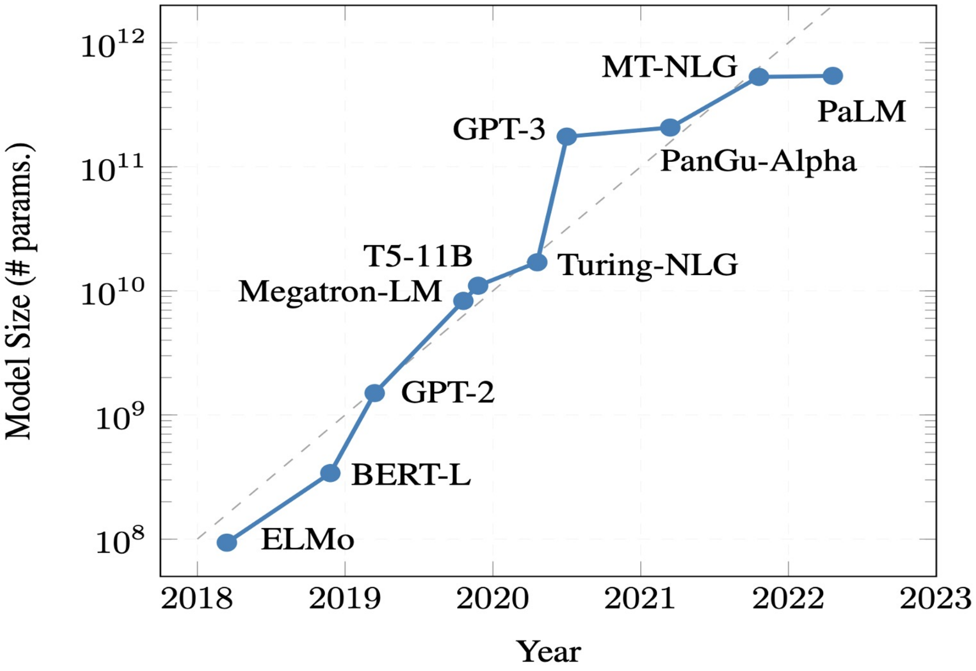
Glot500: Scaling Multilingual Corpora and Language Models to 500 Languages

Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, Hinrich Schütze
CIS/LMU, MCML
Instituto Superior Técnico, Sorbonne/CNRS

The CIS multilingual team

LLMs are getting ever larger



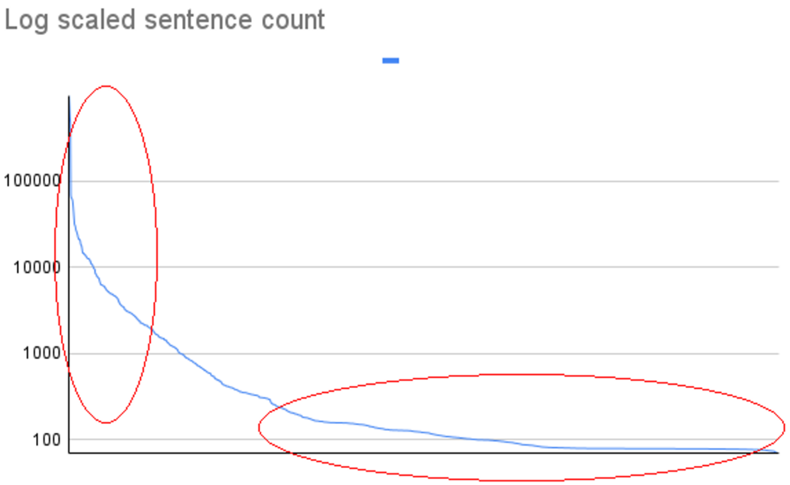
[Treviso et al., 2022]

LLMs: Vertical vs horizontal growth

- Vertical growth: huge model and corpus sizes
  - Only possible for a few languages
  - GPT, Bloom, Bard
- Horizontal growth: more languages
  - Our approach: Glot500

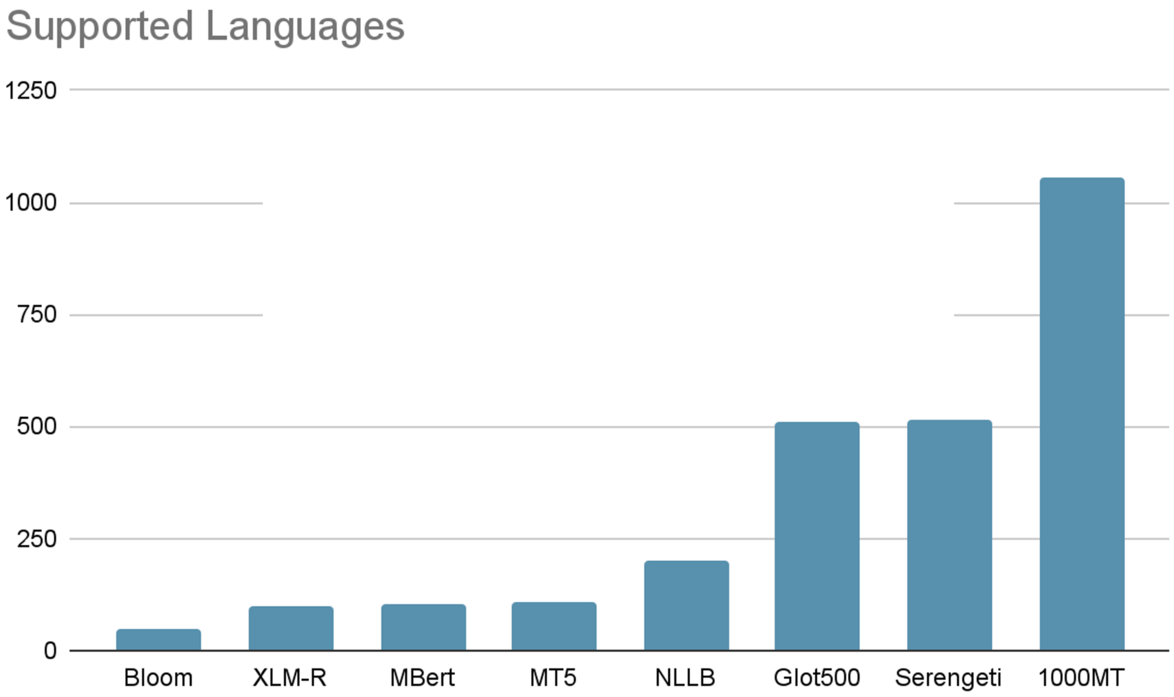
Data available per language

- Typical power law distribution
- About 100 head languages:
  - Large corpora available
  - Covered by main LLMs
- 1000s of (long-)tail languages
  - Little data available
  - Most of it hard to get
  - Our focus in Glot500



Coverage of existing models

Out of +/-7000 languages



Licensing issues

- We are working on making corpora for most languages available.
- But we cannot release the entire corpus due to licensing issues.

Coverage of existing models

- Mostly European
- Plus a few other large national languages
- Primary driver: business

Why multilingual LLMs?

- Preserve culture
- Empower people
- Bread and butter issues
  - Analyze tweets in an emergency

Why multilingual LLMs?

- Making internet accessible
  - Multilingual user base
    - Search, customer support, chatbots
  - Detection of Harmful content in social media
- Translation
- Cross-lingual transfer for standard NLP tasks
  - Text classification
  - Sequence labeling

An LLM for 500 languages: Challenges

- **Collect** good data for tail languages
- **Evaluate** tail languages
- Determine **critical factors** for tail languages

How to collect good data for tail languages

Two corpora: Glot2000 and Glot500

- **Glot2000: >2000 languages**
- **Glot500: subset of Glot2000, >500 languages**
  - **Selection criterion: >=30 000 sentences**
- **Collecting, validating and cleaning the data was (and still is!) a very significant effort**

Challenges with tail languages

- Scarcity of data
- Noise in data
    - Wikipedia is noisy
    - Data leakage
    - Similarity of dialects
    - Macro language / varieties

Challenges: Wikipedia

Magnolia soulangeana

24 languages

Artikulo | Panaghisgot-hisgot | Basaha | Usba | Usba ang wikitext | Tan-awa ang kaagi | Mga galamiton

Gikan sa Wikipedia, ang gawasnong ensiklopedya

Paghimo ni bot Lsjbot.

Kaliwatan sa magnolia ang ***Magnolia soulangeana***.[1] Una ning gihulagway ni Soul.-bod..[2] Ang *Magnolia soulangeana* sakop sa kahenera nga *Magnolia*, ug kabanay nga Magnoliaceae.[1][3]

Kini nga matang hayop na sabwag sa:

- Alabama
- habagatan-sentrong Pangmasang Republika sa Tsina

Walay nalista nga matang nga sama niini.[1]

Magnolia soulangeana

Ang mga gi basihan niini [usba | usba ang wikitext]

1. ↑ 1.0 1.1 1.2 Roskov Y., Kunze T., Orrell T., Abucay L., Paglinawan L., Culham A., Bailly N., Kirk P., Bourgoin T., Baillargeon G., Decock W., De Wever A., Didžiulis V. (ed) (2019). "Species 2000 & ITIS Catalogue of Life: 2019 Annual Checklist". Species 2000: Naturalis, Leiden, the Netherlands. ISSN 2405-884X. TaxonID:

Challenges: Macro vs varieties

Challenges: Leakage

- Example: Swiss German / German, Welsh / English
- Data from high-resource languages leak to low-resource ones
- Made up example
  - 10^7 crawled sentences, mix of H (head) and T (tail)
  - Proportion language T: 10^5 sentences
  - LangID:
    - Accuracy: 99%, false positive rate: 1%
  - Corpus of language T after filtering:
    - Roughly 10^5 in language T
    - 10^5 in language H

Challenges: LangID

How clean are existing multilingual datasets?

	mC4	Oscar	WikiMatrix	ParaCrawl	CCAligned
Source	CC	CC	Wikipedia	Selected websites	CC
Correct (macro F1)	72.40%	87.21%	23.74%	76.14%	29.25%

Data from Kreutzer, et al. “Quality at a glance” 2022

Data from the web: CommonCrawl

- Access for anyone
- Petabytes of data since 2011
- Monthly snapshots (2-3 Billion pages)
- Random sample of URLs
- Noisy web content
- Poor separation of languages
- Bad quality of their LangID

LangID on CommonCrawl

- Domain mismatch with LangID training data
- High false positive rate
- Out-of-model cousins
- So we don’t use CommonCrawl

Our approach: Stand on the shoulders …

- Identify all languages for which some text available
- Our search strategy: publications, low-resource websites (e.g., for Bible), …
- Anything that promises to provide enough volume
- Collect as much as we can
- Analyze, categorize, clean

Our approach: Stand on the shoulders …

- Story:
- Companies doesn’t work
- Crawling the web doesn’t work
- So we decided to rely on academia

Our approach: Stand on the shoulders …

- Story 2:
- Acadmeia: all scattered, no central repository, ELRA: yes,but
- Lrec, elra
- Publications and following all links
- Our knowledge
- Wikipedia multilingual dataset page

Types of sources

- Websites (jw.org) - we crawl them
- Repositories (opus) - we download them
- Datasets published academically - we download them

Repositories / Datasets

- Opus
- LREC publications
- ELRA
- MT-Data
- Hugging Face
- Wikimedia

Data collection: Websites/Datasets

- Websites
  - Jw.org
  - bbc.com
  - lyricstranslate.com
- Datasets (150)
  - Multilingual
    - PBC, Tatoeba, Flores100, TICO, W2C
  - Single language or single family
    - Indic NLP
    - Arabench, Quadi, Shami
    - Afromaft, KinyaSMT

LangID: Reliability and domain issues

- Reliability of language Detection/Verification
  - Automatic (LID)
  - Translator
  - Native speaker or linguist
- Domain
  - News
  - Religious
  - Tweets
  - Radio/TV/Movie transcripts
  - Medical
  - Lyrics

Coverage of existing models

- Mostly European
- Plus a few other large national languages
- Primary driver: business

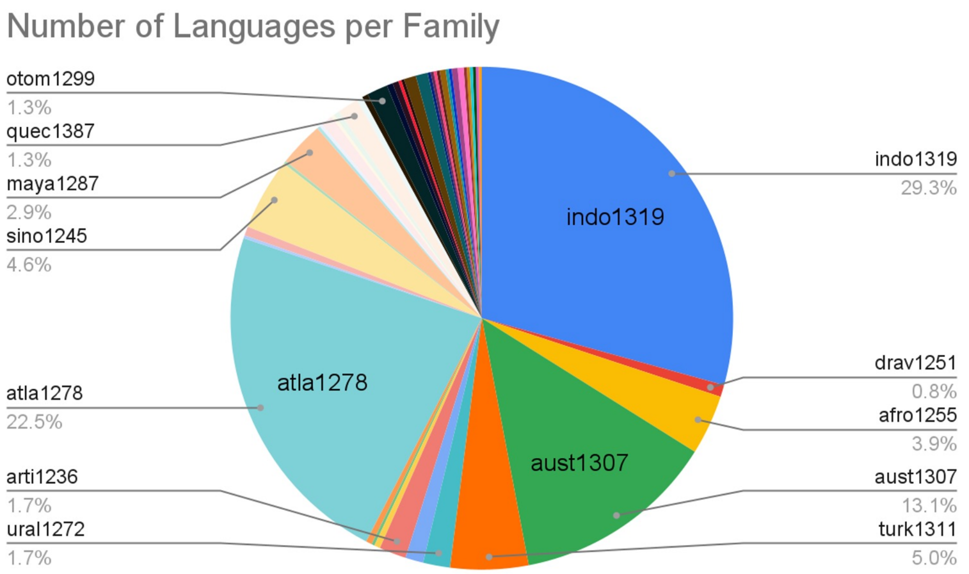
Coverage of existing models

Glot2000: 2266 langs, 728GB

Glot500: Subset of Glot2000

- All language-scripts that had at least 30 000 sentences
- 30 000 is somewhat arbitrary
- Too low for some, too high for others: see last part

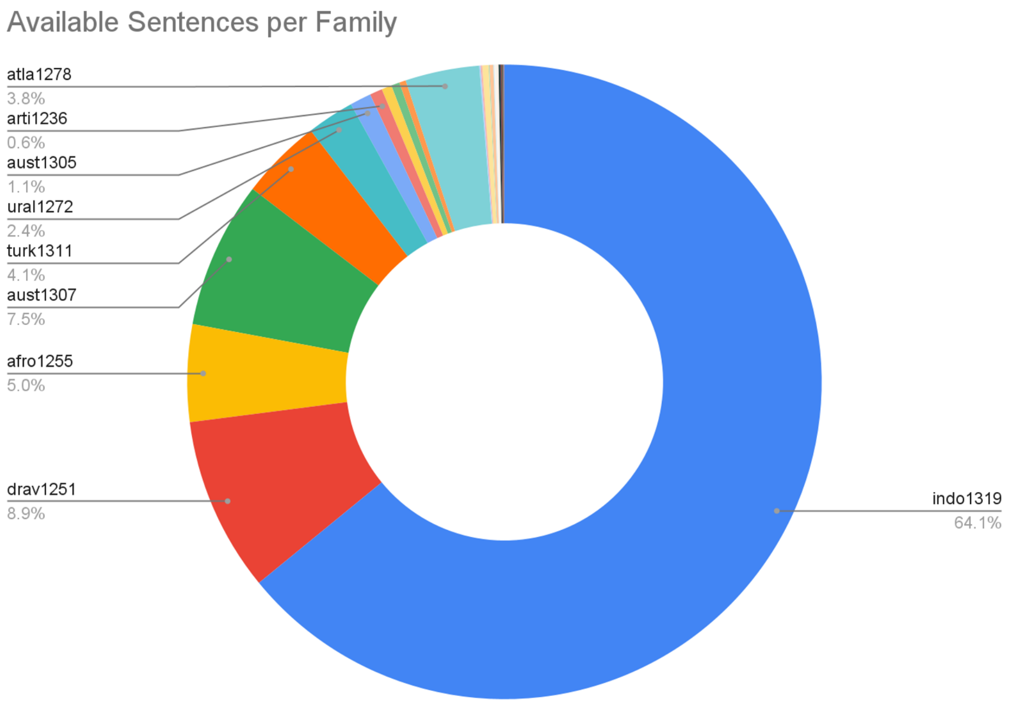
Glot500: Languages per family



Glot500: Languages per family

family	languages	family	languages	family	languages	family	languages
indo1319	152	aust1305	6	choc1280	2	nucl1708	1
atla1278	133	mand1469	5	chib1249	2	guai1249	1
aust1307	74	tupi1275	5	pidg1258	2	book1242	1
sino1245	28	drav1251	5	kart1248	2	tara1323	1
afro1255	25	araw1281	5	mixe1284	2	ticu1244	1
turk1311	20	nucl1709	4	toto1251	2	kore1284	1
maya1287	16	taik1256	3	cent2225	2	mata1289	1
ural1272	12	mong1349	3	tuca1253	2	japo1237	1
arti1236	9	nakh1245	3	gong1255	2	arau1255	1
otom1299	9	abkh1242	2	misu1242	2	atha1245	1
quec1387	8	krua1234	2	hmon1336	2	khoe1240	1
utoa1244	7	eski1264	2	nucl1710	1	tebe1251	1
nilo1247	6	ayma1253	2				

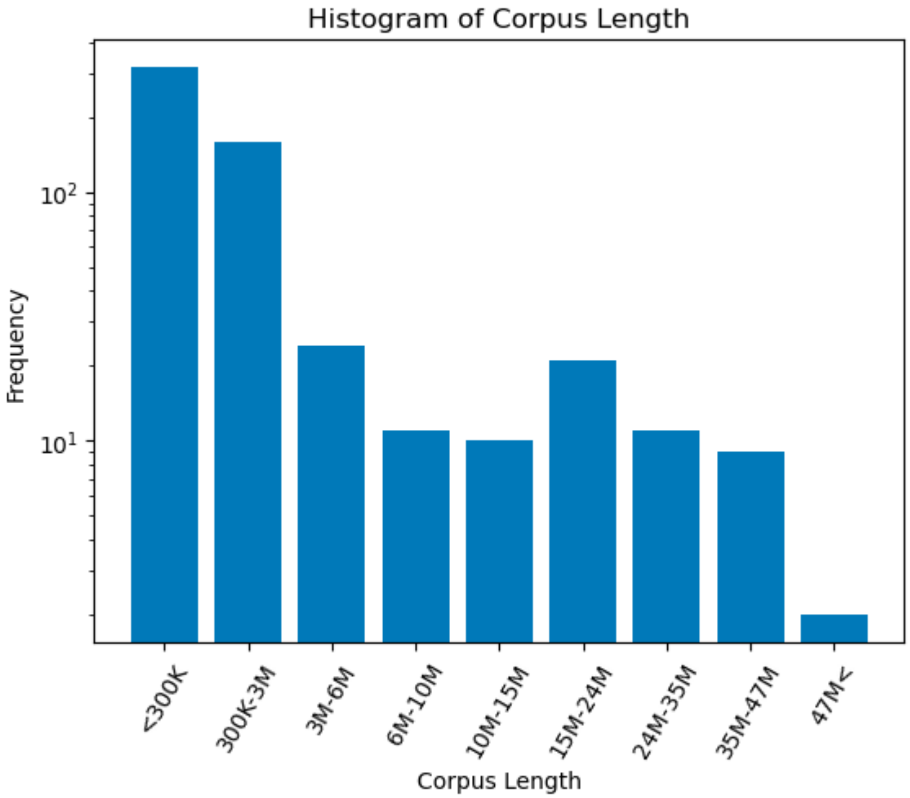
Glot500: Sentences per family



Glot500: Sentences per family

Family	Sentences		Family	Sentences		Family	Sentences		Family	Sentences
indo1319	977086139		maya1287	2892664		abkh1242	389492		cent2225	68472
drav1251	135350643		japo1237	1497574		gong1255	346243		hmon1336	79294
aust1307	1.14E+08		kart1248	1240388		mand1469	324500		tebe1251	50645
afro1255	7.58E+07		quec1387	1194197		chib1249	306124		krua1234	46151
turk1311	63025704		pidg1258	1060411		toto1251	260046		guai1249	44473
atla1278	5.77E+07		otom1299	966777		mixe1284	248719		tuca1253	41681
ural1272	36702676		nakh1245	777504		arau1255	155882		choc1280	39415
aust1305	16747595		utoa1244	735554		atha1245	147702		nucl1708	34349
arti1236	9767069		nilo1247	632011		tara1323	133251		ticu1244	31852
taik1256	8005494		araw1281	551863		misu1242	126118		nucl1710	31765
kore1284	6468444		tupi1275	495319		khoe1240	109747		book1242	30698
mong1349	5107392		eski1264	490504		nucl1709	108755		mata1289	30517
sino1245	4953590		ayma1253	434899						

Corpus size per language: Distribution



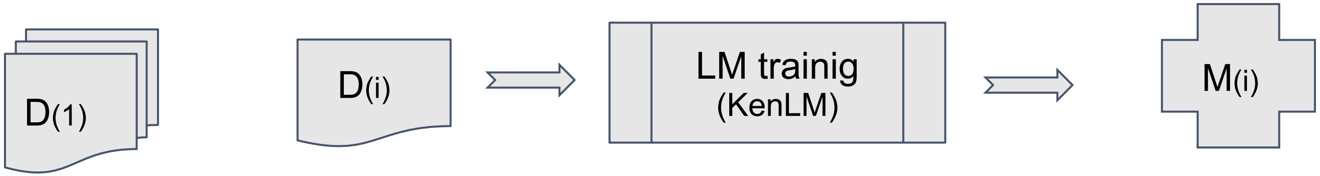
Script detection

- Tajik: Arabic and Cyrillic
- Mongolian: Mongolian, Cyrillic, and Latin

- We detect the script for each sentence
- Treat each language-script as separate entity

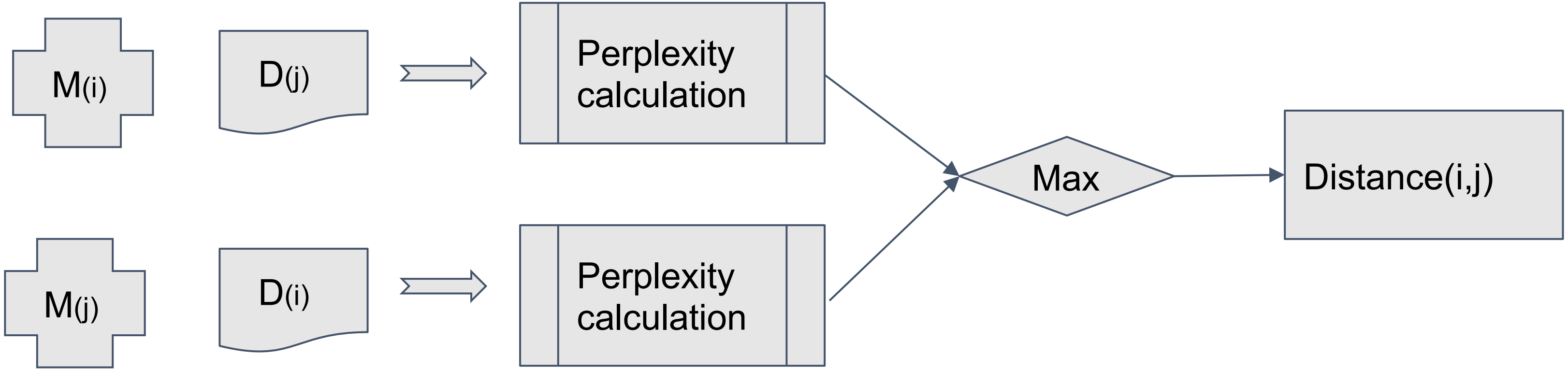
N-gram language models

- D(i): Data for language-script i
- M(i): KenLM Character-level LM using D(i)



Perplexity-based language divergence

$$D(i,j) = \max\big(PP(M(i),D(j)),\ PP(M(j),d(i))\big)$$



Sentence/corpus level filters

- **Sentence level filters**
  - eliminate noisy sentences
- **Corpus level filters:**
  - Drop the whole corpus
  - Majority of the sentences are incorrect
    - Data belongs to another language
    - Non meaningful content from web
- **LangID based filters**
- **Homogeneity Clustering Filters**

Sentence level filters

- Character repetition
- Word repetition
- Special characters
- Small sentences
- Duplicates

Corpus level filters

- Language script mismatch
- Perplexity mismatch
  - Nearest neighbor of L(i) is not a typological family member

LangID filters

- **Out-of-model cousin issue**
- **Combine multiple LangID methods**
  - CLD2 and CLD3
  - LangID.py
  - LangDetect
  - EquilID
  - Fasttext
  - Franc (414 langs)
  - AfroLID (517 langs)
  - CIS-Fasttext (13xx languages from PBS and JW)

LandID for head languages

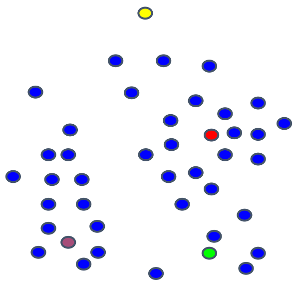
- Works pretty well
- Main issue 1: close languages not covered by LangID
  - E.g., Lombard vs Italian
- Main issue 2: domain, historical text, genre (tweets) etc.

LangID for tail languages (in progress)

- Accept if trusted LIDs agree
- Accept if trusted LIDs agree on macro language
- Accept metadata if confirmed by trusted LIDs
- Accept metadata if macro language confirmed by trusted LID
- Accept metadata if we don’t have LID and i is unique
- Accept metadata if we don’t have LID and i is unique modulo varieties

Is a corpus mono- or bilingual?

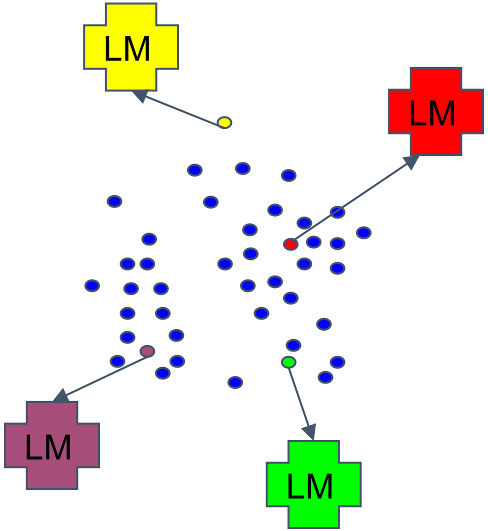
- **Homogeneity Clustering filters**



Pick K cluster seeds

Is a corpus mono- or bilingual?

- **Homogeneity Clustering filters**

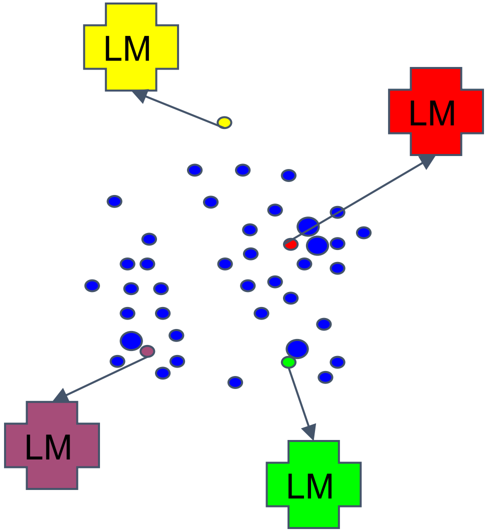


- Train an m-gram LM for each cluster
- For each point find the distance to closest cluster.

Distance = Perplexity of sentence given the language model.

Is a corpus mono- or bilingual?

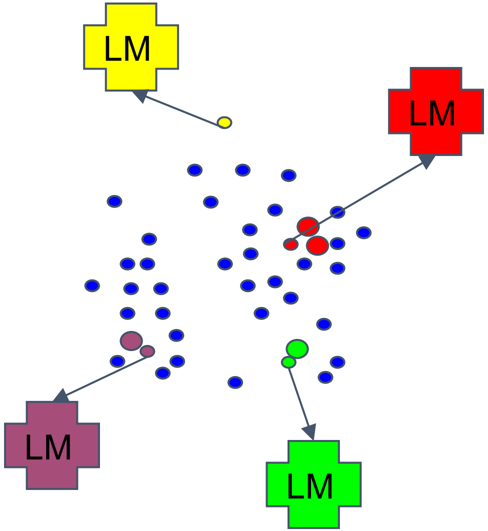
- **Homogeneity Clustering filters**



- Pick the first K samples with least distance to a cluster

Is a corpus mono- or bilingual?

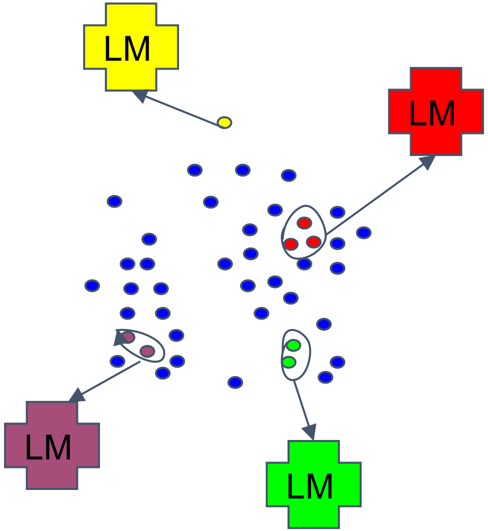
- **Homogeneity Clustering filters**



- Pick the first K samples with least distance to a cluster
- Add them to corresponding cluster

Is a corpus mono- or bilingual?

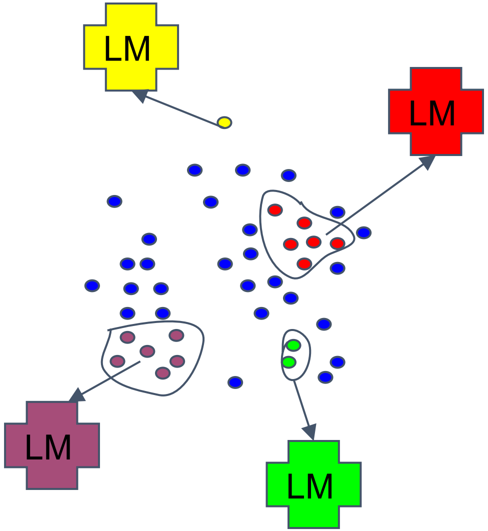
- **Homogeneity Clustering filters**



- Recreate the language models

Is a corpus mono- or bilingual?

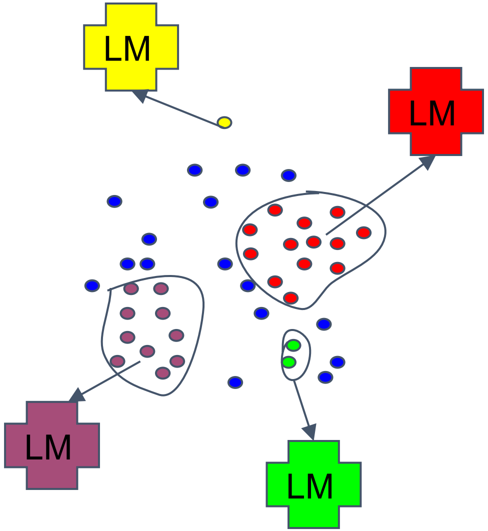
- **Homogeneity Clustering filters**



- Repeat:
  - Find the closest cluster to each sample
  - Add first K samples with the least distance to the corresponding cluster
  - Update language models

Is a corpus mono- or bilingual?

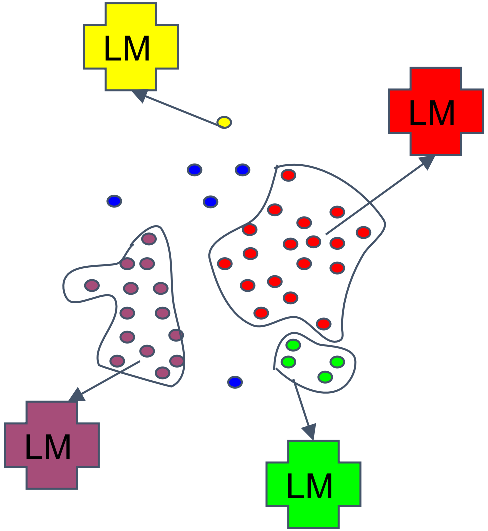
- **Homogeneity Clustering filters**



- Repeat:
  - Find the closest cluster to each sample
  - Add first K samples with the least distance to the corresponding cluster
  - Update language models

Is a corpus mono- or bilingual?

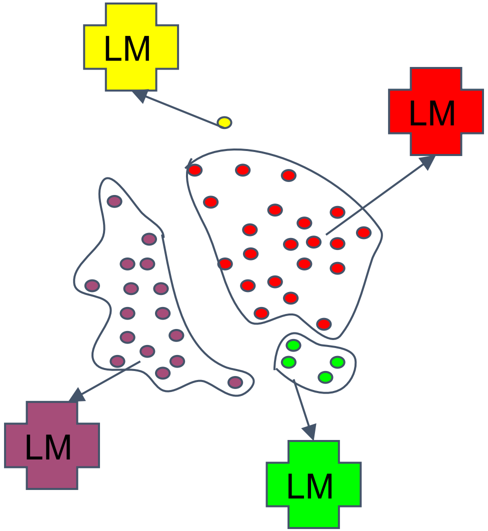
- **Homogeneity Clustering filters**



- Repeat:
  - Find the closest cluster to each sample
  - Add first K samples with the least distance to the corresponding cluster
  - Update language models

Is a corpus mono- or bilingual?

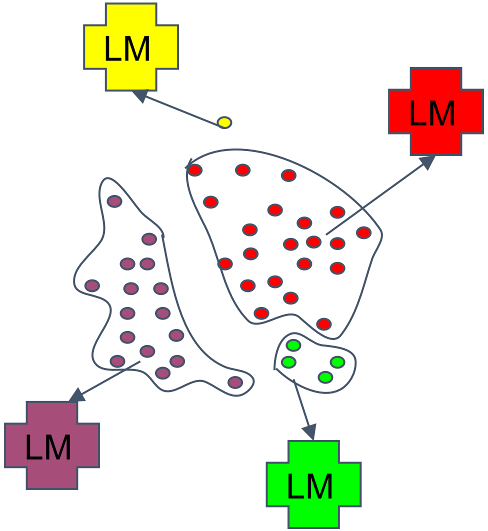
- **Homogeneity Clustering filters**



- Repeat:
  - Find the closest cluster to each sample
  - Add first K samples with the least distance to the corresponding cluster
  - Update language models

Is a corpus mono- or bilingual?

- **If we end up with clusters that highly diverge in terms of perplexity, then we judge the cluster to be multilingual.**



- Repeat:
  - Find the closest cluster to each sample
  - Add first K samples with the least distance to the corresponding cluster
  - Update language models

Glot500 model: Training

Glot500-c: Subset of Glot2000-c

- Language-scripts with at least 30k sentences
- 511 languages
- 534 language-scripts
- 610 GB

Glot500-m: Model trained on Glot500-c

- Continuous pretraining of XLM-R base
- Sampling using multinomial distribution to alleviate bias towards high-resource languages
- Early stopping on average of downstream tasks

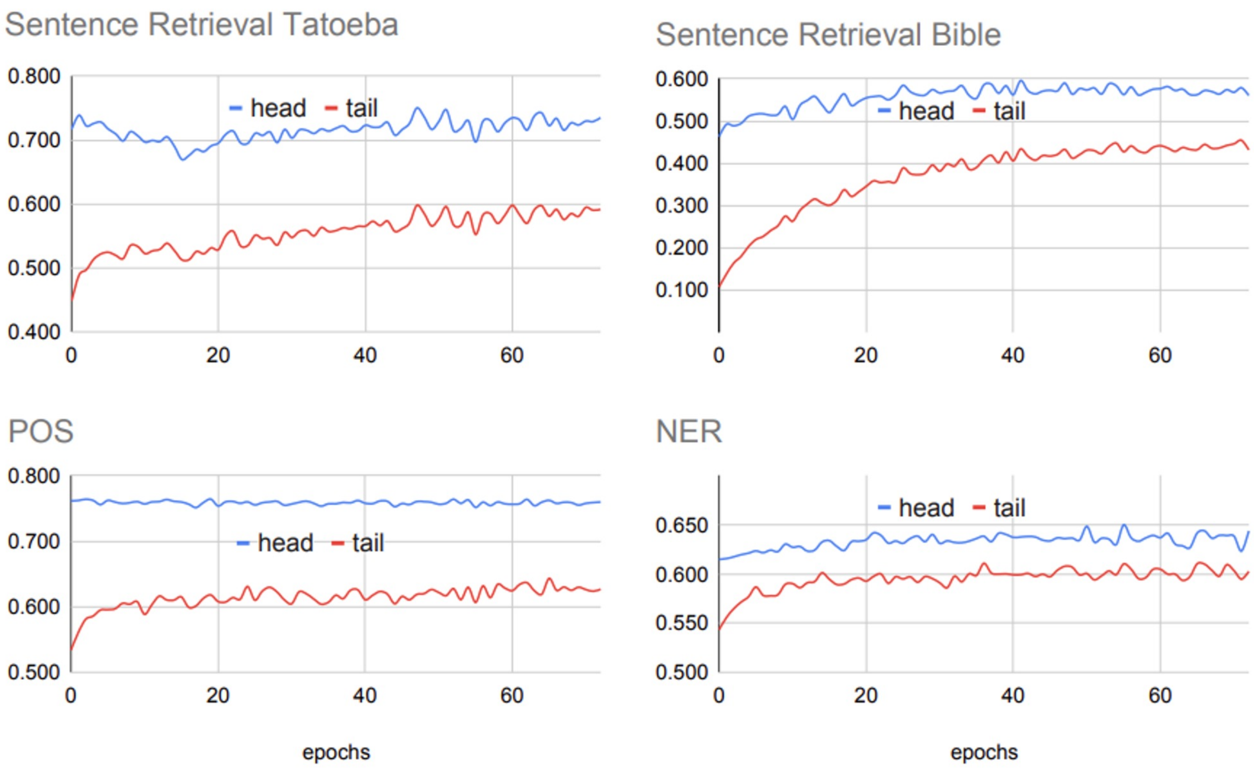
Glot500-m: Vocabulary extension

- Sentence piece with ULM: 250K tokens
- Merge with XLM-R vocabulary
- 150K new tokens
- Vocabulary size 250K + 150K = 400K
- Makes a huge difference for new scripts
- Apart from scripts, makes frustratingly little difference

Glot500: Parameters

	XLM-R-B	XLM-R-L	Glot500-m
Model Size	278M	560M	395M
Vocab Size	250K	250K	401K
Transformer Size	86M	303M	86M

Early stopping



An LLM for 500 languages: Challenges

- **Collect** good data for tail languages
- **Evaluate** tail languages
- Determine **critical factors** for tail languages

How to evaluate tail languages

Tail language evaluation: Challenges

- Most papers claim: we cover N languages
- But for many/most languages there is no quantitative evidence!
- What does coverage mean?

Tail lang evaluation: Challenges

Building Machine Translation Systems for the next 1000 Lang's

- ti tigrinya 4M
- ay aymara 300K
- bm bambara 200K
- ts tsonga ts 1.3M
- lus miso 8M
- Dyula 130K
- We conduct and report the findings from human evaluations of our models (on a subset of 28 languages), confirming that it is possible to build functioning MT systems by following the recipe described in this paper (4.4).
- Impressive. Significant advance over prior work. But how much progress for low-resource?

en→ti	5.4
en→ay	5.1
en→bm	5.0
en→ts	4.9
en→lus	4.9

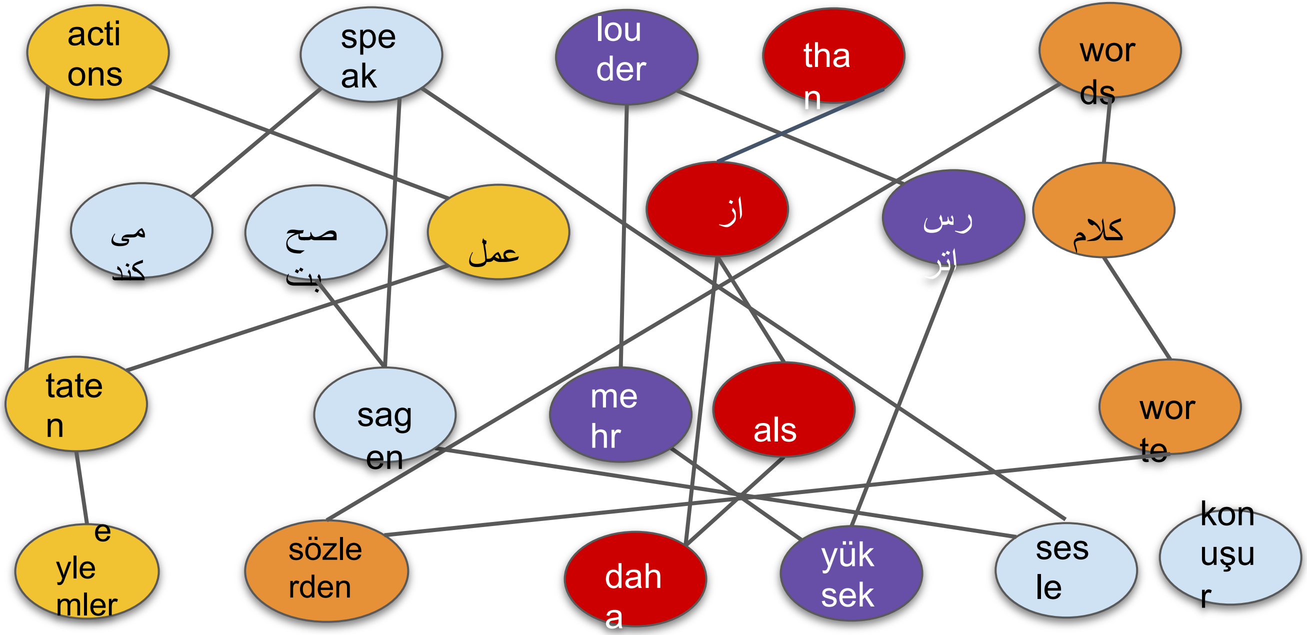
Evaluation tasks

- Pseudo Perplexity
- Round-trip alignment
- Sentence retrieval
  - Bible
  - Tatoeba
- Sequence labeling
  - NER
  - POS
- Text classification

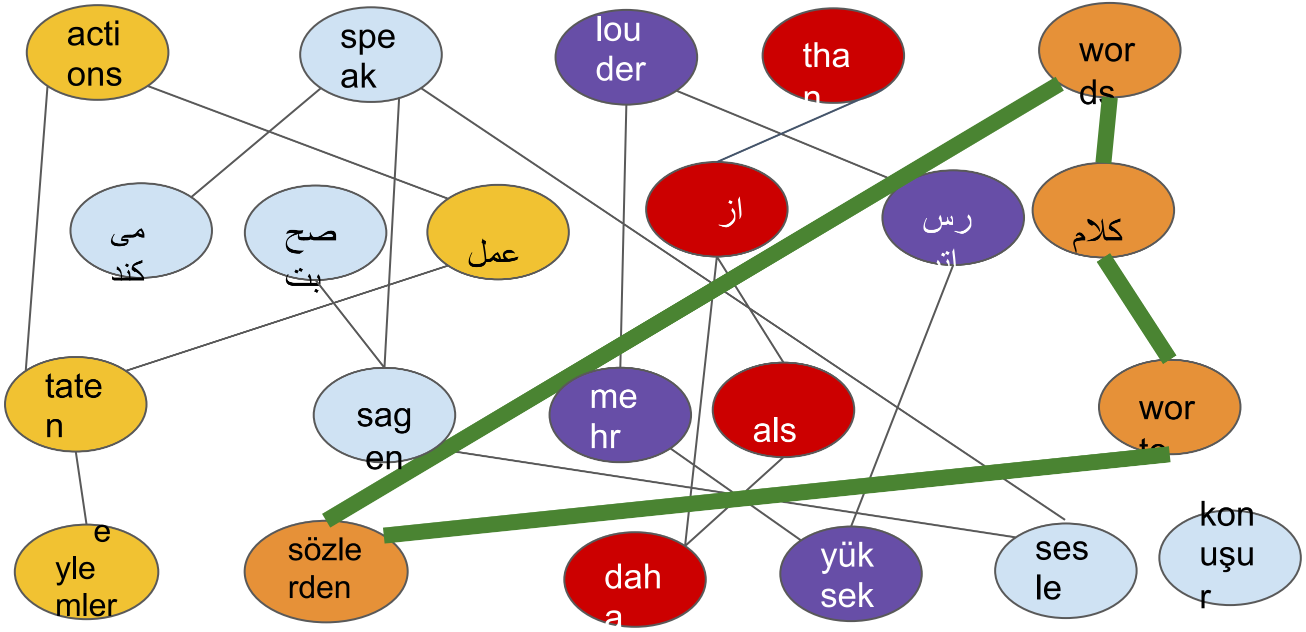
Evaluation tasks

	|head|	|tail|	measure (%)
Sentence Retrieval Tatoeba	70	28	Top10 Acc.
Sentence Retrieval Bible	94	275	Top10 Acc.
Text Classification	90	264	F1
NER	89	75	F1
POS	63	28	F1
Roundtrip Alignment	85	288	Accuracy

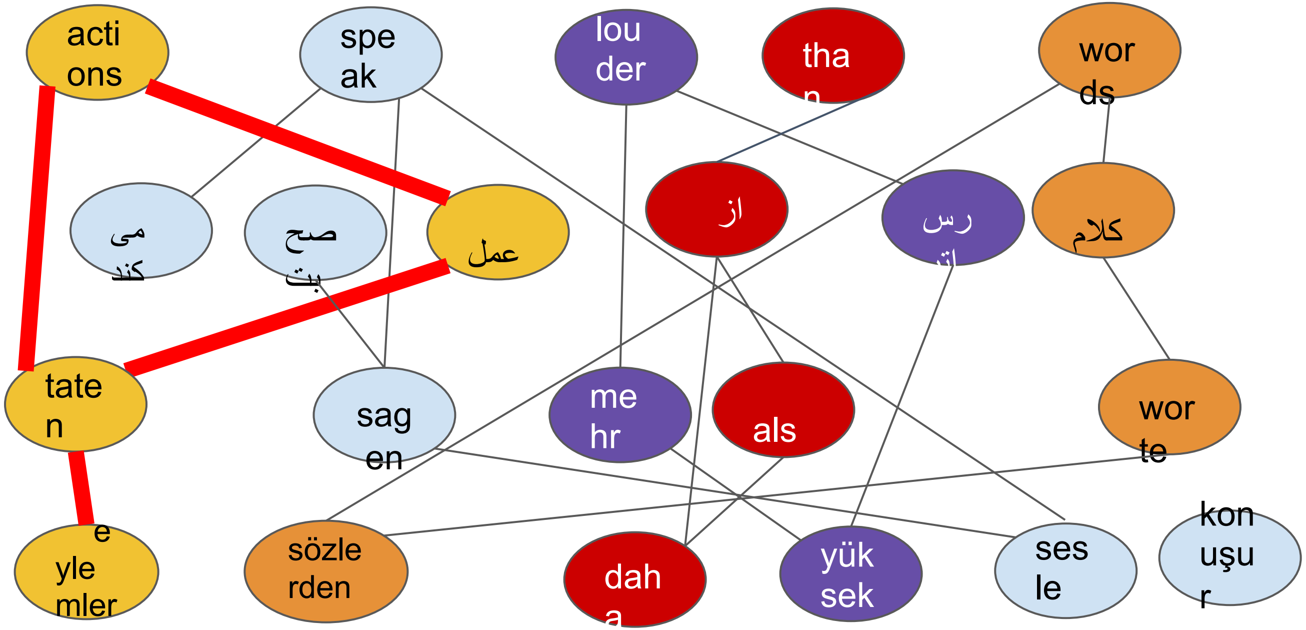
Round Trip Alignment



Round Trip Alignment



Round Trip Alignment



Glot500 results: Average over languages

	tail			head			all		
	XLM-R-B	XLM-R-L	Glot500-m	XLM-R-B	XLM-R-L	Glot500-m	XLM-R-B	XLM-R-L	Glot500-m
Pseudoperplexity	304.2	168.6	**12.2**	12.5	**8.4**	11.8	247.8	136.4	**11.64**
Sentence Retrieval Tatoeba	32.6	33.6	**59.8**	66.2	71.1	**75.0**	56.6	60.4	**70.7**
Sentence Retrieval Bible	7.4	7.1	**43.2**	54.2	58.3	**59.0**	19.3	20.1	**47.3**
Text Classification	13.7	13.9	**46.6**	51.3	**60.5**	54.7	23.3	25.8	**48.7**
NER	47.5	51.8	**60.7**	61.8	**66.0**	63.9	55.3	59.5	**62.4**
POS	41.7	43.5	**62.3**	76.4	**78.4**	76.0	65.8	67.7	**71.8**
Roundtrip Alignment	2.57	3.13	**4.45**	3.42	4.06	**5.46**	2.77	3.34	**4.68**

Glot500 vs XLM-R-Base: Pseudoperplexity

	head languages	tail languages
Glot500-m is better	37	420
XLM-R-B is better	69	8

Glot500 vs XLM-R-Base: Pseudoperplexity

- XML-R-B outperforms Glot500 on 8 langs
- 5 with similar head languages:
  - Standar Estonian -> Estonian
  - Gheg Albanian -> Albanian
  - Norwegian Bokmal -> Norwegian
  - Serbo Croatian -> Serbian
  - Standard Latvian -> Latvian
- 3 with new scripts:
  - Santali -> Ol Chiki script
  - Dhivehi -> Thaana script
  - Inuktitut -> Inuktitut Syllabics
  - Artifact of pseudoperplexity evaluation

	head languages	tail languages
Glot500-m is better	37	420
XLM-R-B is better	69	8

Langs with high pseudoperplexity (up to 94)

- Toki Pona: constructed language, high variability
- Mesopotamian Arabic: tweets
- Three Nilotic languages: Luo, Acoli, Teso
    - Also highly variable?
    - Train/test mismatch?

Glot500 vs XLM-R: Best/worst results

		language-script	XLMR	Glot500	gain			language-script	XLMR	Glot500	gain
high end	SentRetr Tatoeba	tat C Tatar	10.3	70.3	60.0	high end	SentRetr Bible	uzn C Northern Uzbek	5.4	87.0	81.6
		nds L Low German	28.8	77.1	48.3			crs L Seselwa Creole	7.4	80.6	73.2
		tuk L Turkmen	16.3	63.5	47.3			srn L Sranan Tongo	6.8	79.8	73.0
		ile L Interlingue	34.6	75.6	41.0			uzb C Uzbek	6.2	78.8	72.6
		uzb C Uzbek	25.2	64.5	39.3			bcl L Central Bikol	10.2	79.8	69.6
low end		dtp L Kadazan Dusun	5.6	21.1	15.5	low end		xav L Xavánte	2.2	5.0	2.8
		kab L Kabyle	3.7	16.4	12.7			mau L Huautla Mazatec	2.4	3.6	1.2
		pam L Pampanga	4.8	11.0	6.2			ahk L Akha	3.0	3.2	0.2
		lvs L Standard Latvian	73.4	76.9	3.5			aln L Gheg Albanian	67.8	67.6	-0.2
		nob L Bokmål	93.5	95.7	2.2			nob L Bokmål	82.8	79.2	-3.6
high end	NER	div T Dhivehi	0.0	50.9	50.9	high end	POS	mlt L Maltese	21.3	80.3	59.0
		che C Chechen	15.3	61.2	45.9			sah C Yakut	21.9	76.9	55.0
		mri L Maori	16.0	58.9	42.9			sme L Northern Sami	29.6	73.6	44.1
		nan L Min Nan	42.3	84.9	42.6			yor L Yoruba	22.8	64.2	41.4
		tgk C Tajik	26.3	66.4	40.0			quc L K'iche'	28.5	64.1	35.6
low end		zea L Zeeuws	68.1	67.3	-0.8	low end		lzh H Literary Chinese	11.7	18.4	6.7
		vol L Volapük	60.0	59.0	-1.0			nap L Neapolitan	47.1	50.0	2.9
		min L Minangkabau	42.3	40.4	-1.8			hyw A Western Armenian	79.1	81.1	2.0
		wuu H Wu Chinese	28.9	23.9	-5.0			kmr L Northern Kurdish	73.5	75.2	1.7
		lzh H Literary Chinese	15.7	10.3	-5.4			aln L Gheg Albanian	54.7	51.2	-3.5

Languages with multiple scripts

lang-script		XLM-R-B	Glot500	gain
uig_Arab	head	0.458	0.562	0.104
uig_Latn	tail	0.098	0.628	0.530
hin_Deva	head	0.670	0.766	0.096
hin_Latn	tail	0.136	0.432	0.296
uzb_Latn	head	0.548	0.676	0.128
uzb_Cyrl	tail	0.062	0.788	0.726
kaa_Cyrl	tail	0.176	0.738	0.562
kaa_Latn	tail	0.092	0.434	0.342
kmr_Cyrl	tail	0.040	0.424	0.384
kmr_Latn	tail	0.358	0.630	0.272
tuk_Cyrl	tail	0.136	0.650	0.514
tuk_Latn	tail	0.096	0.662	0.566

Major eval result: Poor performance on 10s of langs

		language-script		XLMR	Glot500	gain
SentRetr Tatoeba	high end	tat	C Tatar	10.3	70.3	60.0
		nds	L Low German	28.8	77.1	48.3
		tuk	L Turkmen	16.3	63.5	47.3
		ile	L Interlingue	34.6	75.6	41.0
		uzb	C Uzbek	25.2	64.5	39.3
	low end	dtp	L Kadazan Dusun	5.6	21.1	15.5
		kab	L Kabyle	3.7	16.4	12.7
		pam	L Pampanga	4.8	11.0	6.2
		lvs	L Standard Latvian	73.4	76.9	3.5
		nob	L Bokmål	93.5	95.7	2.2
NER	high end	div	T Dhivehi	0.0	50.9	50.9
		che	C Chechen	15.3	61.2	45.9
		mri	L Maori	16.0	58.9	42.9
		nan	L Min Nan	42.3	84.9	42.6
		tgk	C Tajik	26.3	66.4	40.0
	low end	zea	L Zeeuws	68.1	67.3	-0.8
		vol	L Volapük	60.0	59.0	-1.0
		min	L Minangkabau	42.3	40.4	-1.8
		wuu	H Wu Chinese	28.9	23.9	-5.0
		lzh	H Literary Chinese	15.7	10.3	-5.4

		language-script		XLMR	Glot500	gain
SentRetr Bible	high end	uzn	C Northern Uzbek	5.4	87.0	81.6
		crs	L Seselwa Creole	7.4	80.6	73.2
		srn	L Sranan Tongo	6.8	79.8	73.0
		uzb	C Uzbek	6.2	78.8	72.6
		bcl	L Central Bikol	10.2	79.8	69.6
	low end	xav	L Xavánte	2.2	5.0	2.8
		mau	L Huautla Mazatec	2.4	3.6	1.2
		ahk	L Akha	3.0	3.2	0.2
		aln	L Gheg Albanian	67.8	67.6	-0.2
		nob	L Bokmål	82.8	79.2	-3.6
POS	high end	mlt	L Maltese	21.3	80.3	59.0
		sah	C Yakut	21.9	76.9	55.0
		sme	L Northern Sami	29.6	73.6	44.1
		yor	L Yoruba	22.8	64.2	41.4
		quc	L K'iche'	28.5	64.1	35.6
	low end	lzh	H Literary Chinese	11.7	18.4	6.7
		nap	L Neapolitan	47.1	50.0	2.9
		hyw	A Western Armenian	79.1	81.1	2.0
		kmr	L Northern Kurdish	73.5	75.2	1.7
		aln	L Gheg Albanian	54.7	51.2	-3.5

At least one measure for each covered language

Glot500-m	Language-Script	XLM-R-B	XLM-R-L	Glot500-m	Language-Script	XLM-R-B	XLM-R-L	Glot500-m
8.8	tsn_Latn	264.7	137.8	**12.5**	orm_Latn	23.4	**8.6**	16
7.2	pon_Latn	928.4	181.9	**19.2**	luo_Latn	699.4	258.5	**85.1**
18.3	nmf_Latn	297.6	310.6	**44.9**	pcm_Latn	38.3	169.6	**3.6**
15.2	ajg_Latn	147.1	149.5	**22.6**	nnb_Latn	364.1	95	**28.6**
6.4	tir_Ethi	28.3	15.7	**4.4**	kaz_Cyrl	**4.3**	5.4	9.6
7.6	bhw_Latn	411.2	126.2	**21.6**	dzo_Tibt	8.5	**3.3**	5.7
17.6	mhr_Cyrl	122.9	168.4	**5.8**	sun_Latn	23.6	**11.9**	17
5.8	swe_Latn	4.8	**3.5**	12.7	vec_Latn	40.6	21.1	**9.2**
9.7	scn_Latn	117	64.9	**7.8**	ayr_Latn	261.1	237.6	**27.7**
4.3	udm_Cyrl	356.7	224.9	**6.7**	oke_Latn	209.2	220.1	**13.0**
11.9	ifb_Latn	246.3	177.9	**5.1**	kur_Latn	14.2	**6.8**	10.3
19.5	naq_Latn	136.8	60.2	**15.7**	mgh_Latn	680	272.8	**23.7**
37.7	zlm_Latn	5.6	**3.3**	4.6	tgk_Cyrl	181.3	153	**4.5**
7.2	hrx_Latn	478.1	679.1	**14.9**	sop_Latn	607.5	228.2	**29.5**
9.4	lzh_Hani	70	58	**21.8**	mos_Latn	272.6	118.3	**13.2**
5.2	pap_Latn	674.4	149.3	**18.1**	rap_Latn	36.1	31.1	**2.8**
17.5	cfm_Latn	235.1	155	**14.0**	prk_Latn	69.4	45.9	**7.1**
19.6	chv_Cyrl	122.5	73.8	**5.4**	uzb_Cyrl	236.2	138.4	**4.9**
17.3	tdt_Latn	641.9	78.6	**9.7**	tog_Latn	821.1	777.7	**13.4**
14.3	pan_Guru	4.4	**2.5**	4.3	mal_Mlym	5	**3.7**	6.2

Major eval result: Poor performance on 10s of langs

ceb_Latn	28	30	**49**	lhu_Latn	6	6	**30**	sot_Latn	11	8	**45**
ces_Latn	50	**65**	53	lin_Latn	10	7	**49**	spa_Latn	61	**69**	60
cfm_Latn	8	8	**55**	lit_Latn	54	**66**	53	sqi_Latn	57	**68**	60
che_Cyrl	11	6	**20**	loz_Latn	10	10	**48**	srm_Latn	10	9	**53**
chv_Cyrl	8	7	**52**	ltz_Latn	22	30	**52**	srn_Latn	10	9	**53**
cmn_Hani	53	**62**	56	lug_Latn	16	9	**45**	srp_Latn	55	**67**	56
cnh_Latn	7	8	**56**	luo_Latn	12	10	**39**	ssw_Latn	14	17	**40**
crh_Cyrl	22	31	**57**	lus_Latn	11	7	**52**	sun_Latn	40	**47**	**47**
crs_Latn	14	17	**61**	lzh_Hani	46	**55**	**55**	suz_Deva	15	13	**53**
csy_Latn	9	7	**52**	mad_Latn	23	28	**56**	swe_Latn	60	**66**	56
ctd_Latn	9	8	**56**	mah_Latn	6	6	**42**	swh_Latn	47	**59**	56
ctu_Latn	15	14	**51**	mai_Deva	34	39	**59**	sxn_Latn	11	8	**46**
cuk_Latn	15	7	**44**	mal_Mlym	56	**64**	60	tam_Taml	56	**61**	60
cym_Latn	46	**51**	48	mam_Latn	10	6	**31**	tat_Cyrl	21	28	**64**
dan_Latn	51	**62**	50	mar_Deva	55	**63**	60	tbz_Latn	6	6	**43**
deu_Latn	56	**65**	53	mau_Latn	5	5	**6**	tca_Latn	5	5	**47**
djk_Latn	12	10	**46**	mbb_Latn	11	7	**48**	tdt_Latn	16	13	**56**
dln_Latn	10	5	**52**	mck_Latn	15	10	**41**	tel_Telu	55	**65**	60
dtp_Latn	9	8	**39**	mcn_Latn	13	9	**43**	teo_Latn	12	8	**26**
dyu_Latn	6	8	**52**	mco_Latn	6	7	**28**	tgk_Cyrl	10	7	**55**
dzo_Tibt	6	5	**55**	mdy_Ethi	6	7	**47**	tgl_Latn	48	**60**	56

Major eval result: Poor performance on 10s of langs

- Key methodology requirement for low-resource papers
- Minimum sanity check on actual coverage

An LLM for 500 languages: Challenges

- **Collect** good data for tail languages
- **Evaluate** tail languages
- Determine **critical factors** for tail languages

Critical factors for tail language performance

Non-Factor: Tokenization?

- Character-based representation: performance for scripts that are not covered is terrible
- Byte-based representation: tokenization is only a minor factor?

Factor corpus size

- Other things being equal, corpus size is the key factor that determines performance.
- But things are not equal in many cases!

Factor script

lang-script		XLM-R-B	Glot500	gain
uig_Arab	head	0.458	0.562	0.104
uig_Latn	tail	0.098	0.628	0.530
hin_Deva	head	0.670	0.766	0.096
hin_Latn	tail	0.136	0.432	0.296
uzb_Latn	head	0.548	0.676	0.128
uzb_Cyrl	tail	0.062	0.788	0.726
kaa_Cyrl	tail	0.176	0.738	0.562
kaa_Latn	tail	0.092	0.434	0.342
kmr_Cyrl	tail	0.040	0.424	0.384
kmr_Latn	tail	0.358	0.630	0.272
tuk_Cyrl	tail	0.136	0.650	0.514
tuk_Latn	tail	0.096	0.662	0.566

Factor family

The more langs from a family we support the better performance. (SentRetrB)

family	$\lvert L_G \rvert$	$\lvert L_X \rvert$	XLM-R-B	Glot500-m	gain
indo1319	91	50	41.5	61.4	19.9
atla1278	69	2	5.5	45.2	39.6
aust1307	53	6	13.7	47.0	33.2
turk1311	22	7	20.1	62.9	42.8
sino1245	22	2	7.6	38.9	31.3
maya1287	15	0	3.8	20.3	16.4
afro1255	12	5	13.0	34.3	21.4

Factor related langs

- Glot+1: Adapt to only 1 new language
- Top 3 langs: no “cousin”
- Bottom 3: related lang in Glot500

lang-script	Glot+1	Glot500-m
rug_Latn, Roviana	**51.0**	49.0
yan_Latn, Mayangna/Sumo	**46.4**	31.8
wbm_Latn, Wa/Va	**49.6**	46.4
ctd_Latn, Tedim Chin	47.4	**59.4**
quh_Latn, Southern Quechua	33.4	**56.2**
tat_Cyrl, Tatar	58.8	**67.2**

- Is there really a curse of multilinguality?
- There definitely is a **blessing** of multilinguality!

Summary

An LLM for 500 languages: Challenges

- **Collect** good data for tail languages
- **Evaluate** tail languages
- Determine **critical factors** for tail languages