Multilingual NLP

Author: Matthieu Futeral

- **Neural Machine Translation**
  - Evaluating NMT: from BLEU to COMET
  - Sequence-to-Sequence models from scratch
  - Leveraging monolingual data
  - Zero-shot Machine translation
  - Data-centric approach
  - Document-level Machine translation
- NLP tasks beyond English
  - Multilingual vs Monolingual models

- **Neural Machine Translation**
  - **Evaluating NMT: from BLEU to COMET**
  - Sequence-to-Sequence models from scratch
  - Leveraging monolingual data
  - Zero-shot Machine translation
  - Data-centric approach
  - Document-level Machine translation
- NLP tasks beyond English
  - Multilingual vs Monolingual models

Evaluating NMT - BLEU score

English

I cannot make it today.

French

Je ne peux pas venir aujourd'hui.

How to evaluate if French translation is a good one?

Evaluating NMT - BLEU score

English

I cannot make it today.

French

Je ne peux pas venir aujourd'hui.

How to evaluate if French translation is a good one?
BLEU Score

Evaluating NMT - BLEU score

English

I cannot make it today.

French (possible translations)

Je ne peux pas venir aujourd'hui.
Je ne peux pas me déplacer aujourd'hui.
Je ne peux pas être présent aujourd'hui.

BLEU Score = Bilingual Evaluation Understudy score

How does it work?

Evaluating NMT - BLEU score

BLEU Score evaluates outputs of a translation system comparing them to the references (regardless of the input).

- Given an output translation, first count the number of 1-grams in the output translation that are in the reference translations (threshold by the maximum number of times the 1-gram is in one ref.)
- Do the same for 2, 3 and 4 grams.
- Compute the geometric mean of BLEU-1 … BLEU-4 to get the MEAN BLEU Score
- Multiply by a penalty length to get the final BLEU score.

Evaluating NMT - BLEU score

Example 1:

Ref 1: There is a cat on the mat.
Ref 2: The cat is on the mat.
Output: The cat is on the couch.

BLEU-1: The cat is on the couch . => **6/7**

BLEU-2: The cat cat is is on on the the couch couch . => **4/6**

BLEU-3: The cat is cat is on is on the on the couch the couch . => **3/6**

BLEU-4: The cat is on cat is on the is on the couch on the couch . => **2/4**

BLEU = BP * exp(1 * log(6/7) + ½ * log(4/6) + ⅓ * log(3/6) + ¼ * log(2/4)) (BP = 1)

Evaluating NMT - BLEU score

Example 2:

Ref 1: There is a cat on the mat.
Ref 2: The cat is on the mat.
Output: Cat on mat.

BLEU-1: Cat on mat . => **4/4**

BLEU-2: Cat on on mat mat . => **3/3**

BLEU-3: Cat on mat on mat . => **2/2**

BLEU-4: Cat on mat . => **1/1**

BLEU = BP * 1 Here, length(Output) = 4 < length(min(ref)) = 7 => BP = exp(1 - 7/4)
BLEU = 0.4723

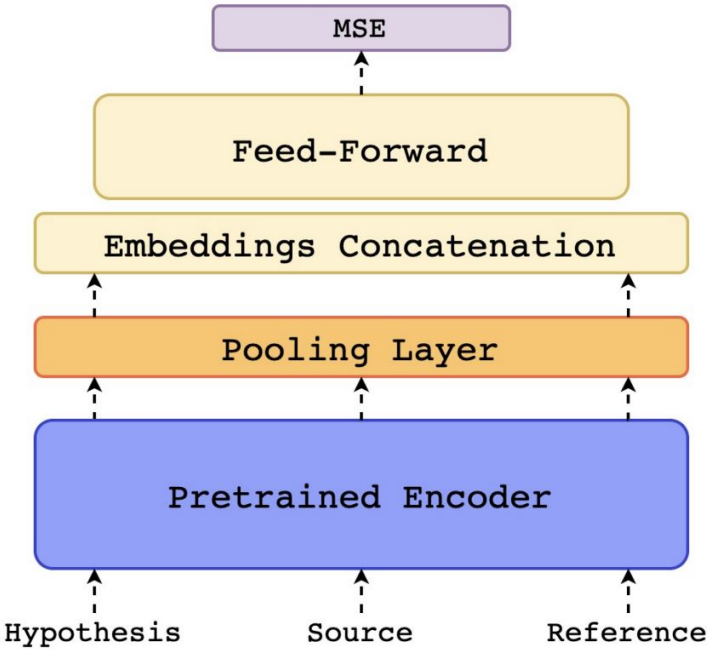
Evaluating NMT - BLEU score

Limitations:

- Higher score for long sentences than short (BP = 1)
- Just a matching word evaluation => a synonym not included in the references is considered an error in the same way as any other words.
- No notion of grammar.

Solution: **COMET**

Evaluating NMT - COMET score, a neural based metric



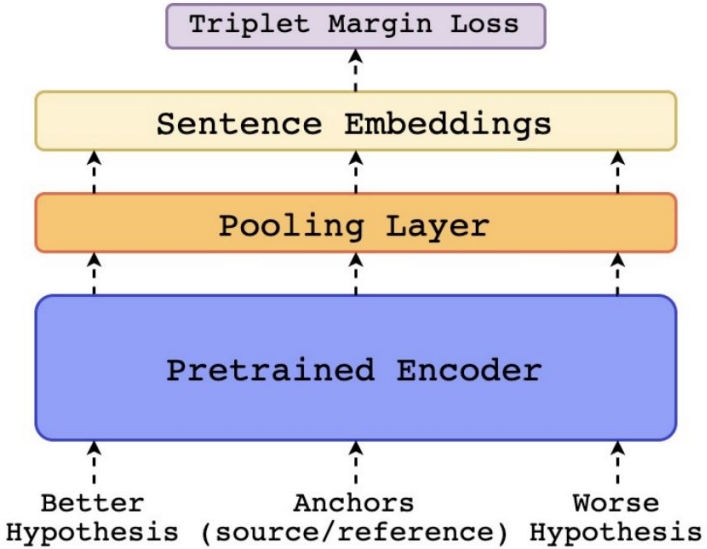


Illustration from Rei et al. (2020)

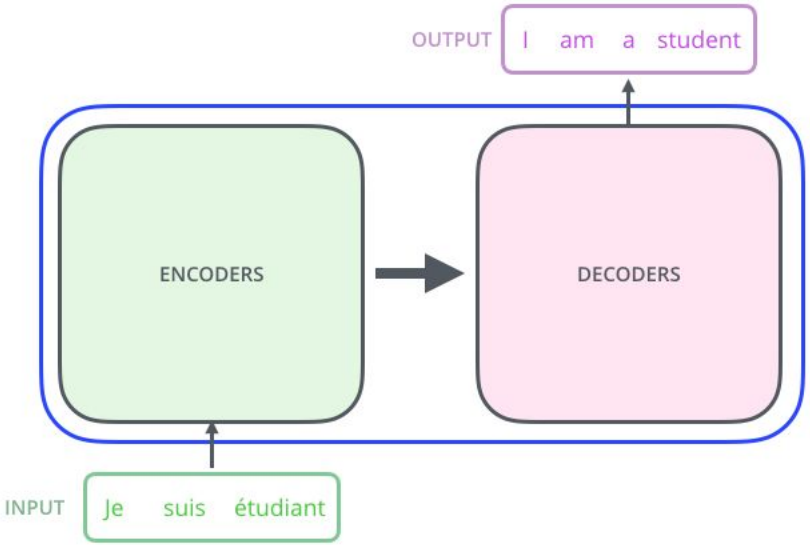
Evaluating NMT - COMET score, a neural based metric

Metric	en-cs	en-de	en-fi	en-gu	en-kk	en-lt	en-ru	en-zh
BLEU	0.364	0.248	0.395	0.463	0.363	0.333	0.469	0.235
CHRF	0.444	0.321	0.518	0.548	0.510	0.438	0.548	0.241
YISI-1	0.475	0.351	0.537	0.551	0.546	0.470	0.585	0.355
BERTSCORE (default)	0.500	0.363	0.527	0.568	0.540	0.464	0.585	0.356
BERTSCORE (xlmr-base)	0.503	0.369	0.553	0.584	0.536	0.514	0.599	0.317
COMET-HTER	0.524	0.383	0.560	0.552	0.508	0.577	0.539	0.380
COMET-MQM	0.537	0.398	0.567	0.564	0.534	0.574	**0.615**	0.378
COMET-RANK	**0.603**	**0.427**	**0.664**	**0.611**	**0.693**	**0.665**	0.580	**0.449**

Comparing different NMT metric on the WMT 2019 Metrics Shared task. Score is Kendall's tau. Table from Rei et al. (2020)

- **Neural Machine Translation**
  - Evaluating NMT: from BLEU to COMET
  - **Sequence-to-Sequence models from scratch**
  - Leveraging monolingual data
  - Zero-shot Machine translation
  - Data-centric approach
  - Document-level Machine translation
- NLP tasks beyond English
  - Multilingual vs Monolingual models

Sequence-to-Sequence models - Bilingual



- Requires large amount of parallel data
- Needs one model for each language pair & direction

Illustration from
https://jalammar.github.io/illustrated-transformer/

Sequence-to-Sequence models - Multilingual

- Still requires large amount of parallel data
- Single model for all language directions
- Joint SentencePiece vocabulary

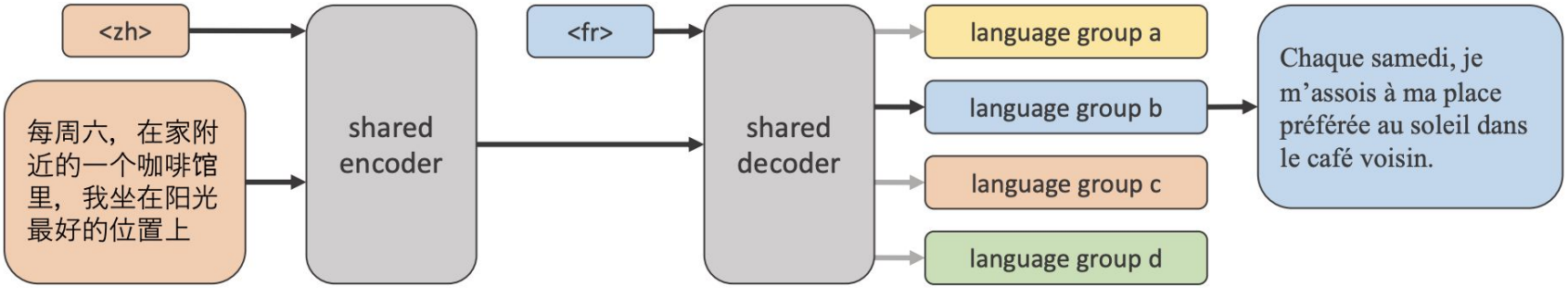


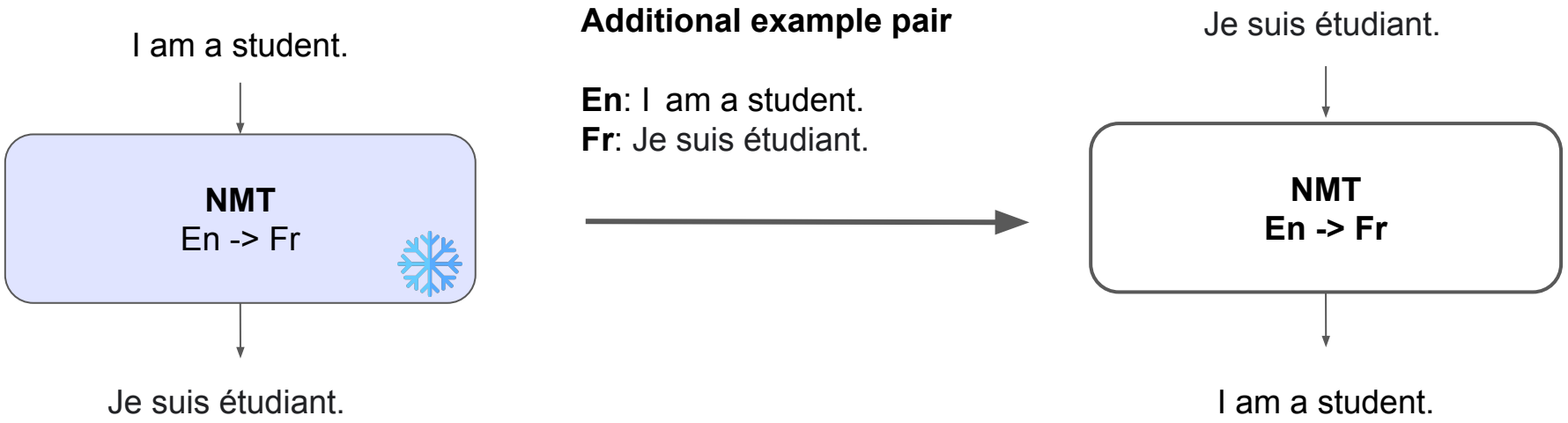
Illustration of the M2M-100 model from Fan et al. (2020)

- **Neural Machine Translation**
  - Evaluating NMT: from BLEU to COMET
  - Sequence-to-Sequence models from scratch
  - **Leveraging monolingual data**
  - Zero-shot Machine translation
  - Data-centric approach
  - Document-level Machine translation
- NLP tasks beyond English
  - Multilingual vs Monolingual models

Leveraging Monolingual data - Backtranslation

- Parallel data is scarce

What if we could use monolingual (cheaper to get)? **Backtranslation**



Leveraging Monolingual data - Backtranslation

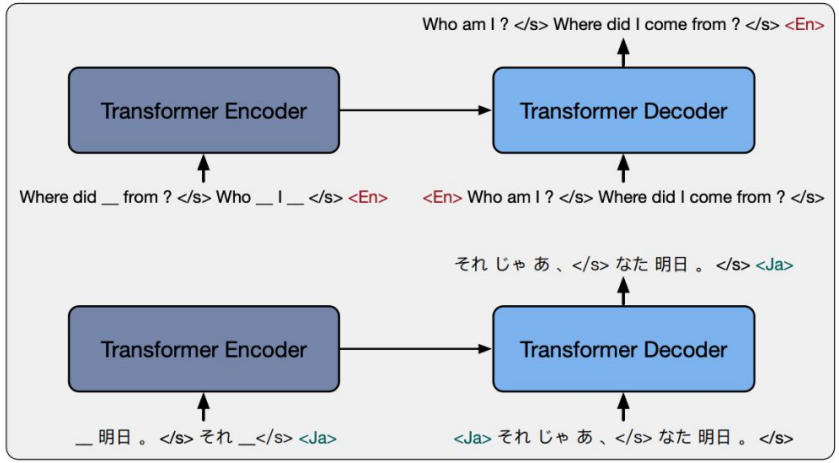
Backtranslation gives significant improvement!

Model	BLEU score
Transformer (baseline)	36.3
Transformer + Back-Translation	38.7

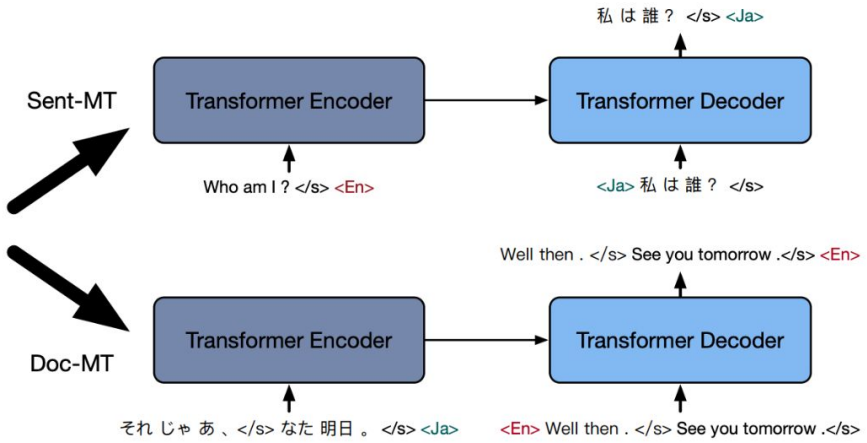
BLEU Results on IWSLT 2015 English -> Turkish from Cuong et al. (2020)

Leveraging Monolingual data - Multilingual Pretraining mBART

- Pretraining on unlabelled data proved to be efficient in many tasks (e.g. BERT)



Multilingual Denoising Pre-Training (mBART)



Fine-tuning on Machine Translation

Illustration from Liu et al. (2020)

Leveraging Monolingual data - Multilingual Pretraining XLM

- Another type of pretraining approach XLM

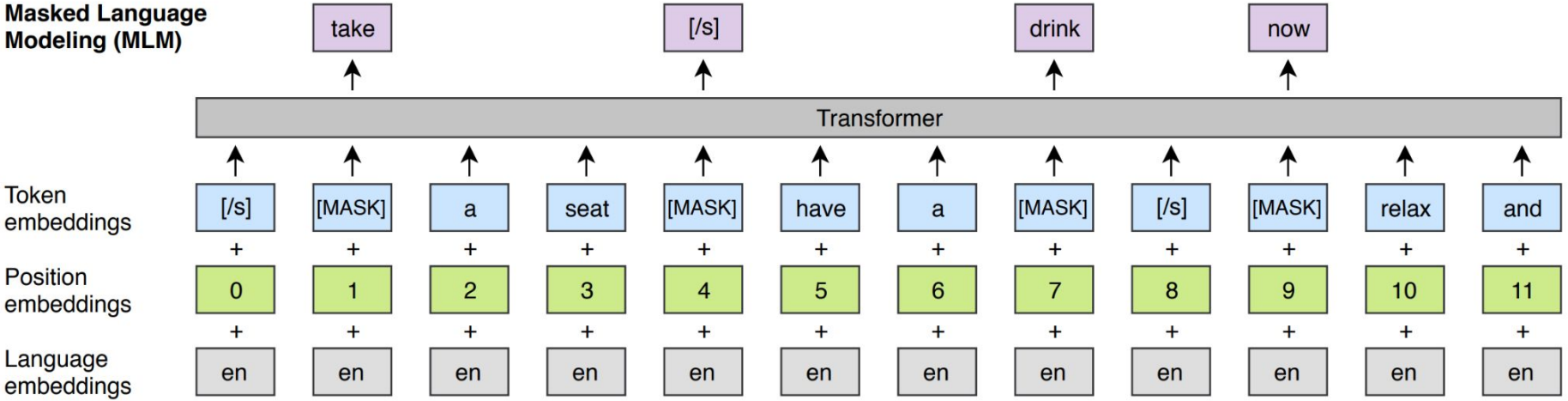


Illustration from Lample et al. (2019)

Leveraging Monolingual data - Multilingual Pretraining

- Performances of pretraining

Pre-training		Fine-tuning		
Model	**Data**	**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

BLEU score comparison of different pretraining approaches on En <-> Ro WMT 2016 test set from Liu et al. 2020

Leveraging Monolingual data - Multilingual Pretraining

- Huge (resp. small) gains for low (resp. high) resource languages

Languages	En-Gu		En-Kk		En-Vi		En-Tr		En-Ja		En-Ko	
Data Source	WMT19		WMT19		IWSLT15		WMT17		IWSLT17		IWSLT17	
Size	10K		91K		133K		207K		223K		230K	
Direction	←	→	←	→	←	→	←	→	←	→	←	→
Random	0.0	0.0	0.8	0.2	23.6	24.8	12.2	9.5	10.4	12.3	15.3	16.3
mBART25	**0.3**	**0.1**	**7.4**	**2.5**	**36.1**	**35.4**	**22.5**	**17.8**	**19.1**	**19.4**	**24.6**	**22.6**

Low/Medium resource MT - BLEU score comparison between mBART pretraining and random init from Liu et al. (2020)

Languages	Cs	Es	Zh	De	Ru	Fr
Size	11M	15M	25M	28M	29M	41M
Random	16.5	33.2	**35.0**	**30.9**	**31.5**	**41.4**
mBART25	**18.0**	**34.0**	33.3	30.5	31.3	41.0

High resource MT - BLEU score comparison between mBART pretraining and random init on WMT test sets from Liu et al. (2020)

Leveraging Monolingual data - Multilingual Pretraining

- Generalization to **Unseen** languages during pretraining

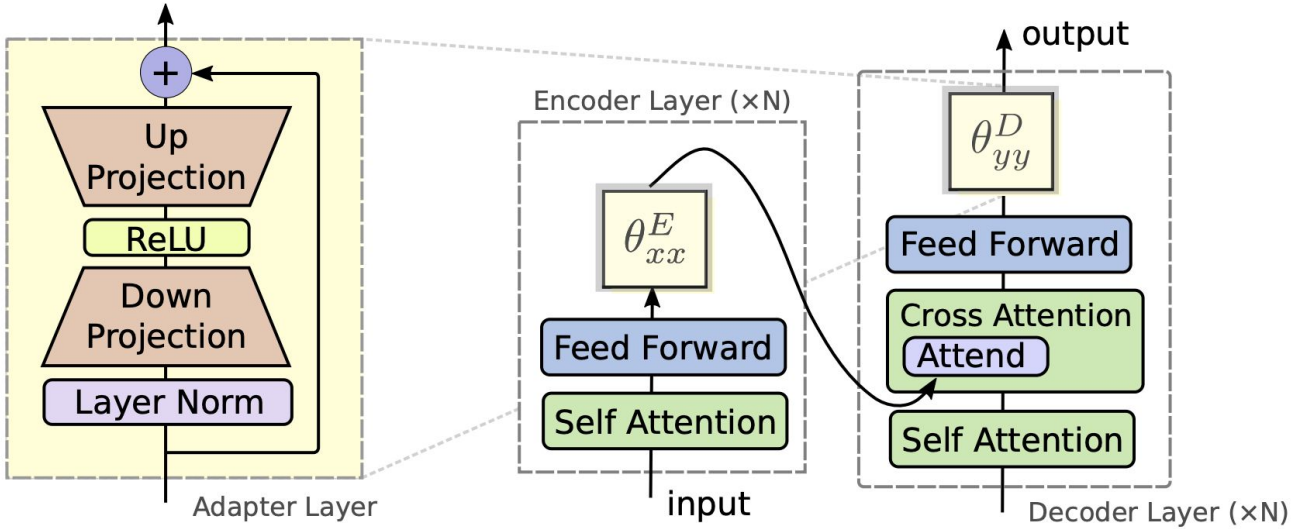
	Monolingual	Nl-En	En-Nl	Ar-En	En-Ar	Nl-De	De-Nl
Random	None	34.6 (-8.7)	29.3 (-5.5)	27.5 (-10.1)	16.9 (-4.7)	21.3 (-6.4)	20.9 (-5.2)
mBART02	En Ro	41.4 (-2.9)	34.5 (-0.3)	34.9 (-2.7)	21.2 (-0.4)	26.1 (-1.6)	25.4 (-0.7)
mBART06	En Ro Cs It Fr Es	43.1 (-0.2)	34.6 (-0.2)	37.3 (-0.3)	21.1 (-0.5)	26.4 (-1.3)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

Generalization to unseen languages. BLEU scores comparison between mBART pretraining and random init. From Liu et al. (2020)

- **Neural Machine Translation**
  - Evaluating NMT: from BLEU to COMET
  - Sequence-to-Sequence models from scratch
  - Leveraging monolingual data
  - **Zero-shot Machine translation**
  - Data-centric approach
  - Document-level Machine translation
- NLP tasks beyond English
  - Multilingual vs Monolingual models

Zero-shot Machine Translation - Adapters

Monolingual adapters to perform NMT on new language directions



Monolingual adapters are inserted in each layer. xx denotes a source language and yy a target language.

Illustration from Philip et al. (2020)

“Zero-shot” Machine Translation - Performances of LLM

The case of CHATGPT (GPT4) from Jiao et al. (2023)

	Translation Prompt
TP1	Translate these sentences from [SRC] to [TGT]:
TP2	Answer with no quotes. What do these sentences mean in [TGT]?
TP3	Please provide the [TGT] translation for these sentences:

Translation prompts used

System	BLEU↑	ChrF++↑	TER↓
Google	31.66	57.09	56.21
DeepL	31.22	56.74	57.84
Tencent	29.69	56.24	57.16
ChatGPT w/ TP1	23.25	53.07	66.03
ChatGPT w/ TP2	24.54	53.05	63.79
ChatGPT w/ TP3	**24.73**	**53.71**	**62.84**

BLEU scores for the different prompts
FLORES-101 En->Zh

“Zero-shot” Machine Translation - Performances of LLM

The case of CHATGPT (GPT4) from Jiao et al. (2023)

Table 5: Performance of ChatGPT for translation robustness on domain-specific or noisy test data.

System	W19 Bio	W20 Rob2		W20 Rob3
	De⇒En	En⇒Ja	Ja⇒En	De⇒En
Google	37.83	29.72	19.21	42.91
DeepL	37.13	26.25	19.83	41.29
ChatGPT	33.22	22.36	18.34	44.59

BLEU scores

“Zero-shot” Machine Translation - Performances of LLM

The case of CHATGPT (GPT4) from Jiao et al. (2023)

System	De-En		Ro-En		Zh-En	
	⇒	⇐	⇒	⇐	⇒	⇐
Google	45.04	41.16	50.12	46.03	31.66	43.58
DeepL	49.23 (+9.3%)	41.46 (+0.7%)	50.61 (+0.9%)	48.39 (+5.1%)	31.22 (-1.3%)	44.31 (+1.6%)
Tencent	n/a	n/a	n/a	n/a	29.69 (-6.2%)	46.06 (+5.6%)
ChatGPT	43.71 (-2.9%)	38.87 (-5.5%)	44.95 (-10.3%)	24.85 (-46.0%)	24.73 (-21.8%)	38.27 (-12.1%)

System	De-Zh		Ro-Zh		De-Ro	
	⇒	⇐	⇒	⇐	⇒	⇐
Google	38.71	21.68	39.05	25.59	33.31	32.27
DeepL	40.46 (+4.5%)	22.82 (+5.2%)	38.95 (-0.2%)	25.39 (-0.7%)	35.19 (+5.6%)	34.27 (+6.1%)
Tencent	40.66 (+5.0%)	19.44 (-10.3%)	n/a	n/a	n/a	n/a
ChatGPT	34.46 (-10.9%)	19.80 (-8.6%)	30.84 (-21.0%)	19.17 (-25.0%)	33.38 (+0.2%)	29.89 (-7.3%)

BLEU scores on FLORES-101

- **Neural Machine Translation**
  - Evaluating NMT: from BLEU to COMET
  - Sequence-to-Sequence models from scratch
  - Leveraging monolingual data
  - Zero-shot Machine translation
  - **Data-centric approach**
  - Document-level Machine translation
- NLP tasks beyond English
  - Multilingual vs Monolingual models

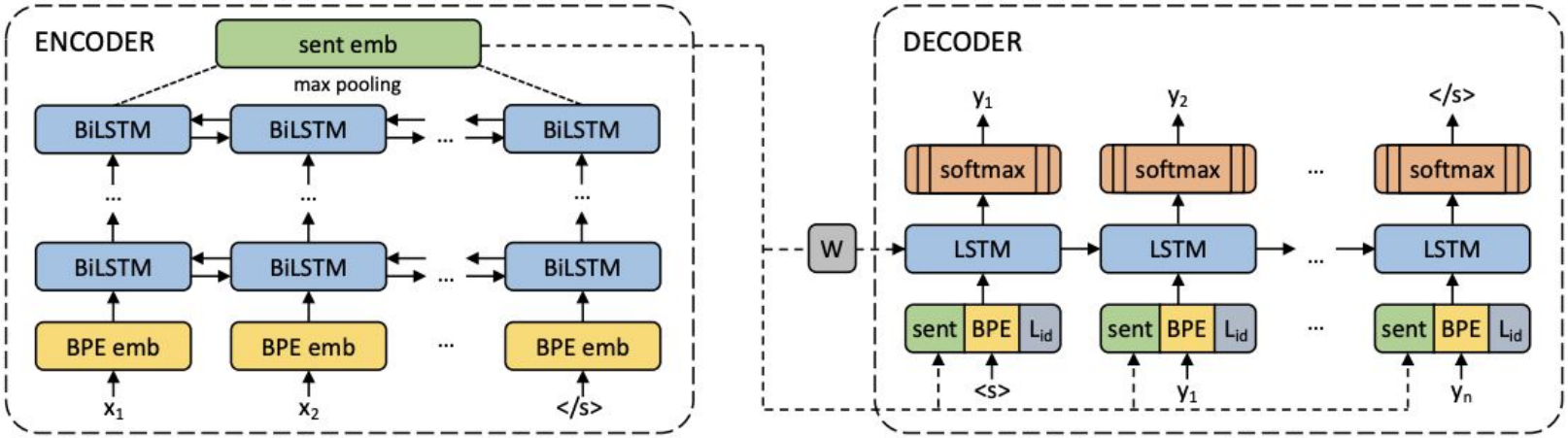
Data-centric approach

- Previous methods led to improvement in medium/low resource settings but performances are still limited by the lack of data.

- Could we mine translation pairs for any language pairs to increase the amount of parallel data available?

Data-centric approach - LASER

Language Agnostic SEntence Representation - LASER

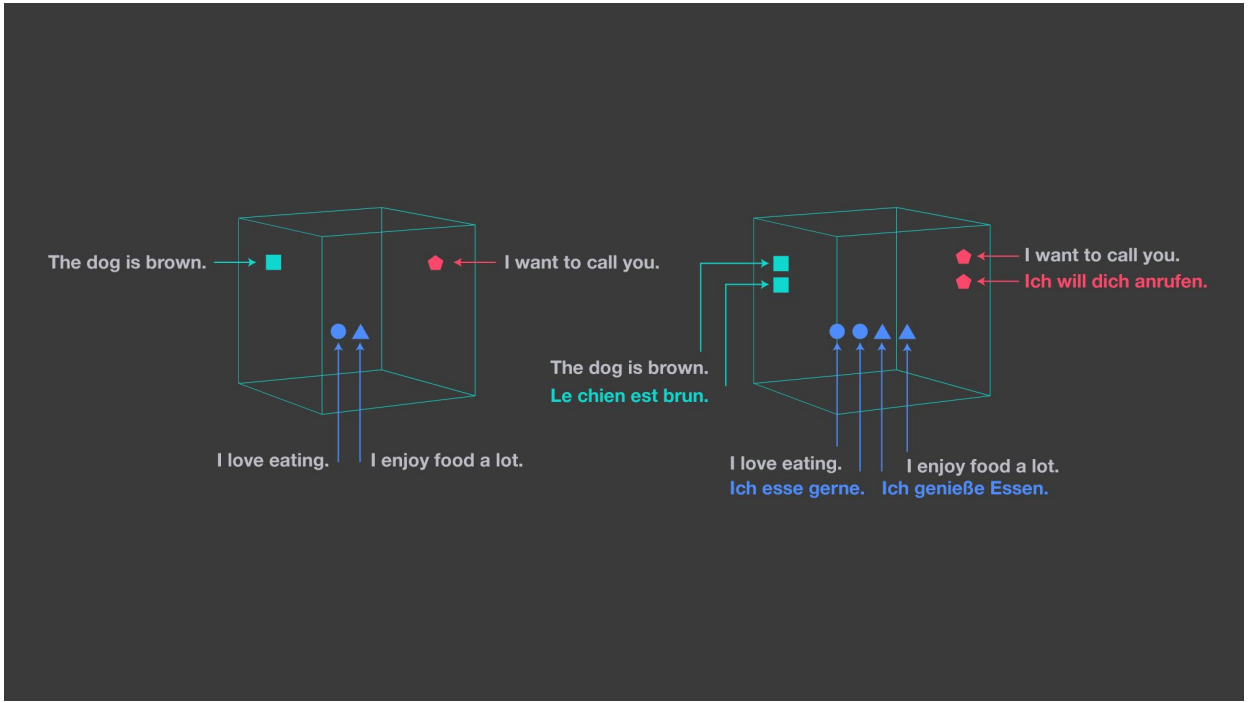


LASER modelling. The sentence embedding is what matters here.

Illustration from Artexte et al. (2019)

Data-centric approach - LASER

Language **A**gnostic **SE**ntence **R**epresentation - LASER

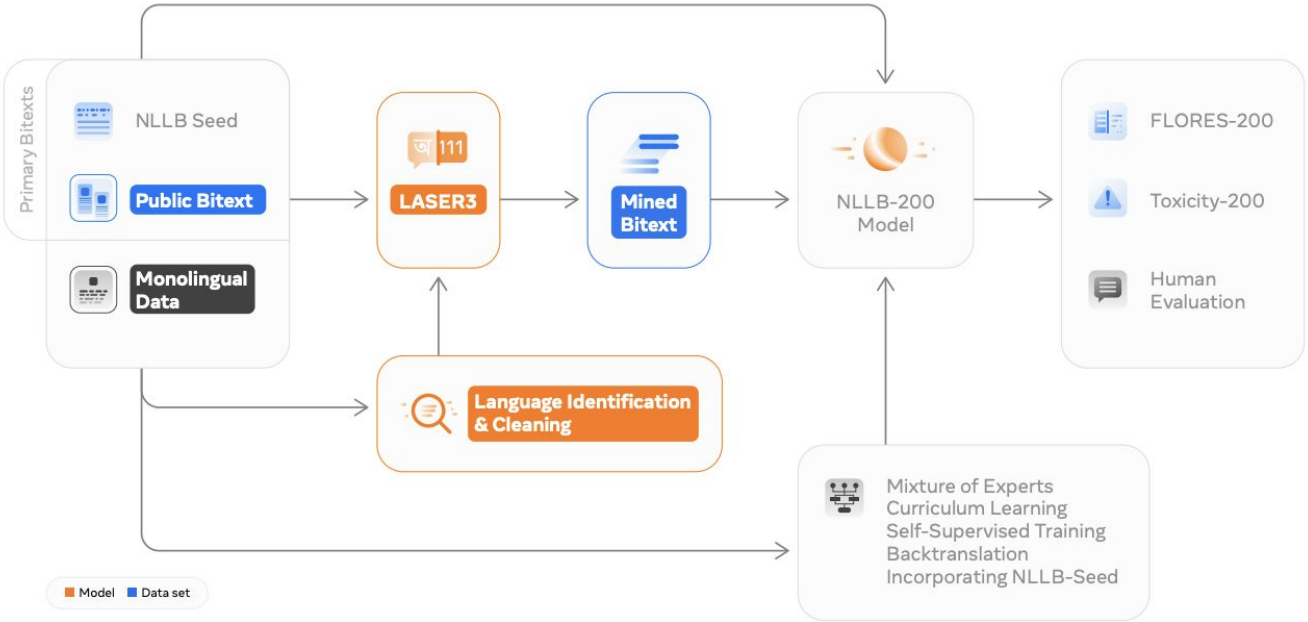


Neural Machine Translation

Illustration from FB blogpost

Data-centric approach - NLLB

No **L**anguage **L**eft **B**ehind (NLLB) approach



NLLB approach - Using LASER to mine bitext data from CommonCrawl

Illustration from NLLB Team et al. (2022)

- **Neural Machine Translation**
  - Evaluating NMT: from BLEU to COMET
  - Sequence-to-Sequence models from scratch
  - Leveraging monolingual data
  - Zero-shot Machine translation
  - Data-centric approach
  - **Document-level Machine translation**
- NLP tasks beyond English
  - Multilingual vs Monolingual models

Document-level Machine Translation

- Most works on NMT focuses on the translation of (short) sentences.
- Many tasks require to translate a whole document.
- 2 simple baselines:
  - **Translate sentences independently** => Low document coherence
  - **Concatenate sentences** => High computational cost + performance downgrades with long context

Document-level Machine Translation - Concatenation methods

Different ways to concatenate contexts:

- Concatenation + self-attention:

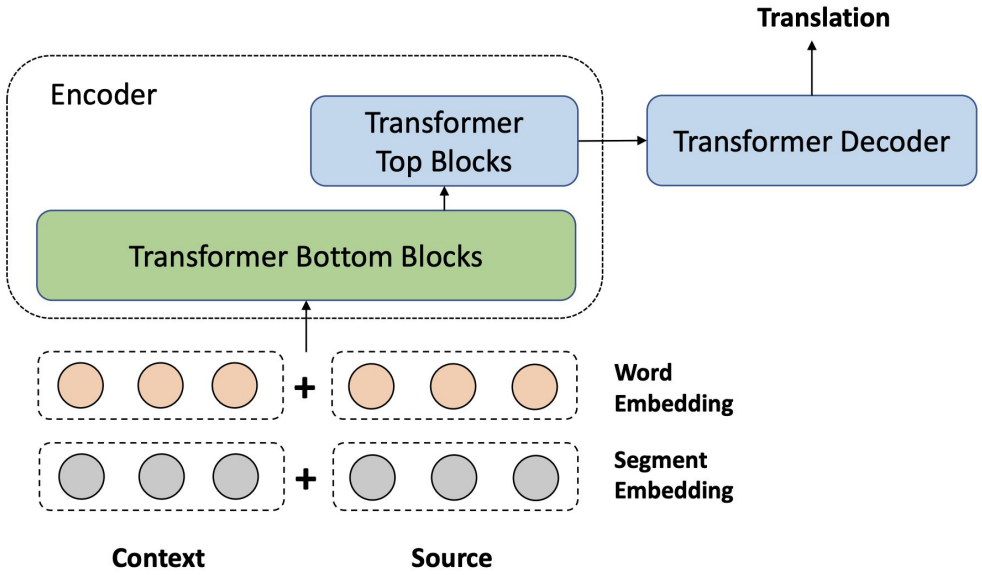


Illustration from Ma et al. (2020)

Document-level Machine Translation - Concatenation methods

Different ways to concatenate contexts:

- Long-Short term attention:

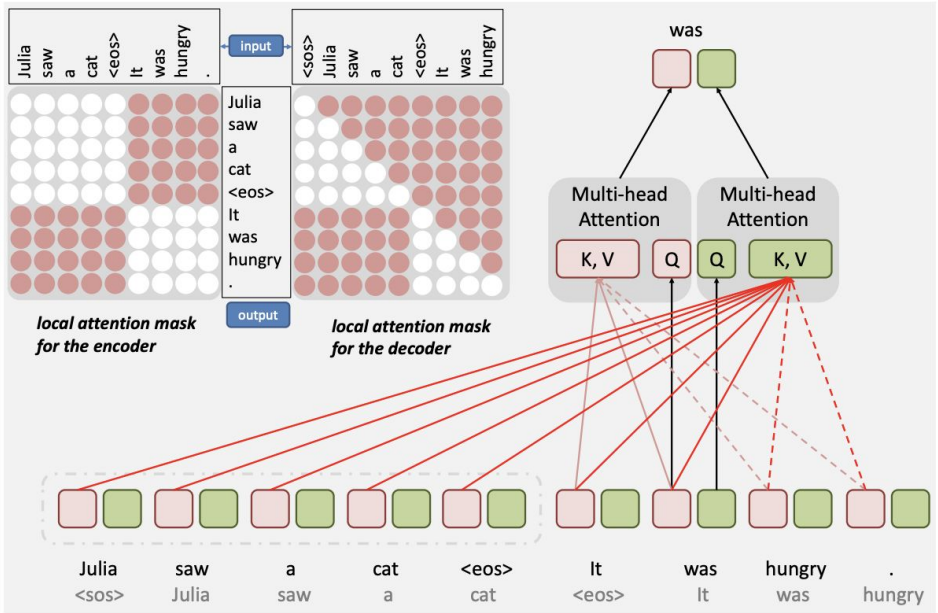


Illustration from Zhang et al. (2020)

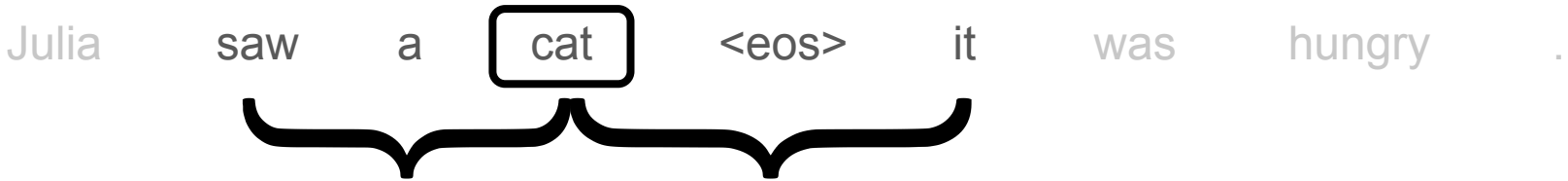
Document-level Machine Translation - Modifying Attention

Herold et al. (2023) proposes **window attention**.

Julia saw a cat <eos> it was hungry .

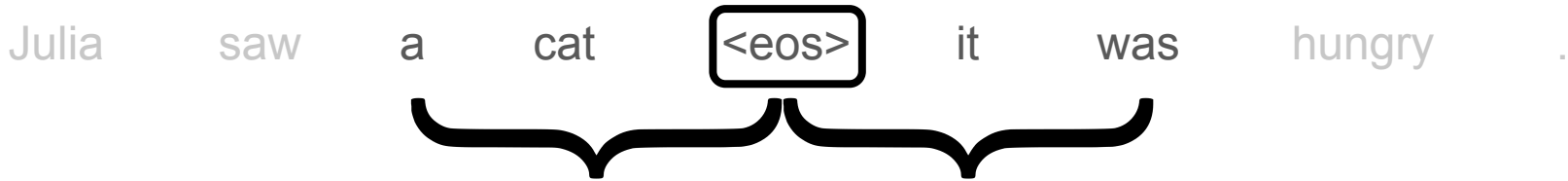
Document-level Machine Translation - Modifying Attention

Herold et al. (2023) proposes **window attention**.



Document-level Machine Translation - Modifying Attention

Herold et al. (2023) proposes **window attention**.



Document-level Machine Translation - Results

Model	Context	NEWS		TED		OS	
		newstest2018		tst2017		test	
		BLEU	TER	BLEU	TER	BLEU	TER
sent.-level (external)	0 sent.	†32.3	-	‡33.4	-	*37.3	-
sent.-level (ours)		32.8	49.0	34.2	46.3	37.1	**43.8**
concat adj.	2 sent.	**33.4**	48.6	34.3	46.3	38.2	43.9
	1000 tok.	29.5	53.7	32.1	48.4	38.1	46.0
LST-attn	1000 tok.	30.0	53.1	29.8	54.5	**38.5**	45.1
window-attn	1000 tok.	33.1	**48.1**	**34.6**	**45.8**	38.3	44.4

Illustration from Herold et al. (2023)

- Neural Machine Translation
  - Evaluating NMT: from BLEU to COMET
  - Sequence-to-Sequence models from scratch
  - Leveraging monolingual data
  - Zero-shot Machine translation
  - Data-centric approach
  - Document-level Machine translation
- **NLP tasks beyond English**
  - Multilingual vs Monolingual models

- Neural Machine Translation
  - Evaluating NMT: from BLEU to COMET
  - Sequence-to-Sequence models from scratch
  - Leveraging monolingual data
  - Zero-shot Machine translation
  - Data-centric approach
  - Document-level Machine translation
- **NLP tasks beyond English**
  - **Multilingual vs Monolingual models**

Multilingual vs Monolingual models

Current NLP systems are English-centric (ChatGPT, BARD, Gemini etc.)

- What about the thousands spoken languages left?

Before diving into multilingual models, let's consider 2 baselines:

1) Translate-Train
2) Translate-Test

Multilingual vs Monolingual models - Translate Train & Translate Test

Starting from a pretrained monolingual or multilingual LM

- **Zero-shot**:

Multilingual LM finetuned on English data and evaluated on other languages.

- **Translate Train**:

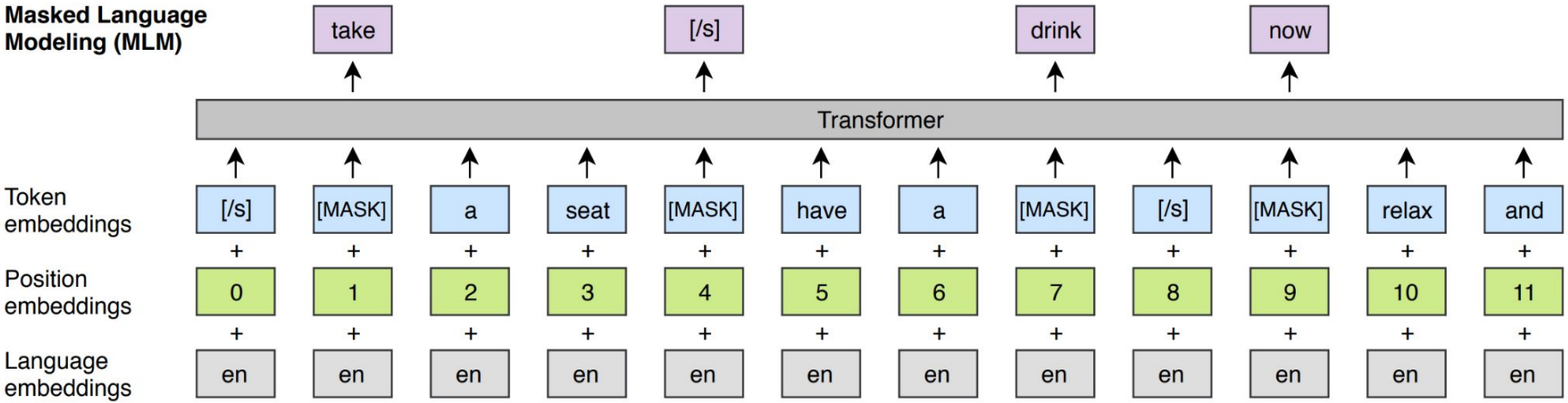
English training data is translated into target languages and the Multilingual LM is finetuned on the translated data

- **Translate Test**:

Translate the test sets into English and evaluate the multilingual LM or monolingual LM on the translated test sets.

Multilingual vs Monolingual models - XLM-R

Multilingual MLM objective - trained on 100 languages



NLP tasks beyond English

Illustration from Lample et al. (2019)

Multilingual vs Monolingual models - XLM-R

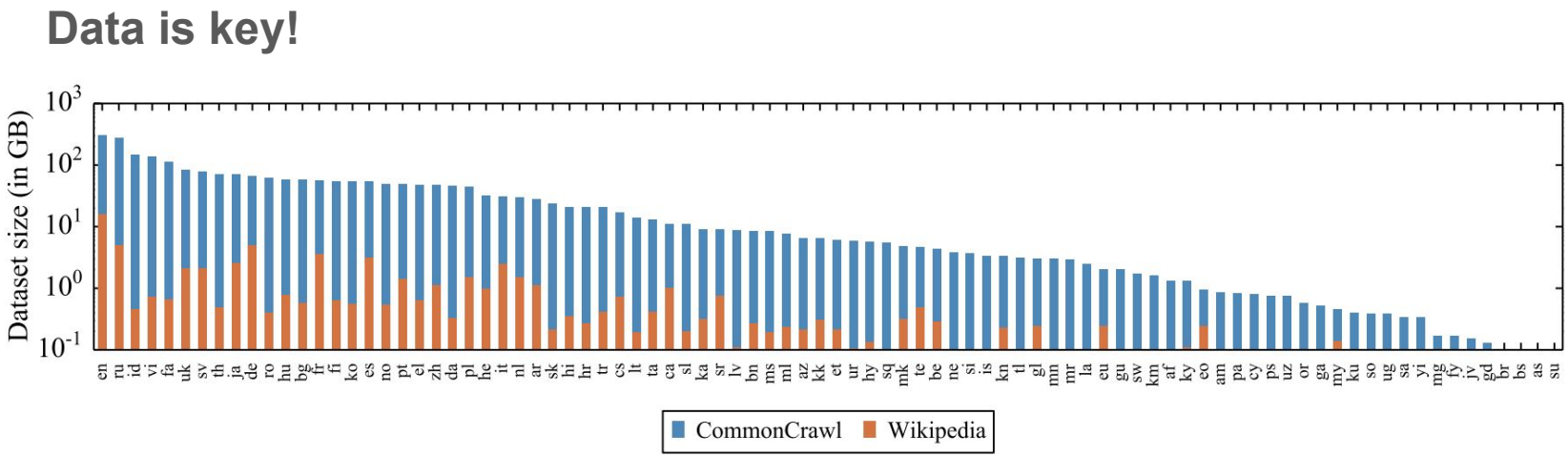


Figure 1: Amount of data in GiB (log-scale) for the 88 languages that appear in both the Wiki-100 corpus used for mBERT and XLM-100, and the CC-100 used for XLM-R. CC-100 increases the amount of data by several orders of magnitude, in particular for low-resource languages.

NLP tasks beyond English

Figure from Conneau et al. (2020)

Multilingual vs Monolingual models - XLM-R results

Model	D	#M	#lg	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Avg
Fine-tune multilingual model on English training set (Cross-lingual Transfer)																			
Lample and Conneau (2019)	Wiki+MT	N	15	85.0	78.7	78.9	77.8	76.6	77.4	75.3	72.5	73.1	76.1	73.2	76.5	69.6	68.4	67.3	75.1
Huang et al. (2019)	Wiki+MT	N	15	85.1	79.0	79.4	77.8	77.2	77.2	76.3	72.8	73.5	76.4	73.6	76.2	69.4	69.7	66.7	75.4
Devlin et al. (2018)	Wiki	N	102	82.1	73.8	74.3	71.1	66.4	68.9	69.0	61.6	64.9	69.5	55.8	69.3	60.0	50.4	58.0	66.3
Lample and Conneau (2019)	Wiki	N	100	83.7	76.2	76.6	73.7	72.4	73.0	72.1	68.1	68.4	72.0	68.2	71.5	64.5	58.0	62.4	71.3
Lample and Conneau (2019)	Wiki	1	100	83.2	76.7	77.7	74.0	72.7	74.1	72.7	68.7	68.6	72.9	68.9	72.5	65.6	58.2	62.4	70.7
XLM-R$_{\text{Base}}$	CC	1	100	85.8	79.7	80.7	78.7	77.5	79.6	78.1	74.2	73.8	76.5	74.6	76.7	72.4	66.5	68.3	76.2
XLM-R	CC	1	100	**89.1**	**84.1**	**85.1**	**83.9**	**82.9**	**84.0**	**81.2**	**79.6**	**79.8**	**80.8**	**78.1**	**80.2**	**76.9**	**73.9**	**73.8**	**80.9**
Translate everything to English and use English-only model (TRANSLATE-TEST)																			
BERT-en	Wiki	1	1	88.8	81.4	82.3	80.1	80.3	80.9	76.2	76.0	75.4	72.0	71.9	75.6	70.0	65.8	65.8	76.2
RoBERTa	Wiki+CC	1	1	**91.3**	82.9	84.3	81.2	81.7	83.1	78.3	76.8	76.6	74.2	74.1	77.5	70.9	66.7	66.8	77.8
Fine-tune multilingual model on each training set (TRANSLATE-TRAIN)																			
Lample and Conneau (2019)	Wiki	N	100	82.9	77.6	77.9	77.9	77.1	75.7	75.5	72.6	71.2	75.8	73.1	76.2	70.4	66.5	62.4	74.2
Fine-tune multilingual model on all training sets (TRANSLATE-TRAIN-ALL)																			
Lample and Conneau (2019)†	Wiki+MT	1	15	85.0	80.8	81.3	80.3	79.1	80.9	78.3	75.6	77.6	78.5	76.0	79.5	72.9	72.8	68.5	77.8
Huang et al. (2019)	Wiki+MT	1	15	85.6	81.1	82.3	80.9	79.5	81.4	79.7	76.8	78.2	77.9	77.1	80.5	73.4	73.8	69.6	78.5
Lample and Conneau (2019)	Wiki	1	100	84.5	80.1	81.3	79.3	78.6	79.4	77.5	75.2	75.6	78.3	75.7	78.3	72.1	69.2	67.7	76.9
XLM-R$_{\text{Base}}$	CC	1	100	85.4	81.4	82.2	80.3	80.4	81.3	79.7	78.6	77.3	79.7	77.9	80.2	76.1	73.1	73.0	79.1
XLM-R	CC	1	100	**89.1**	**85.1**	**86.6**	**85.7**	**85.3**	**85.9**	**83.5**	**83.2**	**83.1**	**83.7**	**81.5**	**83.7**	**81.6**	**78.0**	**78.1**	**83.6**

Table 1: **Results on cross-lingual classification.** We report the accuracy on each of the 15 XNLI languages and the average accuracy. We specify the dataset D used for pretraining, the number of models #M the approach requires and the number of languages #lg the model handles. Our *XLM-R* results are averaged over five different seeds. We show that using the translate-train-all approach which leverages training sets from multiple languages, *XLM-R* obtains a new state of the art on XNLI of 83.6% average accuracy. Results with † are from Huang et al. (2019).

NLP tasks beyond English

Figure from Conneau et al. (2020)

Multilingual vs Monolingual models in the target language
The French case

CamemBERT, a Tasty French language model trained on OSCAR (Martin et al. 2020)

Model	F1
SEM (CRF) (Dupont, 2017)	85.02
LSTM-CRF (Dupont, 2017)	85.57
mBERT (fine-tuned)	87.35
CamemBERT (fine-tuned)	<u>89.08</u>
LSTM+CRF+CamemBERT (embeddings)	**89.55**

Table 3: **NER** scores on the FTB (best model selected on validation out of 4). Best scores in bold, second best underlined.

Model	Acc.	#Params
mBERT (Devlin et al., 2019)	76.9	175M
$XLM_{MLM-TLM}$ (Lample and Conneau, 2019)	<u>80.2</u>	250M
$XLM\text{-}R_{BASE}$ (Conneau et al., 2019)	80.1	270M
CamemBERT (fine-tuned)	**82.5**	110M
Supplement: LARGE models		
$XLM\text{-}R_{LARGE}$ (Conneau et al., 2019)	<u>85.2</u>	550M
$CamemBERT_{LARGE}$ (fine-tuned)	**85.7**	335M

Table 4: **NLI** accuracy on the French XNLI test set (best model selected on validation out of 10). Best scores in bold, second best underlined.

Figures from Martin et al. (2020)

Multilingual vs Monolingual models in English - Closer look to Translate TEST

Do we need these huge multilingual LMs?

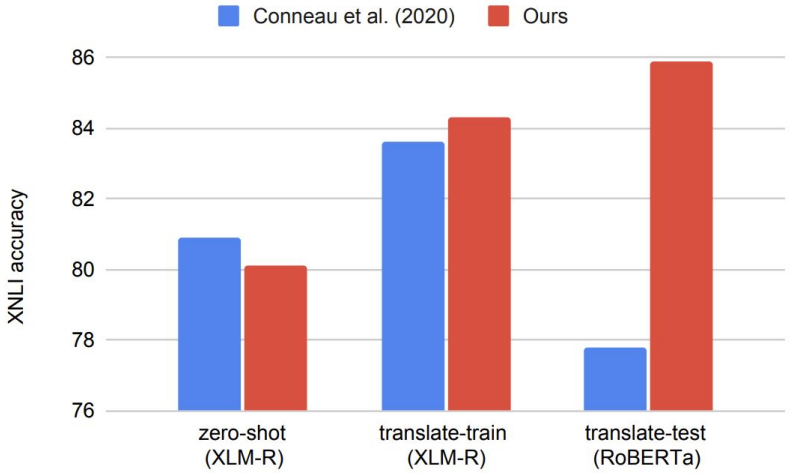


Figure 1: **XNLI accuracy.** We show that *translate-test* can do substantially better than previously reported.

Translate-TEST is highly sensitive to the choice of MT system. With SOTA MT (here NLLB 3.3B), monolingual model outperforms multilingual ones

Figure from Artexte et al. (2023)

Multilingual vs Monolingual models in English - Closer look to Translate TEST

			xnli	pwsx	marc	xcop	xsto	exm	avg
XLM-R	zero-shot		80.1	87.1	60.6	69.1	84.6	36.0	69.6
	translate-train		84.3	**90.7**	**60.8**	67.6	86.7	35.2	70.9
	translate-test	vanilla	79.3	86.9	58.0	68.6	84.8	34.9	68.8
		ours	84.6	89.3	58.8	69.0	87.9	35.0	70.8
RoBERTa	translate-test	vanilla	79.9	87.3	57.6	72.9	89.3	36.3	70.6
		ours	85.9	89.3	59.1	75.7	91.2	36.4	72.9
DeBERTa	translate-test	vanilla	81.0	87.1	58.2	77.7	92.1	**46.1**	73.7
		ours	**86.7**	90.3	59.2	**81.3**	**93.8**	46.0	**76.2**

Table 4: Main results. All systems use NLLB for MT. Best model results underlined, best overall results in **bold**.

Figure from Artexte et al. (2023)

Multilingual vs Monolingual models in English - Closer look to Translate TEST

How to achieve this result?

Reducing mismatch between training data (human generated) and test data (MT generated) in translate-TEST

1) MT adaptation:
   - Finetune NLLB on training data (in English) with backtranslation (domain transfer)
   - Adapt NLLB to documents by concatenating all sentences

2) Training data adaptation:
   - Translate training data into another language + back into English. Then combine with the original training data.

Figure from Artexte et al. (2023)