

Мультиязычные языковые модели

Артем Стрельцов

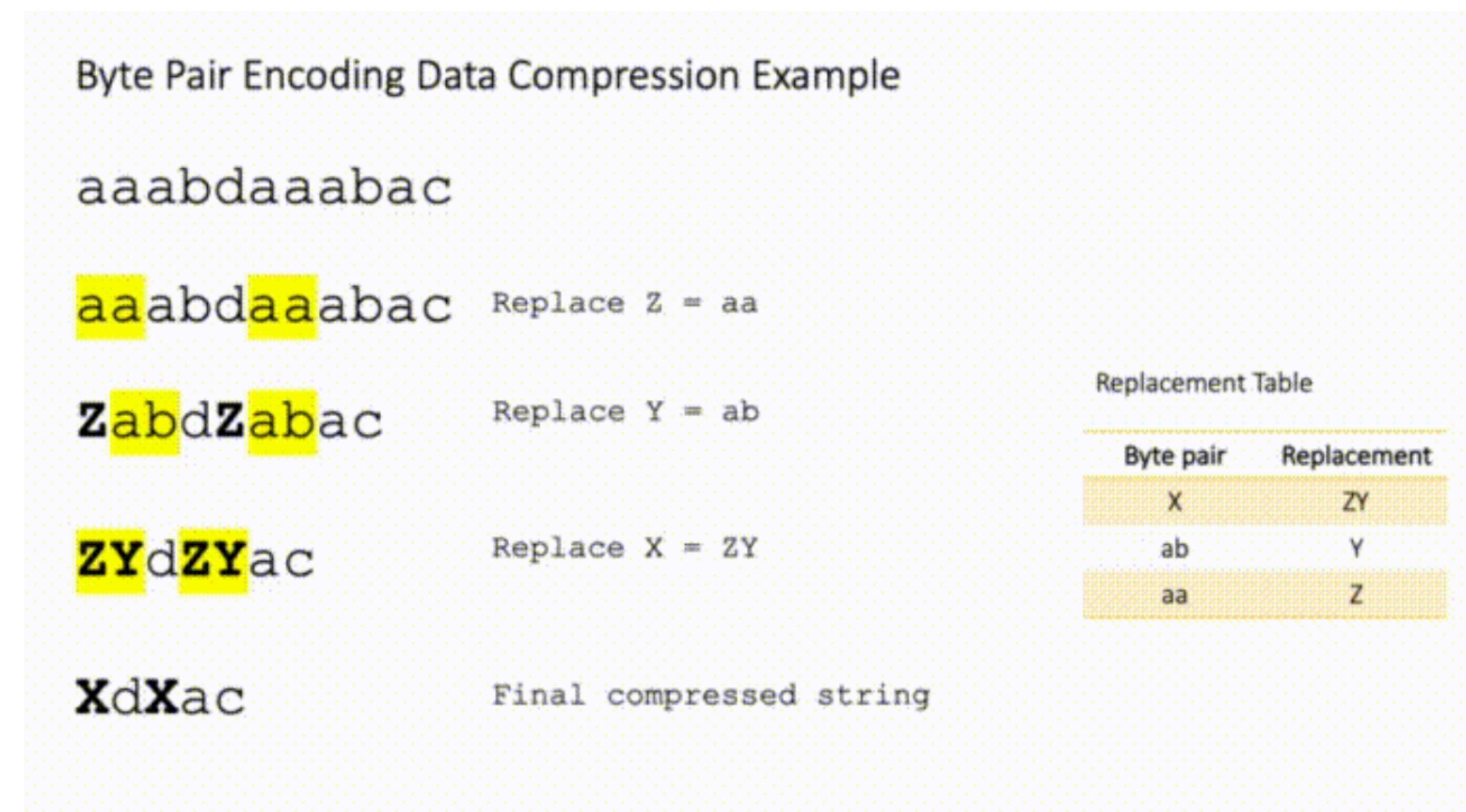
НИС МОП 2020

Что имеется на данный момент времени?

- В NLP хотелось бы иметь системы, поддерживающие как можно больше языков
- Для NLP не так много данных, однако ситуацию спасает то, что некоторые языки синтаксически похожи
- Сейчас имеется тенденция к созданию как можно более «общих» моделей (применимых к как можно большему количеству задач).
- Проблема в том, как их оценивать – обычно это очень разрозненные относительно друг друга задачи, преимущественно **перевода, классификации или типологической схожести языков**, и их не так много

Byte-Pair Encoding

- Вместо слов/символов – Byte-Pair Encoding (BPE). В оригинальном BPE наиболее часто встречающиеся пары байт заменяются на некоторый не встречающийся в последовательности.



Byte-Pair Encoding в NLP:

- Каждое слово – набор символов + </w> (специальный символ конца токена)
- Считаем аналогично оригинальному ВРЕ частоту встречаемости пар «байт».
- Вместо замены – объединяем (но мысленно считаем, что произвели замену)
- Количество таких итераций – гиперпараметр

abacabadaba



- 1) ab a c ab a d ab a </w>
- 2) aba c aba d aba </w>
- 3) abac aba d aba </w>
- 4) abacaba d aba </w>
- 5) abacabad aba </w>

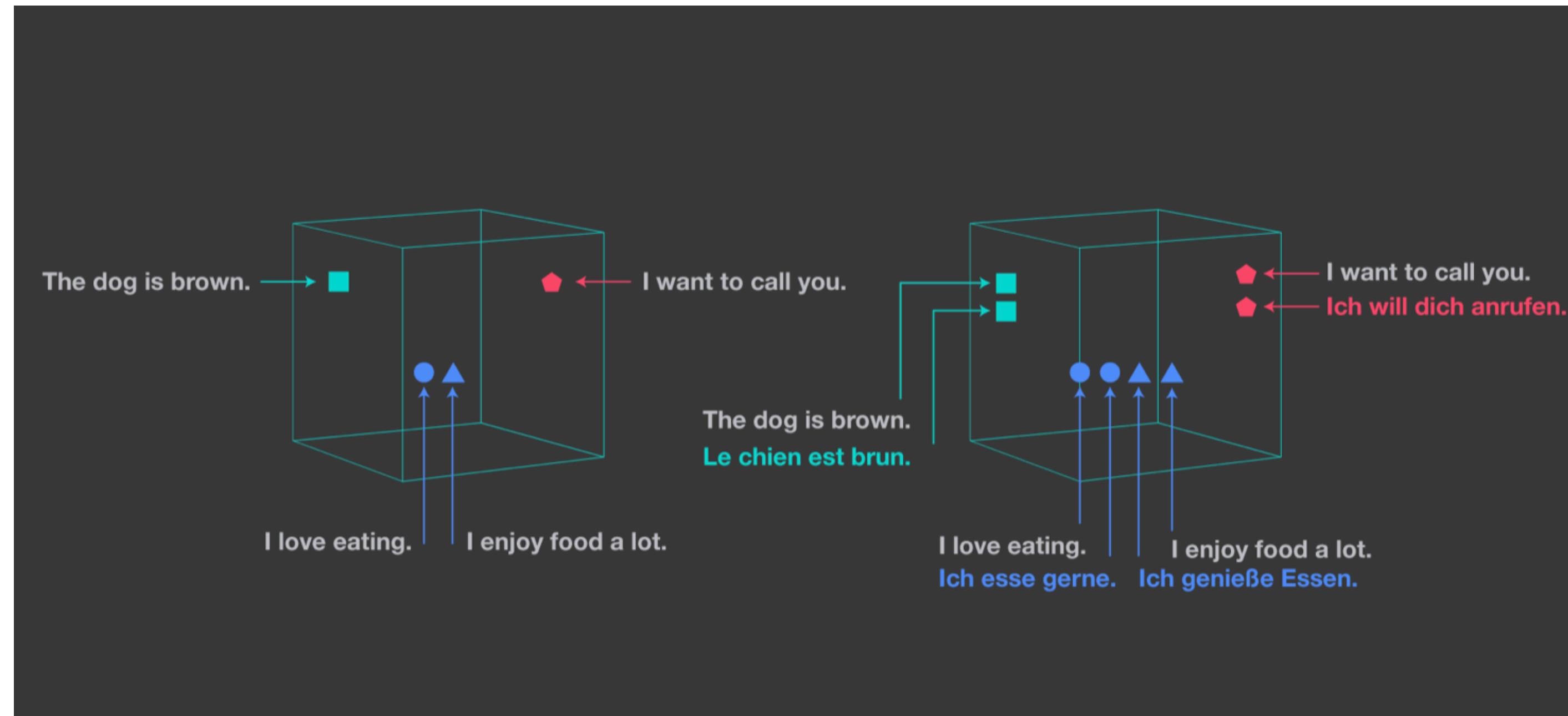
LASER

Language-Agnostic SEntence Representations

- Эмбеддинги, получаемые с помощью LASER, общие для всех языков, а не для каждого свои по-отдельности
- => Zero-shot перевод с одного языка в эмбеддинг другого
- => Одна модель на все языки, в том числе и с ограниченными ресурсами вроде уйгурского или языка у (китайский диалект)
- Можно обучать классификатор на одном языке, а потом в момент отмасштабировать модель на большее количество языков

LASER

- Легко работает на GPU, 2000 предложений в секунду. Энкодер аписан на PyTorch с минимумов зависимостей
- Хорошие результаты на языках с ограниченными ресурсами, так как у нас тренировка идет для всех языков сразу
- Модель поддерживает несколько языков даже в рамках одного предложения
- Модель умеет распознавать семейства языков. Больше языков -> лучше результат



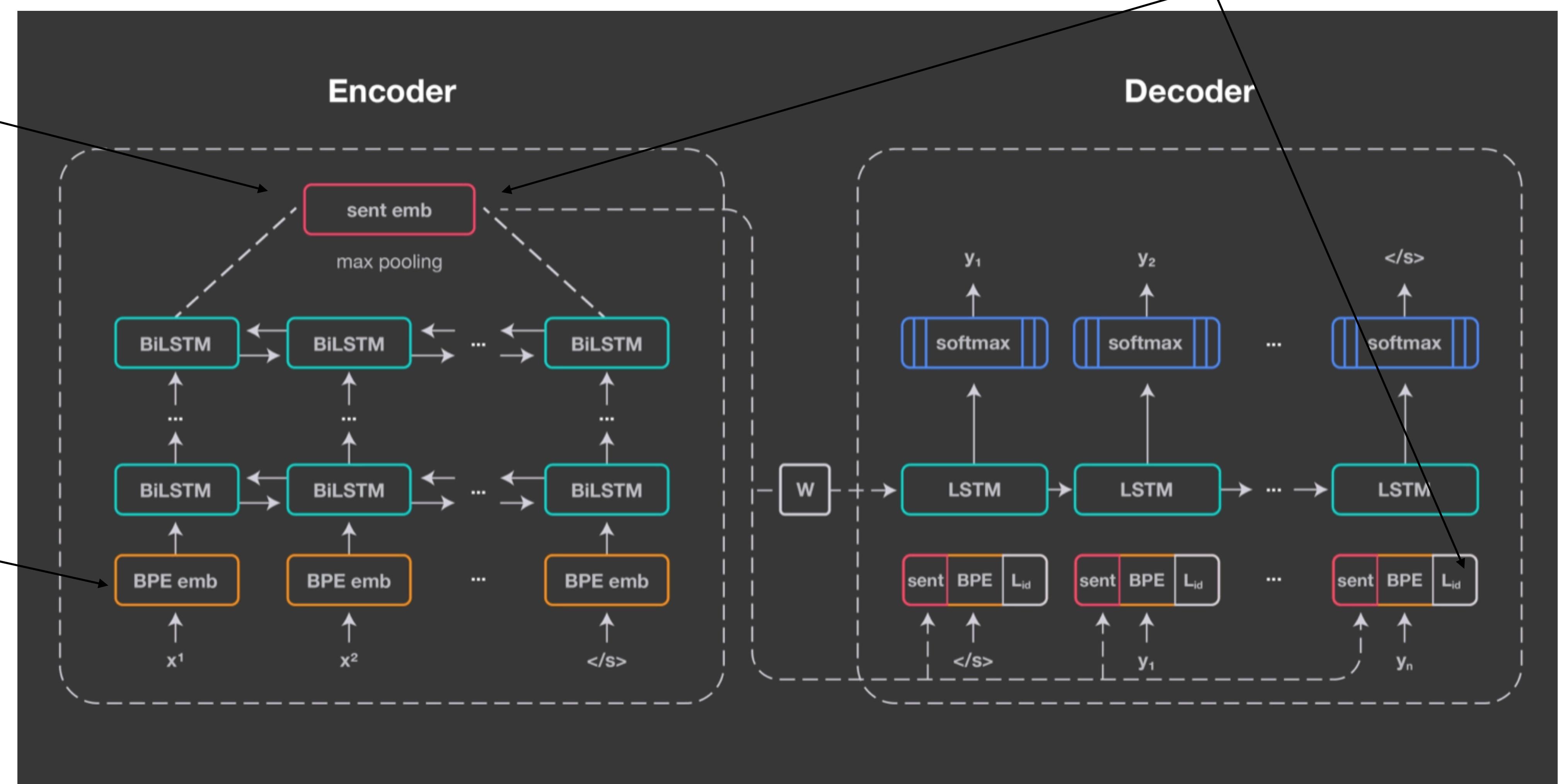
LASER: архитектура (~ seq2seq)

Так как энкодер никак не индицирует язык, то получаются не зависящие от языка представления. Декодеру же подается на вход id языка, в который нужно декодировать

На выходе — вектор размера 1024 за счет пулинга

5 слоев LSTM, без attention

BPE общее для всех языков

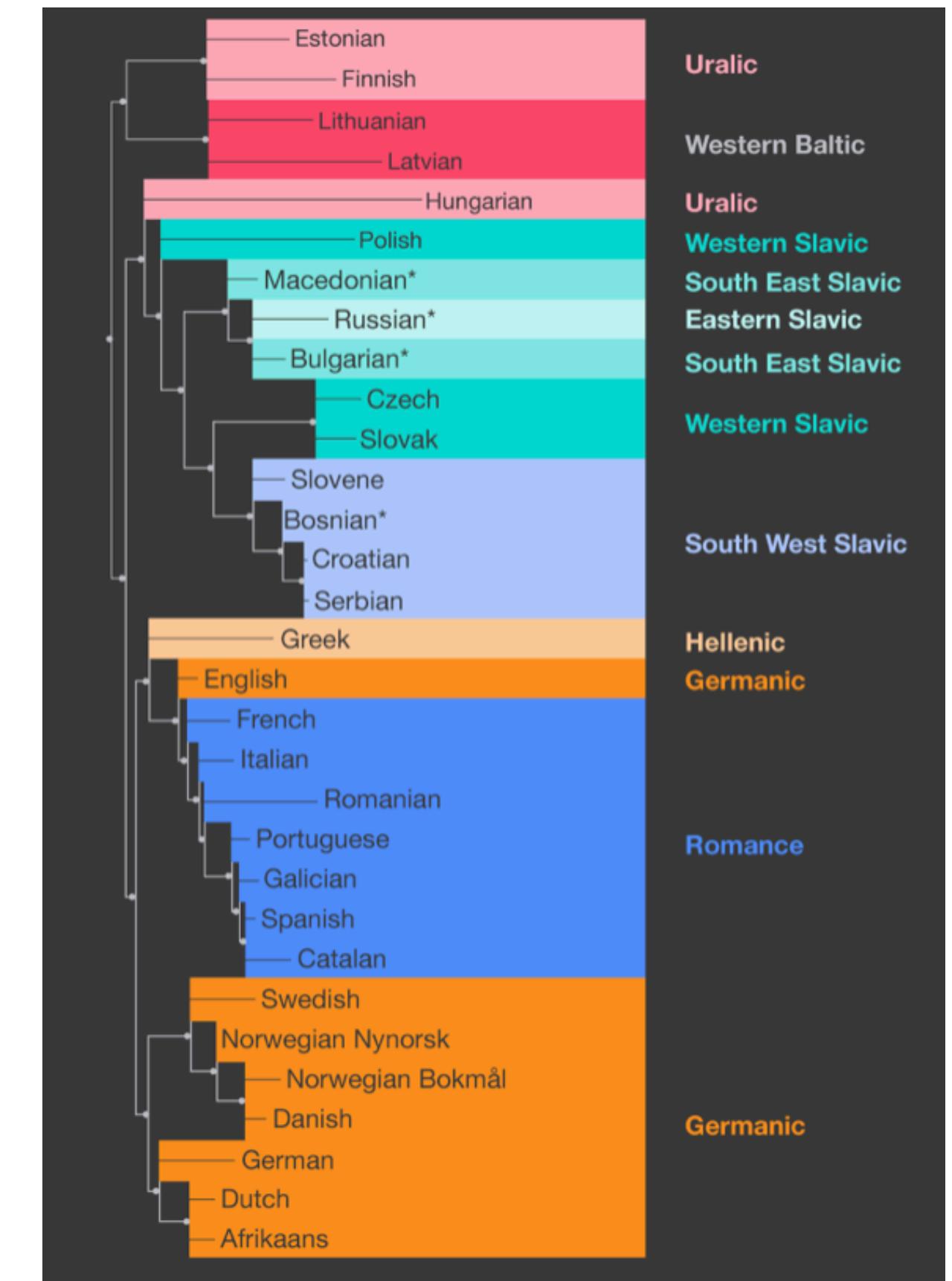


LASER: результаты

Zero-Shot Transfer, one NLI system for all languages		EN		EN → XX												
		fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	
Conneau et. al. (2018c)	X-BiLSTM	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
	X-CBOW	64.5	60.3	60.7	61.0	60.5	60.4	57.8	58.7	57.6	58.8	56.9	58.8	56.3	50.4	52.2
BERT uncased*	Transformer	81.4	—	74.3	70.5	—	—	—	—	62.1	—	—	63.8	—	—	58.3
Proposed method	BiLSTM	74.7	72.3	73.2	72.5	72.7	73.4	71.1	69.8	70.5	71.9	69.2	71.4	66.0	62.1	61.8

LASER: результаты

Premise	Hypothesis	Relation
Bulgarian Никой не знаеше къде отидаха. <i>Their destination was a secret.</i>	Hindi उनका गंतव्य गुप्त था। <i>Nobody knew where they went.</i>	Related (line 210)
Arabic عَمْ ، وَمَذْثُمَ انتَقَلَنَا إِلَى مَنْزَلٍ جَدِيدٍ . <i>Um, then we moved to a new house.</i>	Swahili Tuliishi kwa nyumba moja maisha yetu yote. <i>We stayed in the same house our whole lives.</i>	Opposite (line 393)
Thai สัปดาห์ต่อมา, หลานชายของฉันขอ 기타ร์ อะคูสติ กในวันเกิดของเขา <i>The next week, my nephew asked for an acoustic guitar for his birthday.</i>	Spanish Aprender a tocar la guitarra y comenzar una banda era todo lo que hablaba mi sobrino. <i>Learning to play guitar and starting a band was all that my nephew talked about.</i>	Neutral (line 4702)



Выводы

LASER

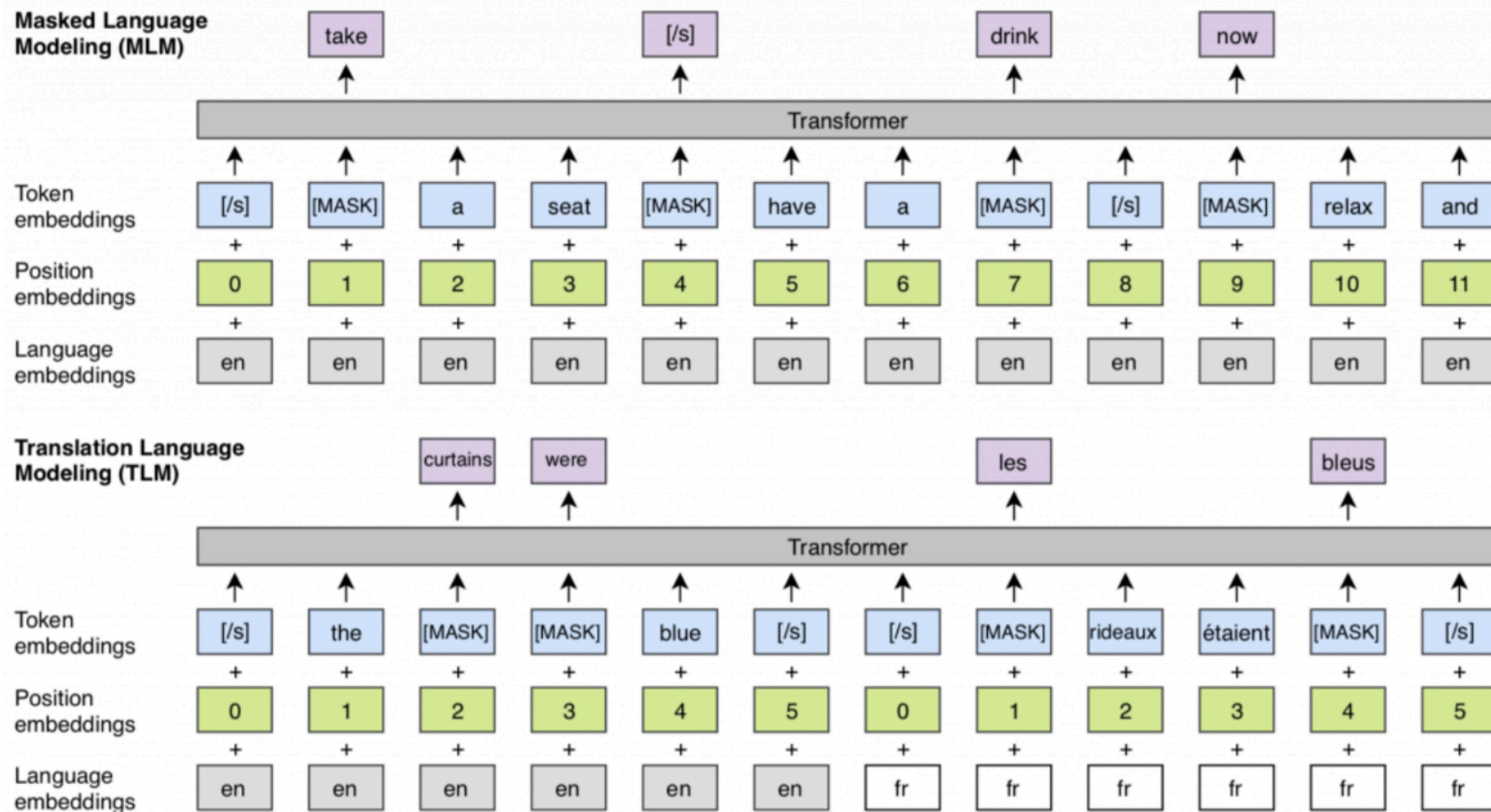
- Эмбеддинги, которые позволяют быстро переключаться между языками
- Архитектура почти аналогичная seq2seq
- Впечатляющие результаты, в том числе в сравнении с BERT
- Сайд-эффект в виде кластеризации по группе языков, если кластеризовать распределения ВРЕ по дивергенции Кульбака-Лейблера

XLM

Как устроен?

- BERT, хоть и предобучен на 100 языках, но плохо умеет в мультиязычность, поскольку словари никак между собой не согласуются
- Каждый объект для тренировки – это текст на двух языках, на каждом варианте из двух обучен BERT
- Так как для BERT нужно скрывать маской некоторые токены (в каждом из двух текстов – разные), то модель может использовать контекст одного текста, чтобы предсказать замаскированные слова в другой
- Используем Language ID и порядок токенов, то есть выполняем позиционное кодирование, чтобы модель лучше понимала взаимосвязь между языками

XLM



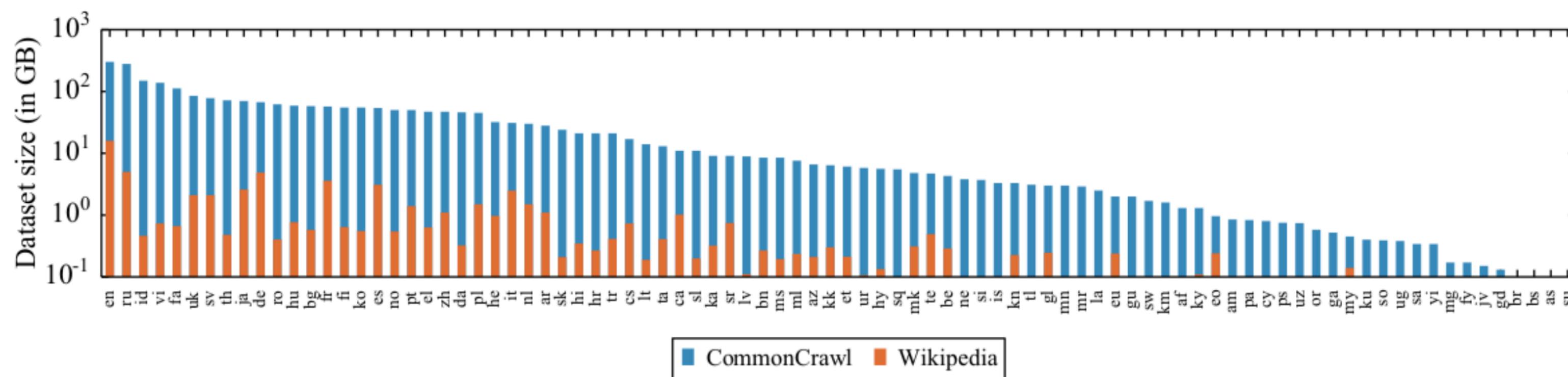
XLM

	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Δ
<i>Machine translation baselines (TRANSLATE-TRAIN)</i>																
Devlin et al. (2018)	81.9	-	77.8	75.9	-	-	-	-	70.7	-	-	76.6	-	-	61.6	-
XLM (MLM+TLM)	<u>85.0</u>	<u>80.2</u>	<u>80.8</u>	<u>80.3</u>	<u>78.1</u>	<u>79.3</u>	<u>78.1</u>	<u>74.7</u>	<u>76.5</u>	<u>76.6</u>	<u>75.5</u>	<u>78.6</u>	<u>72.3</u>	<u>70.9</u>	63.2	<u>76.7</u>
<i>Machine translation baselines (TRANSLATE-TEST)</i>																
Devlin et al. (2018)	81.4	-	74.9	74.4	-	-	-	-	70.4	-	-	70.1	-	-	62.1	-
XLM (MLM+TLM)	<u>85.0</u>	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
<i>Evaluation of cross-lingual sentence encoders</i>																
Conneau et al. (2018b)	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. (2018)	81.4	-	74.3	70.5	-	-	-	-	62.1	-	-	63.8	-	-	58.3	-
Artetxe and Schwenk (2018)	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)	<u>85.0</u>	<u>78.7</u>	<u>78.9</u>	<u>77.8</u>	<u>76.6</u>	<u>77.4</u>	<u>75.3</u>	<u>72.5</u>	<u>73.1</u>	<u>76.1</u>	<u>73.2</u>	<u>76.5</u>	<u>69.6</u>	<u>68.4</u>	<u>67.3</u>	<u>75.1</u>

XLM-R(оBERTа)

Отличия от XLM

- 100 языков (XLM – 15), более 2 тб данных с CommonCrawl (обычно брали Википедию)
- Fine-tuning (когда увеличивается количество языков, то модель хуже понимает каждый из них, это было отчасти исправлено)
- Лучше обработка языков с ограниченными ресурсами



XLM-R: оценивание

XNLI – Cross-lingual Natural Language Inference

- 15 языков
- В обучающей выборке ground-truth на английском языке, переведенный на 14 других языков
- Оценивается перевод с английского на остальные языки
- Сравнение с несколькими бейзлайнами

XLM-R: оценивание

Другие задачи

- NER – Named Entity Recognition. Оценивалась F1-мера и сравнивалась с бейзлайнами (как для языков по отдельности, так и в совокупности, а также оценивался перенос с языка на язык)
- Cross-lingual Question Answering. Оценивались F1-мера и EM (exact match)
- GLUE benchmark. В нем собраны разные задачи классификации. Бейзайном выступает BERT

XLM-R: результаты

+ Transfer-dilution trade-off и Curse of Multilinguality

Есть общий словарь, а также подмешиваем слова из каждого языка пропорционально количеству предложений в корпусе

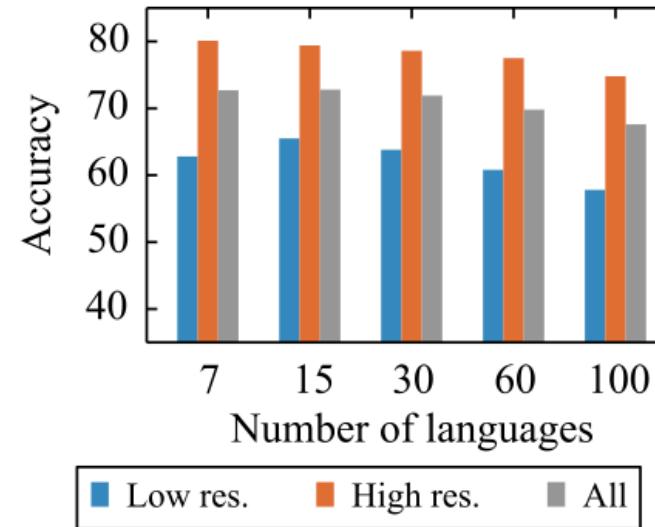


Figure 2: The transfer-interference trade-off: Low-resource languages benefit from scaling to more languages, until dilution (interference) kicks in and degrades overall performance.

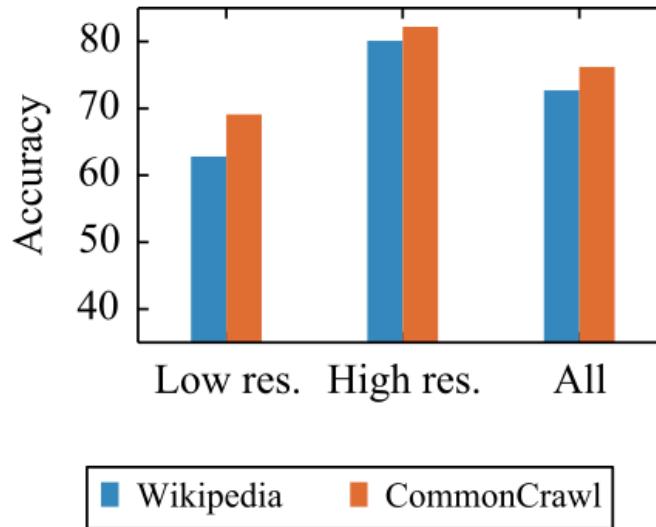


Figure 3: Wikipedia versus CommonCrawl: An XLM-7 obtains significantly better performance when trained on CC, in particular on low-resource languages.

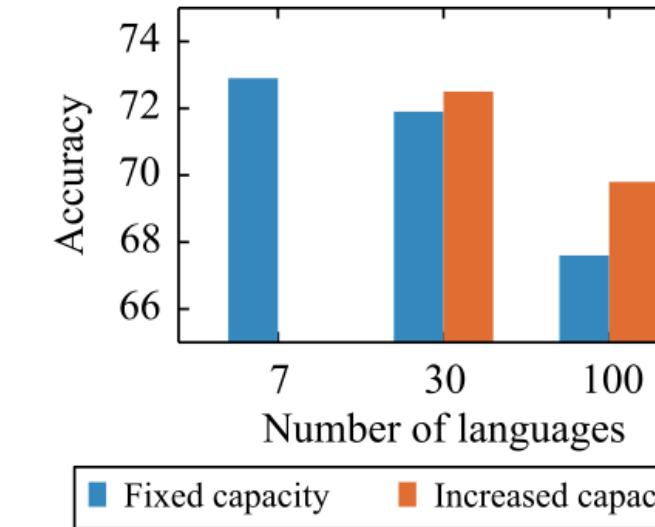


Figure 4: Adding more capacity to the model alleviates the curse of multilinguality, but remains an issue for models of moderate size.

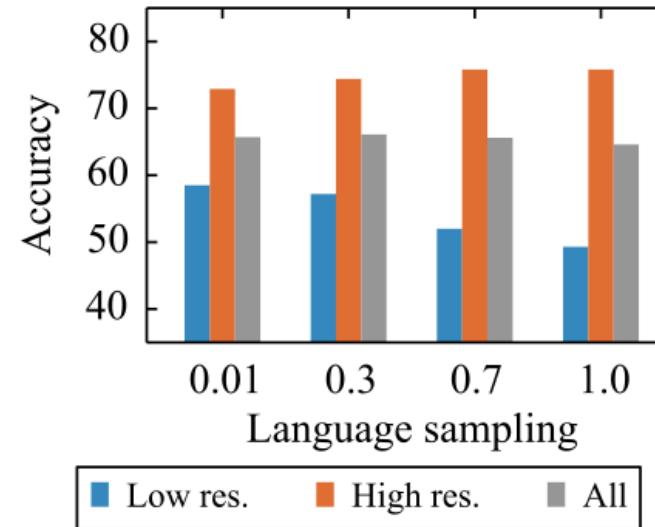


Figure 5: On the high-resource versus low-resource trade-off: impact of batch language sampling for XLM-100.

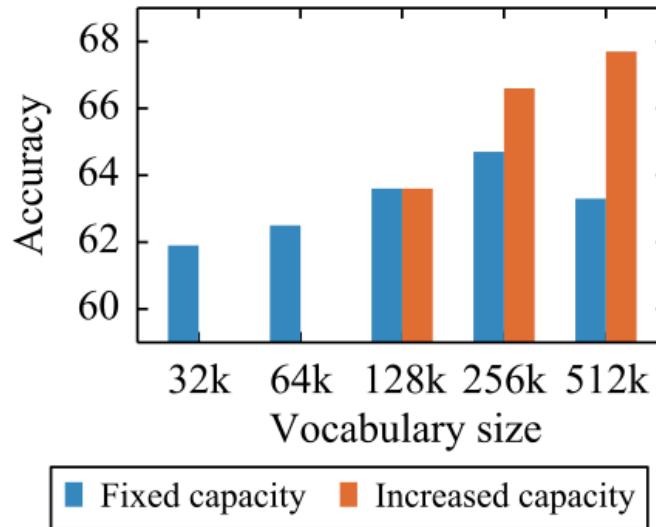


Figure 6: On the impact of vocabulary size at fixed capacity and with increasing capacity for XLM-100.

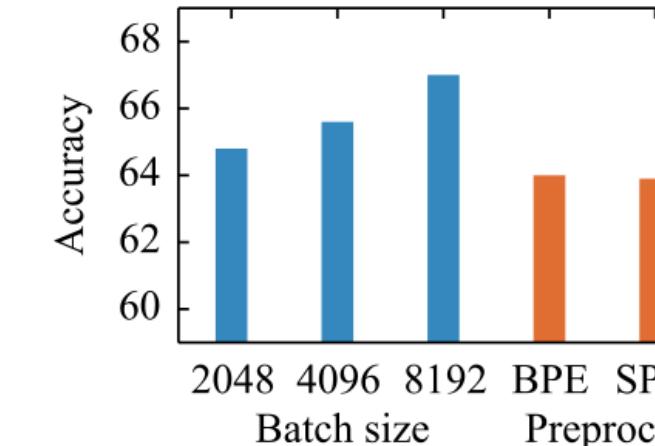


Figure 7: On the impact of large-scale training, and preprocessing simplification from BPE with tokenization to SPM on raw text data.

- Идейно: больше языков – хуже качество на отдельном из-за ограниченности ресурсов

- Языки с ограниченными ресурсами на этом выигрывают, но общее качество по всем языкам ухудшается

XLM-R: результаты

Cross-lingual Classification

Model	D	#M	#lg	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Avg
<i>Fine-tune multilingual model on English training set (Cross-lingual Transfer)</i>																			
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_{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
<i>Translate everything to English and use English-only model (TRANSLATE-TEST)</i>																			
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
<i>Fine-tune multilingual model on each training set (TRANSLATE-TRAIN)</i>																			
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
<i>Fine-tune multilingual model on all training sets (TRANSLATE-TRAIN-ALL)</i>																			
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_{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

XLM-R: результаты

MLQA, NER

Model	train	#M	en	nl	es	de	Avg
Lample et al. (2016)	each	N	90.74	81.74	85.75	78.76	84.25
Akbik et al. (2018)	each	N	93.18	90.44	-	88.27	-
mBERT [†]	each	N	91.97	90.94	87.38	82.82	88.28
	en	1	91.97	77.57	74.96	69.56	78.52
XLM-R _{Base}	each	N	92.25	90.39	87.99	84.60	88.81
	en	1	92.25	78.08	76.53	69.60	79.11
	all	1	91.08	89.09	87.28	83.17	87.66
XLM-R	each	N	92.92	92.53	89.72	85.81	90.24
	en	1	92.92	80.80	78.64	71.40	80.94
	all	1	92.00	91.60	89.52	84.60	89.43

Model	train	#lgs	en	es	de	ar	hi	vi	zh	Avg
BERT-Large [†]	en	1	80.2 / 67.4	-	-	-	-	-	-	-
mBERT [†]	en	102	77.7 / 65.2	64.3 / 46.6	57.9 / 44.3	45.7 / 29.8	43.8 / 29.7	57.1 / 38.6	57.5 / 37.3	57.7 / 41.6
XLM-15 [†]	en	15	74.9 / 62.4	68.0 / 49.8	62.2 / 47.6	54.8 / 36.3	48.8 / 27.3	61.4 / 41.8	61.1 / 39.6	61.6 / 43.5
XLM-R _{Base}	en	100	77.1 / 64.6	67.4 / 49.6	60.9 / 46.7	54.9 / 36.6	59.4 / 42.9	64.5 / 44.7	61.8 / 39.3	63.7 / 46.3
XLM-R	en	100	80.6 / 67.8	74.1 / 56.0	68.5 / 53.6	63.1 / 43.5	69.2 / 51.6	71.3 / 50.9	68.0 / 45.4	70.7 / 52.7

XLM-R: результаты

GLUE, Multilingual vs. Monolingual



Model	#lgs	MNLI-m/mm	QNLI	QQP	SST	MRPC	STS-B	Avg
BERT _{Large} [†]	1	86.6/-	92.3	91.3	93.2	88.0	90.0	90.2
XLNet _{Large} [†]	1	89.8/-	93.9	91.8	95.6	89.2	91.8	92.0
RoBERTa [†]	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

Model	D	#vocab	en	fr	de	ru	zh	sw	ur	Avg
<i>Monolingual baselines</i>										
BERT	Wiki	40k	84.5	78.6	80.0	75.5	77.7	60.1	57.3	73.4
	CC	40k	86.7	81.2	81.2	78.2	79.5	70.8	65.1	77.5
<i>Multilingual models (cross-lingual transfer)</i>										
XLM-7	Wiki	150k	82.3	76.8	74.7	72.5	73.1	60.8	62.3	71.8
	CC	150k	85.7	78.6	79.5	76.4	74.8	71.2	66.9	76.2
<i>Multilingual models (translate-train-all)</i>										
XLM-7	Wiki	150k	84.6	80.1	80.2	75.7	78	68.7	66.7	76.3
	CC	150k	87.2	82.5	82.9	79.7	80.4	75.7	71.5	80.0

Выходы

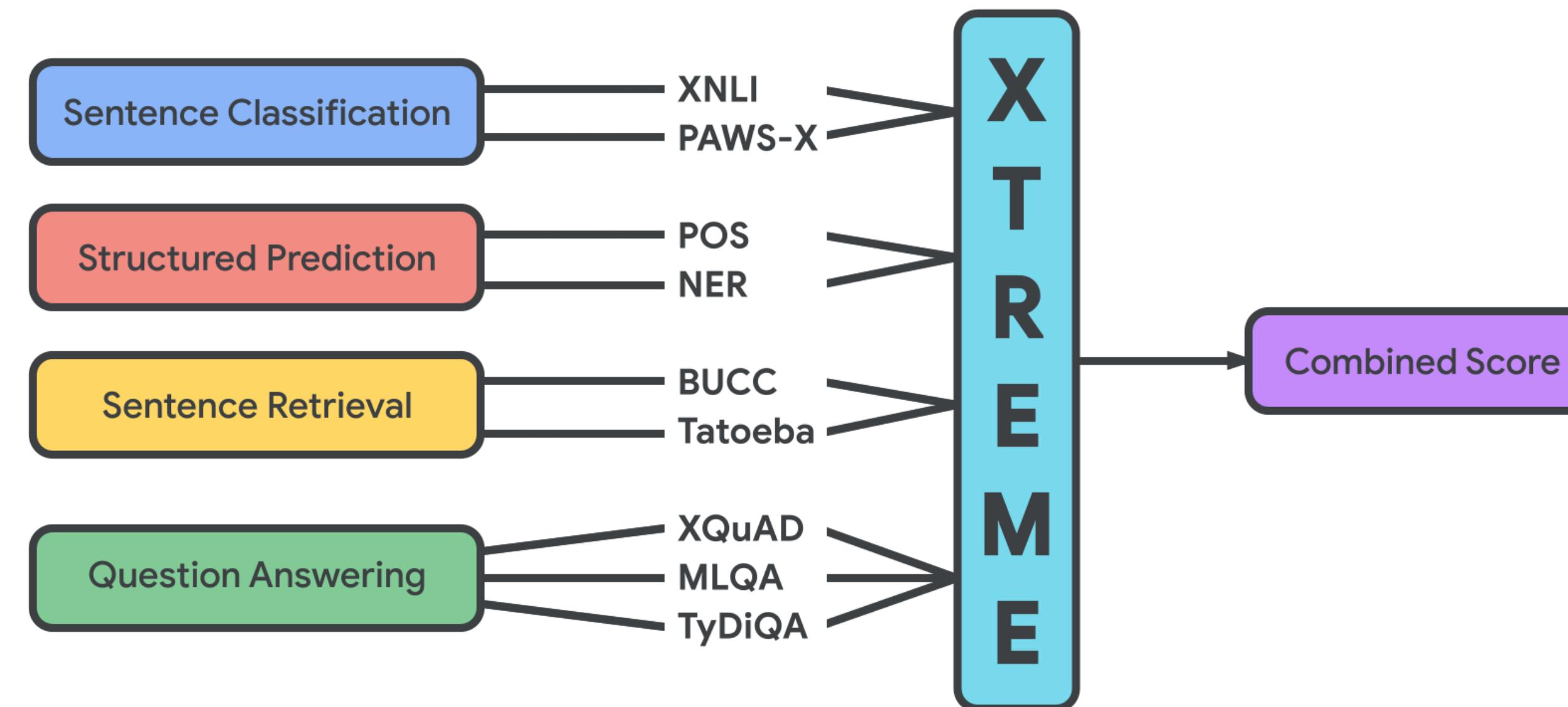
XLM-R

- XLM-R натренирована на 100 языках и > 2 тб данных с CommonCrawl
- XLM-R умеет в языки с ограниченными ресурсами
- Сильнее чем mBERT и XLM на некоторых задачах, к примеру в классификации или Q-A
- Из-за ограниченности ресурсов приходится искать компромисс между количеством языков и качеством на каждом из них
- Мультиязычные модели могут оказаться сильнее моноязычных даже на моноязычных задачах

XTREME

Cross-lingual TRansfer Evaluation of Multilingual Encoders

- 40 типологически различных языков, 12 языковых семейстv, 9 задач
- Служит для оценки модели zero-shot cross-lingual transfer scenario (в которой не производим тюнинг модели после этапа мета-обучения, обучаемся на английском)



XTREME

Особенности бенчмарка

- Сложные и разнообразные задачи
- При этом задачи должны тренироваться не более суток на одном GPU
- Каждая задача должна покрывать как можно больше языков
- При этом должно быть достаточно много одноязычных данных
- Нет быть проблем с правами – данные из источников с подходящей для распространения лицензией

XTREME: задачи

Task	Corpus	Train	Dev	Test	Test sets	Lang.	Task	Metric	Domain
Classification	XNLI	392,702	2,490	5,010	translations	15	NLI	Acc.	Misc.
	PAWS-X	49,401	2,000	2,000	translations	7	Paraphrase	Acc.	Wiki / Quora
Struct. pred.	POS	21,253	3,974	47-20,436	ind. annot.	33 (90)	POS	F1	Misc.
	NER	20,000	10,000	1,000-10,000	ind. annot.	40 (176)	NER	F1	Wikipedia
QA	XQuAD	87,599	34,726	1,190	translations	11	Span extraction	F1 / EM	Wikipedia
	MLQA			4,517–11,590	translations	7	Span extraction	F1 / EM	Wikipedia
	TyDiQA-GoldP			323–2,719	ind. annot.	9	Span extraction	F1 / EM	Wikipedia
Retrieval	BUCC	-	-	1,896–14,330	-	5	Sent. retrieval	F1	Wiki / news
	Tatoeba	-	-	1,000	-	33 (122)	Sent. retrieval	Acc.	misc.

XTREME: языки

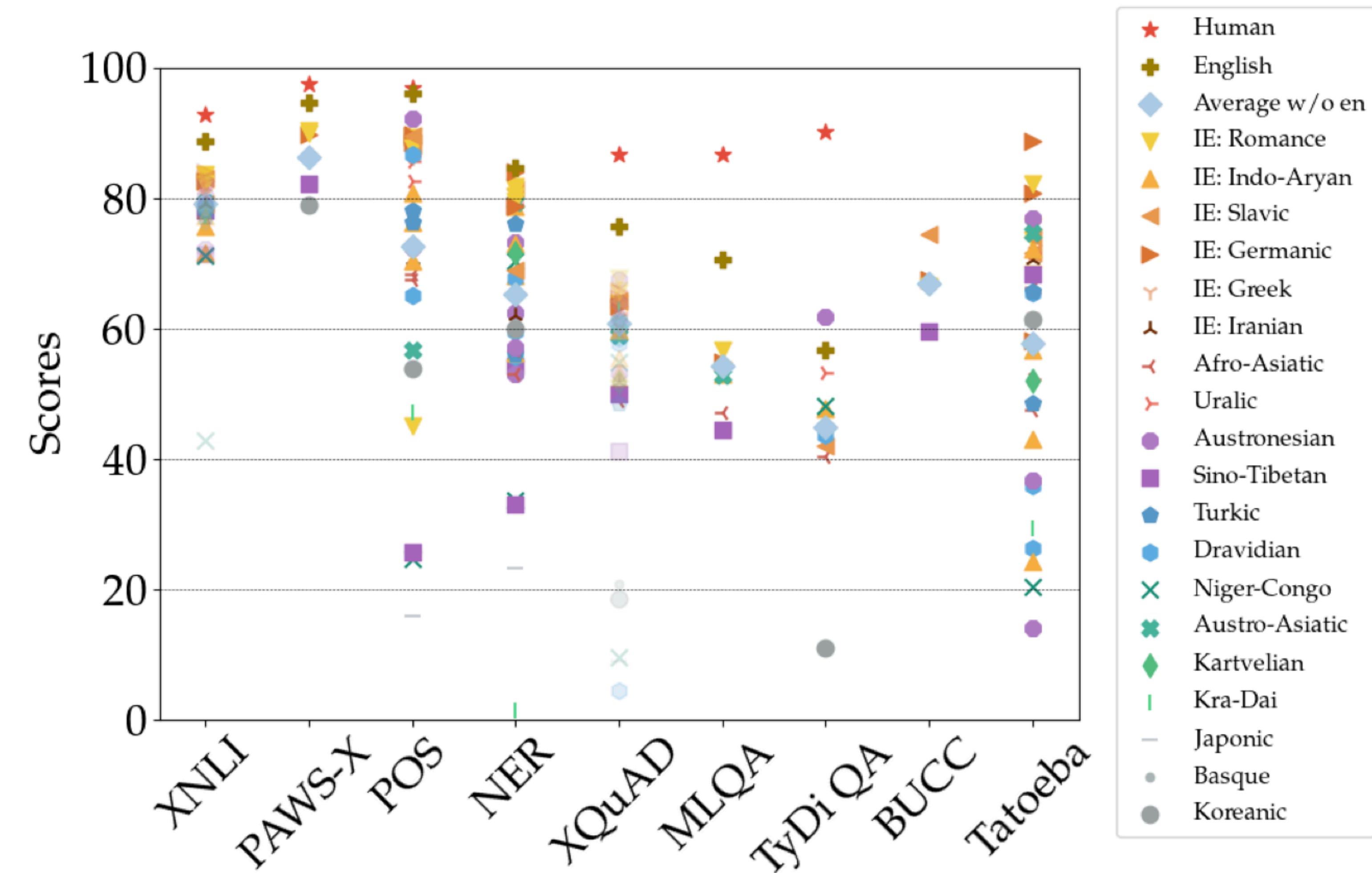
Language	ISO 639-1 code	# Wikipedia articles (in millions)	Script	Language family	Diacritics / special characters	Extensive compound-ing	Bound words / clitics	Inflec-tion	Deriva-tion	# datasets with language
Afrikaans	af	0.09	Latin	IE: Germanic		X				3
Arabic	ar	1.02	Arabic	Afro-Asiatic	X		X	X		7
Basque	eu	0.34	Latin	Basque	X		X	X	X	3
Bengali	bn	0.08	Brahmic	IE: Indo-Aryan	X	X	X	X	X	3
Bulgarian	bg	0.26	Cyrillic	IE: Slavic	X		X	X		4
Burmese	my	0.05	Brahmic	Sino-Tibetan	X	X				1
Dutch	nl	1.99	Latin	IE: Germanic		X				3
English	en	5.98	Latin	IE: Germanic						9
Estonian	et	0.20	Latin	Uralic	X	X		X	X	3
Finnish	fi	0.47	Latin	Uralic				X	X	3
French	fr	2.16	Latin	IE: Romance	X		X			6
Georgian	ka	0.13	Georgian	Kartvelian				X	X	2
German	de	2.37	Latin	IE: Germanic		X		X		8
Greek	el	0.17	Greek	IE: Greek	X	X		X		5
Hebrew	he	0.25	Hebrew	Afro-Asiatic				X		3
Hindi	hi	0.13	Devanagari	IE: Indo-Aryan	X	X	X	X	X	6
Hungarian	hu	0.46	Latin	Uralic	X	X		X	X	4
Indonesian	id	0.51	Latin	Austronesian			X	X	X	4
Italian	it	1.57	Latin	IE: Romance	X		X			3
Japanese	ja	1.18	Ideograms	Japonic			X	X		4
Javanese	јv	0.06	Brahmic	Austronesian	X		X			1
Kazakh	kk	0.23	Arabic	Turkic	X			X	X	1
Korean	ko	0.47	Hangul	Koreanic		X		X	X	5
Malay	ms	0.33	Latin	Austronesian			X	X		2
Malayalam	ml	0.07	Brahmic	Dravidian	X	X	X	X		2
Mandarin	zh	1.09	Chinese ideograms	Sino-Tibetan		X				8
Marathi	mr	0.06	Devanagari	IE: Indo-Aryan			X	X		3
Persian	fa	0.70	Perso-Arabic	IE: Iranian		X		X		2
Portuguese	pt	1.02	Latin	IE: Romance	X		X			3
Russian	ru	1.58	Cyrillic	IE: Slavic				X		7
Spanish	es	1.56	Latin	IE: Romance	X		X			7
Swahili	sw	0.05	Latin	Niger-Congo			X	X	X	3
Tagalog	tl	0.08	Brahmic	Austronesian	X		X	X		1
Tamil	ta	0.12	Brahmic	Dravidian	X	X	X	X	X	3
Telugu	te	0.07	Brahmic	Dravidian	X	X	X	X	X	4
Thai	th	0.13	Brahmic	Kra-Dai	X					4
Turkish	tr	0.34	Latin	Turkic	X	X		X	X	5
Urdu	ur	0.15	Perso-Arabic	IE: Indo-Aryan	X	X	X	X	X	4
Vietnamese	vi	1.24	Latin	Astro-Asiatic	X					6
Yoruba	yo	0.03	Arabic	Niger-Congo	X					1

XTREME: baseline-модели

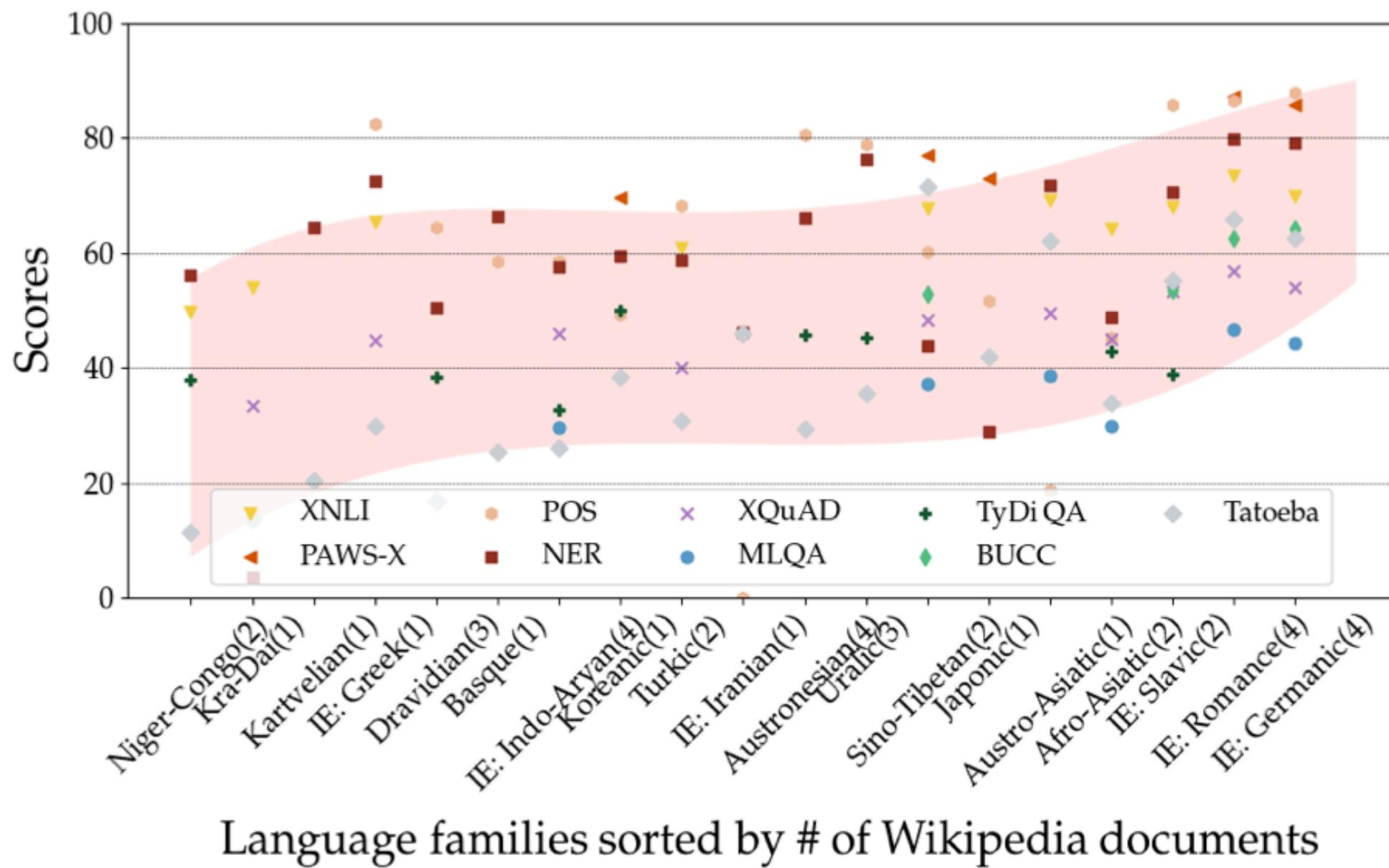
Model	Avg	Pair sentence		Structured prediction		XQuAD	Question answering		Sentence retrieval	
		XNLI	PAWS-X	POS	NER		MLQA	TyDiQA-GoldP	BUCC	Tatoeba
Metrics		Acc.	Acc.	F1	F1	F1 / EM	F1 / EM	F1 / EM	F1	Acc.
<i>Cross-lingual zero-shot transfer (models are trained on English data)</i>										
mBERT	59.8	65.4	81.9	71.5	62.2	64.5 / 49.4	61.4 / 44.2	59.7 / 43.9	56.7	38.7
XLM	55.7	69.1	80.9	71.3	61.2	59.8 / 44.3	48.5 / 32.6	43.6 / 29.1	56.8	32.6
XLM-R Large	68.2	79.2	86.4	73.8	65.4	76.6 / 60.8	71.6 / 53.2	65.1 / 45.0	66.0	57.3
MMTE	59.5	67.4	81.3	73.5	58.3	64.4 / 46.2	60.3 / 41.4	58.1 / 43.8	59.8	37.9
<i>Translate-train (models are trained on English training data translated to the target language)</i>										
mBERT	-	74.6	86.3	-	-	70.0 / 56.0	65.6 / 48.0	55.1 / 42.1	-	-
mBERT, multi-task	-	75.1	88.9	-	-	72.4 / 58.3	67.6 / 49.8	64.2 / 49.3	-	-
<i>Translate-test (models are trained on English data and evaluated on target language data translated to English)</i>										
BERT-large	-	76.8	84.4	-	-	76.3 / 62.1	72.9 / 55.3	72.1 / 56.0	-	-
<i>In-language models (models are trained on the target language training data)</i>										
mBERT, 1000 examples	-	-	-	87.6	77.9	-	-	58.7 / 46.5	-	-
mBERT	-	-	-	89.8	88.3	-	-	74.5 / 62.7	-	-
mBERT, multi-task	-	-	-	91.5	89.1	-	-	77.6 / 68.0	-	-
Human	-	92.8	97.5	97.0	-	91.2 / 82.3	91.2 / 82.3	90.1 / -	-	-

XTREME: результаты

Оценка XLM-R



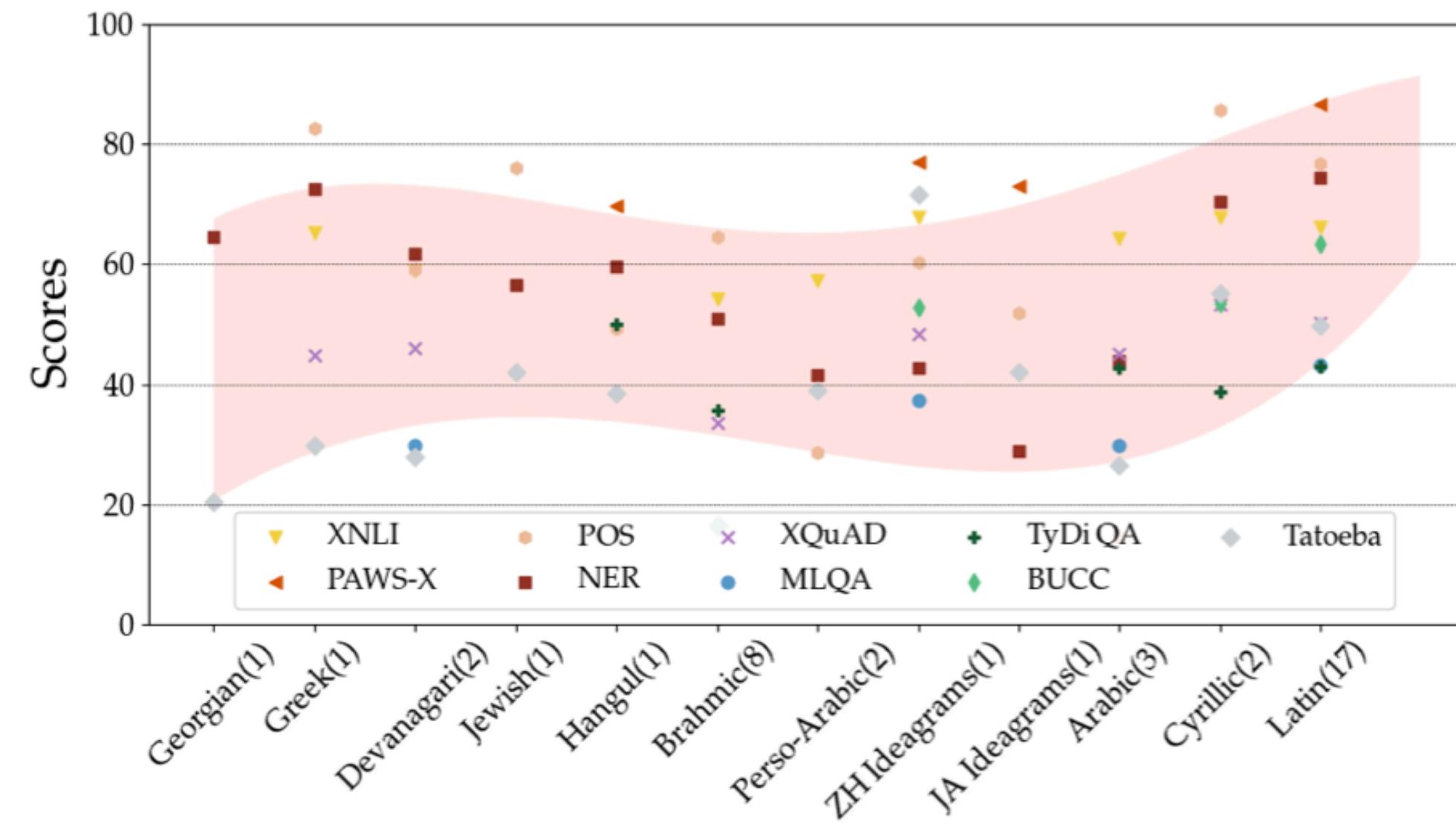
XTREME: результаты



Language families sorted by # of Wikipedia documents

Low-resource

High-resource



Language scripts sorted by # of Wikipedia documents

High-resource

Low-resource

Выходы XTREME

- XTRME позволяет с очень большим охватом различных задач оценить мультиязычные модели
- Бенчмарк подходит для использования даже на одном GPU
- Покрывает 40 языков, в том числе с ограниченными ресурсами
- Задачи разнообразные по типам и сложности

Итог

- Мультиязычные модели сейчас получаются достаточно хорошего качества
- Более того, иногда они превосходят по качеству моноязычные (XLM-7 vs BERT)
- Уже есть возможность обучать cross-lingual эмбеддинги для предложений (LASER)
- Есть достаточно много способов проверять качество мультиязычных моделей. XTREME, к примеру, покрывает большое количество задач и языков

Список источников

- <https://towardsdatascience.com/xlm-enhancing-bert-for-cross-lingual-language-model-5aeed9e6f14b>
- <https://www.aclweb.org/anthology/2020.emnlp-main.368.pdf>
- <https://arxiv.org/pdf/2003.11080.pdf>
- <https://arxiv.org/pdf/1911.02116.pdf>
- <https://github.com/google-research/xtreme>
- <https://towardsdatascience.com/byte-pair-encoding-the-dark-horse-of-modern-nlp-eb36c7df4f10>
- <https://ai.facebook.com/blog/-xlm-r-state-of-the-art-cross-lingual-understanding-through-self-supervision/>
- <https://engineering.fb.com/2019/01/22/ai-research/laser-multilingual-sentence-embeddings/>
- <https://arxiv.org/pdf/1812.10464.pdf>