Multilingual Language Modeling

Ling 575j: Deep Learning for NLP

C.M. Downey

Spring 2023

Roadmap

- Modern multilingual models
  - Motivation
  - Architecture (XLM)
  - Zero-shot transfer
- Evaluation
  - How do they work? (spoiler: we don’t really know)
  - How cross-lingual are they?
  - Benchmarks

Roadmap cont.

- Representation alignment
- Transfering monolingual models
- Newer work

Motivation

- NLP applications are deployed to **large varities of languages/localities**
- **Prohibitively expensive** to train a new model for every language/variety
- **Translation** is especially intractable
    - n languages leads to **n^2 language pairs**
    - Introducing a “hub” language more likely to result in translation artifacts
- Idea: train a model that can encode **all languages you plan to use**

Modeling

XLM

Cross-lingual Language Model Pretraining

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

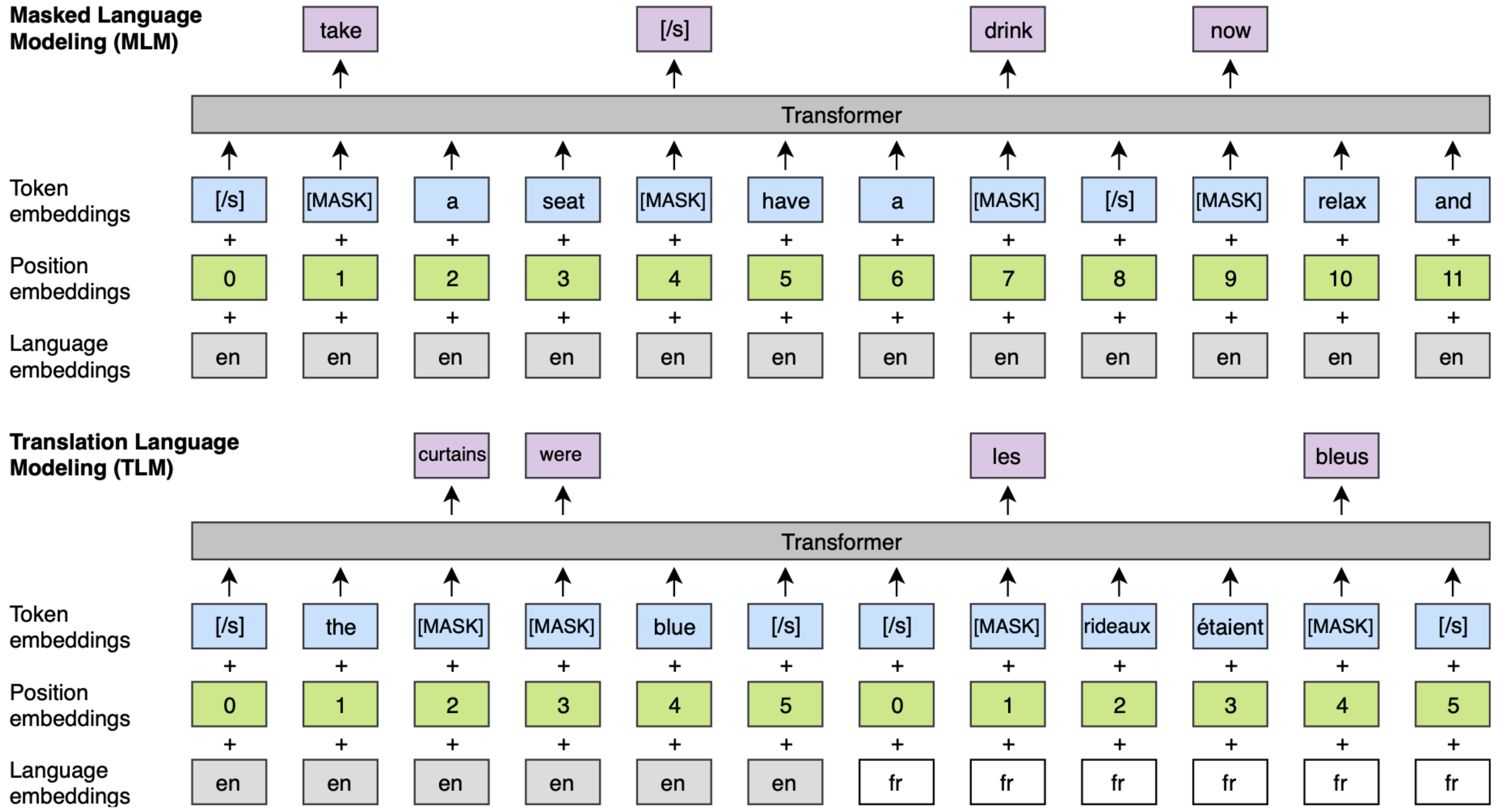
Guillaume Lample*
Facebook AI Research
Sorbonne Universités
glample@fb.com

Abstract

Recent studies have demonstrated the efficiency of generative pretraining for English natural language understanding. In this work, we extend this approach to multiple languages and show the effectiveness of cross-lingual pretraining. We propose two methods to learn cross-lingual language models (XLMs): one unsupervised that only relies on monolingual data, and one supervised that leverages parallel data with a new cross-lingual language model objective. We obtain state-of-the-art results on cross-lingual classification, unsupervised and supervised machine translation. On XNLI, our approach pushes the state of the art by an absolute gain of 4.9% accuracy. On unsupervised machine translation, we obtain 34.3 BLEU on WMT'16 German-English, improving the previous state of the art by more than 9 BLEU. On supervised machine translation, we obtain a new state of the art of 38.5 BLEU on WMT'16 Romanian-English, outperforming the previous best approach by more than 4 BLEU. Our code and pretrained models are publicly available[1].

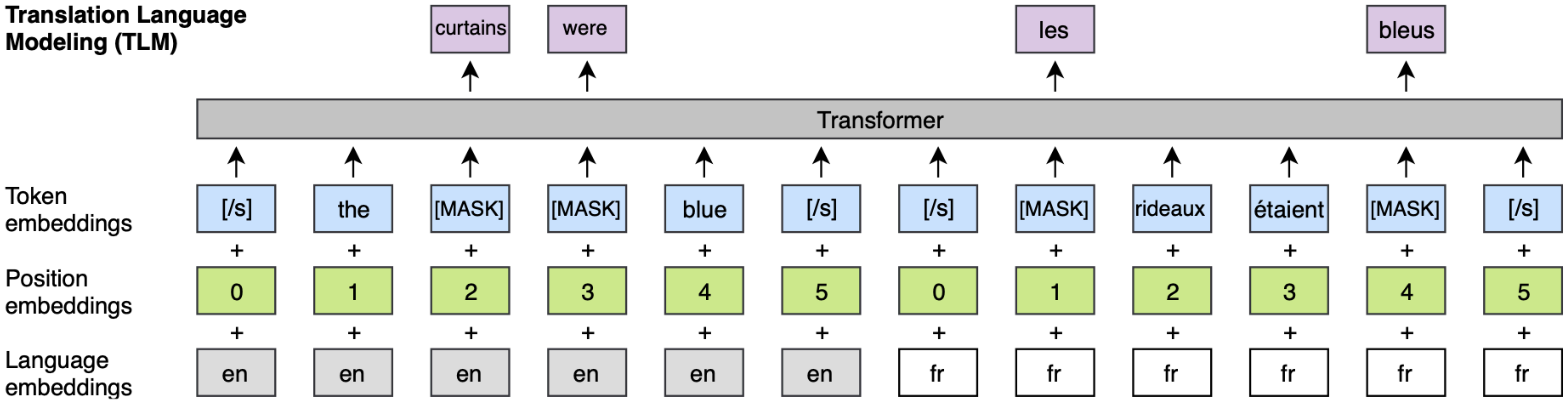
XLM

- Key Ideas
  - Use a **shared subword vocabulary** across languages
  - Do normal language modeling on the **combined language sets**
  - If parallel data is available, do **Translation Language Modeling (TLM)**



XLM: TLM

- TLM == MLM with **concatenated parallel sentences**
- Idea: use each language to help predict the other



“language embeddings” added along with positional

XLM: Results

- Racks up improvements. Better initializations for:
  - Crosslingual classification (XLNI)
  - Translation
  - Low-resource LMs
  - Crosslingual word embeddings

		en-fr	fr-en	en-de	de-en	en-ro	ro-en
Previous state-of-the-art - Lampple et al. [26]							
NMT		25.1	24.2	17.2	21.0	21.2	19.4
PBSMT		28.1	27.2	17.8	22.7	21.3	23.0
PBSMT + NMT		27.6	27.7	20.2	25.2	25.1	23.9
Our results for different encoder and decoder initializations							
-	-	13.0	15.8	6.7	15.3	18.9	18.3
EMB	EMB	29.4	29.4	21.3	27.3	27.5	26.6
CLM	CLM	30.4	30.0	22.7	30.5	29.0	27.8
MLM	MLM	**33.4**	**33.3**	**26.4**	**34.3**	**33.3**	**31.8**
CLM	-	28.7	28.2	24.4	30.3	29.2	28.0
MLM	-	31.6	32.1	**27.0**	33.2	31.8	30.5
-	CLM	25.3	26.4	19.2	26.0	25.7	24.6
-	MLM	29.2	29.1	21.6	28.6	28.2	27.3
CLM	MLM	32.3	31.6	24.3	32.5	31.6	29.8
MLM	CLM	**33.4**	32.3	24.9	32.9	31.7	30.4

Table 2: **Results on unsupervised MT.** BLEU scores on WMT'14 English-French, WMT'16 German-English and WMT'16 Romanian-English. For our results, the first two columns indicate the model used to pretrain the encoder and the decoder. " - " means the model was randomly initialized. EMB corresponds to pretraining the lookup table with cross-lingual embeddings, CLM and MLM correspond to pretraining with models trained on the CLM or MLM objectives.

	Cosine sim.	L2 dist.	SemEval'17
MUSE	0.38	5.13	0.65
Concat	0.36	4.89	0.52
XLM	**0.55**	**2.64**	**0.69**

Table 5: **Unsupervised cross-lingual word embeddings** Cosine similarity and L2 distance between source words and their translations. Pearson correlation on SemEval'17 cross-lingual word similarity task of Camacho-Collados et al. [8].

XLM: XNLI Results

- XNLI = Cross-lingual Natural Language Inference
  - i.e. does sentence A *entail* sentence B, *contradict* it, or *neither*?

	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Δ
Machine translation baselines (TRANSLATE-TRAIN)																
Devlin et al. [14]	81.9	-	77.8	75.9	-	-	-	-	70.7	-	-	76.6	-	-	61.6	-
XLM (MLM+TLM)	85.0	80.2	80.8	80.3	78.1	79.3	78.1	74.7	76.5	76.6	75.5	78.6	72.3	70.9	63.2	76.7
Machine translation baselines (TRANSLATE-TEST)																
Devlin et al. [14]	81.4	-	74.9	74.4	-	-	-	-	70.4	-	-	70.1	-	-	62.1	-
XLM (MLM+TLM)	85.0	79.0	79.5	78.1	77.8	77.6	75.5	73.7	73.7	70.8	70.4	73.6	69.0	64.7	65.1	74.2
Evaluation of cross-lingual sentence encoders																
Conneau et al. [12]	73.7	67.7	68.7	67.7	68.9	67.9	65.4	64.2	64.8	66.4	64.1	65.8	64.1	55.7	58.4	65.6
Devlin et al. [14]	81.4	-	74.3	70.5	-	-	-	-	62.1	-	-	63.8	-	-	58.3	-
Artetxe and Schwenk [4]	73.9	71.9	72.9	72.6	73.1	74.2	71.5	69.7	71.4	72.0	69.2	71.4	65.5	62.2	61.0	70.2
XLM (MLM)	83.2	76.5	76.3	74.2	73.1	74.0	73.1	67.8	68.5	71.2	69.2	71.9	65.7	64.6	63.4	71.5
XLM (MLM+TLM)	**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**

XLM: XNLI Baselines

- Translate-Train: translate **English** training data into the **target language**
- Translate-Test: translate **target** test set into **English**

	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Δ
Machine translation baselines (TRANSLATE-TRAIN)																
Devlin et al. [14]	81.9	-	77.8	75.9	-	-	-	-	70.7	-	-	76.6	-	-	61.6	-
XLM (MLM+TLM)	85.0	80.2	80.8	80.3	78.1	79.3	78.1	74.7	76.5	76.6	75.5	78.6	72.3	70.9	63.2	76.7
Machine translation baselines (TRANSLATE-TEST)																
Devlin et al. [14]	81.4	-	74.9	74.4	-	-	-	-	70.4	-	-	70.1	-	-	62.1	-
XLM (MLM+TLM)	85.0	79.0	79.5	78.1	77.8	77.6	75.5	73.7	73.7	70.8	70.4	73.6	69.0	64.7	65.1	74.2
Evaluation of cross-lingual sentence encoders																
Conneau et al. [12]	73.7	67.7	68.7	67.7	68.9	67.9	65.4	64.2	64.8	66.4	64.1	65.8	64.1	55.7	58.4	65.6
Devlin et al. [14]	81.4	-	74.3	70.5	-	-	-	-	62.1	-	-	63.8	-	-	58.3	-
Artetxe and Schwenk [4]	73.9	71.9	72.9	72.6	73.1	74.2	71.5	69.7	71.4	72.0	69.2	71.4	65.5	62.2	61.0	70.2
XLM (MLM)	83.2	76.5	76.3	74.2	73.1	74.0	73.1	67.8	68.5	71.2	69.2	71.9	65.7	64.6	63.4	71.5
XLM (MLM+TLM)	**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**

Zero-shot Transfer

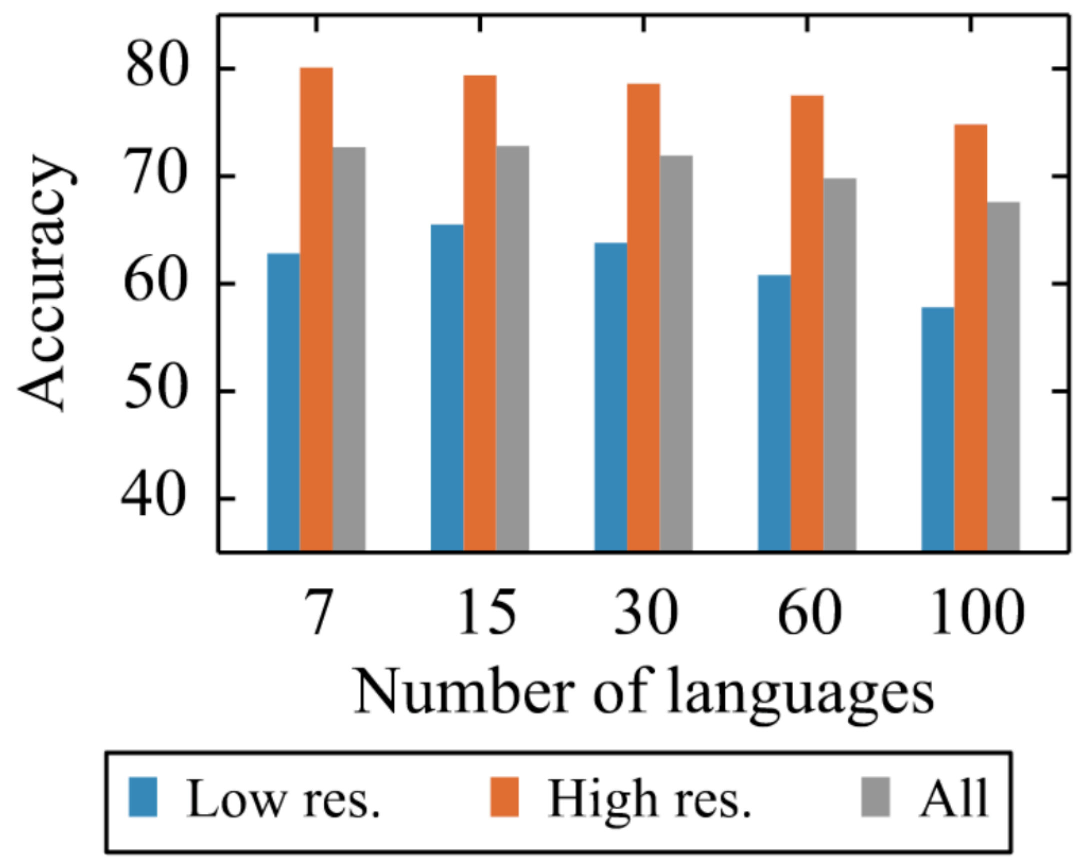
- The ability to do **zero-shot transfer** is probably the greatest strength of crosslingual models
- This setting assumes
  - Training set of **plain text in several languages** OR a **pre-trained multilingual model**
  - Training data for **downstream task**, but only in **English / other high-resource language**
- Process: get crosslingual model, *fine-tune* it on English task data, then **directly apply it to the task in a new language**

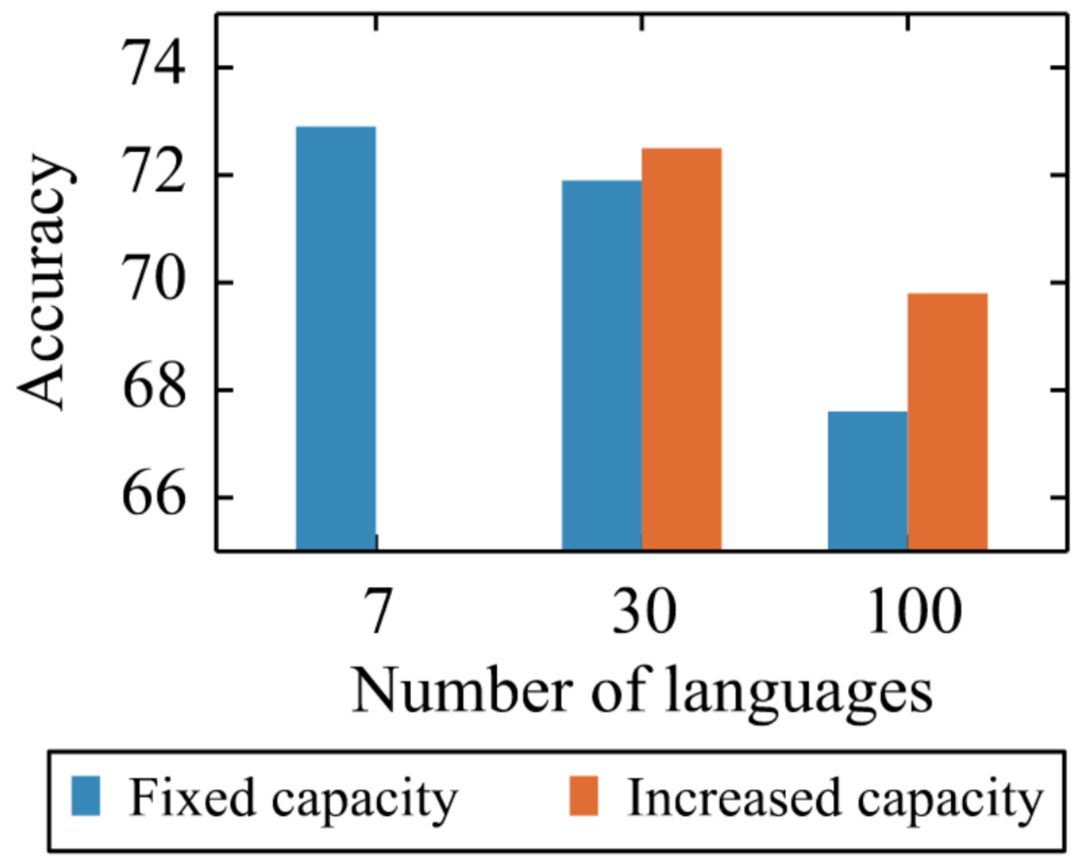
Since XLM

- A lot has happened since XLM
    - XLM-R, mBART, XGLM, BLOOM
- May even be considered an “old” model at this point
- **However**
    - Most subsequent models have re-used the same basic ideas
    - Understanding this paper is a good way to understand others
    - (TLM has stopped being used)

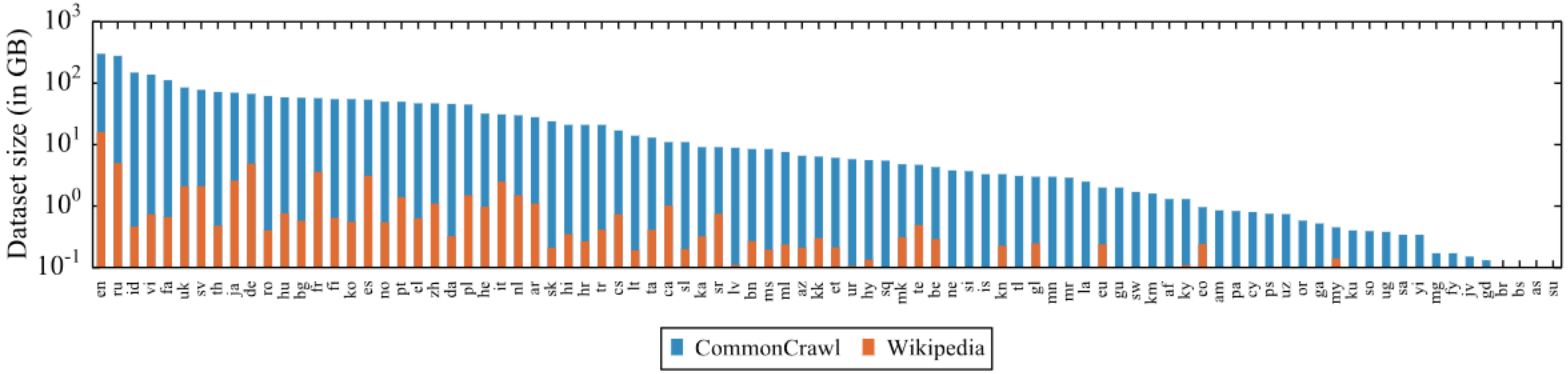
"Curse of Multilinguality"

- The more languages a model covers, the worse it performs for individual languages
- "Crosslingual" models have become huge
- Best performance still comes when you have enough data to train a monolingual model
  - Most languages do not have enough





Language Up/Down-sampling



Questions after XLM

- How do multilingual models work?
- How much data do you need for each language?
- How do you evaluate multilingual models?
- Do these work well for truly low-resource languages?

Analysis

Conneau et al. (2020)

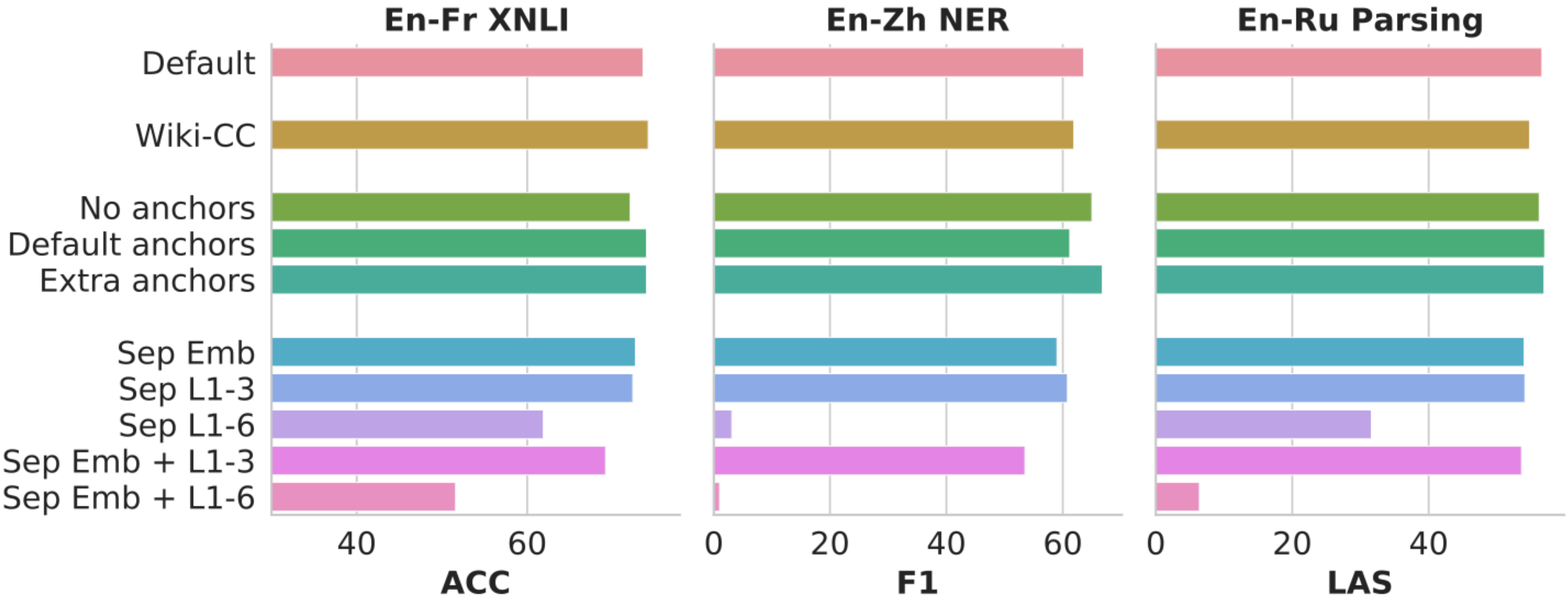
Emerging Cross-lingual Structure in Pretrained Language Models

Alexis Conneau[♡*] **Shijie Wu**[♠*]
Haoran Li[♡] **Luke Zettlemoyer**[♡] **Veselin Stoyanov**[♡]
[♠]Department of Computer Science, Johns Hopkins University
[♡]Facebook AI
aconneau@fb.com, shijie.wu@jhu.edu
{aimeeli,lsz,ves}@fb.com

Conneau et al. (2020)

- A great paper which I recommend, but somewhat involved
- Takeaways
  - Languages do **not need to share vocabulary** to get good performance
  - Only **about half the layers** need to be shared between languages
  - Monolingual BERTs trained for different languages create **similar embeddings** (especially at lower layers)
  - Similar languages have **similar BERT embeddings**

Conneau et al. (2020)



Conneau et al. (2020)

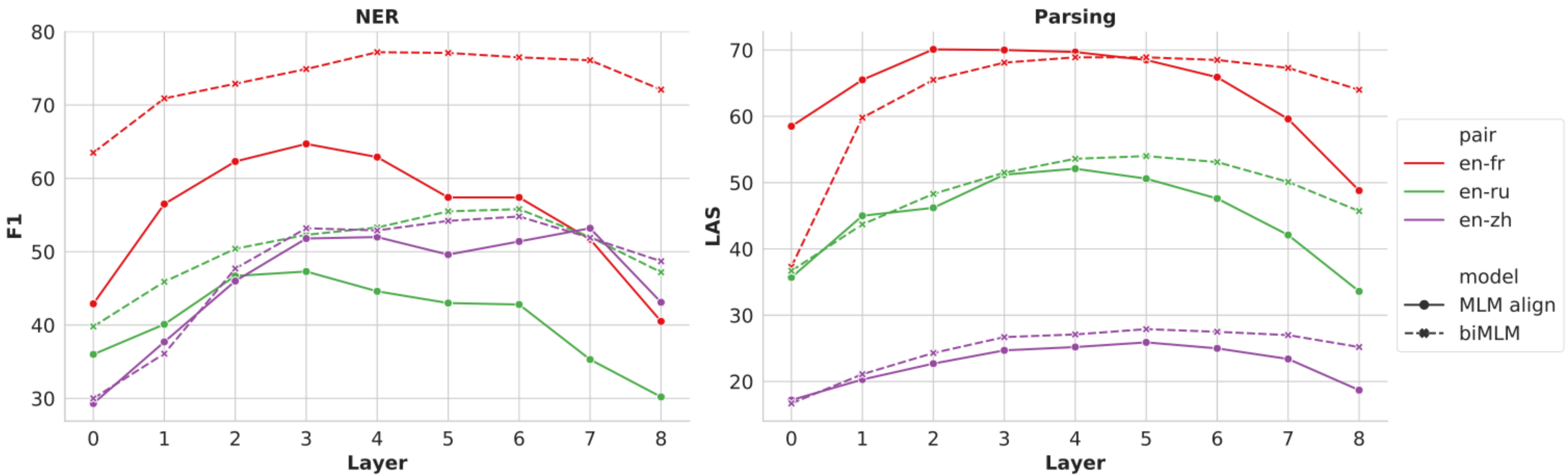


Figure 5: Contextual representation alignment of different layers for zero-shot cross-lingual transfer.

Conneau et al. (2020)

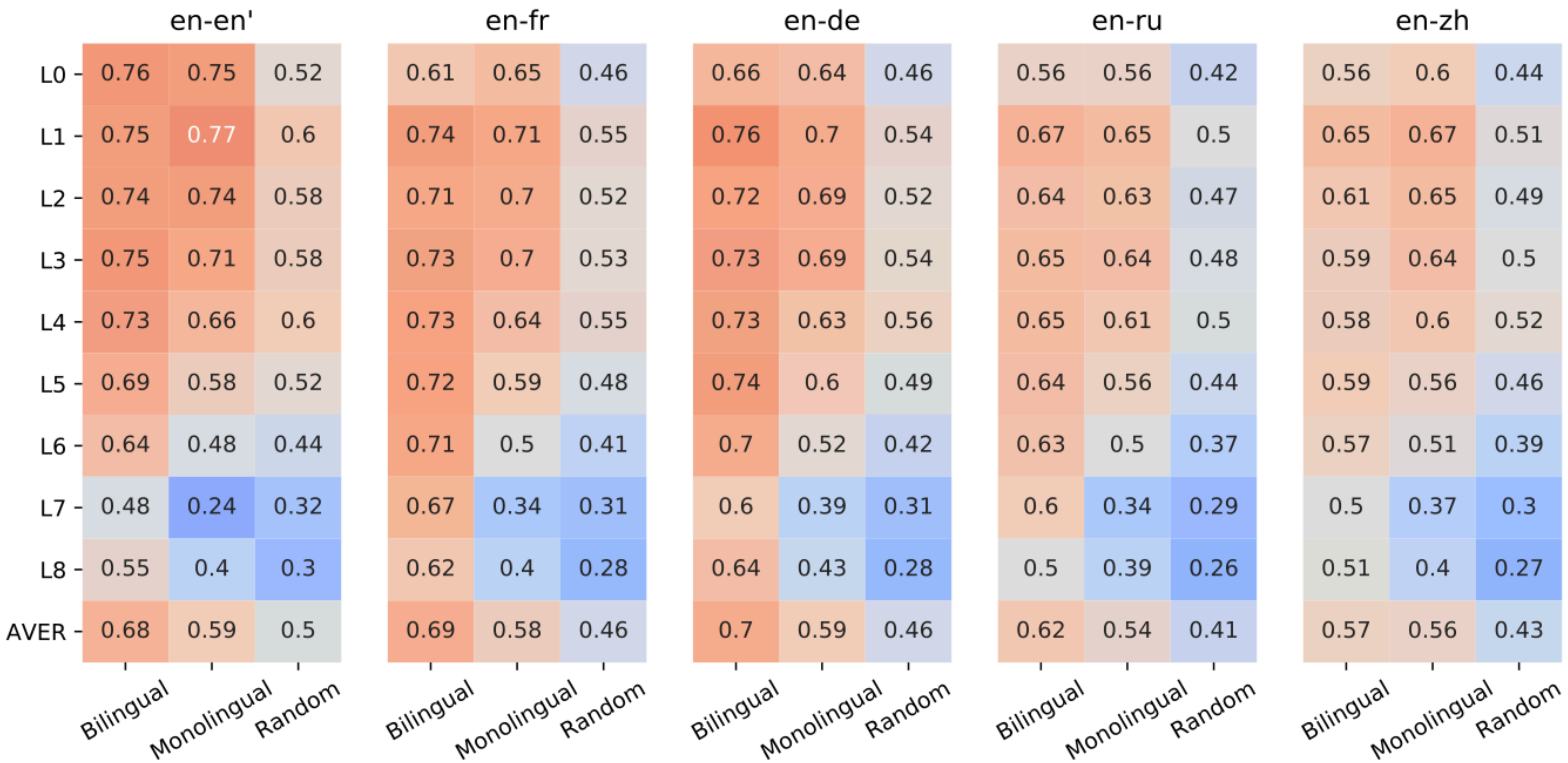


Figure 7: CKA similarity of mean-pooled multi-way parallel sentence representation at each layers. Note en′ corresponds to paraphrases of en obtained from back-translation (en-fr-en′). Random encoder is only used by non-Engligh sentences. L0 is the embeddings layers while L1 to L8 are the corresponding transformer layers. The average row is the average of 9 (L0-L8) similarity measurements.

Wu and Dredze (2020)

Are All Languages Created Equal in Multilingual BERT?

Shijie Wu and **Mark Dredze**
Department of Computer Science
Johns Hopkins University
shijie.wu@jhu.edu, mdredze@cs.jhu.edu

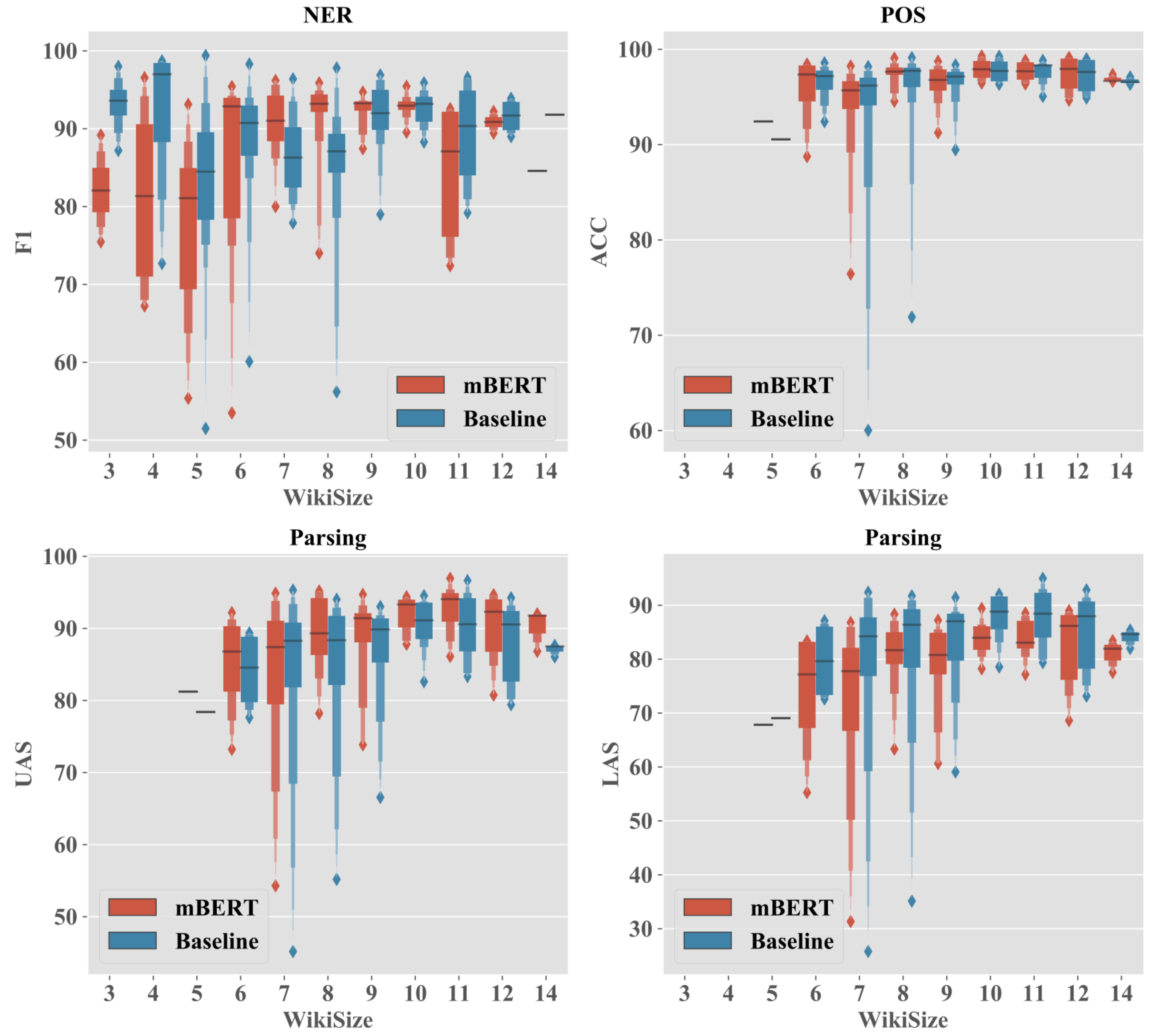
Wu and Dredze (2020)

- “Are all languages created equal in mBERT?”
- Short answer: **no**
- **“mBERT does better than or comparable to baselines on high resource languages but does much worse on low resource languages”**

WikiSize	Languages	# Languages	Size Range (GB)
3	io, pms, scn, **yo**	4	[0.006, 0.011]
4	cv, lmo, mg, min, su, vo	6	[0.011, 0.022]
5	an, bar, br, ce, fy, ga, gu, is, jv, ky, lb, **mn**, my, nds, ne, pa, pnb, sw, tg	19	[0.022, 0.044]
6	**af**, ba, cy, kn, la, mr, oc, sco, sq, tl, tt, uz	12	[0.044, 0.088]
7	az, bn, bs, eu, hi, ka, kk, lt, **lv**, mk, ml, nn, ta, te, ur	15	[0.088, 0.177]
8	ast, be, bg, da, el, et, gl, hr, hy, ms, sh, sk, sl, th, war	15	[0.177, 0.354]
9	fa, fi, he, id, ko, no, ro, sr, tr, vi	10	[0.354, 0.707]
10	ar, ca, cs, hu, nl, sv, uk	7	[0.707, 1.414]
11	ceb, it, ja, pl, pt, zh	6	[1.414, 2.828]
12	de, es, fr, ru	4	[2.828, 5.657]
14	en	1	[11.314, 22.627]

Table 1: List of 99 languages we consider in mBERT and its pretraining corpus size. Languages in **bold** are the languages we consider in §5.

Wu and Dredze (2020)



Evaluation

XTREME

XTREME

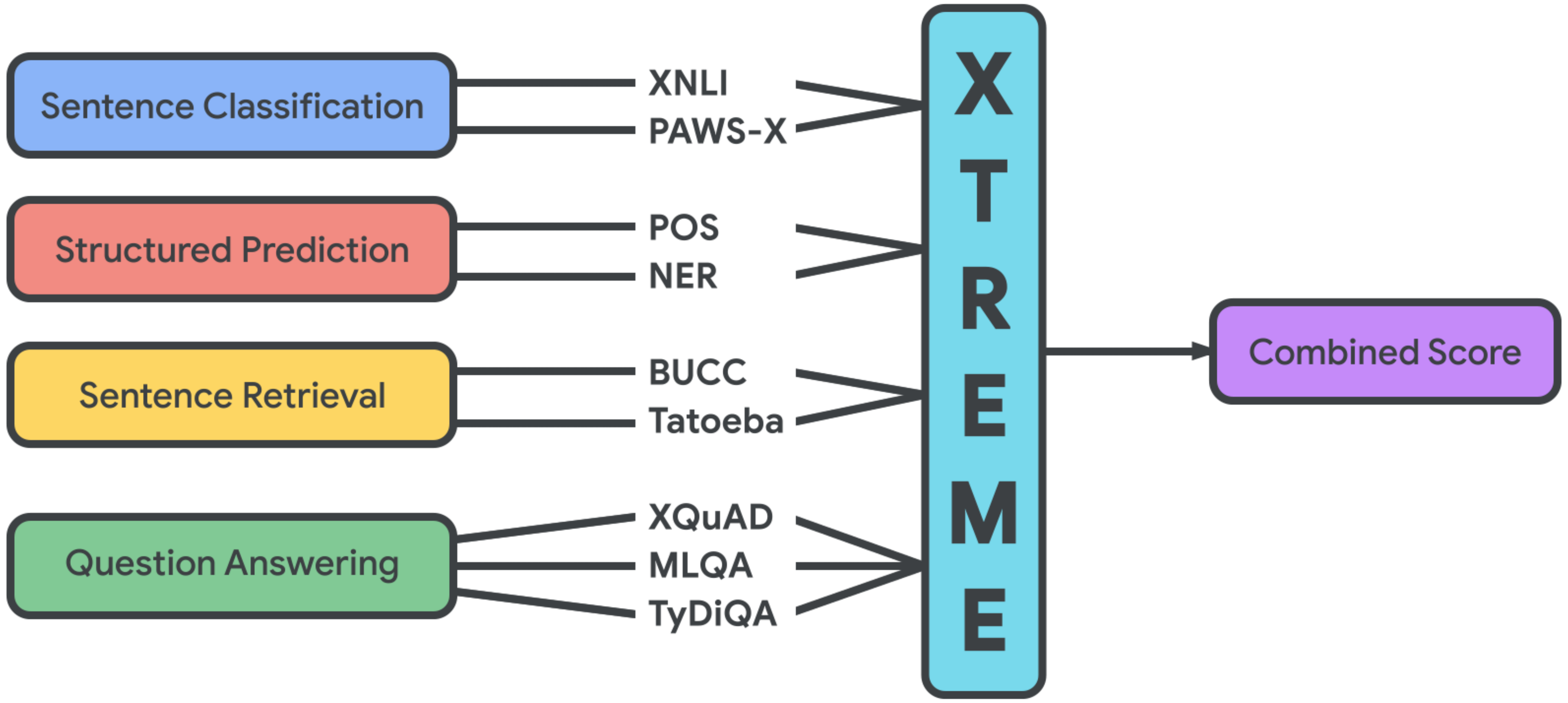
(X) Cross-Lingual **Tr**ansfer **E**valuation of **M**ultilingual **E**ncoders

A comprehensive benchmark for cross-lingual transfer learning on a diverse set of languages and tasks.

(I hate names like
this but oh well)

XTREME

- Like GLUE but for multilingual models
- Nine tasks
  - 3 Question-Answering
  - XNLI
  - Paraphrase detection (PAWS-X)
  - POS
  - NER
  - 2 Bitext mining (BUCC and Tatoeba)



Representation Alignment

Alignment Motivation

- May be desirable to **explicitly align** a model's vector representations between languages
- e.g. classification
  - If the representation in language A gives the correct outcome, logical that **having similar representations for other languages** should also give the correct outcome
- For tasks like bitext mining, paraphrase detection, and dictionary induction, **alignment is the whole point**

Xing et al. (2015)

Normalized Word Embedding and Orthogonal Transform for Bilingual Word Translation

Chao Xing
CSLT, Tsinghua University
Beijing Jiaotong University
Beijing, P.R. China

Dong Wang*
CSLT, RIIT, Tsinghua University
TNList, China
Beijing, P.R. China

Chao Liu
CSLT, RIIT, Tsinghua University
CS Department, Tsinghua University
Beijing, P.R. China

Yiye Lin
CSLT, RIIT, Tsinghua University
Beijing Institute of Technology
Beijing, P.R. China

Xing et al. (2015)

- Common hypothesis that vector spaces should be **approximately isomorphic** between languages
  - This implies an **invertible linear mapping W** between the space of one language and another
  - Xing et al. argue that this transformation should be an **orthogonal** one (i.e. a rotation or reflection of space)
  - An orthogonal transformation W can be computed with the **Orthogonal Procrustes** method
- Alignment work is often centered on learning and refining W

Conneau et al. (2018)

WORD TRANSLATION WITHOUT PARALLEL DATA

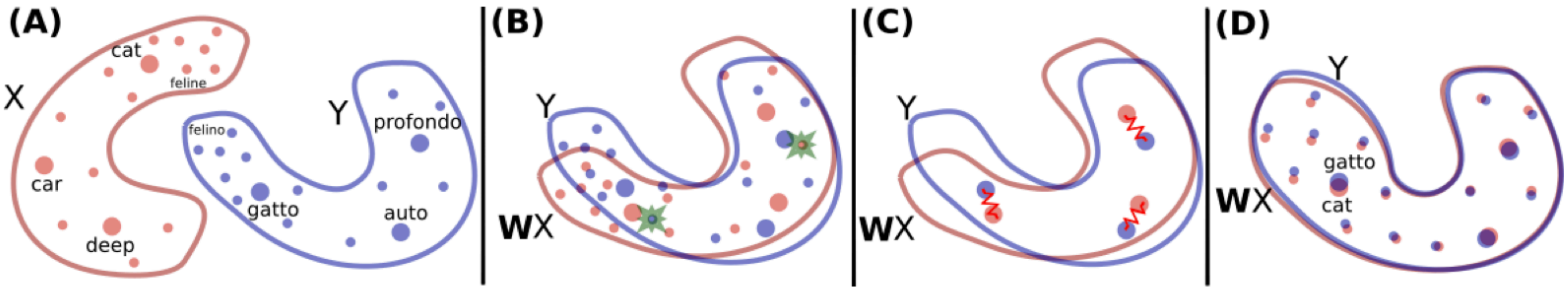
Alexis Conneau[*†‡], **Guillaume Lample**[*†§],
Marc'Aurelio Ranzato[†], **Ludovic Denoyer**[§], **Hervé Jégou**[†]
{aconneau,glample,ranzato,rvj}@fb.com
ludovic.denoyer@upmc.fr

ABSTRACT

State-of-the-art methods for learning cross-lingual word embeddings have relied on bilingual dictionaries or parallel corpora. Recent studies showed that the need for parallel data supervision can be alleviated with character-level information. While these methods showed encouraging results, they are not on par with their supervised counterparts and are limited to pairs of languages sharing a common alphabet. In this work, we show that we can build a bilingual dictionary between two languages without using any parallel corpora, by aligning monolingual word embedding spaces in an unsupervised way. Without using any character information, our model even outperforms existing supervised methods on cross-lingual tasks for some language pairs. Our experiments demonstrate that our method works very well also for distant language pairs, like English-Russian or English-Chinese. We finally describe experiments on the English-Esperanto low-resource language pair, on which there only exists a limited amount of parallel data, to show the potential impact of our method in fully unsupervised machine translation. Our code, embeddings and dictionaries are publicly available[1].

Conneau et al. (2018)

- A: Monolingual vector spaces
- B: Adversarial methods to bring distributions closer
- C: Orthogonal Procrustes
- D: Final aligned vector spaces



Tien and Steinert-Threlkeld (2021)

Bilingual alignment transfers to multilingual alignment for unsupervised parallel text mining

Chih-chan Tien
University of Washington
cctien@uw.edu

Shane Steinert-Threlkeld
University of Washington
shanest@uw.edu

Tien and Steinert-Threlkeld (2021)

- **Cycle Consistency Loss**: how **invertible** is the mapping between one language and another?
- Adversarial loss: can a discriminator tell the difference between language representations?

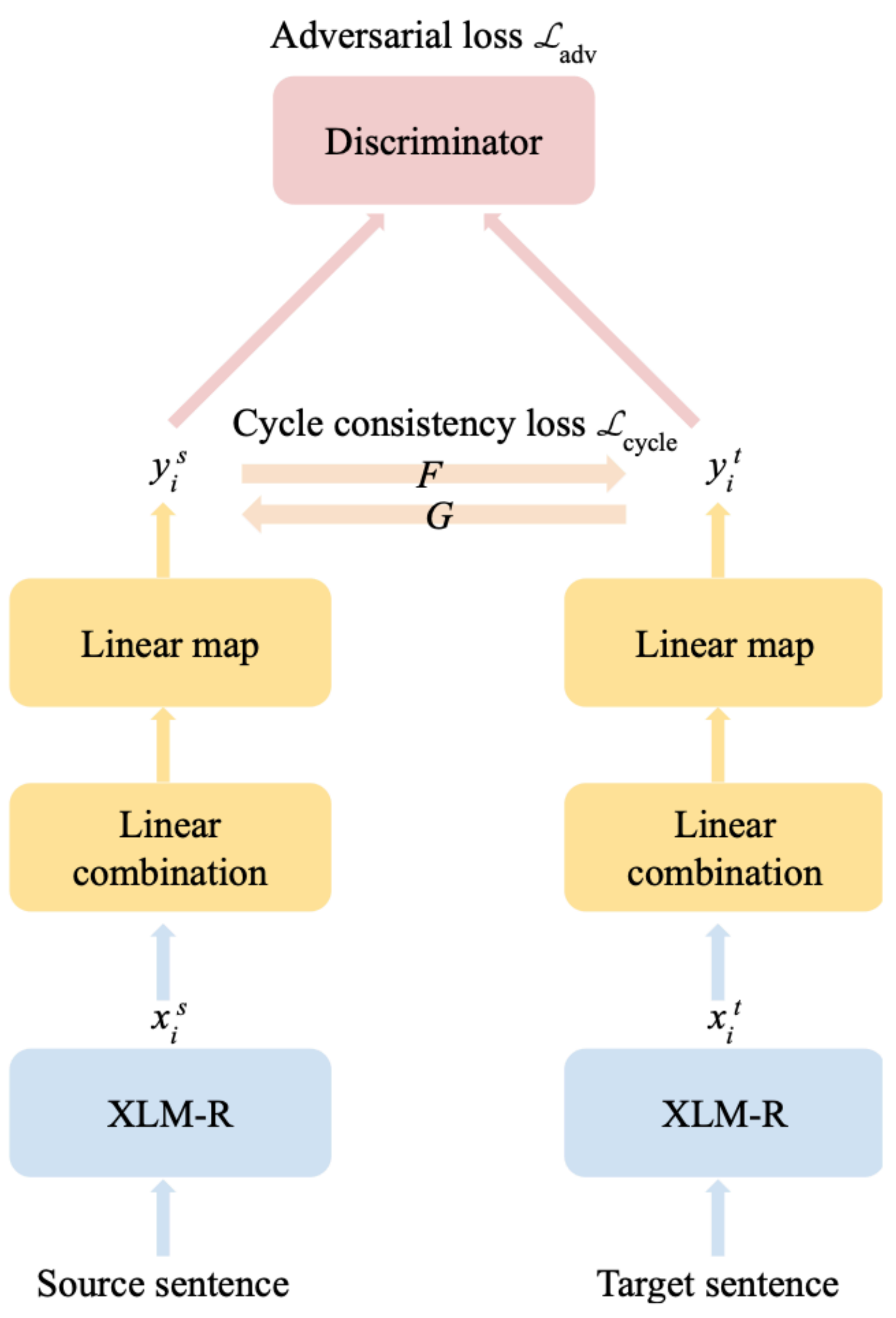


Figure 1: Schematic representation of the unsupervised model with the adversarial loss and the cycle consistency loss.

Alignment Final Thoughts

- Can also just add **difference** between language embeddings as **loss term**
- **Batch normalization** has been shown to be helpful
- Alignment is tricky in general. Often does not work as expected

Monolingual Transfer

Artetxe, Ruder, and Yogatama (2020)

- How transferable to other languages is a **monolingual model?**
- Main idea
  - Train a model on a **high-resource language**
  - **Freeze** transformer layers, initialize **new embeddings/vocab**, train on new language
  - Add in small “adapter layers” between transformer blocks
- Works **strangely well**

Artetxe, Ruder, and Yogatama (2020)

		en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	avg
Prev work	mBERT	81.4	-	74.3	70.5	-	-	-	-	62.1	-	-	63.8	-	-	58.3	-
	XLM (MLM)	83.2	76.5	76.3	74.2	73.1	74.0	73.1	67.8	68.5	71.2	69.2	71.9	65.7	64.6	63.4	71.5
CLWE	300d ident	82.1	67.6	69.0	65.0	60.9	59.1	59.5	51.2	55.3	46.6	54.0	58.5	48.4	35.3	43.0	57.0
	300d unsup	82.1	67.4	69.3	64.5	60.2	58.4	59.2	51.5	56.2	36.4	54.7	57.7	48.2	36.2	33.8	55.7
	768d ident	**82.4**	**70.7**	71.1	**67.6**	**64.2**	61.4	**63.3**	**55.0**	**58.6**	**50.7**	**58.0**	**60.2**	54.8	34.8	**48.1**	**60.1**
	768d unsup	**82.4**	70.4	**71.2**	67.4	63.9	**62.8**	**63.3**	54.8	58.3	49.1	57.2	55.7	**54.9**	**35.0**	33.9	58.7
JOINT MULTI	32k voc	79.0	71.5	72.2	68.5	66.7	66.9	66.5	58.4	64.4	66.0	62.3	66.4	59.1	50.4	56.9	65.0
	64k voc	80.7	72.8	73.0	69.8	69.6	69.5	68.8	63.6	66.1	67.2	64.7	66.7	63.2	52.0	59.0	67.1
	100k voc	81.2	74.5	74.4	72.0	72.3	71.2	70.0	65.1	69.7	68.9	66.4	68.0	64.2	55.6	62.2	69.0
	200k voc	**82.2**	**75.8**	**75.7**	**73.4**	**74.0**	**73.1**	**71.8**	**67.3**	**69.8**	**69.8**	**67.7**	**67.8**	**65.8**	**60.9**	**62.3**	**70.5**
JOINT PAIR	Joint voc	82.2	74.8	76.4	73.1	72.0	71.8	70.2	67.9	68.5	**71.4**	**67.7**	70.8	64.5	**64.2**	**60.6**	70.4
	Disjoint voc	**83.0**	**76.2**	**77.1**	**74.4**	**74.4**	**73.7**	**72.1**	**68.8**	**71.3**	70.9	66.2	**72.5**	**66.0**	62.3	58.0	**71.1**
MONO TRANS	Token emb	83.1	73.3	73.9	71.0	70.3	71.5	66.7	64.5	66.6	68.2	63.9	66.9	61.3	58.1	57.3	67.8
	+ pos emb	**83.8**	74.3	75.1	71.7	72.6	72.8	68.8	66.0	68.6	**69.8**	65.7	69.7	61.1	58.8	58.3	69.1
	+ noising	81.7	74.1	75.2	72.6	**72.9**	73.1	70.2	68.1	70.2	69.1	**67.7**	**70.6**	62.5	**62.5**	**60.2**	**70.0**
	+ adapters	81.7	**74.7**	**75.4**	**73.0**	72.0	**73.7**	**70.4**	**69.9**	**70.6**	69.5	65.1	70.3	**65.2**	59.6	51.7	69.5

Artetxe, Ruder, and Yogatama (2020)

- Advantages
  - **Very cheap** — can take a model off the shelf and just re-train embeddings
  - Does **comparably to crosslingual models**
- Caveats
  - Not so many replicating studies
  - This paper transferred to *fairly* high-resource languages

Recent Work

mBART

Multilingual Denoising Pre-training for Neural Machine Translation

Yinhan Liu[‡*], Jiatao Gu[†*], Naman Goyal[†*], Xian Li[†], Sergey Edunov[†], Marjan Ghazvininejad[†], Mike Lewis[†], and Luke Zettlemoyer[‡]

[†]Facebook AI
[‡]Birch Technology
[†]{jgu,naman,xianl,edunov,ghazvini,mikelewis,lsz}@fb.com
[‡]yinhan@birch.ai

mBART

- Seq2Seq transformer
  - Trained to reconstruct a corrupted/masked sentence
  - Multilingual, but no “crosslingual signal” w/ parallel sentences during pre-training
- Very good for initializing translation systems

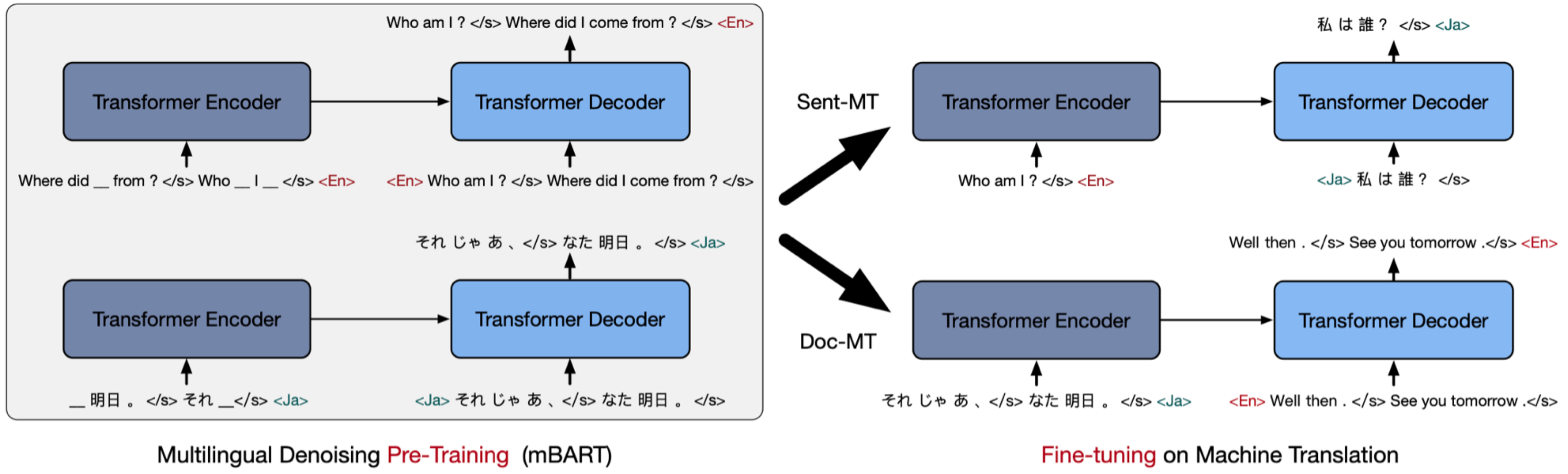


Figure 1: Framework for our Multilingual Denoising Pre-training (left) and fine-tuning on downstream MT tasks (right), where we use (1) sentence permutation (2) word-span masking as the injected noise. A special language id token is added at both the encoder and decoder. One multilingual pre-trained model is used for all tasks.

XGLM

- Decoder-only ("causal/generative") tranformer LM
- 564M-7.5B parameters
- Emphasis on doing the type of in-context learning seen with GPT-3
- From Meta

Few-shot Learning with Multilingual Generative Language Models

Xi Victoria Lin*, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, Xian Li*
Meta AI

BLOOM

- Very large decoder-only LM (176B parameters)
- Open access (from Huggingface, kinda)
- Also a strong emphasis on in-context learning

a BigScience initiative

BLOOM

176B params · 59 languages · Open-access

Questions?