Chapter 13: Multilingual Contextualized Models

Hinrich Schütze

February 10, 2021

Outline

1. mBERT

2. XLM-R

3. Investigating the mystery

Outline

1 mBERT

2 XLM-R

3 Investigating the mystery

Recap: BERT

- Transformer
- Training: Masked language modeling (MLM)
- BERT learns an enormous amount of knowledge about language and the world through MLM training on large corpora.
- Applications: finetune on a particular task
- Combines: (i) leveraging pretraining on large corpora and (ii) supervised training on specific task
- Great performance!
- In this lecture: how can we make BERT multilingual?

mBERT: Multilingual BERT

- https://github.com/google-research/bert/blob/master/multilingual.md (no publication)
- Trained on top 100 languages with largest Wikipedias
- For training and vocab generation:
  oversample low-resource, undersample high-resource
- 110K shared WordPiece vocabulary
- There is no marking of the language (e.g., no special symbol to indicate that a sentence is an English sentence).
  - makes zero-shot training possible
- accent removal, punctuation splitting, whitespace tokenization
- BERT-Base, Multilingual Cased: 104 languages, 12 layers, 768-hidden, 12 heads, 110M parameters

Languages covered by mBERT

Afrikaans Albanian Arabic Aragonese Armenian Asturian Azerbaijani Bashkir Basque Bavarian Belarusian Bengali Bishnupriya Manipuri Bosnian Breton Bulgarian Burmese Catalan Cebuano Chechen Chinese (Simplified) Chinese (Traditional) Chuvash Croatian Czech Danish Dutch English Estonian Finnish French Galician Georgian German Greek Gujarati Haitian Hebrew Hindi Hungarian Icelandic Ido Indonesian Irish Italian Japanese Javanese Kannada Kazakh Kirghiz Korean Latin Latvian Lithuanian Lombard Low Saxon Luxembourgish Macedonian Malagasy Malay Malayalam Marathi Minangkabau Nepali Newar Norwegian (Bokmal) Norwegian (Nynorsk) Occitan Persian (Farsi) Piedmontese Polish Portuguese Punjabi Romanian Russian Scots Serbian Serbo-Croatian Sicilian Slovak Slovenian South Azerbaijani Spanish Sundanese Swahili Swedish Tagalog Tajik Tamil Tatar Telugu Turkish Ukrainian Urdu Uzbek Vietnamese Volapük Waray-Waray Welsh West Frisian Western Punjabi Yoruba Thai Mongolian

Evaluation: XNLI

- `https://cims.nyu.edu/~sbowman/xnli/`
- Derived from MultiNLI
- `https://cims.nyu.edu/~sbowman/multinli/`
- Crowd-sourced collection of 433k sentence pairs annotated with textual entailment information
- Multigenre
- Task: for a sentence pair, classify it as neutral, contradiction or entailment

Example for neutral

Your gift is appreciated by each and every student who will benefit from your generosity. $<?>$ Hundreds of students will benefit from your generosity.

(genre: letters)

Example for contradiction

if everybody like in August when everybody's on vacation or something we can dress a little more casual <? > August is a black out month for vacations in the company.

(genre: telephone speech)

Example for entailment

At the other end of Pennsylvania Avenue, people began to line up for a White House tour. $<? >$ People formed a line at the end of Pennsylvania Avenue.

(genre: 9/11 report)

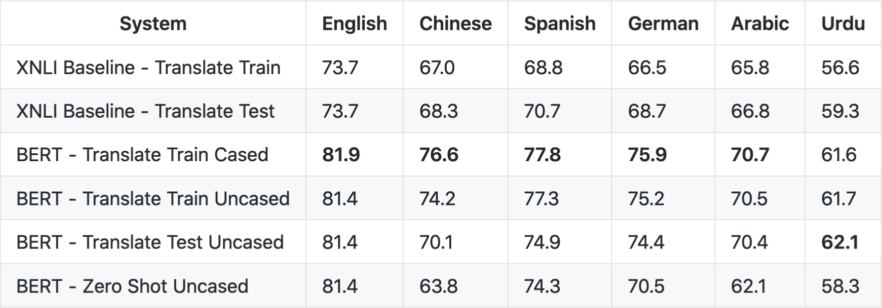
XNLI

- 5000 test pairs and 2500 dev pairs from MNLI
- Translated (by crowd sourcing) into French, Spanish, German, Greek, Bulgarian, Russian, Turkish, Arabic, Vietnamese, Thai, Chinese, Hindi, Swahili, Urdu
- “The corpus is made to evaluate how to perform inference in any language (including low-resources ones like Swahili or Urdu) when only English NLI data is available at training time.”

XNLI evaluation setup

- Train on large English training set
- Evaluate on 14 Non-English languages
- This is zero-shot: there is training data for a related task (English), but zero training data for the task that is evaluated (Urdu, Swahili etc.).
- XNLI is frequently used for the evaluation of multilingual representations, i.e., a common representational space for multiple languages.

Performance of mBERT on XNLI



System	English	Chinese	Spanish	German	Arabic	Urdu
XNLI Baseline - Translate Train	73.7	67.0	68.8	66.5	65.8	56.6
XNLI Baseline - Translate Test	73.7	68.3	70.7	68.7	66.8	59.3
BERT - Translate Train Cased	**81.9**	**76.6**	**77.8**	**75.9**	**70.7**	61.6
BERT - Translate Train Uncased	81.4	74.2	77.3	75.2	70.5	61.7
BERT - Translate Test Uncased	81.4	70.1	74.9	74.4	70.4	**62.1**
BERT - Zero Shot Uncased	81.4	63.8	74.3	70.5	62.1	58.3

- translate train = training set translation into foreign
- translate test = test set translation into foreign
- zero shot = no translation
- advantage of mBERT: you only need one model, you don't need translation

Big mystery

- Why does this model learn a multilingual representation even though it has zero multilingual signal?
- Recall that mBERT is trained on a multlingual corpus – but there are no alignments of words or even sentences. In fact, the corpora are not parallel.
- Maybe the shared vocabulary between languages is crucial?
  - E.g., names are often the same across languages
- We will answer this in the last section today.

Big mystery

- Why does this model learn a multilingual representation even though it has zero multilingual signal?
- Recall that mBERT is trained on a multlingual corpus – but there are no alignments of words or even sentences. In fact, the corpora are not parallel.
- Maybe the shared vocabulary between languages is crucial?
    - E.g., names are often the same across languages
- We will answer this in the last section today.

Big mystery

- Why does this model learn a multilingual representation even though it has zero multilingual signal?
- Recall that mBERT is trained on a multlingual corpus – but there are no alignments of words or even sentences. In fact, the corpora are not parallel.
- Maybe the shared vocabulary between languages is crucial?
  - E.g., names are often the same across languages
- We will answer this in the last section today.

Big mystery

- Why does this model learn a multilingual representation even though it has zero multilingual signal?
- Recall that mBERT is trained on a multlingual corpus – but there are no alignments of words or even sentences. In fact, the corpora are not parallel.
- Maybe the shared vocabulary between languages is crucial?
  - E.g., names are often the same across languages
- We will answer this in the last section today.

Big mystery

- Why does this model learn a multilingual representation even though it has zero multilingual signal?
- Recall that mBERT is trained on a multlingual corpus – but there are no alignments of words or even sentences. In fact, the corpora are not parallel.
- Maybe the shared vocabulary between languages is crucial?
  - E.g., names are often the same across languages
- We will answer this in the last section today.

Big mystery

- Why does this model learn a multilingual representation even though it has zero multilingual signal?
- Recall that mBERT is trained on a multlingual corpus – but there are no alignments of words or even sentences. In fact, the corpora are not parallel.
- Maybe the shared vocabulary between languages is crucial?
  - E.g., names are often the same across languages
- We will answer this in the last section today.

Big mystery

- Why does this model learn a multilingual representation even though it has zero multilingual signal?
- Recall that mBERT is trained on a multlingual corpus – but there are no alignments of words or even sentences. In fact, the corpora are not parallel.
- Maybe the shared vocabulary between languages is crucial?
  - E.g., names are often the same across languages
- We will answer this in the last section today.

Summary: mBERT advantages

- Single model in multilingual setting
- Supports easy transfer learning
- In particular:
  transfer learning for low-resource languages

Outline

1 mBERT

2 XLM-R

3 Investigating the mystery

XLM-R

Unsupervised Cross-lingual Representation Learning at Scale

Alexis Conneau* Kartikay Khandelwal*

Naman Goyal Vishrav Chaudhary Guillaume Wenzek Francisco Guzmán

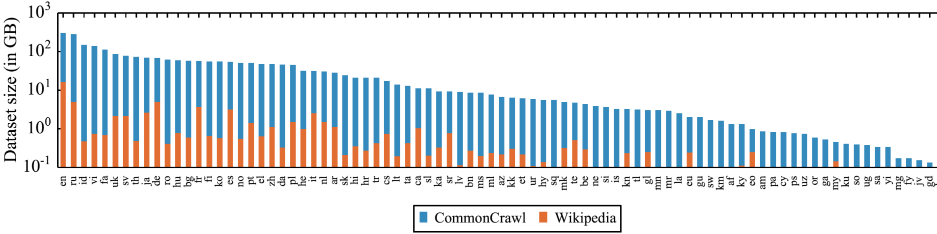
Edouard Grave Myle Ott Luke Zettlemoyer Veselin Stoyanov

Facebook AI

XLM-R

- XLM-R = XLM-RoBERTa
- Quite similar to mBERT
- Trained on 100 languages
- Much larger training corpus
  (2 terabytes CommonCrawl vs. Wikipedia)
- Better performance
- Claim: performance competitive with monolingual model
- As in the case of mBERT: no parallel data is used.

Dataset sizes Wikipedia vs. CommonCrawl



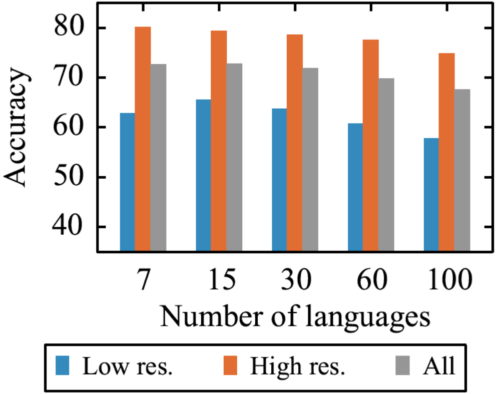
Curse of multilinguality

"more languages leads to better cross-lingual performance on low-resource languages up until a point, after which the overall performance on monolingual and cross-lingual benchmarks degrades."

However, it's easy to address this, simply by increasing the capacity of the model.

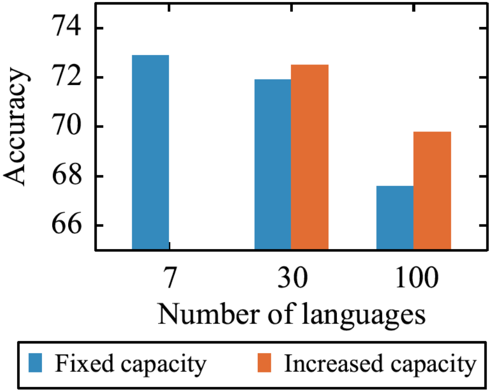
Curse of multilinguality

The transfer-interference trade-off: Low-resource languages benefit from scaling to more languages, until dilution (interference) kicks in and degrades overall performance.



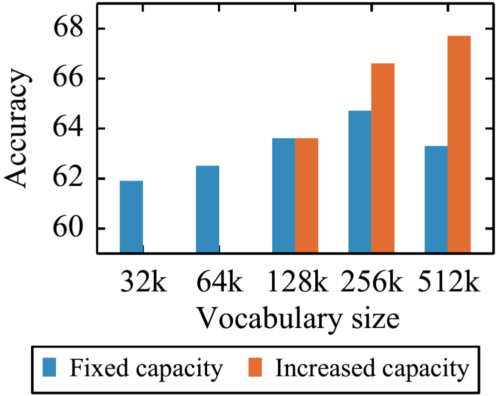
How to fix curse of multilinguality

Adding more capacity to the model alleviates the curse of multilinguality, but remains an issue for models of moderate size.



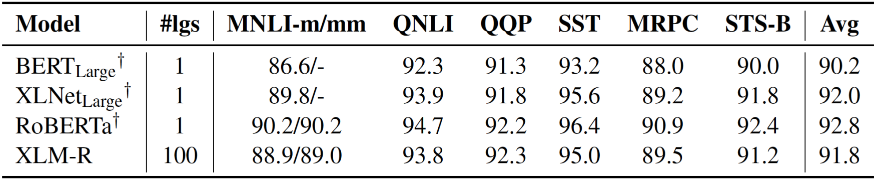
Effect of vocabulary size

Multilingual models can benefit from allocating a higher proportion of the total number of parameters to the embedding layer even though this reduces the size of the Transformer.



XLM-R competitive with monolingual models

What is a fair comparison? Number of languages vs. number of parameters.



Model	#lgs	MNLI-m/mm	QNLI	QQP	SST	MRPC	STS-B	Avg
$\text{BERT}_{\text{Large}}^{\dagger}$	1	86.6/-	92.3	91.3	93.2	88.0	90.0	90.2
$\text{XLNet}_{\text{Large}}^{\dagger}$	1	89.8/-	93.9	91.8	95.6	89.2	91.8	92.0
$\text{RoBERTa}^{\dagger}$	1	90.2/90.2	94.7	92.2	96.4	90.9	92.4	92.8
XLM-R	100	88.9/89.0	93.8	92.3	95.0	89.5	91.2	91.8

Challenges (for mBERT and XLM-R)

- Vocabulary coverage
  - 250K not enough for 100 languages?
  - No language-specific preprocessing!
- Low-resource transfer doesn't work well for very different languages?
  - How to evaluate?
  - XNLI?

Outline

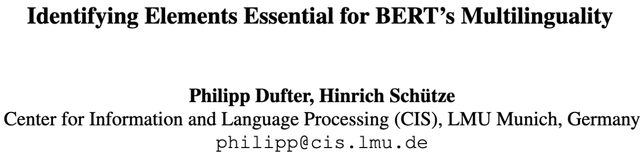
1. mBERT

2. XLM-R

3. Investigating the mystery

Big mystery

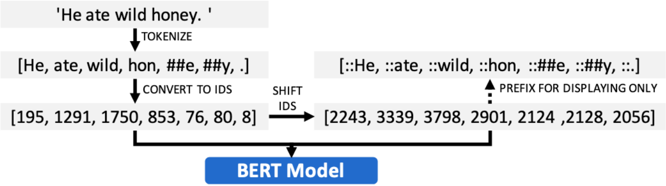
- Why does this model learn a multilingual representation even though it has zero multilingual signal?
- Recall that mBERT is trained on a multlingual corpus – but there are no alignments of words or even sentences. In fact, the corpora are not parallel.
- Maybe the shared vocabulary between languages is crucial?
  - E.g., names are often the same across languages
- We will answer this in the last section today.



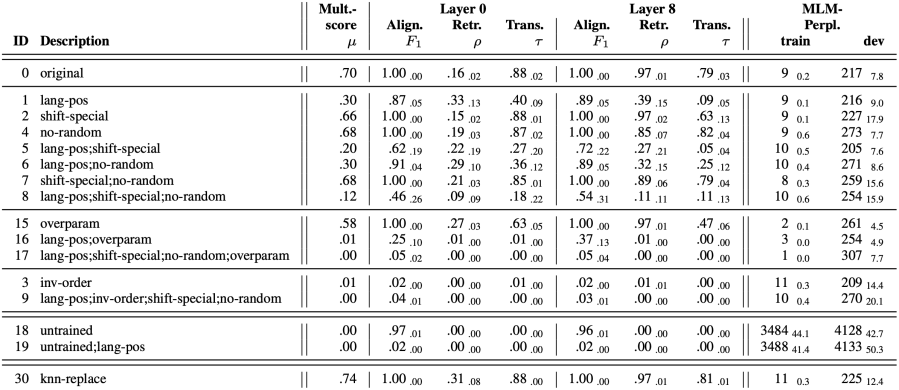
Identifying Elements Essential for BERT's Multilinguality

Philipp Dufter, Hinrich Schütze
Center for Information and Language Processing (CIS), LMU Munich, Germany
philipp@cis.lmu.de

Fake-English: No vocabulary overlap with English

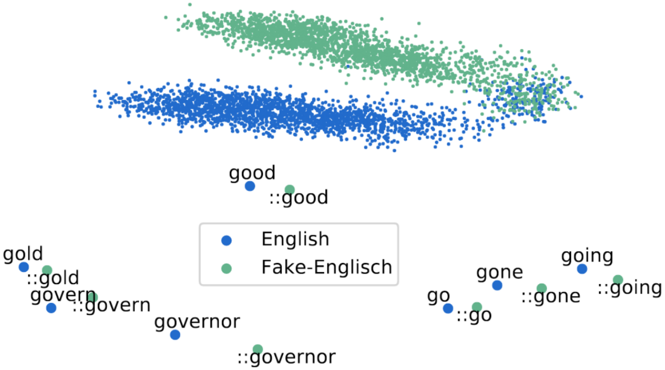


English vs Fake-English:
Good performance without vocabulary overlap



		Mult.-		Layer 0			Layer 8		MLM-	
		score	Align.	Retr.	Trans.	Align.	Retr.	Trans.	Perpl.	
ID	Description	μ	F_1	ρ	τ	F_1	ρ	τ	train	dev
0	original	.70	1.00$_{.00}$	.16$_{.02}$	.88$_{.02}$	1.00$_{.00}$	.97$_{.01}$	.79$_{.03}$	9$_{0.2}$	217$_{7.8}$
1	lang-pos	.30	.87$_{.05}$	.33$_{.13}$	.40$_{.09}$	.89$_{.05}$	.39$_{.15}$	.09$_{.05}$	9$_{0.1}$	216$_{9.0}$
2	shift-special	.66	1.00$_{.00}$	.15$_{.02}$	.88$_{.01}$	1.00$_{.00}$	.97$_{.02}$	.63$_{.13}$	9$_{0.1}$	227$_{17.9}$
4	no-random	.68	1.00$_{.00}$	.19$_{.03}$	.87$_{.02}$	1.00$_{.00}$	.85$_{.07}$	.82$_{.04}$	9$_{0.6}$	273$_{7.7}$
5	lang-pos;shift-special	.20	.62$_{.19}$	.22$_{.19}$	.27$_{.20}$	.72$_{.22}$	.27$_{.21}$	.05$_{.04}$	10$_{0.5}$	205$_{7.6}$
6	lang-pos;no-random	.30	.91$_{.04}$	.29$_{.10}$	.36$_{.12}$	.89$_{.05}$	.32$_{.15}$	.25$_{.12}$	10$_{0.4}$	271$_{8.6}$
7	shift-special;no-random	.68	1.00$_{.00}$	.21$_{.03}$	.85$_{.01}$	1.00$_{.00}$	.89$_{.06}$	.79$_{.04}$	8$_{0.3}$	259$_{15.6}$
8	lang-pos;shift-special;no-random	.12	.46$_{.26}$	.09$_{.09}$	.18$_{.22}$	.54$_{.31}$	.11$_{.11}$	.11$_{.13}$	10$_{0.6}$	254$_{15.9}$
15	overparam	.58	1.00$_{.00}$	.27$_{.03}$	.63$_{.05}$	1.00$_{.00}$	.97$_{.01}$	.47$_{.06}$	2$_{0.1}$	261$_{4.5}$
16	lang-pos;overparam	.01	.25$_{.10}$	.01$_{.00}$	.01$_{.00}$	.37$_{.13}$	.01$_{.00}$	.00$_{.00}$	3$_{0.0}$	254$_{4.9}$
17	lang-pos;shift-special;no-random;overparam	.00	.05$_{.02}$	.00$_{.00}$	.00$_{.00}$	.05$_{.04}$	.00$_{.00}$	.00$_{.00}$	1$_{0.0}$	307$_{7.7}$
3	inv-order	.01	.02$_{.00}$	.00$_{.00}$	.01$_{.00}$	.02$_{.00}$	.01$_{.01}$	.00$_{.00}$	11$_{0.3}$	209$_{14.4}$
9	lang-pos;inv-order;shift-special;no-random	.00	.04$_{.01}$	.00$_{.00}$	.00$_{.00}$	.03$_{.01}$	.00$_{.00}$	.00$_{.00}$	10$_{0.4}$	270$_{20.1}$
18	untrained	.00	.97$_{.01}$	.00$_{.00}$	.00$_{.00}$	.96$_{.01}$	.00$_{.00}$	.00$_{.00}$	3484$_{44.1}$	4128$_{42.7}$
19	untrained;lang-pos	.00	.02$_{.00}$	.00$_{.00}$	.00$_{.00}$	.02$_{.00}$	.00$_{.00}$	.00$_{.00}$	3488$_{41.4}$	4133$_{50.3}$
30	knn-replace	.74	1.00$_{.00}$	.31$_{.08}$	.88$_{.00}$	1.00$_{.00}$	.97$_{.01}$	.81$_{.01}$	11$_{0.3}$	225$_{12.4}$

English vs Fake-English: Embedding structure



Factors that influence degree of multilinguality (i.e., shared English / Fake-English representations)

- Limited model capacity:
  overparameterization decreases multilinguality
- Shared special tokens and position embeddings contribute to multilinguality
- Extreme linguistic divergence destroys multilinguality
  (experiment in paper: reverse word order)
- Lack of parallelism reduces multilinguality
  (even though parallelism is not directly exploited)
- Recap: shared vocabulary is not necessary.

English vs Fake-English: Other factors: model capacity, shared tokens/pos embeddings, linguistic divergence

		Mult.-		Layer 0			Layer 8		MLM-	
		score	Align.	Retr.	Trans.	Align.	Retr.	Trans.	Perpl.	
ID	**Description**	μ	F_1	ρ	τ	F_1	ρ	τ	**train**	**dev**
0	original	.70	$1.00_{.00}$	$.16_{.02}$	$.88_{.02}$	$1.00_{.00}$	$.97_{.01}$	$.79_{.03}$	$9_{0.2}$	$217_{7.8}$
1	lang-pos	.30	$.87_{.05}$	$.33_{.13}$	$.40_{.09}$	$.89_{.05}$	$.39_{.15}$	$.09_{.05}$	$9_{0.1}$	$216_{9.0}$
2	shift-special	.66	$1.00_{.00}$	$.15_{.02}$	$.88_{.01}$	$1.00_{.00}$	$.97_{.02}$	$.63_{.13}$	$9_{0.1}$	$227_{17.9}$
4	no-random	.68	$1.00_{.00}$	$.19_{.03}$	$.87_{.02}$	$1.00_{.00}$	$.85_{.07}$	$.82_{.04}$	$9_{0.6}$	$273_{7.7}$
5	lang-pos;shift-special	.20	$.62_{.19}$	$.22_{.19}$	$.27_{.20}$	$.72_{.22}$	$.27_{.21}$	$.05_{.04}$	$10_{0.5}$	$205_{7.6}$
6	lang-pos;no-random	.30	$.91_{.04}$	$.29_{.10}$	$.36_{.12}$	$.89_{.05}$	$.32_{.15}$	$.25_{.12}$	$10_{0.4}$	$271_{8.6}$
7	shift-special;no-random	.68	$1.00_{.00}$	$.21_{.03}$	$.85_{.01}$	$1.00_{.00}$	$.89_{.06}$	$.79_{.04}$	$8_{0.3}$	$259_{15.6}$
8	lang-pos;shift-special;no-random	.12	$.46_{.26}$	$.09_{.09}$	$.18_{.22}$	$.54_{.31}$	$.11_{.11}$	$.11_{.13}$	$10_{0.6}$	$254_{15.9}$
15	overparam	.58	$1.00_{.00}$	$.27_{.03}$	$.63_{.05}$	$1.00_{.00}$	$.97_{.01}$	$.47_{.06}$	$2_{0.1}$	$261_{4.5}$
16	lang-pos;overparam	.01	$.25_{.10}$	$.01_{.00}$	$.01_{.00}$	$.37_{.13}$	$.01_{.00}$	$.00_{.00}$	$3_{0.0}$	$254_{4.9}$
17	lang-pos;shift-special;no-random;overparam	.00	$.05_{.02}$	$.00_{.00}$	$.00_{.00}$	$.05_{.04}$	$.00_{.00}$	$.00_{.00}$	$1_{0.0}$	$307_{7.7}$
3	inv-order	.01	$.02_{.00}$	$.00_{.00}$	$.01_{.00}$	$.02_{.00}$	$.01_{.01}$	$.00_{.00}$	$11_{0.3}$	$209_{14.4}$
9	lang-pos;inv-order;shift-special;no-random	.00	$.04_{.01}$	$.00_{.00}$	$.00_{.00}$	$.03_{.01}$	$.00_{.00}$	$.00_{.00}$	$10_{0.4}$	$270_{20.1}$
18	untrained	.00	$.97_{.01}$	$.00_{.00}$	$.00_{.00}$	$.96_{.01}$	$.00_{.00}$	$.00_{.00}$	$3484_{44.1}$	$4128_{42.7}$
19	untrained;lang-pos	.00	$.02_{.00}$	$.00_{.00}$	$.00_{.00}$	$.02_{.00}$	$.00_{.00}$	$.00_{.00}$	$3488_{41.4}$	$4133_{50.3}$
30	knn-replace	.74	$1.00_{.00}$	$.31_{.08}$	$.88_{.00}$	$1.00_{.00}$	$.97_{.01}$	$.81_{.01}$	$11_{0.3}$	$225_{12.4}$

Why are multilingual contextualized model multilingual?

- Question: Why are these models multilingual even though there is no direct multilingual training signal.
- Answer: There are many different factors all of which play a role.
- The most important one seems to be limited model capacity:
  If there are not enough parameters for independent representation of the languages, parameter sharing is forced upon the model during training.