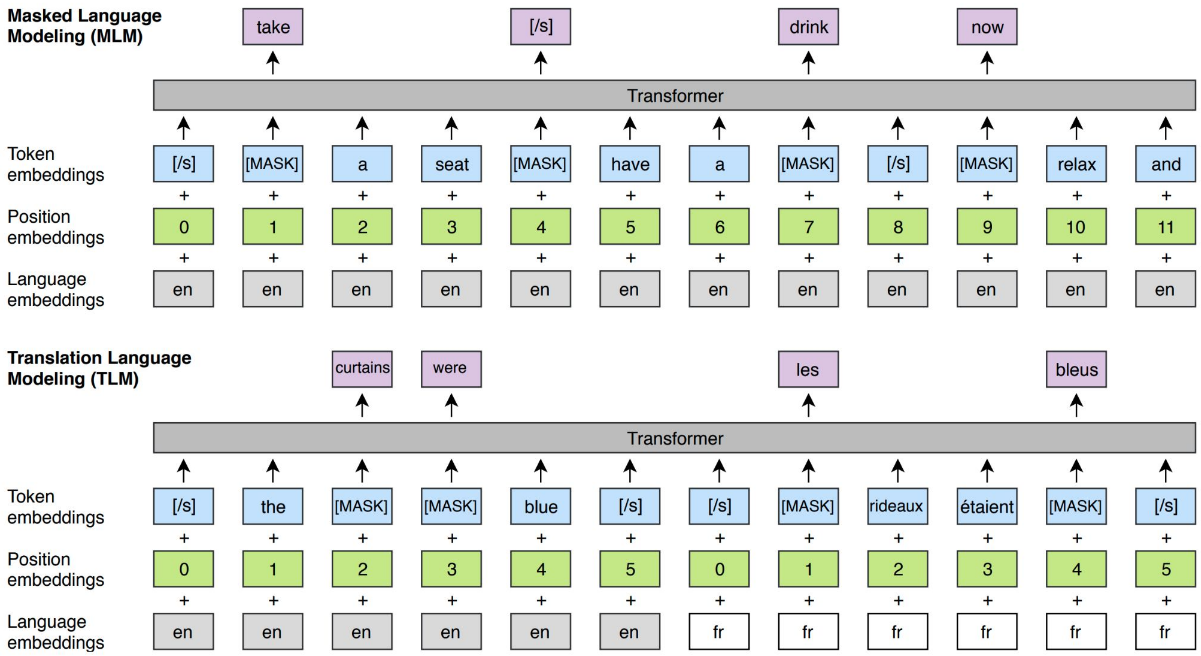
PARADISE: Exploiting Parallel Data for Multilingual Sequence-to-Sequence Pretraining

Machel Reid (UTokyo) and Mikel Artetxe (Facebook AI)

Paper: https://arxiv.org/abs/2108.01887
Code: https://github.com/machelreid/paradise

Machel Reid
October 7th 2021
Facebook NLP Summit

Multilingual Pre-trained LMs (e.g. XLM*, mBERT, etc...)

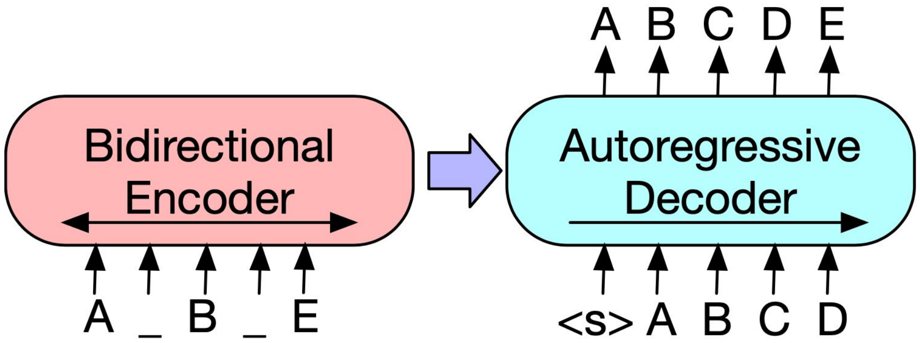


Generally, a variant of BERT-style (MLM pre-training) with a Transformer encoder (and sometimes using parallel data -- as shown above)

BART: Introducing sequence-to-sequence pre-training

- Extends the masked language modeling paradigm of BERT, but to encoder decoder models

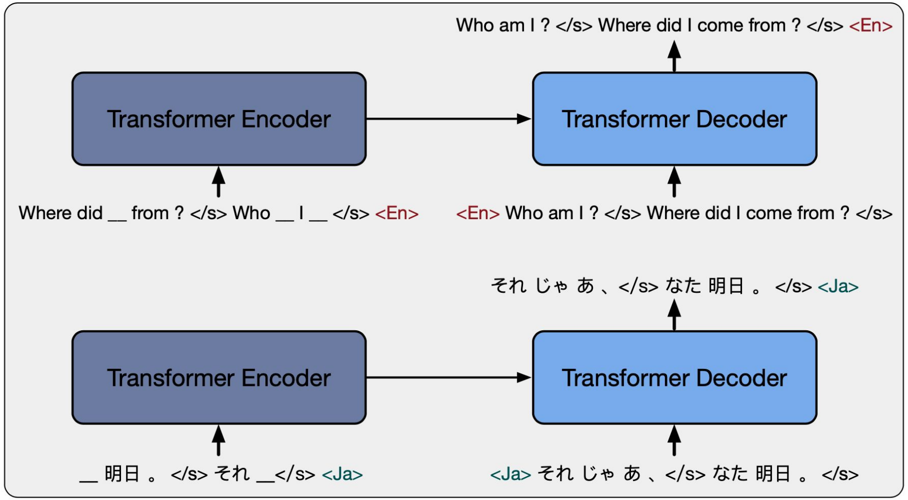
- Improved performance on generative tasks



mBART: Multilingual Sequence-to-Sequence Pre-training

- Extension of BART, however with multilingual data

- Tested on downstream machine translation/showing large gains over random initialization



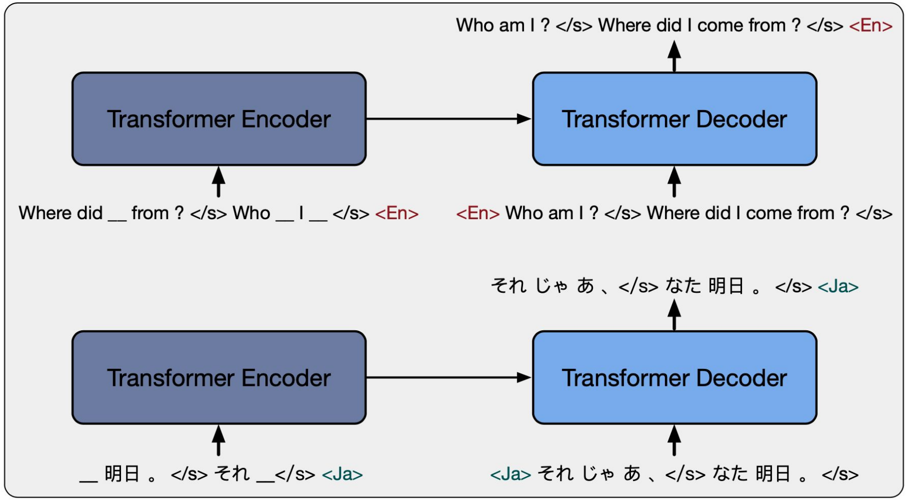
Multilingual Denoising Pre-Training (mBART)

mBART: Multilingual Sequence-to-Sequence Pre-training

- Extension of BART, however with multilingual data

- Tested on downstream machine translation/showing large gains over random initialization

(i.e. parallel information only at fine-tuning)



Multilingual Denoising Pre-Training (mBART)

But, can we use parallel information help sequence-to-sequence **pre-training**?

Enters PARADISE!

Big Ideas

- We look at ways of integrating parallel information/data into the pre-training process

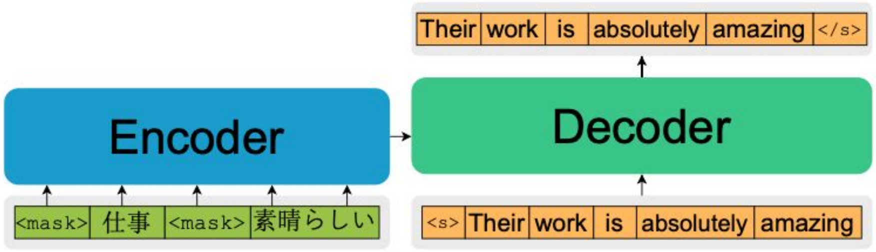
Bitext

Dictionaries

Bitext Denoising

- Using sentence-level machine translation data

- Token masking as noise to prevent overfitting to small datasets (e.g. En-Vi 100k examples)

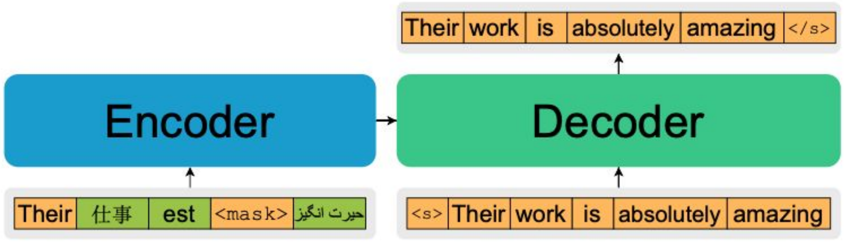


(b) Bitext Denoising

Dictionary Denoising

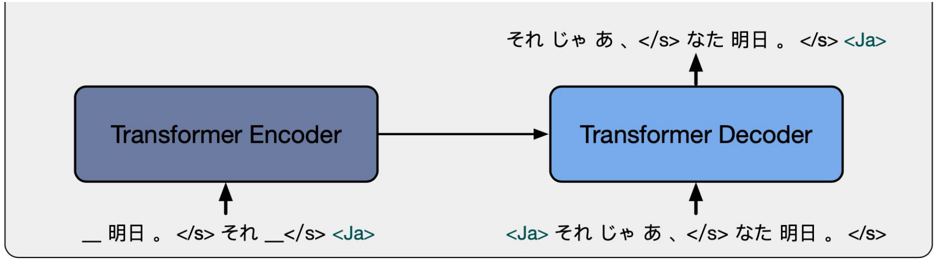
- Uses a multilingual dictionary
  - (which we can construct by using multiple English-XX bilingual dictionaries with English as a pivot language)

- Corrupt input by replacing tokens according to this multilingual dictionary and learning to correct this

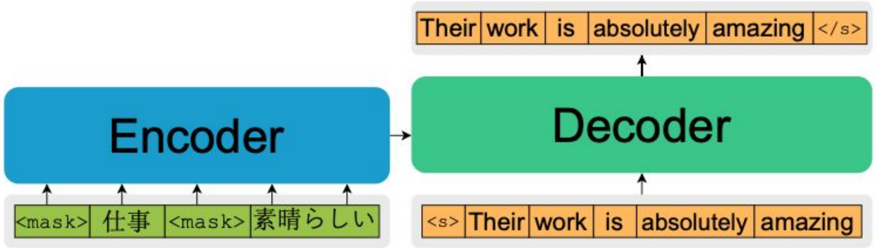


(a) Dictionary Denoising

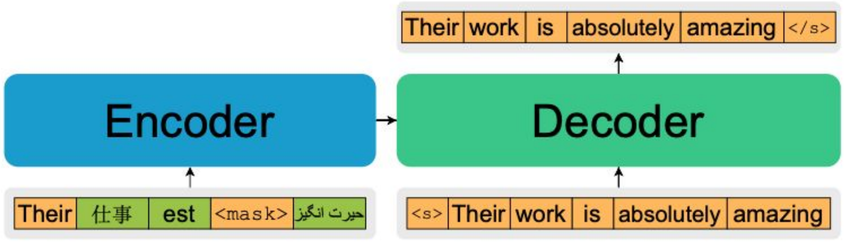
Objectives



Multilingual Denoising Pre-Training (mBART)



(b) Bitext Denoising



(a) Dictionary Denoising

An analogy with human second+ language learning

Learning languages with
- a bunch of books in different languages

Learning language with:
- bunch of books in different languages
- a dictionary
- some example sentences

An analogy with human second+ language learning

Learning language with:
- bunch of books in different languages
- a dictionary
- a some example sentences

Really important! You can add parallel info at scale really cheaply!

An analogy with human second+ language learning

Learning language with:
- bunch of books in different languages
- a dictionary
- a some example sentences

Really important! You can add parallel info at scale really cheaply!

This is why language dictionaries exist!

Experiments & Results

Training details

- Trained on 20 languages
- On 32 V100 (16GB) GPUs for one day [much less compute than prev. methods!] (196M param. model)
- Uses (81-95GB of data -- depending on configuration)
    - With (9-23 GB of parallel data included)

Downstream Tasks

- Machine Translation (10 lang. Pairs, 20 directions)
- PAWS-X
- XNLI

Machine Translation Results

Languages	En-Vi		En-Tr		En-Ja		En-Ar		En-Ne		En-Ro		En-Si		En-Hi		En-Es		En-Fr	
Data Source	**IWSLT15**		**WMT17**		**IWSLT17**		**IWSLT17**		**FLoRes**		**WMT16**		**FLoRes**		**IITB**		**WMT13**		**WMT14**	
Size	133K		207K		223K		250K		564K		608K		647K		1.56M		15M		41M	
Direction	←	→	←	→	←	→	←	→	←	→	←	→	←	→	←	→	←	→	←	→
Random init.	23.6	24.8	12.2	9.5	10.4	12.3	27.5	16.9	7.6	4.3	34.0	34.3	7.2	1.2	10.9	14.2	32.1	31.4	37.0	38.9
mBART (ours)	29.1	31.5	21.3	15.8	15.7	17.3	32.1	19.2	10.3	6.1	34.3	34.9	11.0	2.7	20.2	19.0	29.8	30.4	36.0	38.2
PARADISE	**30.0**	**32.6**	**23.5**	**17.2**	**17.2**	**19.2**	**35.3**	**21.1**	**13.7**	**7.9**	**35.9**	**36.5**	**14.0**	**3.7**	**23.6**	**20.7**	**32.6**	**32.7**	**37.8**	**39.8**

Table 1: Machine translation results. Random initialization numbers taken from Liu et al. (2020).

Despite seeing the same data (incl. finetuning), adding parallel information at pre-training time helps when fine-tuning on machine translation.

MT Ablation

- Increased parallel data helps
- Dictionary noising is important! (+0.5 BLEU)
  - Especially on Hindi, Sinhala (with non-Latin scripts)

Lang. pair (En-XX)	Tr	Ro	Si	Hi	Es	Avg$_\Delta$
mBART (ours)	15.8	34.9	2.7	19.0	30.4	$20.6_{\pm 0.0}$
PARADISE (w/o dict.)	16.8	36.2	3.2	20.5	32.4	$21.8_{+1.2}$
PARADISE	17.2	36.5	3.7	20.7	32.7	$22.2_{+1.6}$
PARADISE++	19.0	37.3	4.2	20.7	33.0	$\mathbf{22.8}_{+2.2}$
Lang. pair (XX-En)	**Tr**	**Ro**	**Si**	**Hi**	**Es**	**Avg$_\Delta$**
mBART (ours)	21.3	34.3	11.0	20.2	29.8	$23.3_{\pm 0.0}$
PARADISE (w/o dict.)	23.2	35.6	13.2	22.3	31.6	$25.2_{+1.9}$
PARADISE	23.5	35.9	14.0	23.6	32.6	$25.9_{+2.6}$
PARADISE++	24.9	36.8	15.1	23.5	32.9	$\mathbf{26.6}_{+3.3}$

Table 2: Ablation results on machine translation.

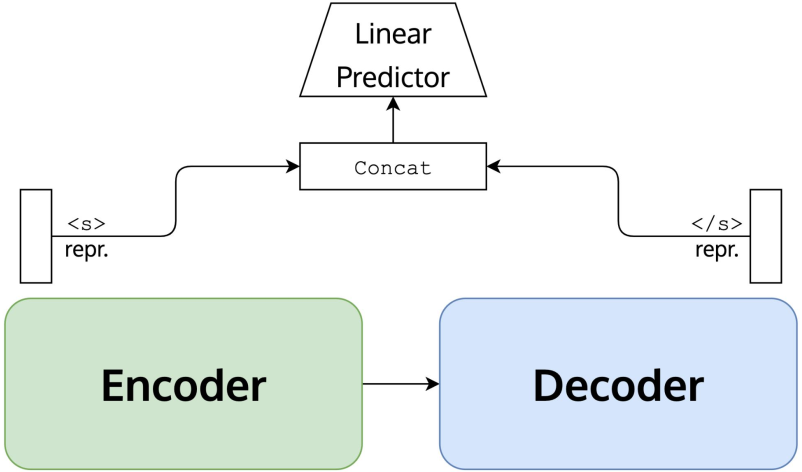
Classification

When fine-tuning on classification we propose a new method:

Concatenate encoder + decoder representations before class prediction

Model	avg	Δ
PARADISE++ (*encoder-decoder*)	74.3	—
decoder-only	73.8	-0.5
encoder-only	72.0	-2.3

Table 4: Ablation of finetuning methods on XNLI.



XNLI

Models	en	zh	es	de	ar	ur	ru	bg	el	fr	hi	sw	th	tr	vi	avg
Finetune a multilingual model on the English training set (ZERO-SHOT)																
mBART (ours)	77.5	68.0	70.7	68.8	66.7	62.2	68.6	72.1	69.6	70.1	63.4	62.6	66.6	65.0	69.7	68.1
PARADISE	**83.4**	73.8	77.6	76.0	**72.4**	65.1	74.0	74.4	73.2	**77.7**	**70.6**	66.2	70.4	72.1	75.3	73.5
PARADISE++ (w/o dict.)	83.3	72.9	77.2	75.7	64.4	**66.9**	73.4	74.8	75.7	**77.7**	68.5	67.4	71.0	73.3	75.0	73.1
PARADISE++	83.0	**74.0**	**79.0**	**76.5**	68.5	66.8	**74.3**	**76.0**	**76.4**	**77.7**	70.2	**70.5**	**72.3**	**74.2**	**75.4**	**74.3**
Finetune a multilingual model on all machine translated training sets (TRANSLATE-TRAIN-ALL)																
mBART (ours)	77.8	72.0	74.0	72.6	69.5	66.5	70.9	74.3	72.7	73.8	68.9	68.2	70.5	70.5	73.9	71.7
PARADISE	84.0	77.6	81.2	79.4	75.9	68.0	76.8	79.1	79.0	79.9	73.4	72.6	75.7	76.2	78.6	77.2
PARADISE++ (w/o dict.)	83.2	77.2	79.7	78.5	72.0	68.3	76.5	78.2	79.2	79.3	73.3	73.3	75.3	77.5	77.3	76.6
PARADISE++	**84.8**	**78.3**	**81.7**	**80.5**	**76.0**	**70.6**	**78.8**	**80.4**	**81.3**	**80.6**	**74.9**	**74.2**	**77.3**	**78.4**	**79.2**	**78.5**

Table 3: Accuracy of zero-shot crosslingual classification on the XNLI dataset.

PAWS-X

Model	de	en	es	fr	zh	Avg
mBERT	85.7	94.0	87.4	87.0	77.0	86.2
MMTE	85.1	93.1	87.2	86.9	75.9	85.6
mT5-small	86.2	92.2	86.1	86.6	77.9	85.8
AMBER	89.4	**95.6**	89.2	**90.7**	80.9	89.2
XLM-15	88.5	94.7	89.3	89.6	78.1	88.0
XLM-100	85.9	94.0	88.3	87.4	76.5	86.4
XLM-R-base	87.0	94.2	88.6	88.7	78.5	87.4
XLM-R-large	**89.7**	94.7	**90.1**	90.4	**82.3**	**89.4**
PARADISE++	89.1	94.3	89.6	90.6	**82.3**	89.2

Table 6: Accuracy of zero-shot cross-lingual classification on PAWS-X. Bold numbers highlight the highest scores across languages on the existing models (upper part) and PARADISE variants (bottom part). We source baseline results from Hu et al. (2020, 2021); Xue et al. (2021).

Almost reaches XLM-R large level performance!

Comparison with popular models ⬇

model	#Langs	Task	Params.	Est. GPU Days	Data (GB)	XNLI	PAWS-X	MT
mBERT (Devlin et al., 2019)†	104	MLM	179M (0.9x)	—	60	65.4	86.2	—
MMTE (Siddhant et al., 2019)†	102	Translation	375M (1.9x)	—	5000	67.4	85.6	—
mT5-small (Xue et al., 2021)	101	Eq. 1	300M (1.5x)	—	27000	67.5	85.8	—
mT6 (Chi et al., 2021a)	94	SC+PNAT+TSC	300M (1.5x)	40 (1.3x)	2120	64.7	86.6	—
AMBER (Hu et al., 2021)	104	MLM+TLM	179M (0.9x)	1000 (31x)	100	71.6	89.2	—
XLM-15 (Conneau and Lample, 2019)‡	15	MLM+TLM	250M (1.3x)	450 (14x)	100	72.6	88.0	—
XLM-100 (Conneau and Lample, 2019)†	100	MLM	570M (2.9x)	640 (20x)	60	69.1	86.4	—
XLM-R-base (Conneau et al., 2020a)‡	100	MLM	270M (1.4x)	13K (406x)	2400	73.4	87.4	—
XLM-R-large (Conneau et al., 2020a)†	100	MLM	550M (2.8x)	27K (844x)	2400	**79.2**	**89.4**	—
mBART (Liu et al., 2020)	25	Eq. 1	680M (3.5x)	4.5K (140x)	2400	—	—	**23.5**
mBART (ours)	20	Eq. 1	196M (1.0x)	32 (1.0x)	72	68.1	85.4	21.1
PARADISE	20	Eq. 1, 2, 3	196M (1.0x)	32 (1.0x)	81	73.5	89.0	23.1
PARADISE++	20	Eq. 1, 2, 3	196M (1.0x)	32 (1.0x)	95	**74.3**	**89.2**	**23.8**

Table 5: Comparison with prior work. † denotes results taken from Hu et al. (2020), and ‡ denotes results taken from Hu et al. (2021). The rest of the numbers are taken from the original papers.

Outperforms XLM-R-base (XTREME baseline) on these tasks using **400x less compute** and mT5 with **much less data**

Comparison with original mBART

- For most pairs PARADISE obtains competitive/better results (despite 140x less compute / 3.5 fewer params.)

- We only show significant losses on En-Es (with 13M pairs) where the architecture size (196M vs 660M params.) may not have had enough capacity (related to scaling laws, etc...)

Lang. Pair	En-Tr	En-Ro	En-Si	En-Hi	En-Es	Tr-En	Ro-En	Si-En	Hi-En
mBART (ours)	15.8	34.9	2.7	19.0	30.4	21.3	34.3	11.0	20.2
PARADISE (w/o dict.)	16.8	36.2	3.2	20.5	32.4	23.2	35.6	13.2	22.3
PARADISE	17.2	36.5	3.7	20.7	32.7	23.5	35.9	14.0	**23.6**
PARADISE++	**19.0**	37.3	**4.2**	20.7	33.0	**24.9**	36.8	**15.1**	23.5
mBART	17.8	**37.7**	3.3	**20.8**	**34.0**	22.5	**37.8**	13.7	23.5

Table 7: Ablation results on machine translation. Note that mBART is trained with 140x more compute and 3.5x more parameters.

Takeaways

Takeaways

- Use parallel information at pre-training (+ don't constrain parallel data to only translation data!)

- With dictionaries, you can add parallel information very cheaply + easy to scale!
  - Even helps at finetuning (in prelim. experiments) with 5% (in the case of our mBART) and 1-2% (for PARADISE) during XNLI finetuning

Interesting Future Questions

Interesting Future Questions

- How much performance is derived from **modeling** with parallel vs the data itself? (e.g. synthetic data vs gold data)

- What exactly changes when including parallel signal at pre-training versus just finetuning -- even with the same data?

- Do these improvements hold at scale? (or do they diminish?)

東京大学
THE UNIVERSITY OF TOKYO

Thank you!

Q&A