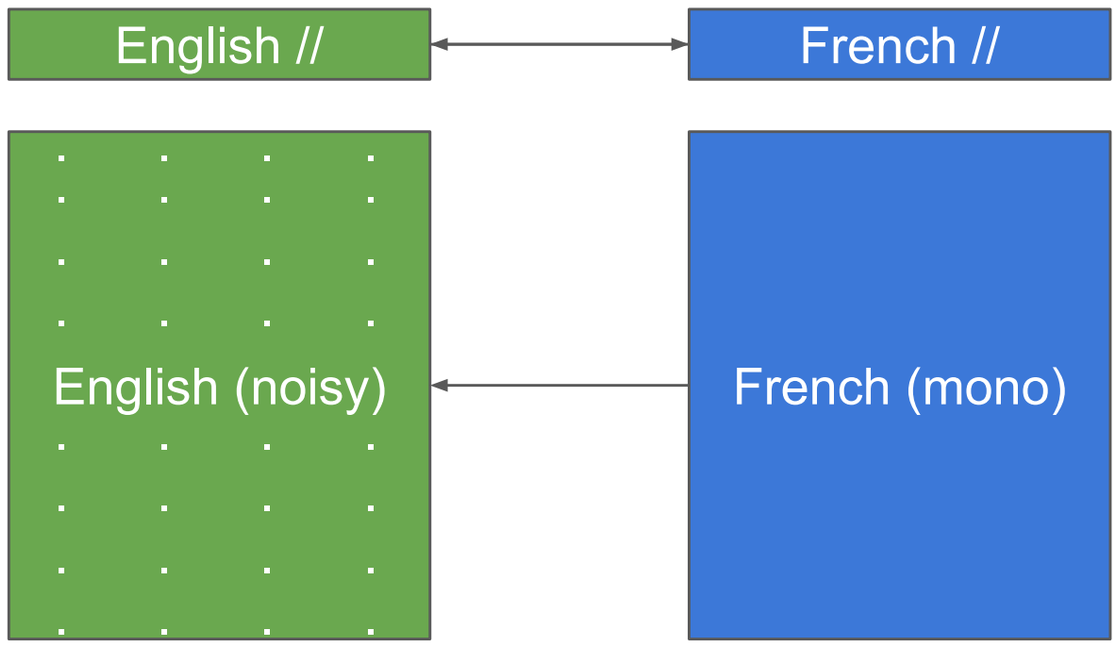
- Small parallel dataset
- Huge monolingual corpus in target language
- Train a (target → source) model $\mathbf{M}_{t2s}$



Back-translation (Sennrich et al. 2016)

Improving Neural Machine Translation Models with Monolingual Data

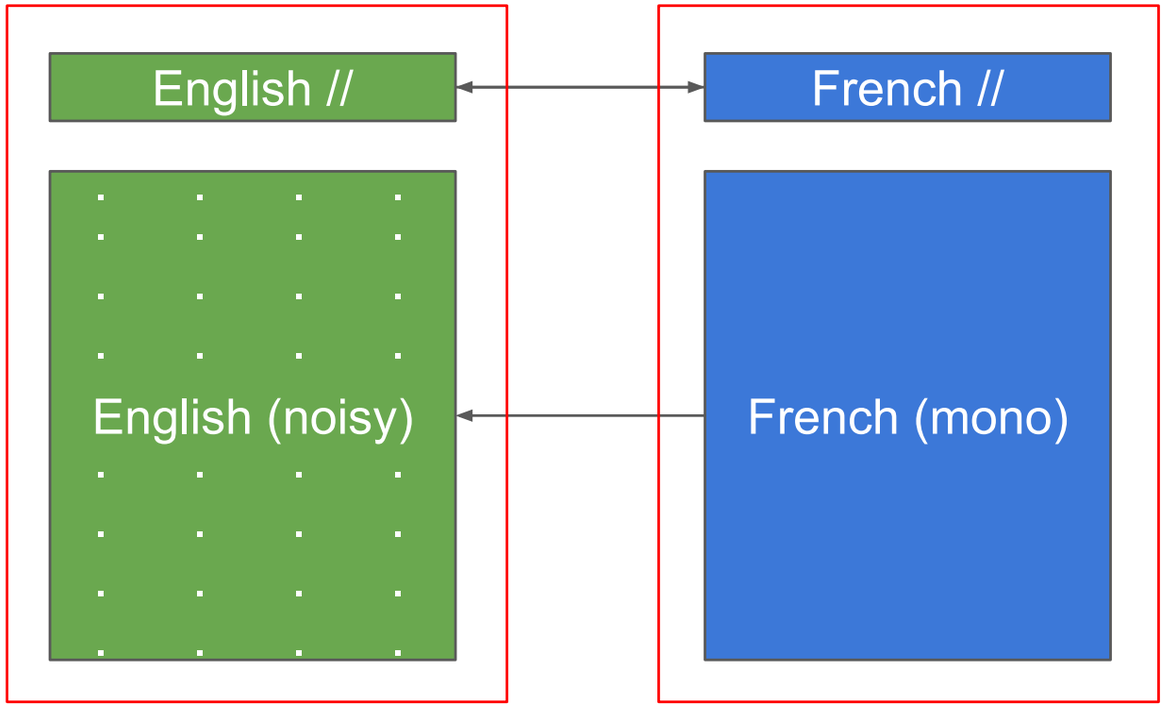
- Small parallel dataset
- Huge monolingual corpus in target language
- Train a (target → source) model $\mathbf{M}_{t2s}$
- Use $\mathbf{M}_{t2s}$ to translate target monolingual corpus



Back-translation (Sennrich et al. 2016)

Improving Neural Machine Translation Models with Monolingual Data

- Small parallel dataset
- Huge monolingual corpus in target language
- Train a (target → source) model $\mathbf{M}_{t2s}$
- Use $\mathbf{M}_{t2s}$ to translate target monolingual corpus
- Use the two parallel datasets to train $\mathbf{M}_{s2t}$



Back-translation (Sennrich et al. 2016)

Improving Neural Machine Translation Models with Monolingual Data

- en-->de WMT14
  - Parallel only: 20.4
  - + back-translation: 23.8
- en-->de WMT15
  - Parallel only: 23.6
  - + back-translation: 26.5

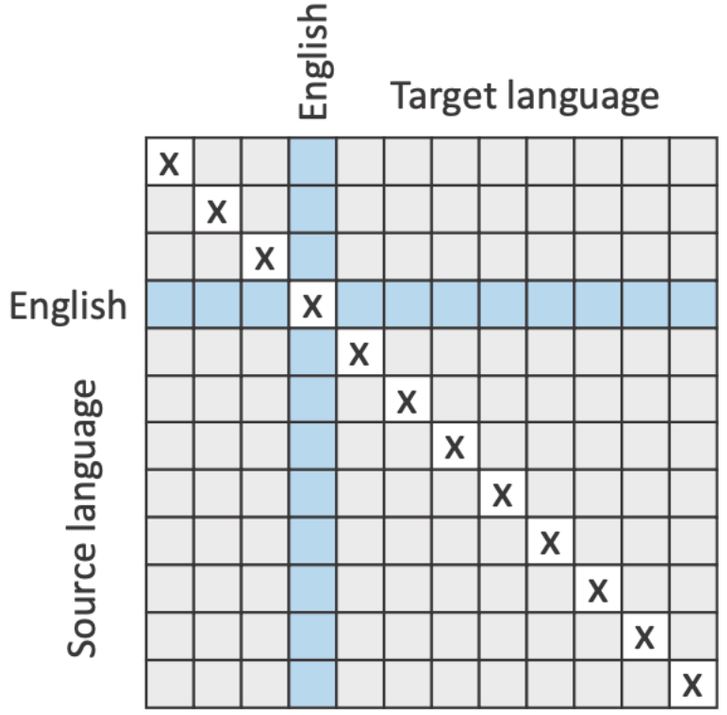
Back-translation (Sennrich et al. 2016)

Improving Neural Machine Translation Models with Monolingual Data

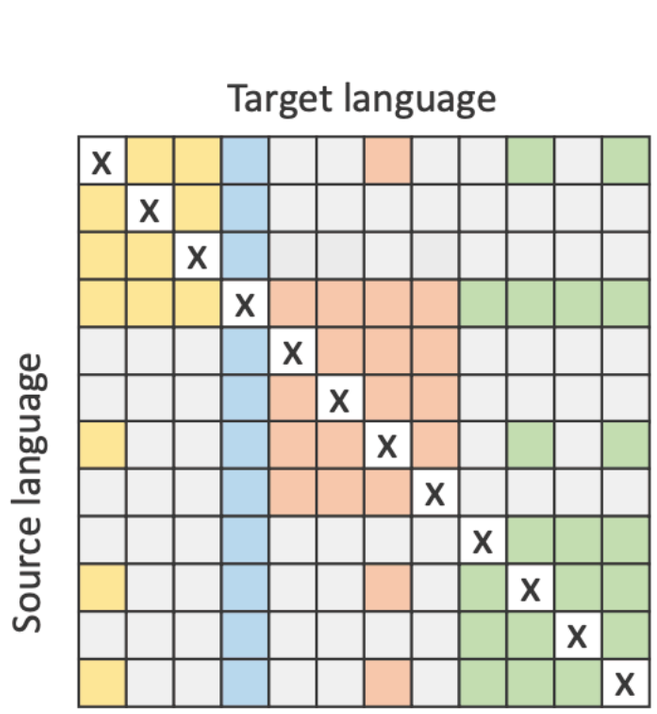
- Back-translation can be used for
  - Semi-supervised machine translation
  - Style transfer
  - Domain transfer
  - (small parallel, large unlabeled data)

Many-to-Many Translation

- A *single model* capable of translating between 100 languages (any of them can be source or target)



(a) English-Centric Multilingual



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

Beyond English-Centric Multilingual MT paper, Fan et al., arXiv 2021

Data preparation

- Selected 100 widely-spoken languages
  - from geographically diverse language families
  - have at least some parallel data available
  - also have larger monolingual data available
- Apply SentencePiece (subword tokenization) to all datasets
- In the end, 7.5 *billion* parallel sentences in 2,200 directions are collected
  - Backtranslation is also used to further augment the data

Results

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**

Table 4: **Comparison of Many-to-Many and English-Centric Systems.** Many-to-Many matches the performance of English-centric on evaluation directions involving English, but is significantly better on non English directions.

Bilingual baseline: trained only on a specific SRC>TGT direction

English-centric: only trained on ENG>X or X>ENG data, at test-time we feed in X>Y data

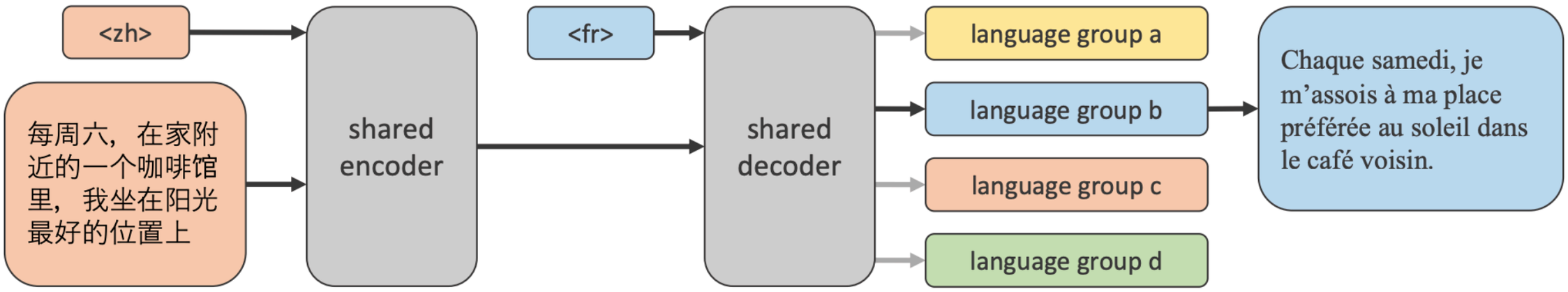
English-centric w/ pivot: only trained on ENG>X or X>ENG data, at test-time we do X>ENG and then ENG>Y

Zero-shot performance

Setting	w/ bitext	w/o bitext
En-Centric	5.4	7.6
En-Centric Piv.	9.8	12.4
M2M	**12.3**	**18.5**

Table 5: **Many-to-Many versus English-Centric on zero-shot directions.** We report performance on language pairs with and without bitext in the Many-to-Many training dataset.

Adding language-specific params can improve further



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

Multilingual instruction tuned LMs: Bactrian-X

LoRa: low rank adaptation

Regular Finetuning

1

Forward pass with original model

2

Obtain weight update via backpropagation

Weight update ΔW

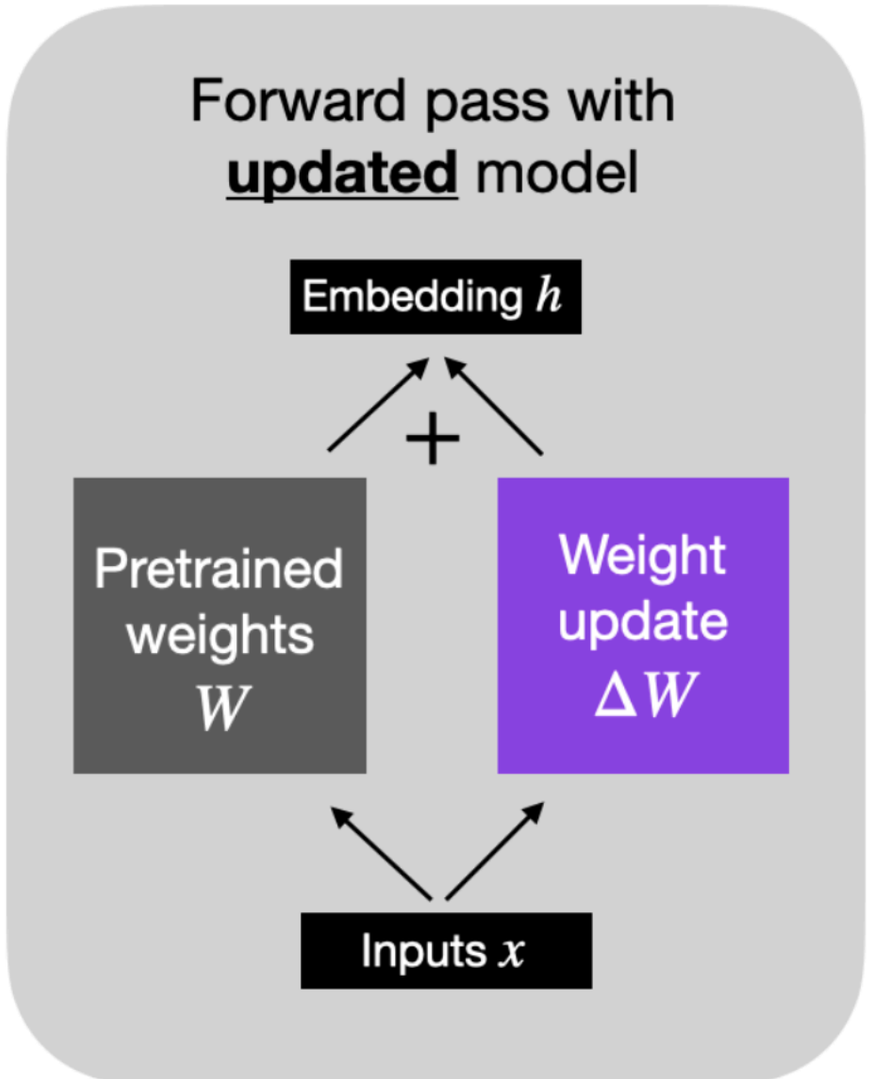
3

Forward pass with **updated** model

* The pretrained model could be any LLM, e.g., an encoder-style LLM (like BERT) or a generative decoder-style LLM (like GPT)

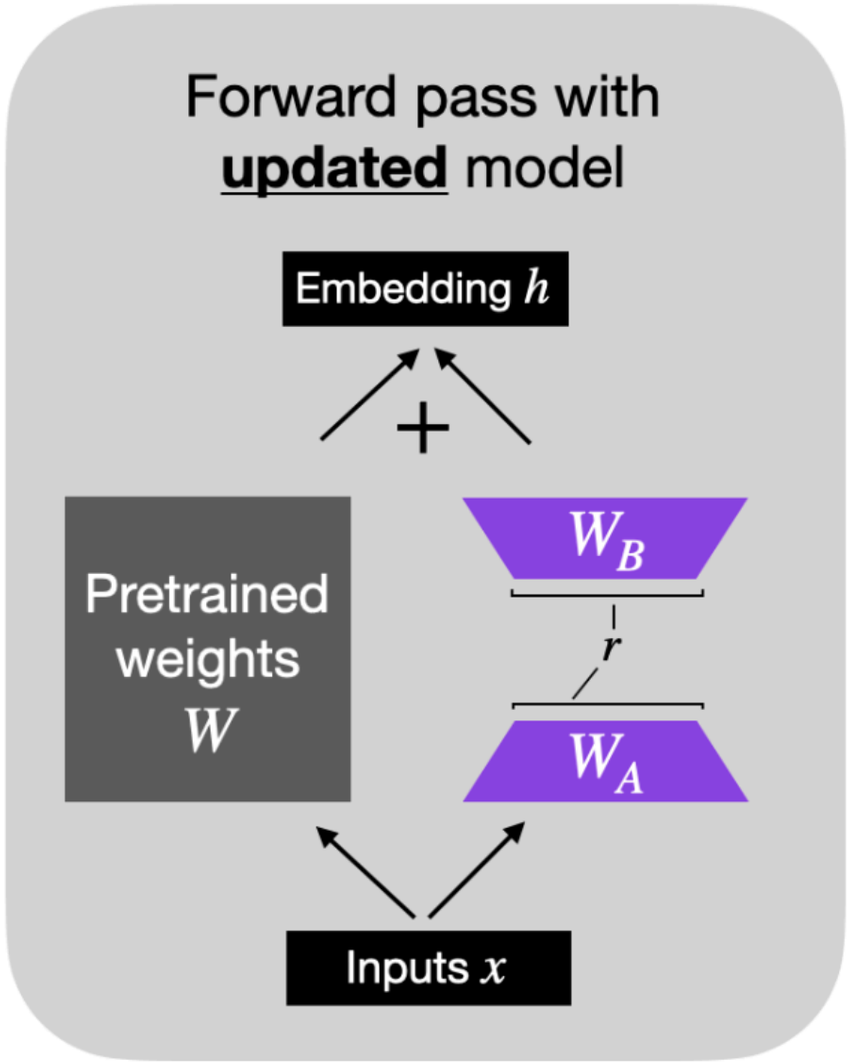
LoRa: low rank adaptation

Alternative formulation (regular finetuning)



LoRa: low rank adaptation

LoRA weights, W_A and W_B, represent ΔW



LoRa: low rank adaptation

LoRA can even outperform full finetuning training only 2% of the parameters

Model&Method	# Trainable Parameters	WikiSQL Acc. (%)	MNLI-m Acc. (%)	SAMSum R1/R2/RL
GPT-3 (FT)	175,255.8M	**73.8**	89.5	52.0/28.0/44.5
GPT-3 (BitFit)	14.2M	71.3	91.0	51.3/27.4/43.5
GPT-3 (PreEmbed)	3.2M	63.1	88.6	48.3/24.2/40.5
GPT-3 (PreLayer)	20.2M	70.1	89.5	50.8/27.3/43.5
GPT-3 ($\text{Adapter}^{\text{H}}$)	7.1M	71.9	89.8	53.0/28.9/44.8
GPT-3 ($\text{Adapter}^{\text{H}}$)	40.1M	73.2	**91.5**	53.2/29.0/45.1
GPT-3 (LoRA)	4.7M	73.4	**91.7**	**53.8/29.8/45.9**
GPT-3 (LoRA)	37.7M	**74.0**	**91.6**	53.4/29.2/45.1

Full finetuning

Only tune bias vectors

Prompt tuning

Prefix tuning

ROUGE scores