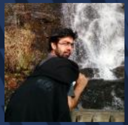
Multilingual Denoising Pre-training for Neural Machine Translation

Jiatao Gu
Facebook AI Research, NYC
July 10, 2020







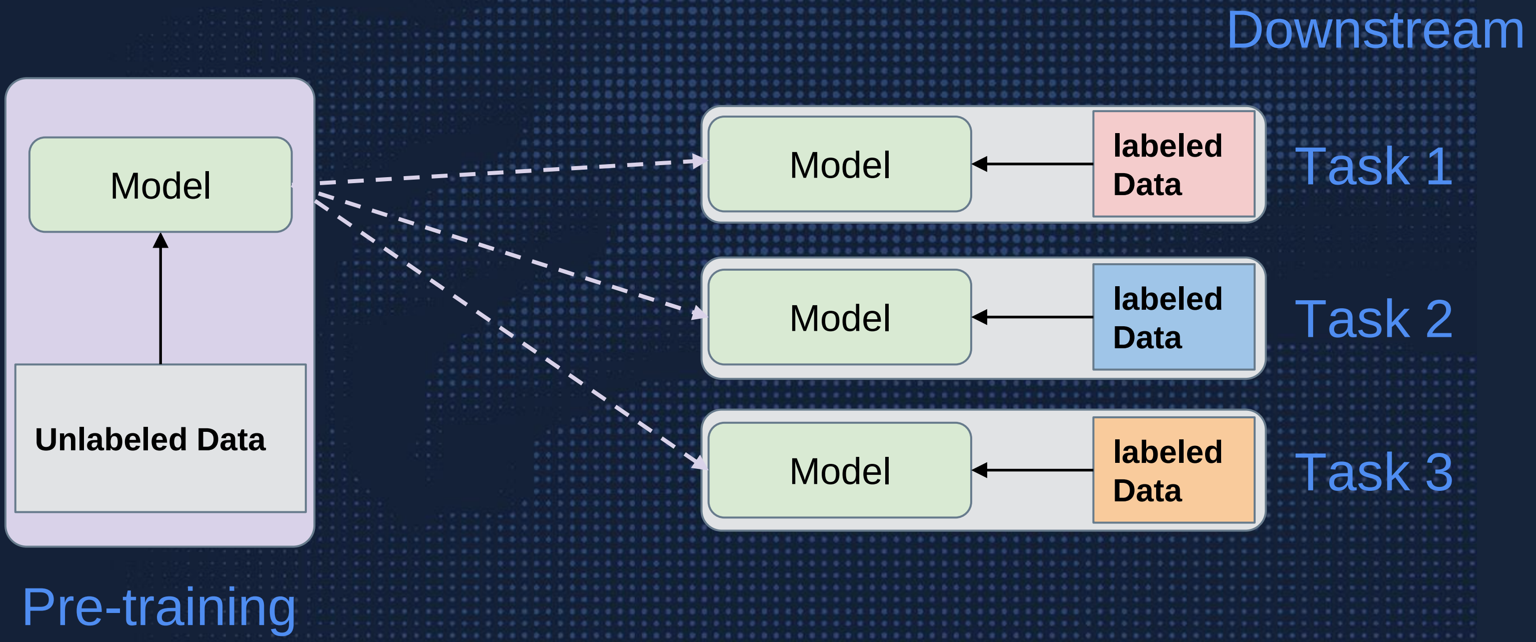

FACEBOOK AI

Low-Resource Machine Translation

- Sequence-to-sequence (Seq2Seq) Learning:
  - Modeled as Encoder-Decoder with Transformers.
- Low resource languages, low resource domains, documents, etc.

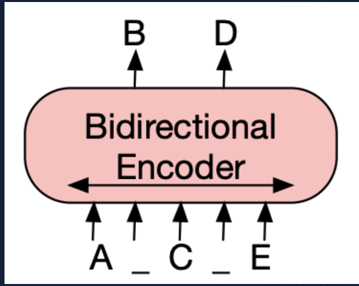
Pre-training for NLP

Recent advances (~2018) on self-supervised pre-training has changed the field of NLP applications dramatically.



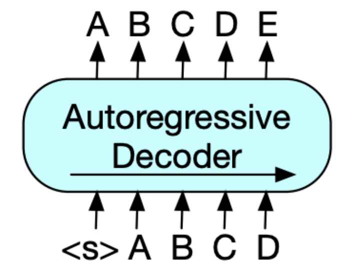
Pre-training for NLP

Encoder Pre-training



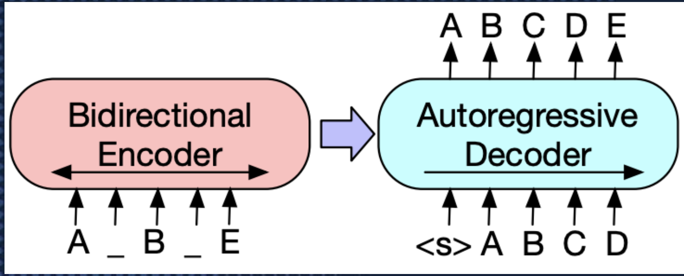
BERT, RoBERTa, Electra

Decoder Pre-training



GPT-1/2/3

Seq2Seq Pre-training



MASS, BART

Architecture is commonly a variant of Transformers.

Lewis et. al (2019)

BART

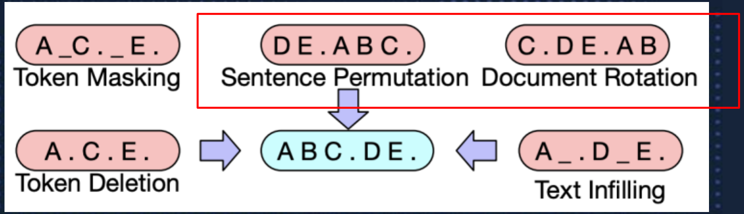
Seq2Seq Denoising Autoencoder

$$\mathbb{E}_{X \sim D}\left[\log P(X|g(X))\right], g(.) \text{ is the noise function}$$

What is BART?

- Bidirectional and Auto-Regressive Transformers
- Encoder-Decoder Pre-training, more specifically, Seq2Seq Decoding Autoencoder.
- DAE is not new for NLP, so what is different?
  - Large Scale Unlabeled Data
  - ~hundreds of million parameters
  - A set of noise functions
- Best for sequence generation tasks, e.g. Summarization.

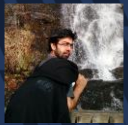
How about Machine Translation?



Lewis et. al (2019)

Multilingual Denoising Pre-training for Neural Machine Translation

Accepted by TACL2020

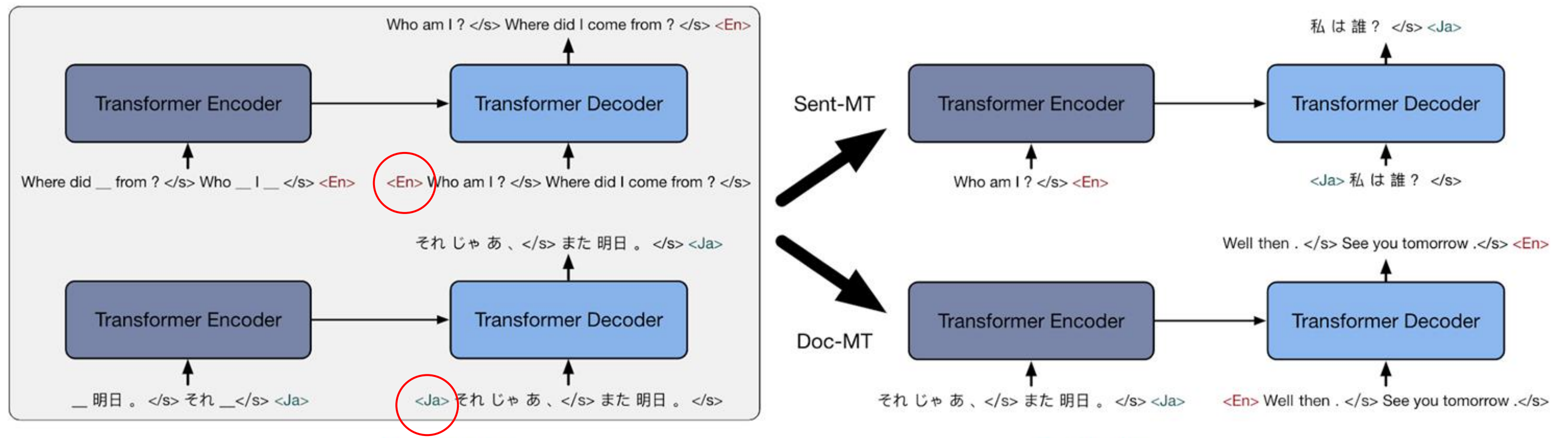






FACEBOOK AI

mBART

- We extend BART pre-training to multilingual -- mBART
  - Pre-train mBART on a multilingual unlabeled corpus (with additional language token);
  - Finetune mBART weights for machine translation.

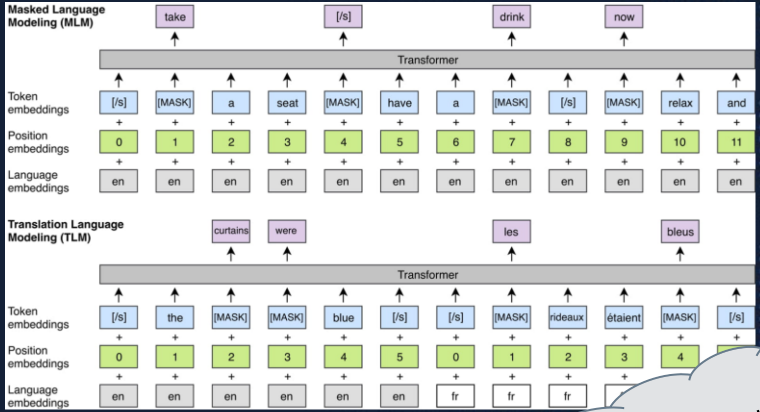


Multilingual Denoising Pre-Training (mBART) | Fine-tuning on Machine Translation

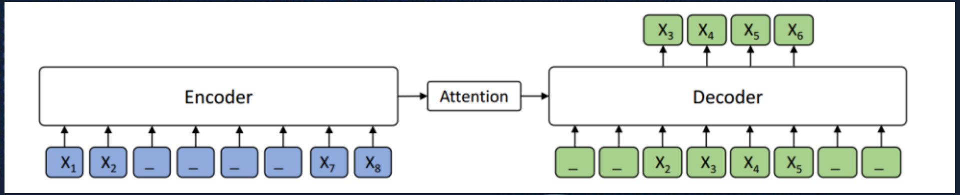
Liu et. al (2020)

mBART

- Closely Related Research:
  - XLM (Lample et.al, 2019) / XLM-R (Conneau et.al, 2019)
  - MASS (Song et.al, 2019)



Encoder-Pretraining

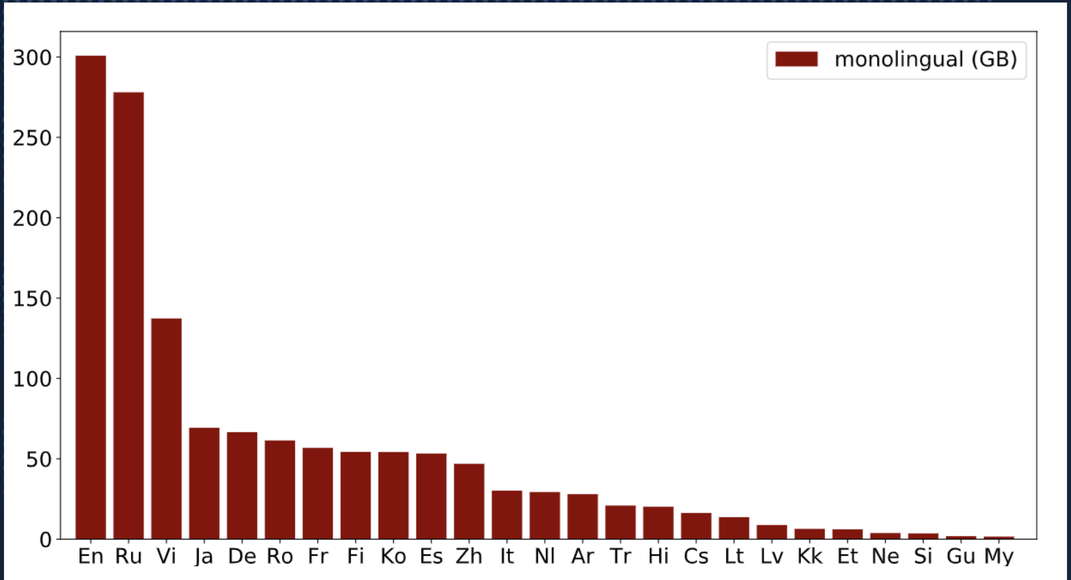


Not a fully seq2seq autoencoder

Data: CC25 Corpus

- Subset of the Common Crawl (CCNet) data on 25 languages;
- Large-scale & Document-level
- Language unbalanced
- Total size ~2TB

- Sentencepiece subwords used by XLM-RoBERTa, with a vocabulary size of **250,000** tokens.

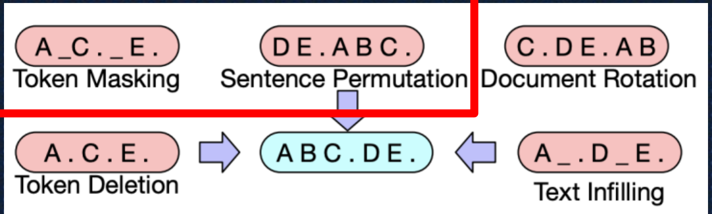


Model

- Transformers
  - **12** Layer Encoder + 12 layer Decoder, following the same architecture as the original BART model.
  - ~610M parameters
    - Bigger than traditional neural translation models (e.g. Transformer-base/big for most language pairs)
    - Much smaller than the recent biggest pre-training models such as T5, etc.

Learning

- Noise functions
  - Whole-word masking
  - Sentence Permutation



- Learning Details
  - To learn a full model on 25 languages (mBART25):
    - 256 V100 GPUs x 2.5 weeks (500K updates)
  - Deal with the language imbalance:
    - Temperature-based Resampling during

$$\lambda_i = \frac{1}{p_i} \cdot \frac{p_i^\alpha}{\sum_i p_i^\alpha},$$

Results

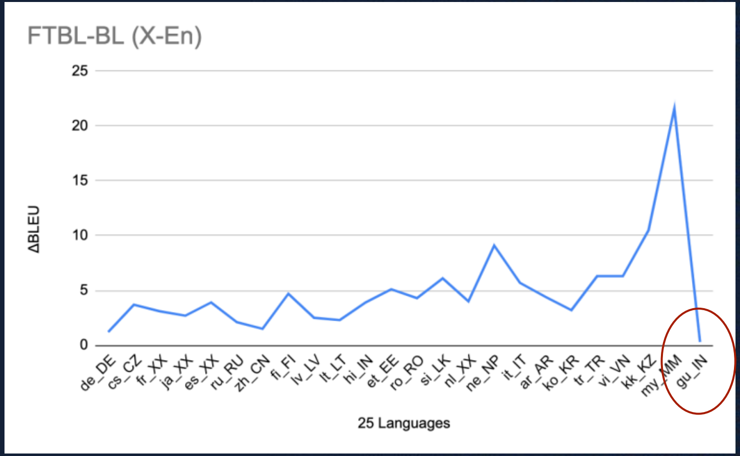
We fine-tune the pre-trained mBART model on three sets of experiments:

- Sentence-level Translation
- Document-level Translation
- Unsupervised Translation

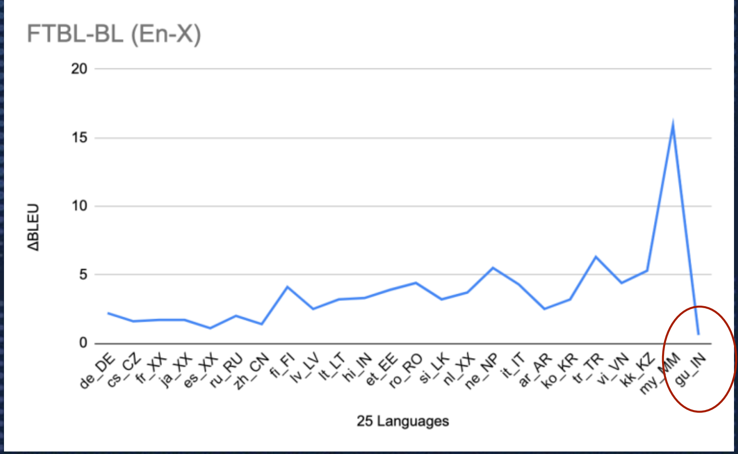
Supervised Translation

Sentence-level MT

- We collect the 24 pairs (X-En) of publicly available parallel corpus, so called ML25 benchmark.
- Improvements on BLEU score compared to best baselines.



High Resource → *Low Resource*



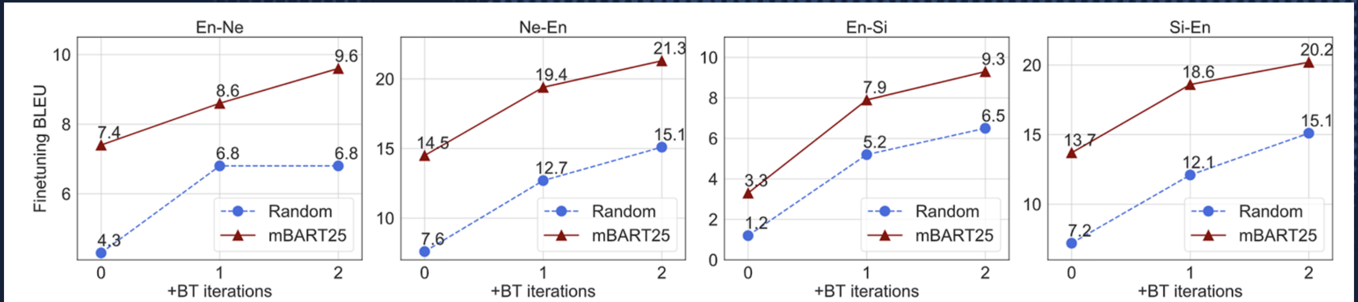
High Resource → *Low Resource*

FACEBOOK AI

***** We recently update a new version of training set, which appears to be slightly different from the arxiv version. *****

Sentence-level MT

- with Back-Translation (BT)



Pre-training and BT can be combined!

Sentence-level MT

- v.s. Other Pre-training Methods

Pre-training Model	Data	Fine-tuning En→Ro	Ro→En	+BT
Random	None	34.3	34.0	36.8
XLM (2019)	En Ro	-	35.6	38.5
MASS (2019)	En Ro	-	-	39.1
BART (2019)	En	-	-	38.0
XLM-R (2019)	CC100	35.6	35.8	-
BART-En	En	36.0	35.8	37.4
BART-Ro	Ro	37.6	36.8	38.1
mBART02	En Ro	**38.5**	**38.5**	**39.9**
mBART25	CC25	37.7	37.8	38.8

Bilingual mBART is better than multilingual here.

Sentence-level MT

- How many pre-training steps are needed?

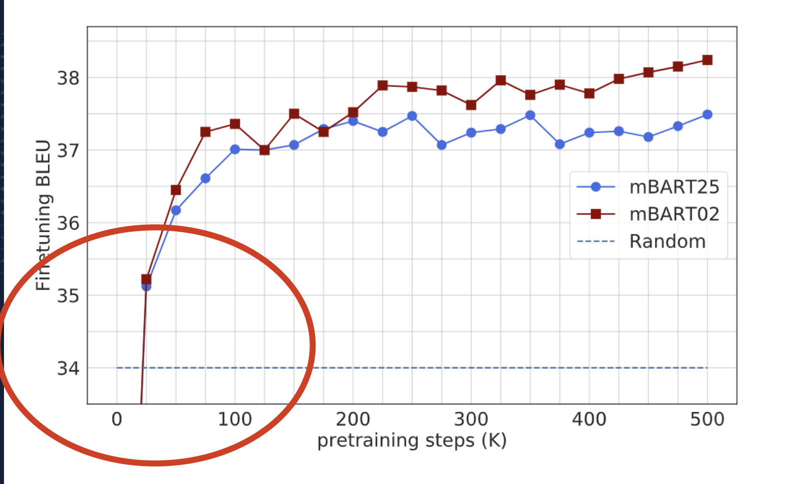


Figure 3: **Fine-tuning curves for Ro-En along with Pre-training steps**. Both mBART25 and mBART02 outperform the best baseline system after 25K steps.

Sentence-level MT

- Generalize to Unseen Languages

	Monolingual	Nl-En	En-Nl	Ar-En	En-Ar	Nl-De	De-Nl
Random	None	34.6 (-8.7)	29.3 (-5.5)	27.5 (-10.1)	16.9 (-4.7)	21.3 (-6.4)	20.9 (-5.2)
mBART02	En Ro	41.4 (-2.9)	34.5 (-0.3)	34.9 (-2.7)	21.2 (-0.4)	26.1 (-1.6)	25.4 (-0.7)
mBART06	En Ro Cs It Fr Es	43.1 (-0.2)	34.6 (-0.2)	37.3 (-0.3)	21.1 (-0.5)	26.4 (-1.3)	25.3 (-0.8)
mBART25	All	**43.3**	**34.8**	**37.6**	**21.6**	**27.7**	**26.1**

Table 7: **Generalization to Unseen Languages** Language transfer results, fine-tuning on language-pairs without pre-training on them. mBART25 uses all languages during pre-training, while other settings contain at least one unseen language pair. For each model, we also show the gap to mBART25 results.

Gap is relatively small as long as we have enough bi-text data for fine-tuning!

Document-level MT

Doc-level MT aims to translate with the document information.

- Doc-MT is relatively a "low-resource" problem:
  - e.g. Zh-En only has ～1.7K training documents.
- Training from scratch tend to produce much shorter translations.

SOURCE	作为一名艺术家， 联系对我来说是非常重要的 。 通过我的艺术作品我试着阐明 人类不是与自然分隔开 而是每一件事都是互相联系的。 大约10年前我第一次去了南极洲， 我也第一次看到了冰山。 我感到敬畏。 我的心快速地砰动，头晕目眩， 试着理解在我面前的这到底是什么。 在我身边的冰山 浮出水面几乎200英尺。 我只能感到很奇怪 这就是一片雪花 覆盖在另一片雪花， 年复一年形成的。 冰山的形成是 当它们从冰川断裂开 或者从冰架上断裂开。 每个冰山都有它们自己的独特个性。 它们 与其周边的环境 和其情况的互动具有一个鲜明的方式。 有些冰山拒绝妥协 坚持到底， 而另一些冰山就不能忍受 在一时剧烈激情喷涌下就水崩冰裂。 当你看到冰山，很容易就想到 它们都是孤立的， 它们是独立的，单独一体的， 更像是我们人类有时候对自身的看法。 但现实远不止这个。 随着冰山融化， 我呼吸到 它古老的气味。 随着冰山融化， 它释放了富有矿物质的鲜水 它们滋养了万物。 我着手拍摄这些冰山 好似我在拍摄我祖先的肖像， 了解到在这些个别的时刻 冰山原是以那样方式存在 但再也不会像那样存在了。 当它们融化时，这绝不是死亡；也绝不是一个终结， 而是 通往生生不息之路的一个延续。 我拍摄过的冰山，有些冰是非常年轻-- 几千年年龄。 有些冰 超过十万年。 我想给大家展示的最后图片 是我在格陵兰岛的 Kekertsuatsiak上拍摄的一个冰山。 这是一个非常难得的机会 大家实际上得以见证 一个冰山翻滚。 所以这就如图所示。 在左边你能看到一个小船。 这是一个约15英尺的船。 我想让你注意 冰山的形状 它在水面上的变形。 在这儿你看到它开始翻滚， 小船移动到另一边，一个男人站在那里。 这是一个平均尺寸的格陵兰冰山。 它浮出水面大约有120英尺高 或者40米高。 这视频是实时拍摄的。 就像这冰山， 它们展示给大家的是其个性的不同方面。 谢谢。
Random DOC-MT	As an artist, as an artist, as an artist, as an artist, as an artist, as an artist, as an artist, as an artist, as an artist, as an artist. I'm going to focus on the glacier and the glacier and the glacier and the glacier. There's a lot of ice in the ice in the ice, and there's a lot of ice in the ice, and there's a lot of ice in the ice, and there's a lot of ice in the ice in the ice. It's a ice that's ice that's melted from the ice of the ice that's melted from the ice of the ice that's melted from the ice of the ice that's melted from the ice of the ice that I've lost. There's a lot of ice that I'm going to show you some pictures that I'm going to show you. And you can see that it's moving to the top of it, and it's moving to the top of it.

- Existing approaches typically work at sentence-level, with document information as additional context(s).

Document-level MT

mBART pre-training enables to train document-level MT directly in seq2seq.

(a) Sentence- and Document-level BLEU scores on **En-De**

Model	Random		mBART25	
	s-BLEU	d-BLEU	s-BLEU	d-BLEU
Sent-MT	34.5	35.9	36.4	38.0
Doc-MT	×	7.7	**37.1**	**38.5**

(b) Document-level BLEU scores on **Zh-En**

Model	Random d-BLEU	mBART25 d-BLEU	HAN (2018) d-BLEU
Sent-MT	22.0	28.4	-
Doc-MT	3.2	**29.6**	24.0

SOURCE

作为一名艺术家， 联系对我来说是非常重要的 。 通过我的艺术作品我试着阐明 人类不是与自然分隔开 而是每一件事都是互相联系的。 大约10年前我第一次去了南极洲， 我也第一次看到了冰山。 我感到敬畏。 我的心快速地砰动，头晕目眩， 试着理解在我面前的这到底是什么。 在我身边的冰山 浮出水面几乎200英尺。 我只能感到很奇怪 这就是一片雪花 覆盖在另一片雪花， 年复一年形成的。 冰山的形成是 当它们从冰川断裂开 或者从冰架上断裂开。 每个冰山都有它们自己的独特个性。 它们 与其周边的环境 和其情况的互动具有一个鲜明的方式。 有些冰山拒绝妥协 坚持到底， 而另一些冰山就不能忍受 在一时剧烈激情喷涌下就水崩冰裂。 当你看到冰山，很容易就想到 它们都是孤立的， 它们是独立的，单独一体的， 更像是我们人类有时候对自身的看法。 但现实远不止这个。 随着冰山融化， 我呼吸到 它古老的气味。 随着冰山融化， 它释放了富有矿物质的鲜水 它们滋养了万物。 我着手拍摄这些冰山 好似我在拍摄我祖先的肖像， 了解到在这些个别的时刻 冰山原是以那样方式存在 但再也不会像那样存在了。 当它们融化时，这绝不是死亡； 也绝不是一个终结， 而是 通往生生不息之路的一个延续。 我拍摄过的冰山，有些冰是非常年轻-- 几千年年龄。 有些冰 超过十万年。 我想给大家展示的最后图片 是我在格陵兰岛的 Kekertsuatsiak上拍摄的一个冰山。 这是一个非常难得的机会 大家实际上得以见证 一个冰山翻滚。 所以这就如图所示。 在左边你能看到一个小船。 这是一个约15英尺的船。 我想让你注意 冰山的形状 它在水面上的变形。 在这儿你看到它开始翻滚， 小船移动到另一边，一个男人站在那里。 这是一个平均尺寸的格陵兰冰山。 它浮出水面大约有120英尺高 或者40米高。 这视频是实时拍摄的。 就像这冰山， 它们展示给大家的是其个性的不同方面。 谢谢。

Random DOC-MT

As an artist, as an artist, as an artist, as an artist, as an artist, as an artist, as an artist, as an artist, as an artist, as an artist. I'm going to focus on the glacier and the glacier and the glacier and the glacier. There's a lot of ice in the ice in the ice, and there's a lot of ice in the ice, and there's a lot of ice in the ice, and there's a lot of ice in the ice in the ice. It's a ice that's ice that's melted from the ice of the ice that's melted from the ice of the ice that's melted from the ice of the ice that's melted from the ice of the ice that I've lost. There's a lot of ice that I'm going to show you some pictures that I'm going to show you. And you can see that it's moving to the top of it, and it's moving to the top of it.

mBART25 DOC-MT

And as an artist, connection is very important to me. Through my artwork, I try to convey the idea that humans are not separated from nature, but that everything is connected to each other. When I first went to Antarctica about 10 years ago, I saw for the first time icebergs. And I felt awe. My heart was shaking, my head was shaking, trying to understand what was in front of me. The icebergs around me were floating almost 200 feet above the surface of the water, and I could only feel how strange it was that this was a snowflake covering another snowflake, forming over and over again and over and over again. And icebergs form when they break off from glaciers or when they break off from ice shelves. And each iceberg has its own unique personality. They interact in a very distinctive way with the environment around them and with the circumstances in which they're located. Some icebergs refuse to settle down, and some icebergs can't stand the heat of passion that pours down and breaks ice. And when you look at icebergs, it's easy to think that they're isolated, that they're independent, that they're individual, that they're more like the way we think about ourselves sometimes. But the reality is much more than that. As the icebergs melt, I breathe in the smell of its ancient past. As the icebergs melt, they release fresh water that is rich in minerals that feed everything. And I'm photographing these icebergs like I'm photographing my ancestors, and I'm learning that in these individual moments, icebergs used to exist in that way and will never be the same again. When they melt, it's not about death; it's not about the end, but it's about a continuation of a lifetime. And the icebergs I've photographed, some of them are very young -- thousands of years old. And some of them are more than 100,000 years old. And the last picture I want to show you is a iceberg that I photographed on Kekertsuatsiak in Greenland. And it's a very difficult opportunity for you to actually witness the rolling of a iceberg. So here it is. On the left you can see a little boat. It's a little boat about 15 feet long. And I want you to notice that the shape of the iceberg changes as it floats over the surface of the water. And here you see it start to roll, and the boat moves to the other side, and a man is standing there. And this is an average size Icelandic iceberg. And it floats about 120 feet above the surface of the water, or 40 meters. And this video was taken in real time. And like these icebergs, they show you different aspects of their personality. Thank you.

FACEBOOK AI

Unsupervised MT

Similar to prior research (XLM, MASS), we also use unsupervised translation as the testing benchmark.

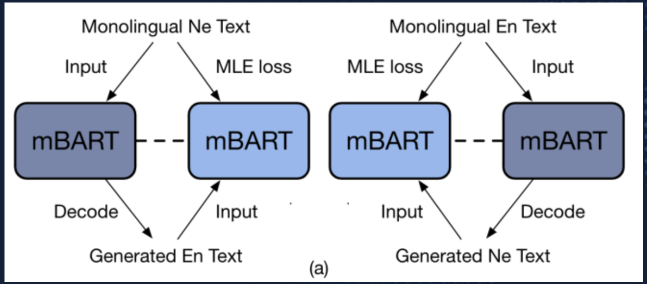
- The goal is to build a translation system for $X \rightarrow Y$ while we don't have direct parallel data between X and Y.
- In practise, unsupervised MT is more meaningful for "real low resource" and "distinct" languages.

In this part, we discuss two types

Kim, Yunsu, Miguel Graça, and Hermann Ney. "When and Why is Unsupervised Neural Machine Translation Useless?." arXiv preprint arXiv:2004.10581 (2020).

Unsupervised MT

Starting from mBART pretrained model, we generate BT data given X/Y monolingual data.



(a)

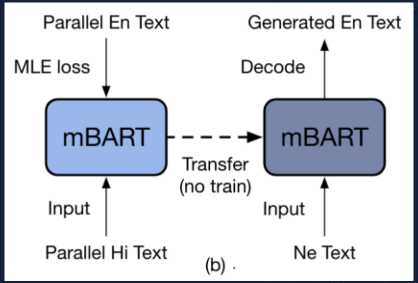
	En-De		En-Ne		En-Si	
	←	→	←	→	←	→
Random	21.0	17.2	0.0	0.0	0.0	0.0
XLM (2019)	34.3	26.4	0.5	0.1	0.1	0.1
MASS (2019)	**35.2**	28.3	-	-	-	-
mBART	34.0	**29.8**	**10.0**	**4.4**	**8.2**	**3.9**

Unsupervised MT

Specifically for "to En" direction, we perform language transfer from another pair.

*As a reference, without mBART pretraining, the transfer BLEU is almost **0** for all pairs.*

Fine-tune on X-En



Directly test on Y-En

Testing Languages	Domain	Zh News	Ja TED	Ko TED	Cs News	Ro News	Nl TED	It TED	Ar TED	Hi News	Ne Wiki	Si Wiki	Gu Wiki
		Fine-tuning Languages											
	Zh	23.7	8.8	**9.2**	2.8	7.8	7.0	6.8	6.2	7.2	4.2	5.9	0.0
	Ja	9.9	19.1	**12.2**	0.9	4.8	6.4	5.1	5.6	4.7	4.2	6.5	0.0
	Ko	5.8	**16.9**	24.6	5.7	8.5	9.5	9.1	8.7	9.6	8.8	11.1	0.0
	Cs	9.3	15.1	17.2	21.6	**19.5**	17.0	16.7	16.9	13.2	15.1	16.4	0.0
	Ro	16.2	18.7	17.9	**23.0**	37.8	22.3	21.6	22.6	16.4	18.5	22.1	0.0
	Nl	14.4	30.4	32.3	21.2	27.0	43.3	**34.1**	31.0	24.6	23.3	27.3	0.0
	It	16.9	25.8	27.8	17.1	23.4	30.2	39.8	**30.6**	20.1	18.5	23.2	0.0
	Ar	5.8	**15.5**	12.8	12.7	12.0	14.7	14.7	37.6	11.6	13.0	16.7	0.0
	Hi	3.2	10.1	9.9	5.8	6.7	6.1	5.0	7.6	23.5	**14.5**	13.0	0.0
	Ne	2.1	6.7	6.5	5.0	4.3	3.0	2.2	5.2	**17.9**	14.5	10.8	0.0
	Si	5.0	5.7	3.8	3.8	1.3	0.9	0.5	3.5	8.1	**8.9**	13.7	0.0
	Gu	8.2	8.5	4.7	5.4	3.5	2.1	0.0	6.2	**13.8**	13.5	12.8	0.3

Cross-lingual Retrieval for Iterative Self-Supervised Training

In submission

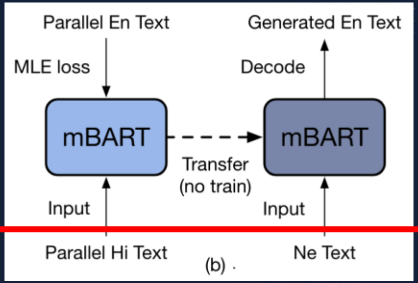


FACEBOOK AI

Emergent Cross-lingual Alignment

Why **language transfer** can work?

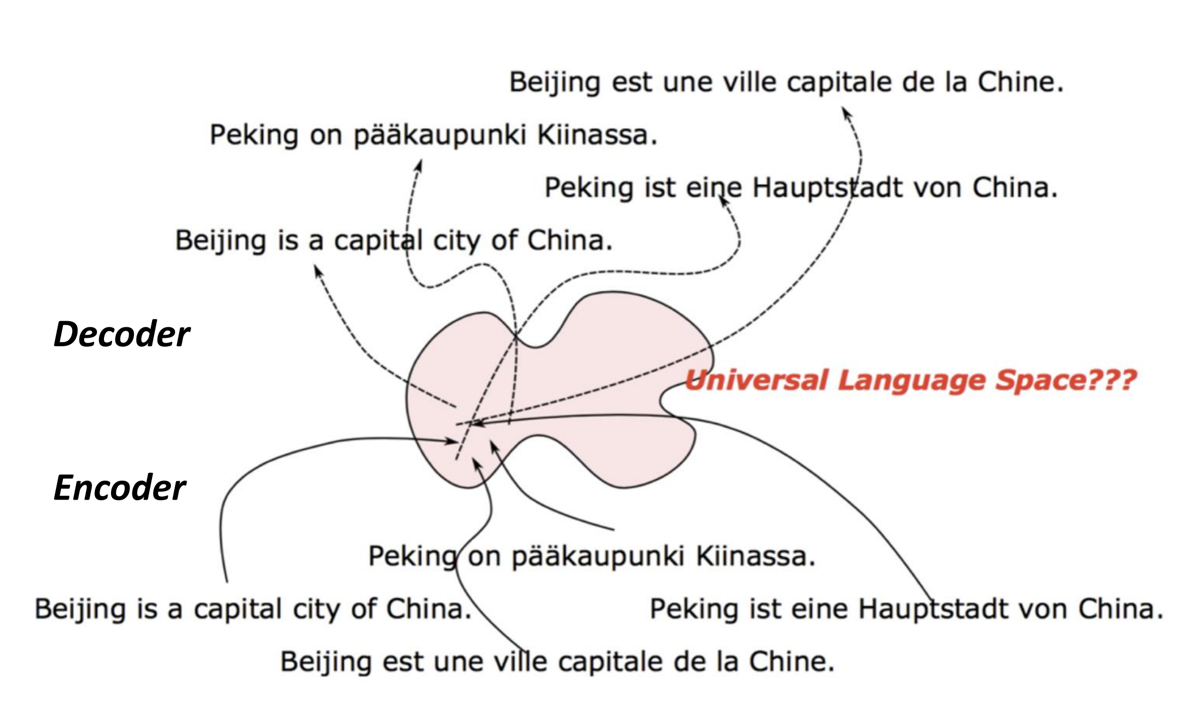
- ***Language agnostic representation*** emerged from the pretrained encoder?
- Bi-text has additional information useful for language transfer?



Similar Representations

Testing Languages	Domain	Fine-tuning Languages Zh News	Ja TED	Ko TED	Cs News	Ro News	Nl TED	It TED	Ar TED	Hi News	Ne Wiki	Si Wiki	Gu Wiki
	Zh	23.7	8.8	**9.2**	2.8	7.8	7.0	6.8	6.2	7.2	4.2	5.9	0.0
	Ja	9.9	19.1	**12.2**	0.9	4.8	6.4	5.1	5.6	4.7	4.2	6.5	0.0
	Ko	5.8	**16.9**	24.6	5.7	8.5	9.5	9.1	8.7	9.6	8.8	11.1	0.0
	Cs	9.3	15.1	17.2	21.6	**19.5**	17.0	16.7	16.9	13.2	15.1	16.4	0.0
	Ro	16.2	18.7	17.9	**23.0**	37.8	22.3	21.6	22.6	16.4	18.5	22.1	0.0
	Nl	14.4	30.4	32.3	21.2	27.0	43.3	**34.1**	31.0	24.6	23.3	27.3	0.0
	It	16.9	25.8	27.8	17.1	23.4	30.2	39.8	**30.6**	20.1	18.5	23.2	0.0
	Ar	5.8	**15.5**	12.8	12.7	12.0	14.7	14.7	37.6	11.6	13.0	16.7	0.0
	Hi	3.2	10.1	9.9	5.8	6.7	6.1	5.0	7.6	23.5	**14.5**	13.0	0.0
	Ne	2.1	6.7	6.5	5.0	4.3	3.0	2.2	5.2	**17.9**	14.5	10.8	0.0
	Si	5.0	5.7	3.8	3.8	1.3	0.9	0.5	3.5	8.1	**8.9**	13.7	0.0
	Gu	8.2	8.5	4.7	5.4	3.5	2.1	0.0	6.2	**13.8**	13.5	12.8	0.3

Emergent Cross-lingual Alignment



Dong et. al (2015)

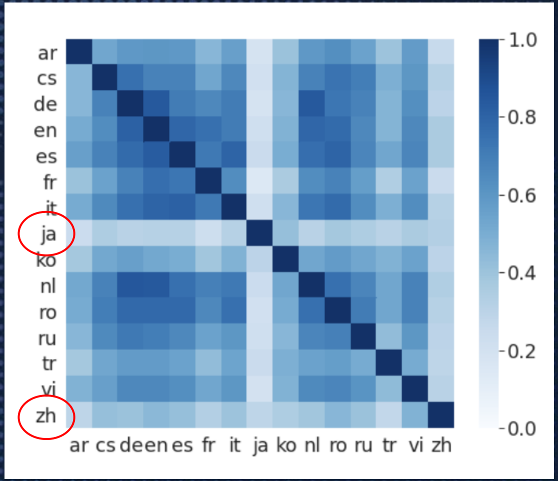
Emergent Cross-lingual Alignment

The pre-trained encoder tends to output similar representations across different languages without parallel supervision.

We verify our assumption based on a sentence retrieval task using TED58 corpus. For each sentence pair:

- Encode sentences with the pre-trained mBART encoder;
- Use the the pooled last layer hidden states to search the nearest neighbor in the target language;
- Report the Top-1 accuracy.

57% (on average) v.s. 0.04% (random)



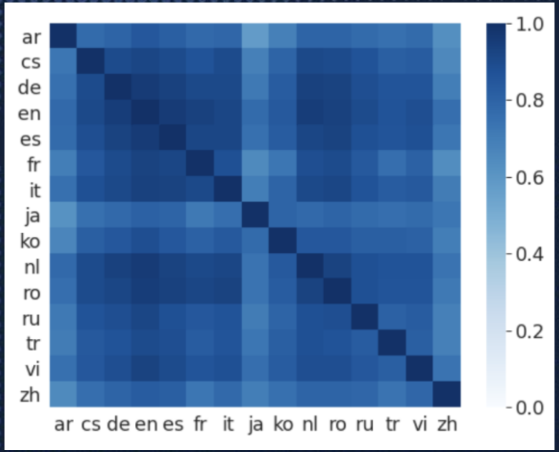
Emergent Cross-lingual Alignment

The alignment gets stronger after fine-tuning on any pair...

We fine-tune the pre-trained mBART on the bitext data of Ja-En of TED58, and REDO the retrieval task:

- The retrieval accuracy of all pairs gets improved significantly!
- **This directly explains why language transfer will work:**
    - ***When fine-tuning on bitext of any language, the model automatically learns to translate all languages because of the aligned representations.***

84% (after) v.s. 57% (before)

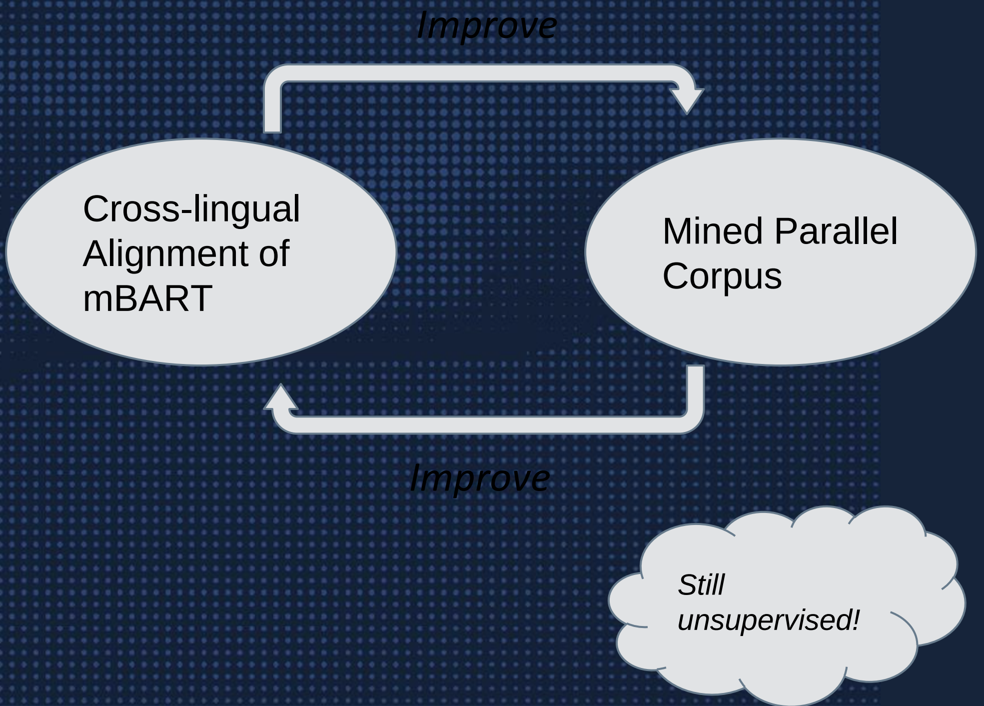


Emergent Cross-lingual Alignment

Inspired from the previous findings….

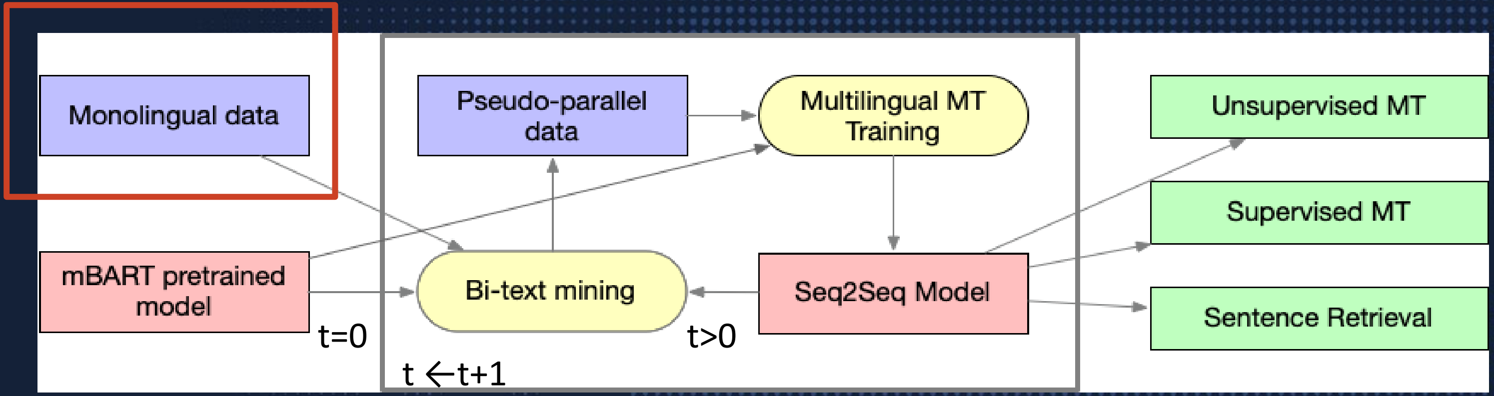
We hypothesize such cross-lingual alignment can be **self-improved** without using real parallel data.

Instead, we replace it with pseudo parallel data mined by the model itself based on sentence retrieval.



CRISS (Cross-lingual Retrieval for Iterative Self-Supervised Training)

The same datasets for training mBART, to speed-up retrieval, we subsample 100M for each language.



CRISS

Mining Stage:

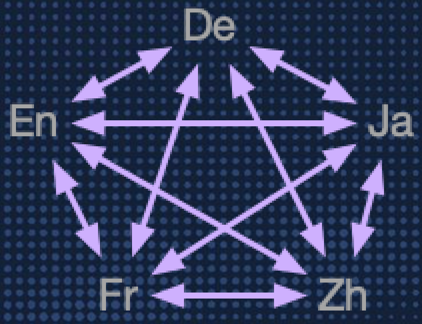
- Apply a score function based on K nearest neighbors (defined by cosine distance) to score and rank pairs;
- Keep pairs with scores larger than certain threshold to create the pseudo corpus;

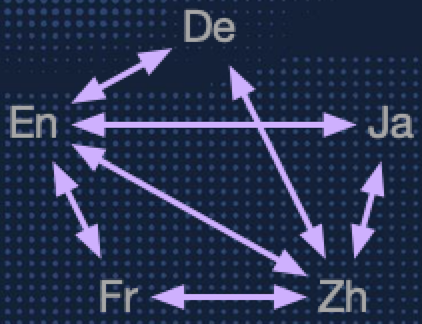
$$\mathrm{score}(x, y) = \frac{\cos(x, y)}{\sum_{z \in N_x} \frac{\cos(x,z)}{2k} + \sum_{z \in N_y} \frac{\cos(z,y)}{2k}}$$

CRISS

Training Stage:

- Merge all pseudo parallel datasets and training over the pre-trained mBART model in multilingual settings.
- Ideally, for N languages (e.g. N=25 for mBART25), we need to mine *(N-1)^2* directions to train.
- In practise, we find that a small number of pivot languages (by default, *English, Spanish, Hindi, Chinese*) are enough to achieve good performance.





CRISS

Another Intuition why CRISS might work….

- mBART helps learn good representations for ***Self-attention*** in both the encoder and decoder side;
- However, because of the nature of auto-encoder, the ***encoder-decoder attention*** is completely not useful in machine translation downstream tasks.
- In contrast, CRISS directly works in a cross-lingual setting, which naturally enables encoder-decoder attention.

Concurrent work which used a similar retrieve and sequence-to-sequence learning:
Lewis, Mike, et al. "Pre-training via Paraphrasing." arXiv preprint arXiv:2006.15020 (2020).

Results

We verify the proposed CRISS compared with mBART on three sets of experiments:

- Unsupervised Translation
- Sentence Retrieval
- Supervised Translation

We directly evaluate on the CRISS model;

We use CRISS model to initialize and fine-tune on supervised datasets.

Unsupervised MT

Comparison with existing methods:

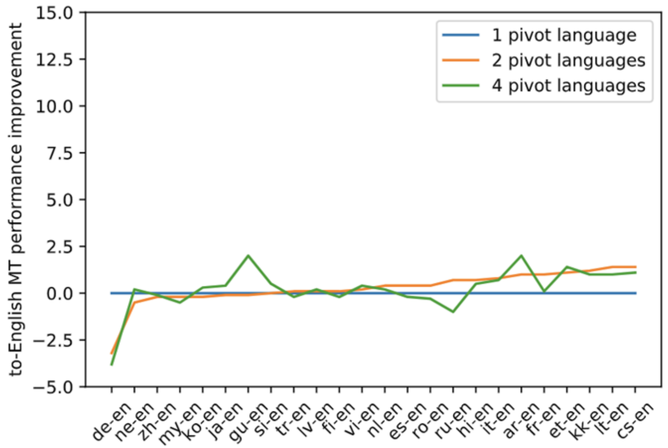
- Unlike mBART and other pre-training methods, CRISS itself is an unsupervised translation system, and do not need additional training steps (e.g. online BT in XLM/MASS/mBART)

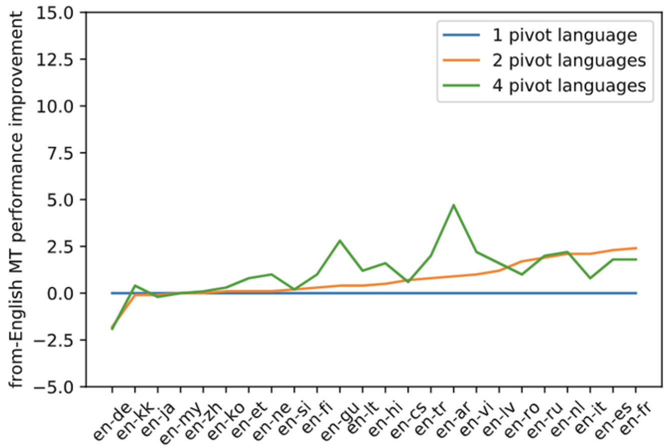
Direction	en-de	de-en	en-fr	fr-en	en-ne	ne-en	en-ro	ro-en	en-si	si-en
CMLM [30]	27.9	35.5	34.9	34.8	-	-	34.7	33.6	-	-
XLM [6]	27.0	34.3	33.4	33.0	0.1	0.5	33.3	31.8	0.1	0.1
MASS [33]	28.3	35.2	37.5	34.9	-	-	35.2	33.1	-	-
D2GPO [19]	28.4	35.6	37.9	34.9	-	-	**36.3**	33.4	-	-
mBART [20]	29.8	34	-	-	4.4	10.0	35.0	30.5	3.9	8.2
CRISS Iter 1	21.6	28.0	27.0	29.0	2.6	6.7	24.9	27.9	1.9	6
CRISS Iter 2	30.8	36.6	37.3	36.2	4.2	12.0	34.1	36.5	5.2	12.9
CRISS Iter 3	**32.1**	**37.1**	**38.3**	**36.3**	**5.5**	**14.4**	35.1	**37.6**	**6.0**	**13.6**

Unsupervised MT

How many pivot languages do we need?

- We compare the unsupervised translation results with 1 (En), 2 (En, Es) and 4 (En, Es, Hi, Zh) pivot languages:





FACEBOOK AI

Sentence Retrieval

We apply CRISS on Tatoeba sentence retrieval task:

- We use the pooled Encoder's hidden states to represent sentences as we did for TED58.

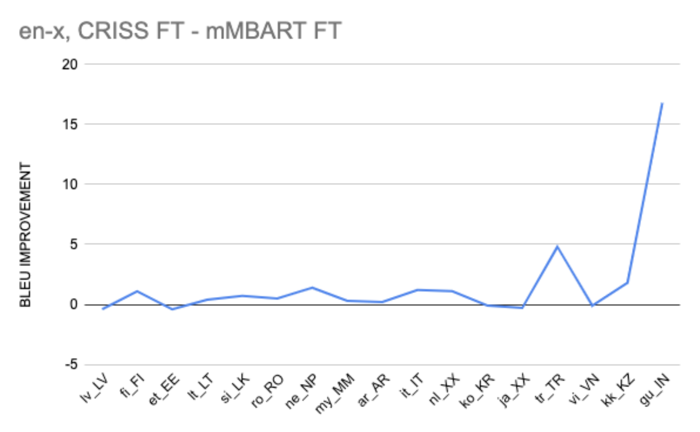
Lang	ar	de	es	et	fi	fr	hi	it
XLMR [5]	47.5	88.8	75.7	52.2	71.6	73.7	72.2	68.3
mBART [20]	39	86.8	70.4	52.7	63.5	70.4	44	68.6
CRISS Iter 1	72	97.5	92.9	85.6	88.9	89.1	86.8	88.7
CRISS Iter 2	76.4	**98.4**	95.4	**90**	92.2	91.8	91.3	91.9
CRISS Iter 3	**78.0**	98.0	**96.3**	89.7	**92.6**	**92.7**	**92.2**	**92.5**
LASER [2]	92.2	99	97.9	96.6	96.3	95.7	95.2	95.2

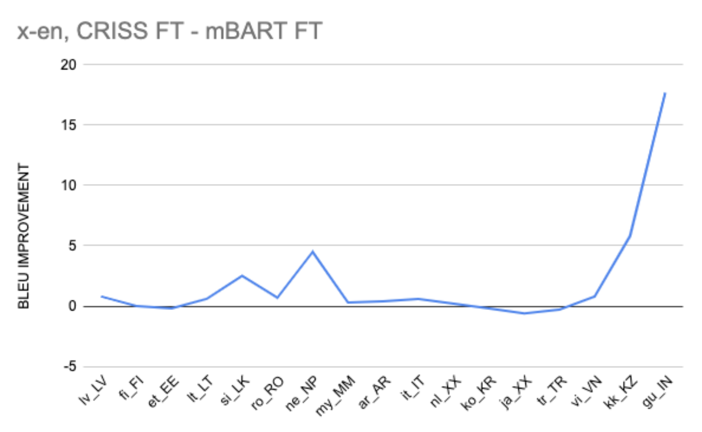
Lang	ja	kk	ko	nl	ru	tr	vi	zh
XLMR [5]	60.6	48.5	61.4	80.8	74.1	65.7	74.7	68.3 (71.6)
mBART [20]	24.9	35.1	42.1	80	68.4	51.2	63.9	14.8
CRISS Iter 1	76.8	67.7	77.4	91.5	89.9	86.9	89.9	69
CRISS Iter 2	**84.8**	74.6	**81.6**	92.8	**90.9**	92	92.5	81
CRISS Iter 3	84.6	**77.9**	81.0	**93.4**	90.3	**92.9**	**92.8**	**85.6**
LASER [2]	94.6	17.39	88.5	95.7	94.1	97.4	97	95

LASER is a supervised approach listed for reference.

Supervised MT

Similar to mBART, we apply CRISS as the initialization on the same benchmark of 25 languages.

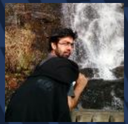




Multilingual Neural Machine Translation with Multilingual Denoising Pretraining

Arxiv and In submission



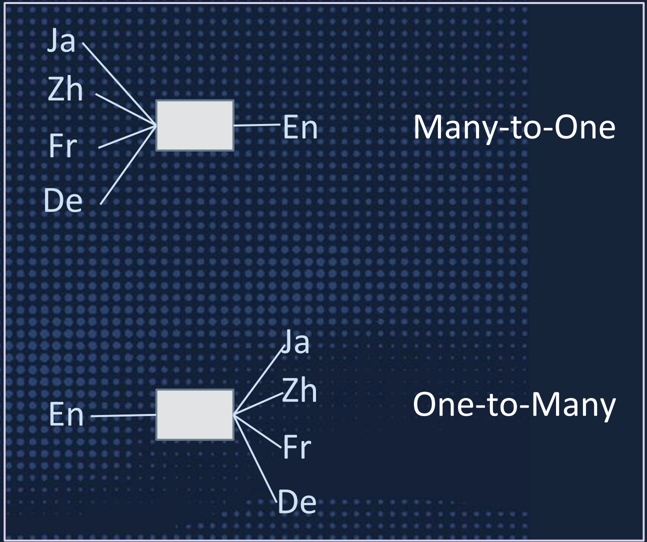
FACEBOOK AI

Low-Resource Machine Translation

- Sequence-to-sequence (Seq2Seq) Learning:
  - Modeled as Encoder-Decoder with Transformers.
- Low resource language translation, Document-level translation, etc.

Multilingual NMT (mNMT)

- Different from mBART, multilingual translation is supervised based on multilingual parallel corpus.
- Typically, we only have English as the common language, resulting in three types of mNMT: *many to one, one to many, and many to many*.
- mNMT can leverage high-resource language data to improve low-resource translation.



Many-to-Many

mBART + mNMT

- When we have both monolingual and multilingual resources, we can first pre-train mBART, and perform multilingual fine-tuning on the trained model.
- Temperature sampling is also applied (the same as pre-training).

Following the previous evaluation, we report translation performance on the same test sets.

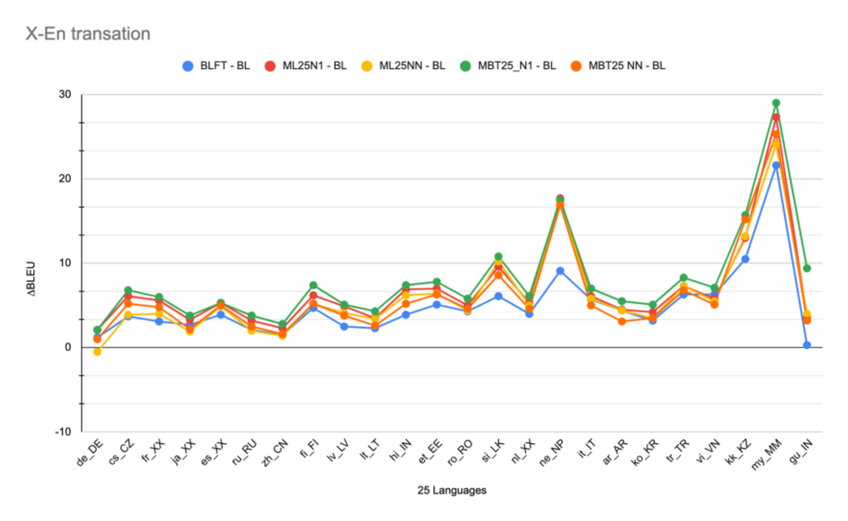
Sentence-level Translation

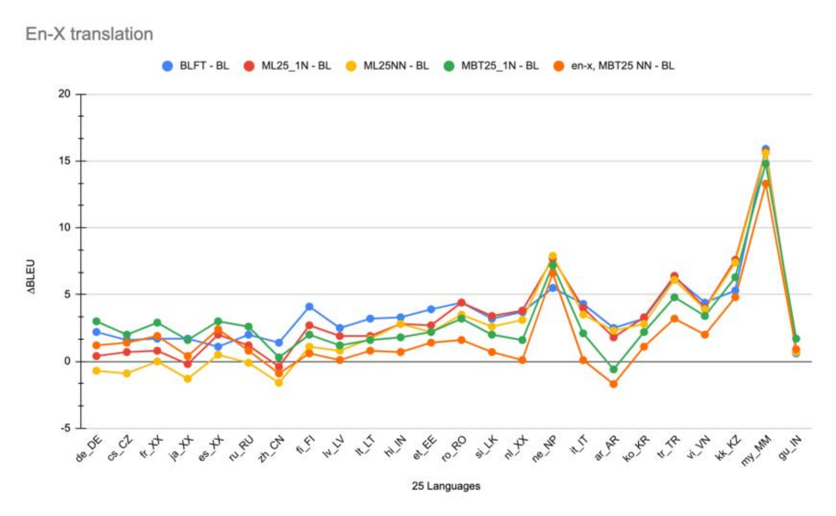
Another option is to jointly train system with both monolingual and parallel data:

Siddhant, Aditya, et al. "Leveraging Monolingual Data with Self-Supervision for Multilingual Neural Machine Translation." *arXiv preprint arXiv:2005.04816* (2020).

Results on 25 Languages:

- Overall Results





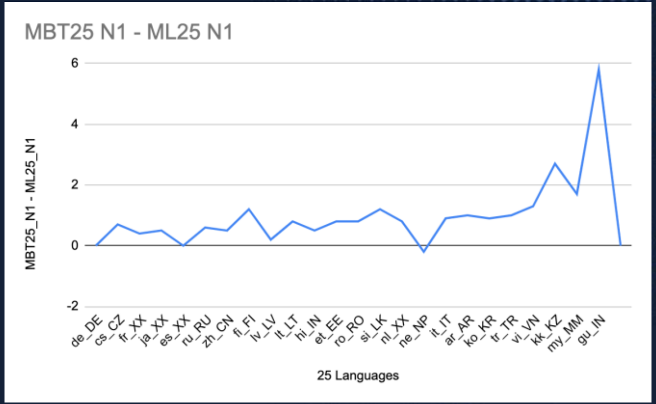
High Resource → *Low Resource*

High Resource → *Low Resource*

Results on 25 Languages:

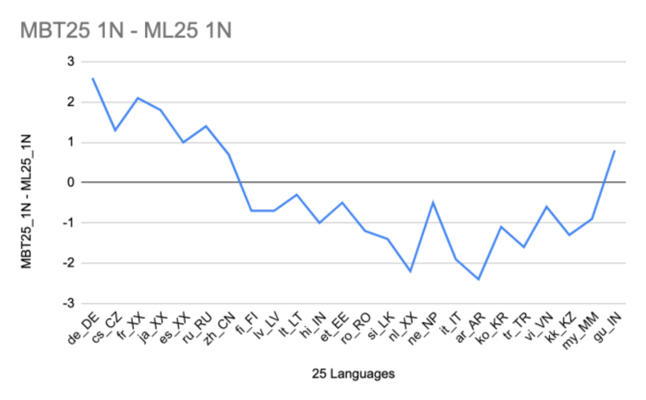
- mNMT with/without mBART Pre-training

X-En



High Resource → *Low Resource*

En-X

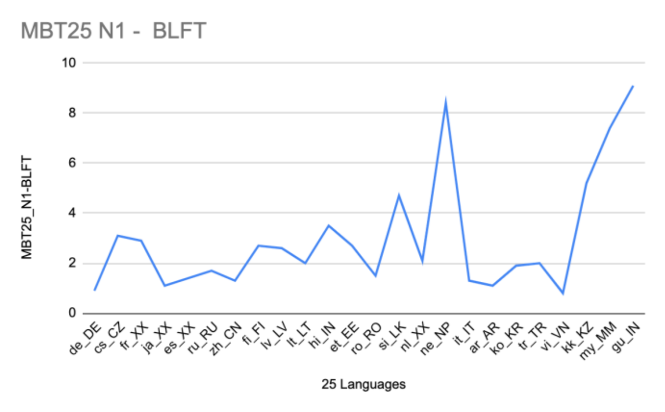


High Resource → *Low Resource*

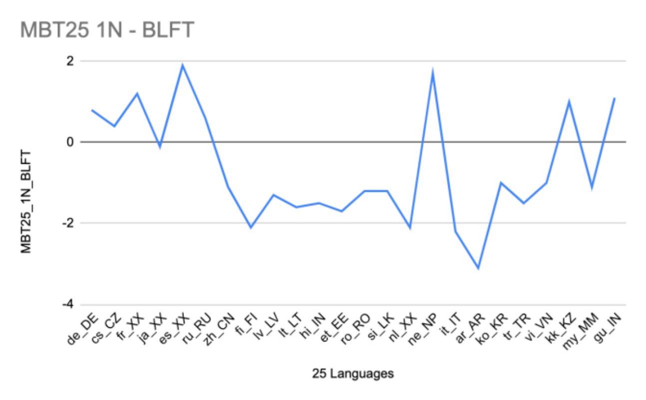
FACEBOOK AI

Results on 25 Languages:

- mBART Pre-training + bilingual or multilingual fine-tuning



High Resource → *Low Resource*



High Resource → *Low Resource*

Extending to 50 Languages

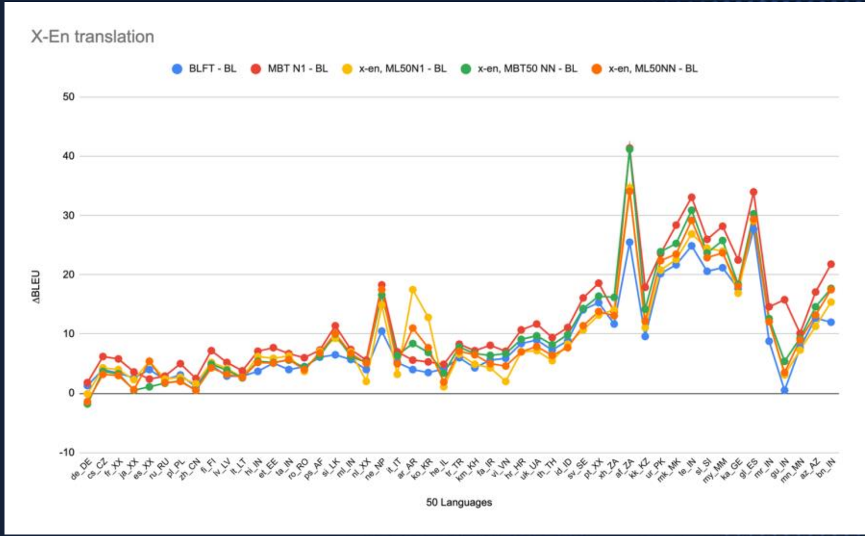
- Up to now, all our discussions are restricted in 25 languages as proposed in the original mBART.
- We gather additional 25 languages, both for monolingual (CC corpus) and parallel (TED talks, WAT, etc) datasets.

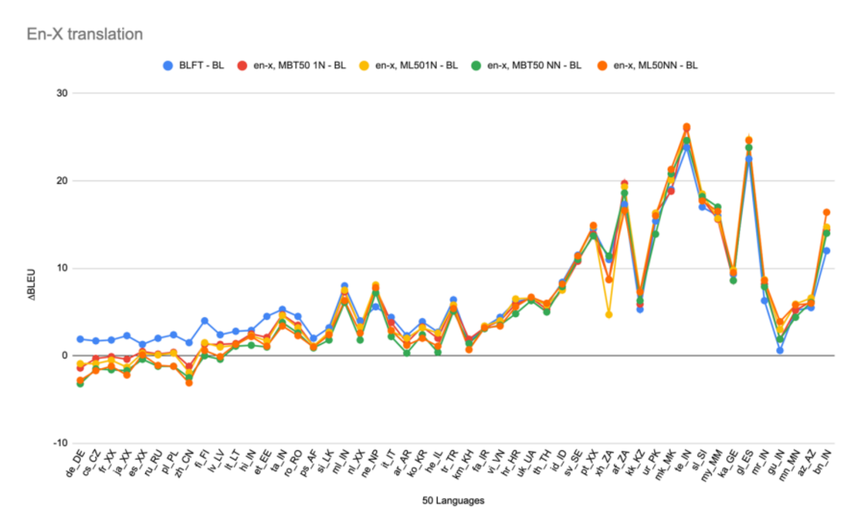
Data size	Languages
10M+	German, Czech, French, Japanese, Spanish, Russian, Polish, Chinese
1M - 10M	Finnish, Latvian, Lithuanian, Hindi, Estonian
100k to 1M	Tamil, Romanian, Pashto, Sinhala, Malayalam, Dutch, Nepali, Italian, Arabic, Korean, Hebrew, Turkish, Khmer, Farsi, Vietnamese, Croatian, Ukrainian
10K to 100K	Thai, Indonesian, Swedish, Portuguese, Xhosa, Afrikaans, Kazakh, Urdu, Macedonian, Telugu, Slovenian, Burmese, Georgia
10K-	Marathi, Gujarati, Mongolian, Azerbaijani, Bengali

- We did not train mBART50 from scratch, instead, we find that it is possible to **simply take the mBART25 checkpoint, and continue training the model with more languages.**

Results on 50 Languages:

Overall Results



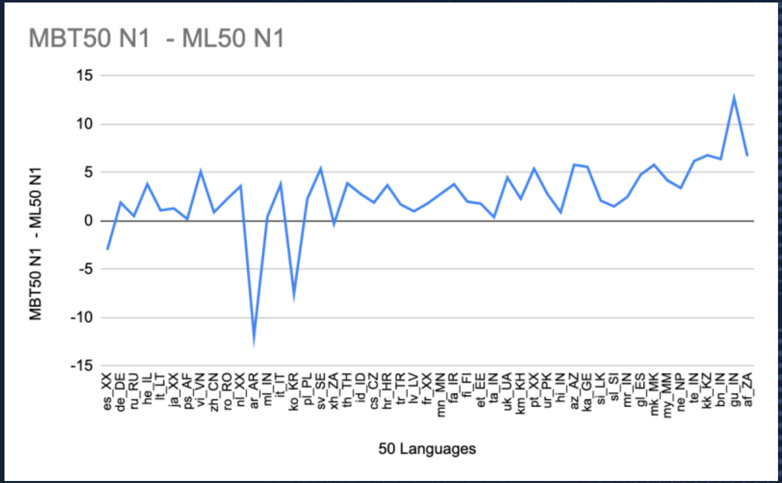


High Resource → *Low Resource* *High Resource* → *Low Resource*

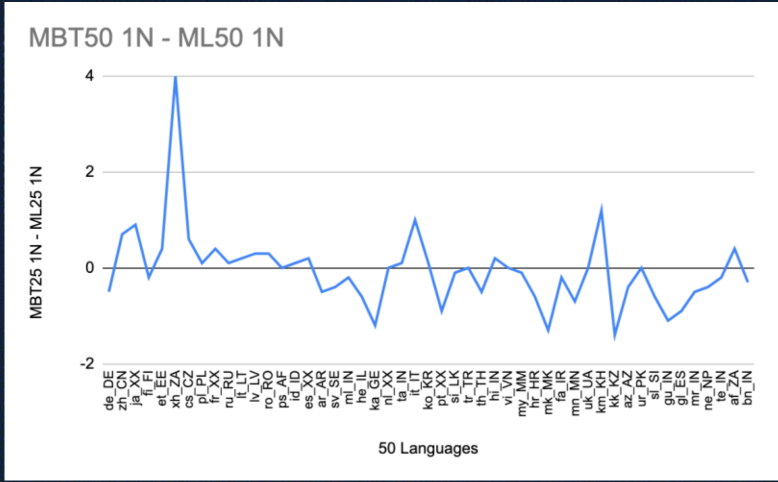
FACEBOOK AI

Results on 50 Languages:

- mNMT with/without mBART Pre-training



High Resource → *Low Resource*

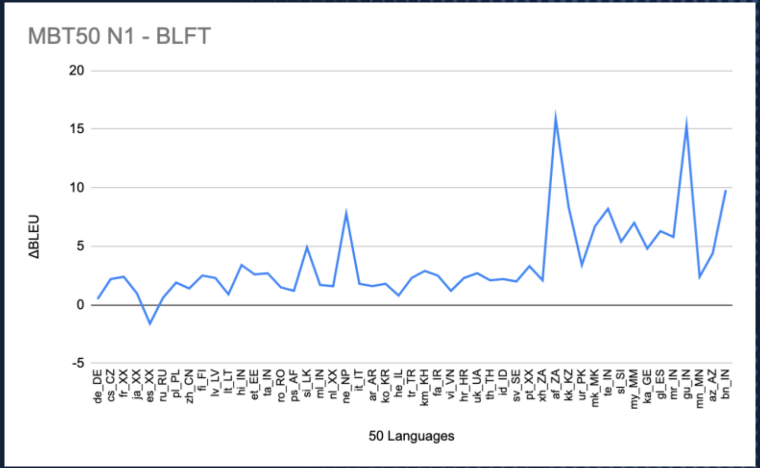


High Resource → *Low Resource*

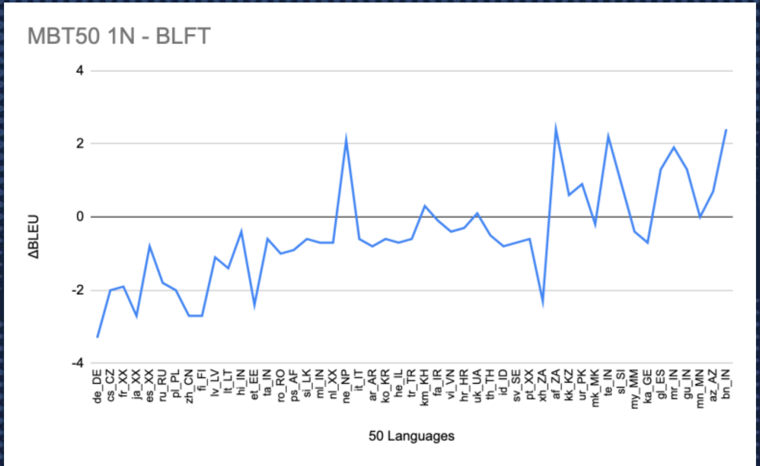
FACEBOOK AI

Results on 50 Languages:

- mBART Pre-training + bilingual or multilingual fine-tuning



High Resource → *Low Resource*



High Resource → *Low Resource*

Results on Zero-shot Translation

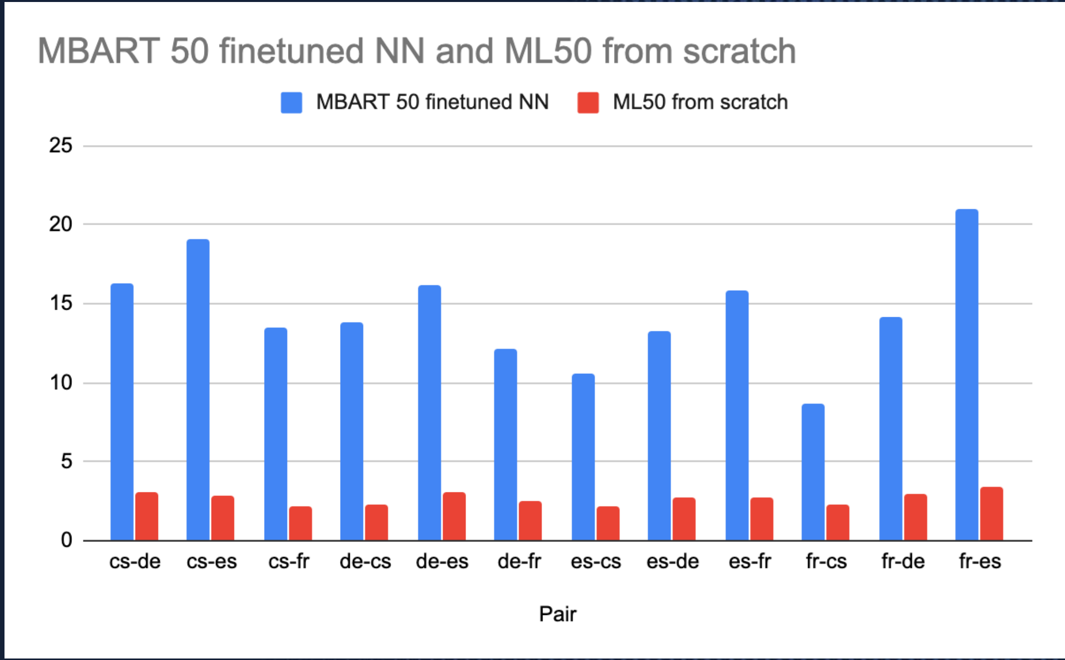
Training many-to-many models naturally enables us to perform zero-shot translation which has already seen in (Johnson, et. al., 2017)

We perform some “initial” experiments on many-to-many models both from scratch and mBART fine-tuning. Both models are trained on ML50 benchmark without any specific modification (e.g. auxiliary loss to encourage language agnostic representations).

We evaluate the learned models on translation test sets of {CS/DE/ES/FR} from WMT data.

Results on Zero-shot Translation

The results are somewhat expected...



- Model trained from scratch is degenerated and only output English.
- In contrast, model fine-tuned from mBART achieves quite stable performance across these languages.
- We suspect two possible reasons:
  - Universal representation;
  - Pretrained Decoder.

Conclusions for mBART + mNMT

With the experiments of both 25 languages and 50 languages, it is clear to draw the following conclusions:

(1) For many-to-one (X-En) translation, mNMT with mBART pre-training almost improves the performance compared to both bilingual fine-tuning and multilingual training from scratch.

(2) For one-to-many (En-X) translation, things get complicated.

   (a) mBART becomes only useful for high-resource or very low resource languages;

   (b) Bilingual fine-tuning is more stable for medium sized languages.

(3) Many-to-many translation follows similar trend in (1) and (2), while mBART pre-training enables “stable” zero-shot machine translation results.

Future Work

- Identify the issue and improve the performance of one-to-many translation with mBART pre-training;
- Efficient Inference for mBART fine-tuned models;
- Extend to 100+ languages + 10B models;
- Online CRISS and v.s. BT
- More...

Open Source & Reference

Dataset:

- Pretraining: CC-Net (https://github.com/facebookresearch/cc_net)
- ML25/ML50 benchmark: TBD

Code: https://github.com/pytorch/fairseq/tree/master/examples/mbart

SCAN ME

Thank You !