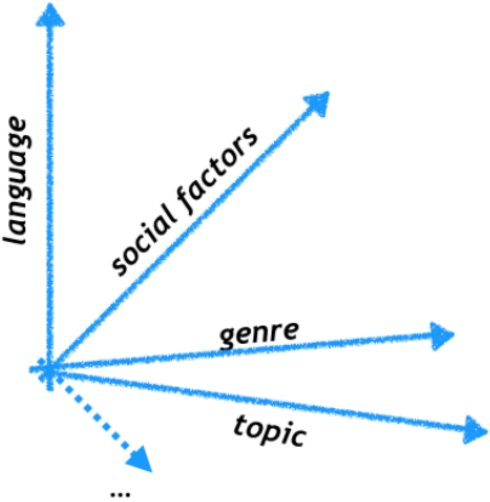
ROBustness in NLP over the years

Lexical normalization

Multi-task learning

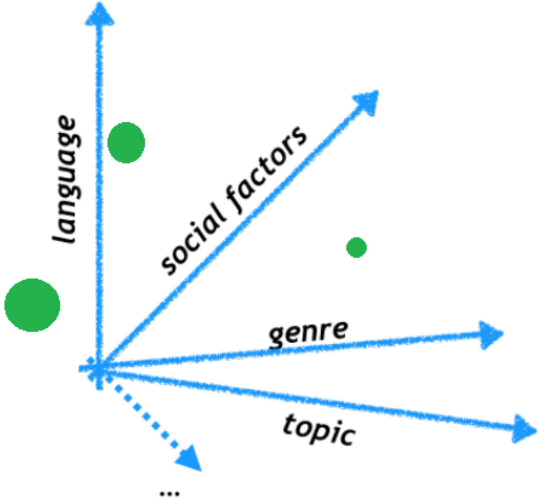
Is X solved?

Problem



From: Barbara Plank. What to do about non-standard (or non-canonical) language in NLP.

Problem



1. Lexical Normalization

u	hve	to	let	ppl	decide	what	dey	want	to	do
you	have	to	let	people	decide	what	they	want	to	do

Lexical Normalization

Situation in 2015:

- Some benchmarks for English: main one LexNorm
- Many models assume gold detection
- Some people working on their own languages
- Differences in models, task definitions and metrics

Lexical Normalization

Situation in 2015:

- Some benchmarks for English: main one LexNorm
- Many models assume gold detection
- Some people working on their own languages
- Differences in models, task definitions and metrics

MoNoise

- First multi-lingual normalization model
- SOTA wherever evaluated
- Outputs top-n; successfully integrated in syntactic parsers.

Parsing

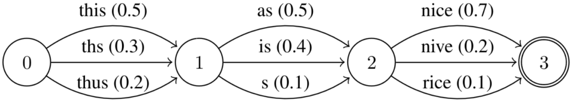


Figure 1: A possible output of the normalization model for the sentence 'ths s nice'.

Parsing

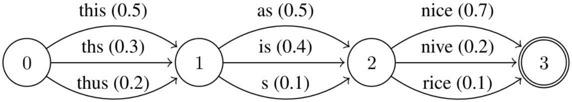


Figure 1: A possible output of the normalization model for the sentence 'ths s nice'.

intersection of a context-free language and a regular language is itself context-free (Bar-Hillel, 1961)

Parsing

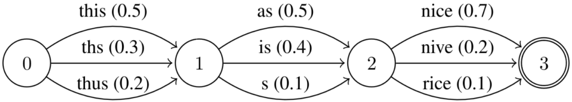


Figure 1: A possible output of the normalization model for the sentence 'ths s nice'.

Performance for syntactic tasks improve when normalizing, even more when integrating the top-n, but still not by a lot.

MultiLexNorm: A Shared Task on Multilingual Lexical Normalization

Rob van der Goot, Alan Ramponi, Arkaitz Zubiaga, Barbara Plank, Benjamin Muller, Iñaki San Vicente Roncal, Nikola Ljubešić, Özlem Çetinoğlu, Rahmad Mahendra, Talha Çolakoğlu, Timothy Baldwin, Tommaso Caselli and Wladimir Sidorenko











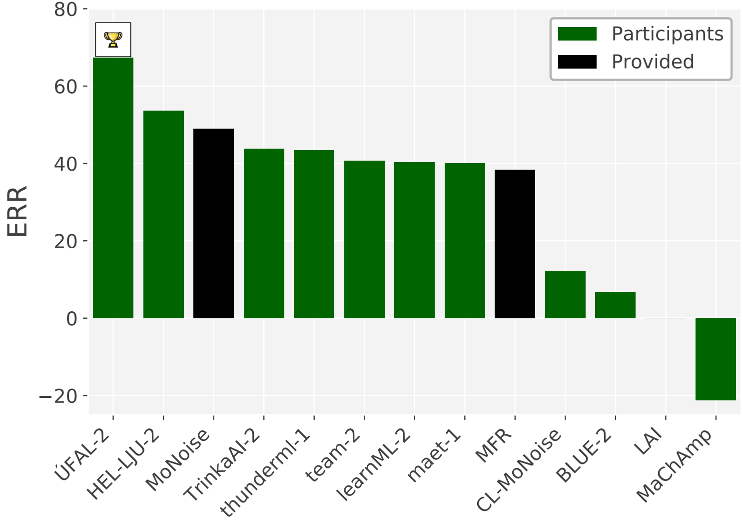

MultiLexNorm

Lang.	Language name	Normalization example
DA	Danish	De skarpe lamper gjorde destromindre ek bedre . De skarpe lamper gjorde destro mindre ikke bedre .
DE	German	ogäj isch hätts auch dwiddern könn Okay ich hätte es auch twittern können
EN	English	u hve to let ppl decide what dey want to do you have to let people decide what they want to do
ES	Spanish	@username cuuxamee sii peroo veen yaa eem @username escúchame sí pero ven ya eh
HR	Croatian	svi frendovi mi nešto rade , veceras san osta sam . svi frendovi mi nešto rade , večeras sam ostao sam .
ID-EN	Indonesian-English	pdhal not fully bcs those ppl jg sih . padahal not fully because those people juga sih .
IT	Italian	a Roma è cosí primavera che sembra gia giov a Roma è così primavera che sembra già giovedì
NL	Dutch	Kga me wss trg rolle vant lachn Ik ga me waarschijnlijk terug rollen van het lachen
SL	Slovenian	jst bi tud najdu kovanec vreden veliko denarja . jaz bi tudi našel kovanec vreden veliko denarja .
SR	Serbian	komunalci kace pocne kaznjavanje ? komunalci kad počne kažnjavanje ?
TR	Turkish	He o dediyin suala cvb verdim He o dediğin suale cevap verdim
TR-DE	Turkish-German	@username Yerimm senii , damkee schatzymm :-* @username Yerim seni , danke Schatzym :-*

MultiLexNorm

- ÚFAL: ByT5 for every word; synthetic data
- HEL-LJU: Pre-classify type of normalization (BERT) $\mapsto$ Char-SMT
- MoNoise: Feature-based, generate candidates and rank
- BLUE: NMT MBart-50
- CL-MoNise: Cross-lingual
- MaChAmp: Normalization as sequence labeling

Results


80
60
40
20
0
−20
ERR
Participants
Provided
ÚFAL-2
HEL-LJU-2
MoNoise
TrinkaAI-2
thunderml-1
team-2
learnML-2
maet-1
MFR
CL-MoNoise
BLUE-2
LAI
MaChAmp

Results

- Include detection in task (= the hardest part)
- Multi-lingual benchmark
- Wide variety of models
- Near-human performance for some datasets (in-lang/in-domain)

Open problems

- Cross-lingual/multi-lingual normalization
- Tokenization
- Limited downstream gains; lexical level might not be enough
- Bias in languages
- Bias in data source

MultiLexNorm 2

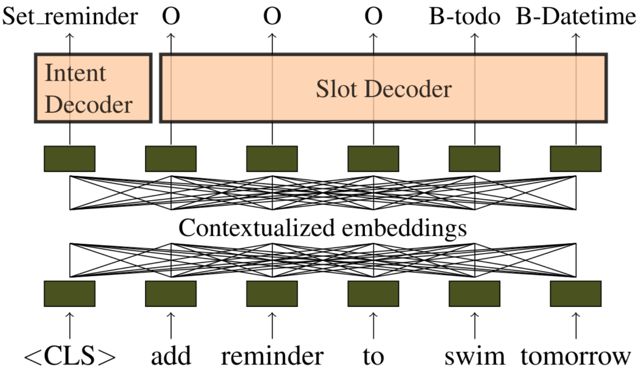
To be held at WNUT2025 (if accepted), including non Indo-European languages!

2. Multi-task learning

Massive Choice, Ample Tasks (MaChAmp):
A Toolkit for Multi-task Learning in NLP

Rob van der Goot **Ahmet Üstün** **Alan Ramponi** **Ibrahim Sharaf**
Barbara Plank
IT University of Copenhagen University of Groningen University of Trento
Fondazione the Microsoft Research - University of Trento COSBI Factmata
robv@itu.dk, a.ustun@rug.nl, alan.ramponi@unitn.it
ibrahim.sharaf@factmata.com, bapl@itu.dk

MaChAmp


Set_reminder
O
O
O
B-todo
B-Datetime
Intent Decoder
Slot Decoder
Contextualized embeddings
<CLS>
add
reminder
to
swim
tomorrow

Multi-task learning

MaChAmp at SemEval-2022 Tasks 2, 3, 4, 6, 10, 11, and 12: Multi-task Multi-lingual Learning for a Pre-selected Set of Semantic Datasets

Rob van der Goot
IT University of Copenhagen
robv@itu.dk

MaChAmp at SemEval-2023 tasks 2, 3, 4, 5, 7, 8, 9, 10, 11, and 12: On the Effectiveness of Intermediate Training on an Uncurated Collection of Datasets.

Rob van der Goot
IT University of Copenhagen
robv@itu.dk

MaChAmp @ SemEval 2022-2023

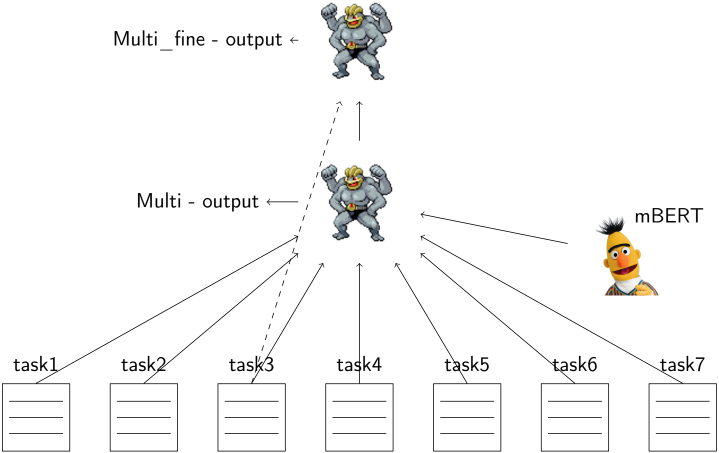
Evaluate effect of:

- Intermediate training with encoder LM's
- Heterogeneous batching
- Dataset smoothing
- Task interactions (correlation study)

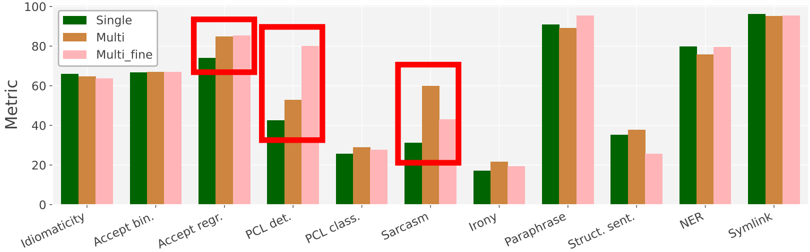
SemEval 2022

SemEval Task	Included sub-tasks	Languages
2: Multilingual Idiomaticity Detection	Idiomaticity detection (1-shot)	EN, PT, GL
3: PreTENS	1: Binary acceptability	EN, IT, FR
	2: Regression acceptability	EN, IT, FR
4: Patronizing and Condescending Language Detection	1: Binary PCL detection	EN
	2: Multi-label PCL classification	EN
6: iSarcasmEval	1: Sarcasm detection	EN, AR
	2: Irony-labeling	EN
	3: Paraphrase sarcasm detection	EN, AR
10: Structured Sentiment Analysis	Expressions, entities and relations	CA, EN, ES, EU, NO
11: MultiCoNER - Multilingual Complex Named Entity Recognition	Named Entity Recognition	BN, DE, EN, ES, FA, HI, KO, MI, NL, RU, TR, ZH
12: Symlink	Entities and relations	EN

Intermediate task finetuning



MaChAmp @ SemEval 2022



Let's do some analysis!

Let's do some analysis!

SemEval 2023

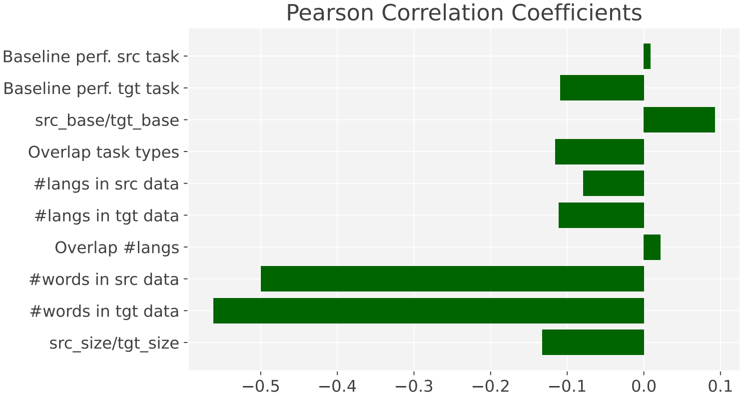
Name	Subtasks	Languages	Size
2. MultiCoNER II	NER	BN, DE, EN, ES, FA, FR, HI, IT, PT, SV, UK, ZH	2,672,490
3. News persuasion	1. News categorization	EN, FR, GE, IT, PO, RU	741,561
	2. Framing classification	EN, FR, GE, IT, PO, RU	725,740
	3. Persuasion technique classification	EN, FR, GE, IT, PO, RU	19,561,550
4. ValueEval	Human value classification	EN	116,294
5. Clickbait spoiling	1. Spoiler type classification	EN	34,520
	2. Spoiler detection	EN	1,647,176
6. LegalEval	1. Rhetorical role detection	EN	755,280
	2. NER	EN	369,205
	3. Legal judgement prediction	EN	5,082
7. Clinical NLI	1. Entailment	EN	21,828
	2. Evidence retrieval	EN	311,687
8. Medical claims	1. Claim identification	EN	549,231
	2. PIO frame extraction	EN	78,864
9. Tweet intimicay	Intimacy Analysis	EN, ES, IT, PT, FR, ZH	73,698
10. Explainable sexism	1. Sexism detection	EN	262,939
	2. Sexism classification	EN	68,043
	3. Fine-grained sexism classification	EN	68,043
11. Le-Wi-Di	1. Hate speech detection*	EN	14,252
	2. Misogyny detection*	AR	12,788
	3. Abuse detection*	EN	64,738
	4. Offensiveness detection*	EN	145,245
12. AfriSenti-SemEval	Sentiment classification	AM, DZ, HA, IG, KR, MA, PCM, PT, SW,	795,449

MaChAmp @ SemEval 2023

	Result	Rank		Result	Rank
task2	73.74	8/18	task8-1	78.40	1/7
task3-1	31.67		task8-2	40.55	1/6
task3-2	38.01		task9	57.47	18/46
task3-3	29.36		task10	?	
task4-1	48	15/42	task11-1	0.69	15/27
task4-2	34	3/20	task11-2	1.11	20/27
task4-2	19	10/12	task11-3	0.47	18/27
task5	?		task11-4	0.61	12/27
task7-1	—		task12	2.26-51.17	33/33
task7-2	75.6	14/19			

Table: Scores and ranking on test data, — means submission failed, and ? means that results are not available yet.

MaChAmp @ SemEval 2023



MaChAmp @ SemEval 2022-2023

Evaluate effect of:

- Intermediate training with encoder LM's: +-
- Heterogeneous batching: -
- Dataset smoothing: -
- Task interactions (correlation study): +-

What else did I learn?

- Don't participate in too many tasks at once
- How to win?
  - Careful tuning
  - Right LM
  - More data
  - Ensembling
  - Download data early
- Most of the time went into obtaining data, understanding data, format conversion
- CRF layer almost always beneficial
- When an instance has 0-n labels, BCE loss and threshold over logits is best
- Conversion of structured task to sequence labeling leads to mediocre performance
- # participants: classification > sequence labeling > others
- # things learned: classification < sequence labeling < others

3. Future

To what extent are these tasks solved? what are the remaining issues?:

- Tokenization
- Language identification
- Cultural awareness

Tokenization

The problem of finding/segmenting tokens (UD):

Input:
If␣momma␣ain't␣happy,␣nobody␣ain't␣happy.

Tokenization:
If␣momma␣ain't␣happy␣,␣nobody␣ain't␣happy␣.

Multi-word expansions:
If␣momma␣is␣not␣happy,␣nobody␣is␣not␣happy.

Subword segmentation:
If␣mo␣##mma␣ai␣##n␣'␣t␣happy␣,␣no␣##body␣ai␣##n␣'␣t␣happy␣.

Methods

1)	Dr. Dron is his backup.							
2)	s=[·][.][])} > ”′]∗*$=\1 \2\3 =g							
3)	biiobiiiobiobiiobiiiiib							
4)	Dr	.	Dro	##n	is	his	backup	.
	b	i	b	i	b	b	b	b

LM for tokenization?

- Finetuning a language model for this task might be overkill
- Multi-task learning can make it efficient: add a decoder head for tokenization
- Adapters used before (costly to train)

Settings

- RB: Rule Based
- ST: Single Task: just tokenization
- MT: Multi-task: UPOS, morph. tagging, lemmatization, dep. parsing
- ML+MT: Multi-lingual Multi-task model

In treebank results

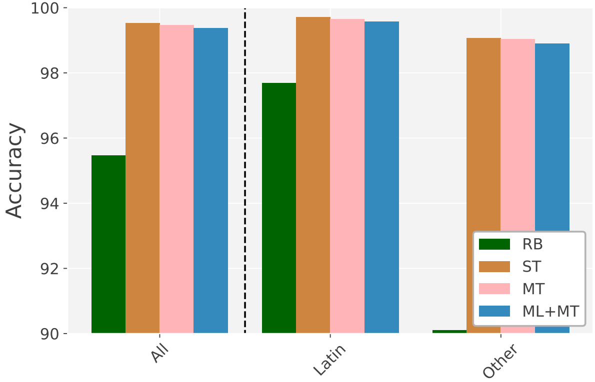


Figure: Results of tokenizers on Latin vs non-Latin languages. RB=RuleBased, ST=SingleTask, MT=MultiTask, ML+MT=Multi-Lingual+MultiTask

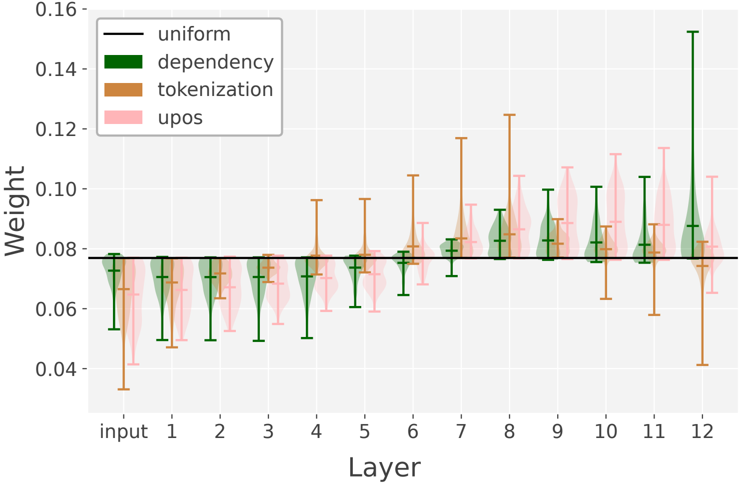
Cross-treebank results

setting	F1 tok.	# treebanks
all	93.23	90
in-language	95.11	34
in-script	94.16	84
new-script	80.11	6

Table: Results on test-only treebanks

More analysis

EACL 2024 findings



Open challenges in language identification

- Many tools/benchmarks available
- When to use which?:

Open challenges in language identification

- Many tools/benchmarks available
- When to use which?:
  - # languages
  - input size
  - # training instances per language
  - scripts
  - language families
  - domains

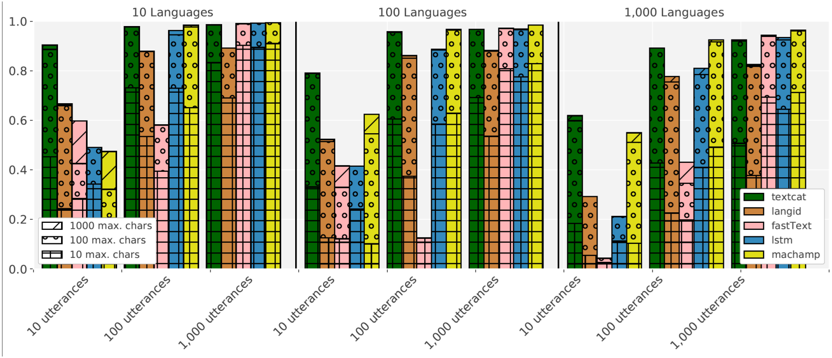
Data

Dataset	langs	scripts	fams	domains
OpenLID	139	25	16	literature, news, wiki, social, grammar, subtitles, spoken
UDHR	397	38	61	rights
LTI LangID	2,110	47	139	wiki, political, religious, grammar
TwitUser	59	20	13	social
MassiveSumm	77	24	13	news
UD2.12	54	11	17	medical, news, academic, wiki, legal, nonfiction, learner-essays, fiction, social, grammar-examples, reviews, religious, spoken
Total	2176/ 7850	51/ 163	145/ 298	

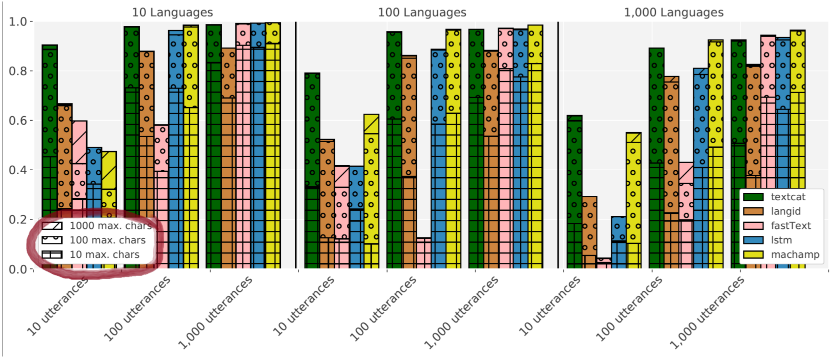
Models

- Heuristics: textcat
- Naive Bayes: langid.py
- Embeddings: FastText
- Neural: BiLSTM
- CLM: Glot500

Size

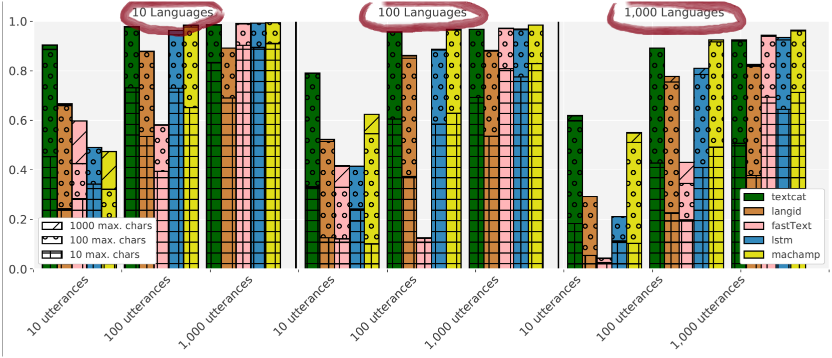

10 Languages
100 Languages
1,000 Languages
1.0
0.8
0.6
0.4
0.2
0.0
10 utterances
100 utterances
1,000 utterances
1000 max. chars
100 max. chars
10 max. chars
textcat
langid
fastText
lstm
machamp

Size



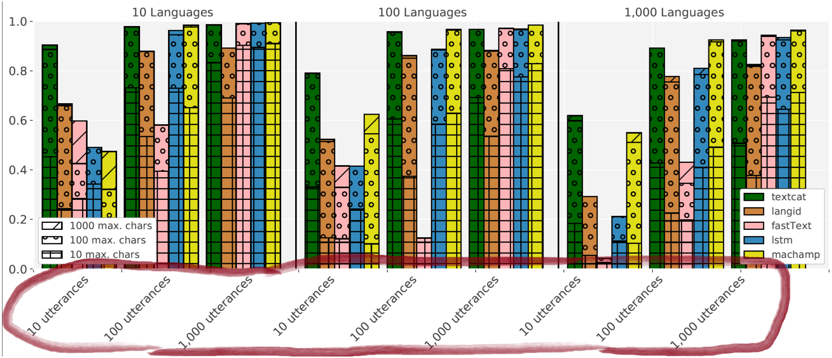
10 Languages
100 Languages
1,000 Languages
1.0
0.8
0.6
0.4
0.2
0.0
1000 max. chars
100 max. chars
10 max. chars
textcat
langid
fastText
lstm
machamp
10 utterances
100 utterances
1,000 utterances
10 utterances
100 utterances
1,000 utterances
10 utterances
100 utterances
1,000 utterances

Size



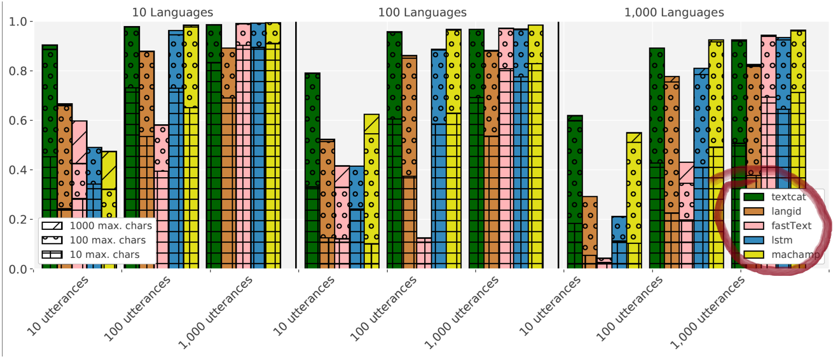
10 Languages
100 Languages
1,000 Languages
1.0
0.8
0.6
0.4
0.2
0.0
1000 max. chars
100 max. chars
10 max. chars
textcat
langid
fastText
lstm
machamp
10 utterances
100 utterances
1,000 utterances
10 utterances
100 utterances
1,000 utterances
10 utterances
100 utterances
1,000 utterances

Size



10 Languages
100 Languages
1,000 Languages
1.0
0.8
0.6
0.4
0.2
0.0
1000 max. chars
100 max. chars
10 max. chars
textcat
langid
fastText
lstm
machamp
10 utterances
100 utterances
1,000 utterances
10 utterances
100 utterances
1,000 utterances
10 utterances
100 utterances
1,000 utterances

Size

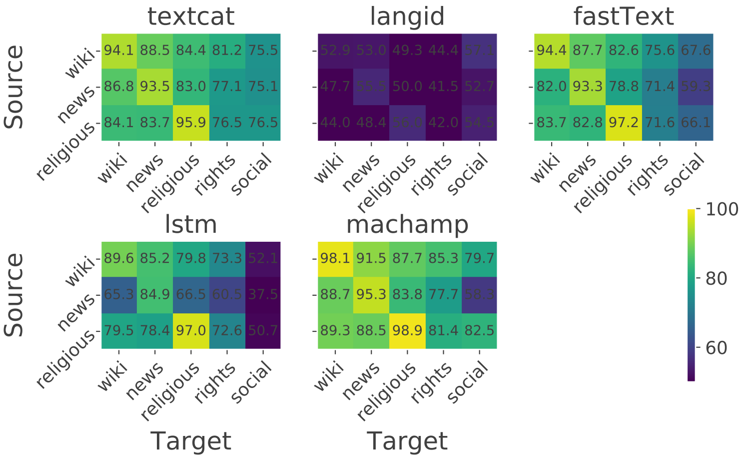


10 Languages
100 Languages
1,000 Languages
1.0
0.8
0.6
0.4
0.2
0.0
1000 max. chars
100 max. chars
10 max. chars
textcat
langid
fastText
lstm
machamp
10 utterances
100 utterances
1,000 utterances
10 utterances
100 utterances
1,000 utterances
10 utterances
100 utterances
1,000 utterances

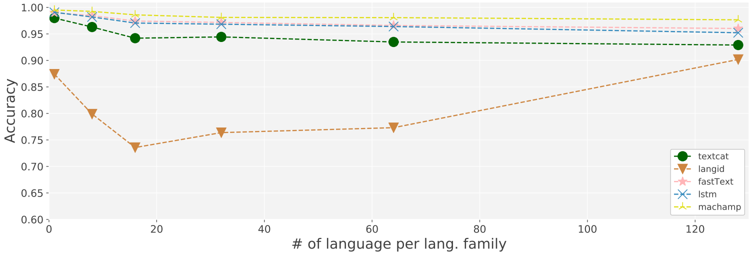
Size: takeaways

- 100 characters is enough
- # of languages is not very influential when there are enough (100) utterances
- Glot500 most robust
- Character n-gram overlap still impressively good

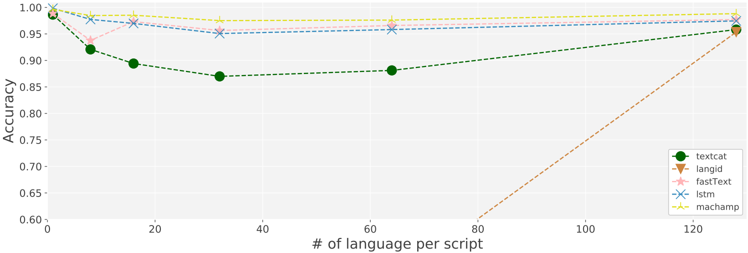
Domains


textcat
langid
fastText
lstm
machamp
Source
Target
wiki
news
religious
rights
social
94.1 88.5 84.4 81.2 75.5
86.8 93.5 83.0 77.1 75.1
84.1 83.7 95.9 76.5 76.5
52.9 53.0 49.3 44.4 57.1
47.7 55.5 50.0 41.5 52.7
44.0 48.4 56.0 42.0 54.5
94.4 87.7 82.6 75.6 67.6
82.0 93.3 78.8 71.4 59.3
83.7 82.8 97.2 71.6 66.1
89.6 85.2 79.8 73.3 52.1
65.3 84.9 66.5 60.5 37.5
79.5 78.4 97.0 72.6 50.7
98.1 91.5 87.7 85.3 79.7
88.7 95.3 83.8 77.7 58.3
89.3 88.5 98.9 81.4 82.5
100
80
60

Language families



Scripts



Accuracy
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0
20
40
60
80
100
120
# of language per script
textcat
langid
fastText
lstm
machamp

How about LLM's

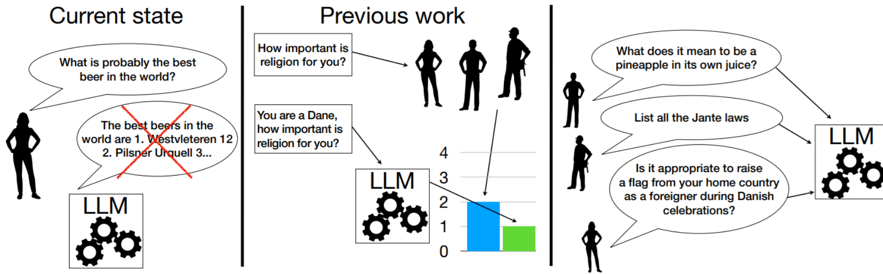
- Tokenization
- Language classification

Cultural evaluation of LLMs

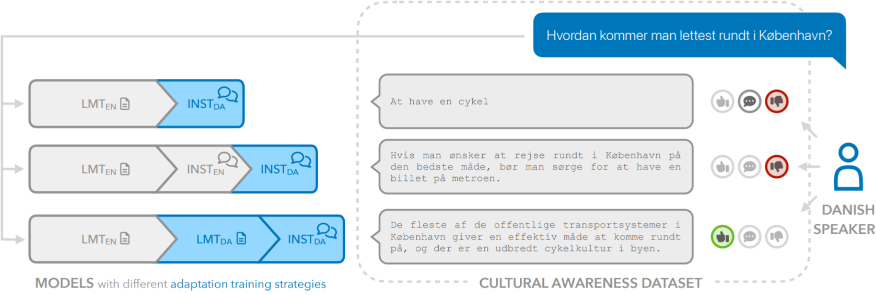
Retrain Llama:

- 13B words LM
- 3M instructions (translated)

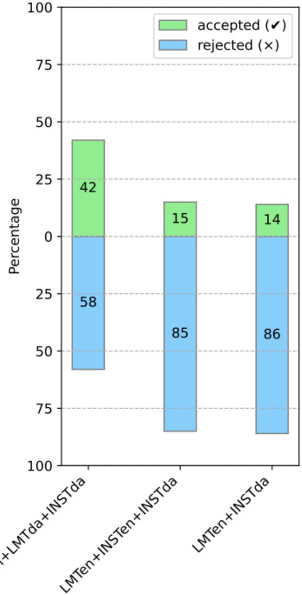
Cultural evaluation of LLMs



Cultural evaluation of LLMs



Cultural evaluation of LLMs



Thanks!

Lexical normalization

Multi-task learning

Is X solved?