Are the First Steps of the Traditional NLP Pipeline Solved?

Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics?

Christopher D. Manning

Departments of Linguistics and Computer Science
Stanford University
353 Serra Mall, Stanford CA 94305-9010
manning@stanford.edu

Abstract. I examine what would be necessary to move part-of-speech tagging performance from its current level of about 97.3% token accuracy (56% sentence accuracy) to close to 100% accuracy. I suggest that

How about:

- Word segmentation
- Language classification

How about:

- Word segmentation
- Language classification
- And let's broaden our scope beyond English!

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)	biiobiiiobiobiiobiiiiiib
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
- Adapters used before (costly to train)
- Let's do multi-task learning! add a decoder head for tokenization task and sum loss

`https://github.com/machamp-nlp/machamp/`

Settings

UD 2.10: 122 languages in 20 different scripts

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

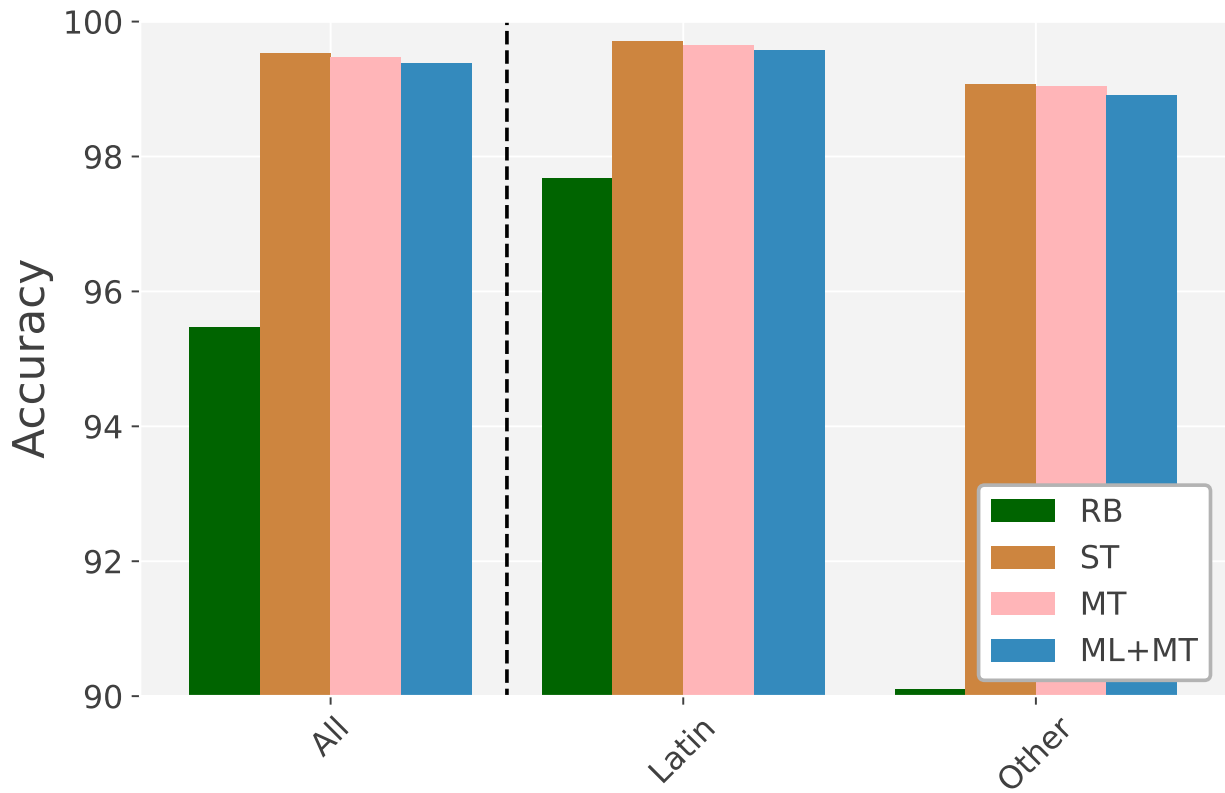
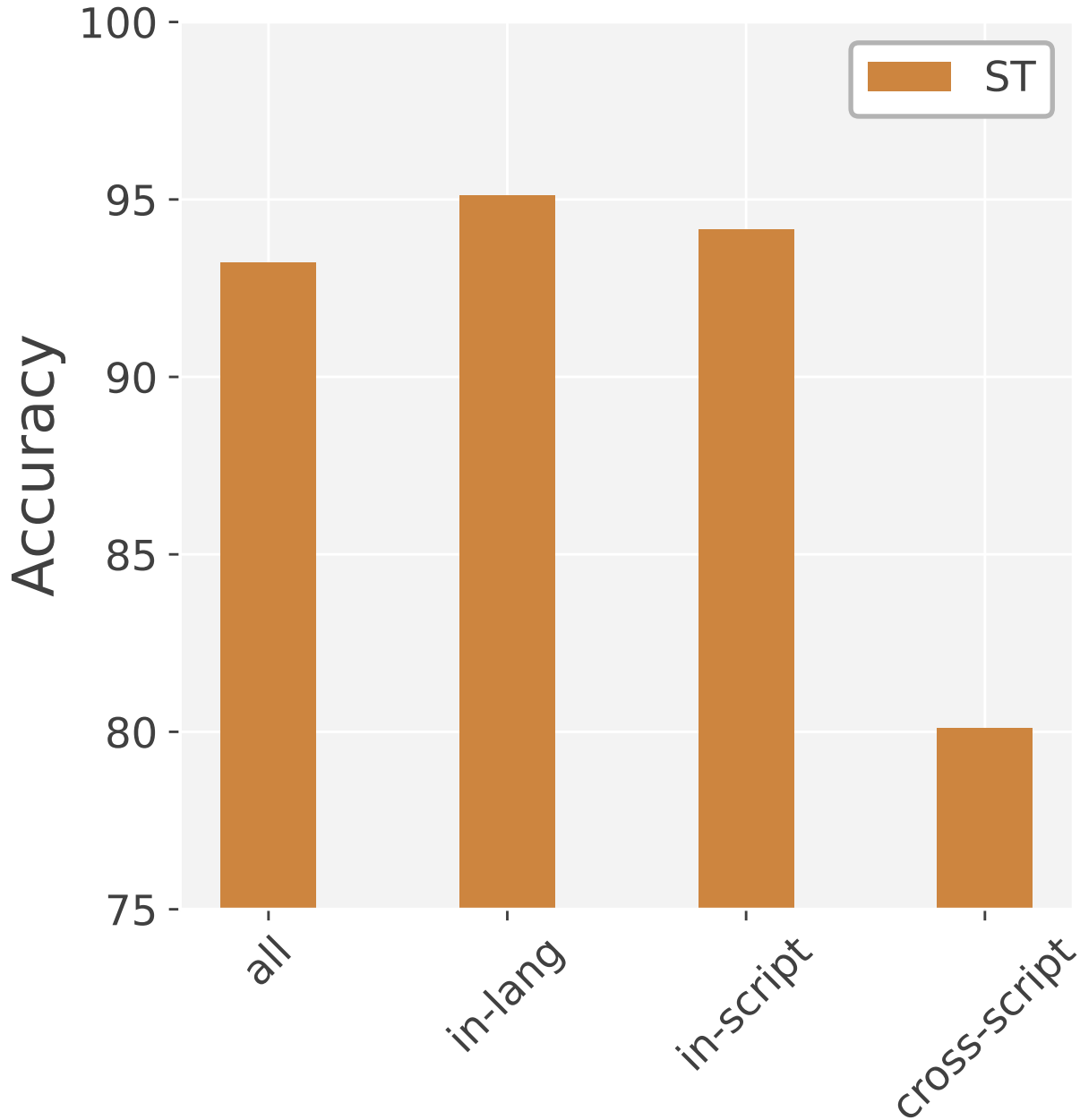


Figure: Results of tokenizers on Latin vs non-Latin languages. RB=RuleBased, ST=SingleTask, MT=MultiTask, ML+MT=Multi-Lingual+MultiTask

Cross-treebank results



Analysis

Main takeaways

- For Latin languages: edge-cases and inconsistencies
- For other scripts: names, adpositions, compound words
- More in the paper

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

Open challenges in language identification

- Many tools/benchmarks available
- When to use which?:
  - # languages
  - input size
  - # training instances per language
  - scripts
  - language families
  - domains
- (Language: ISO 639-3 code)

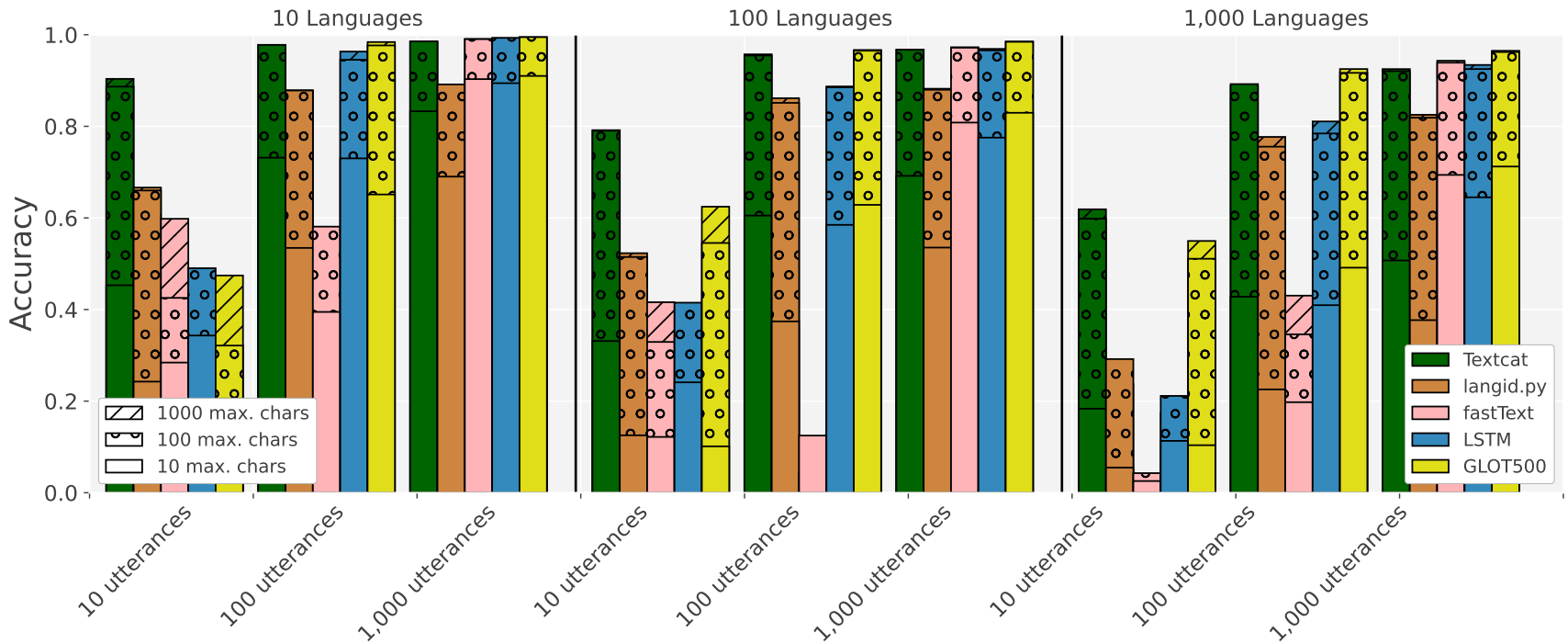
Data

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	

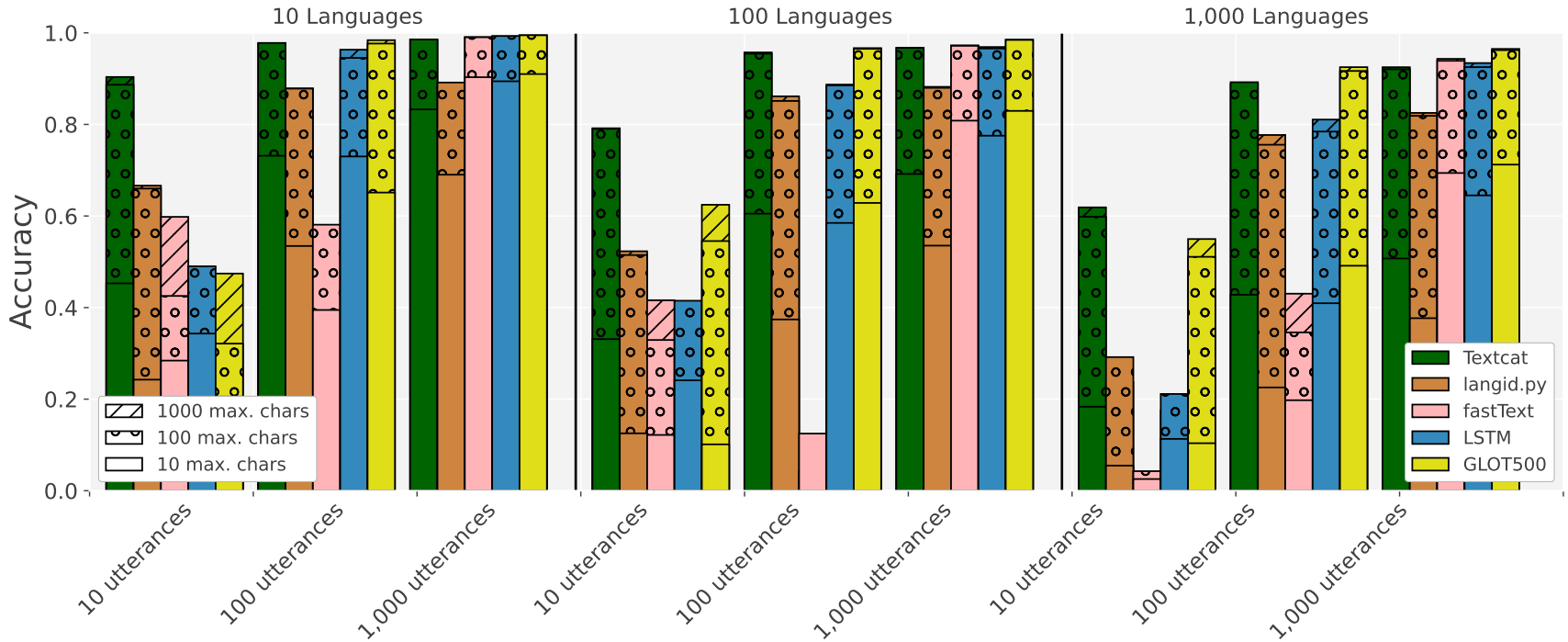
Models

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

Size

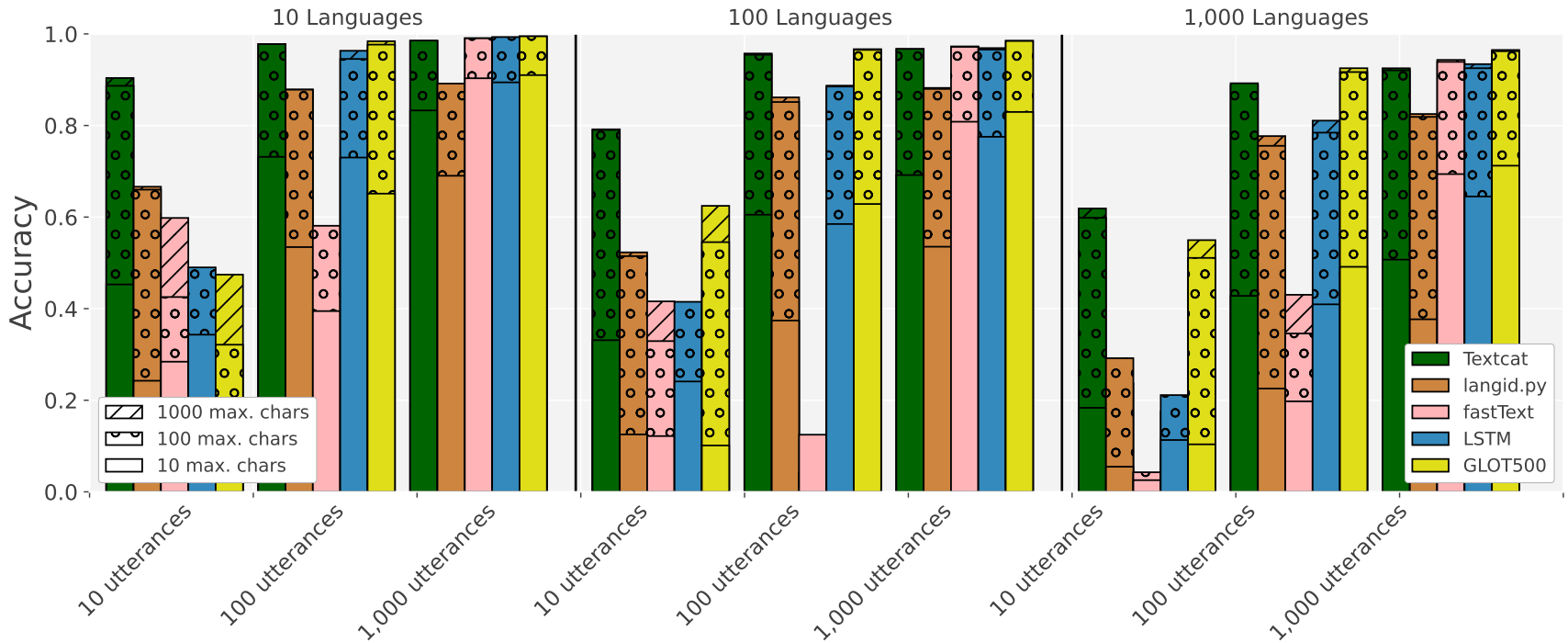

10 Languages
100 Languages
1,000 Languages
Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
1000 max. chars
100 max. chars
10 max. chars
Textcat
langid.py
fastText
LSTM
GLOT500
10 utterances
100 utterances
1,000 utterances

Size



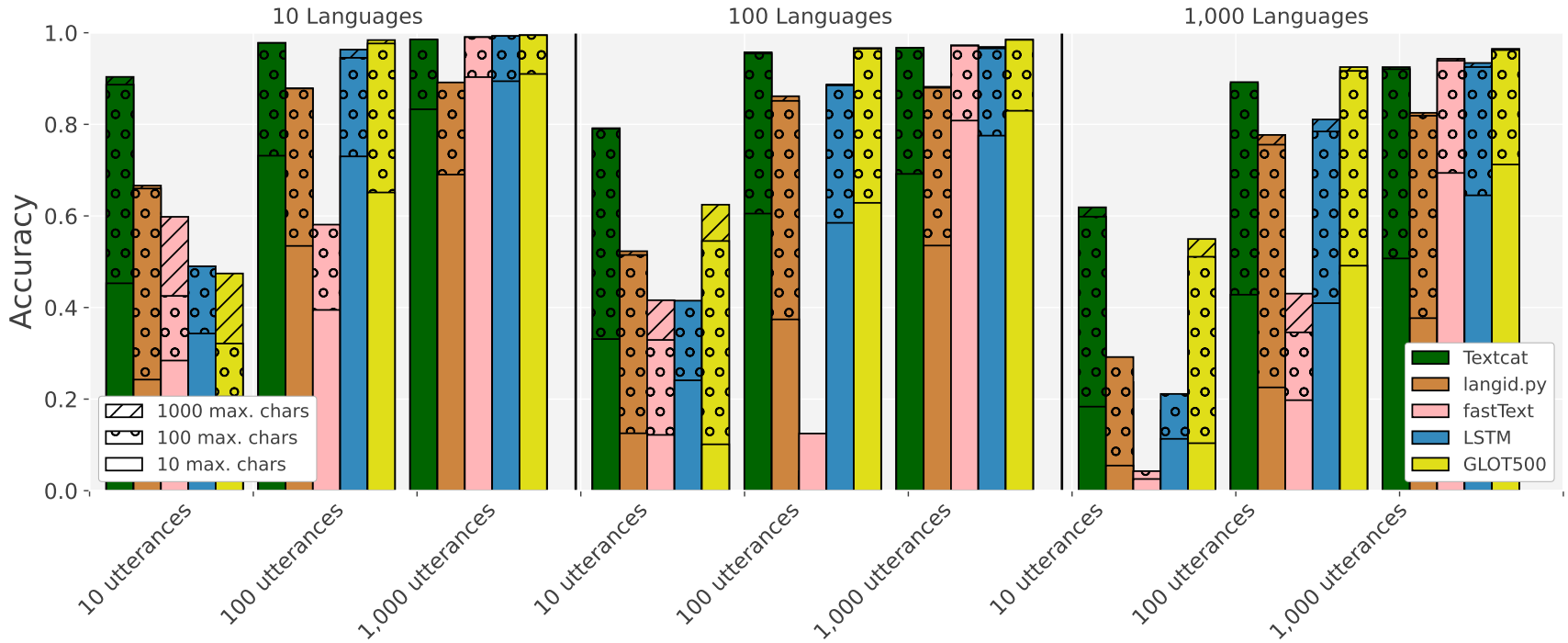
- 100 characters is enough.
- Number of utterances more influential, but 100 already quite good

Size: languages



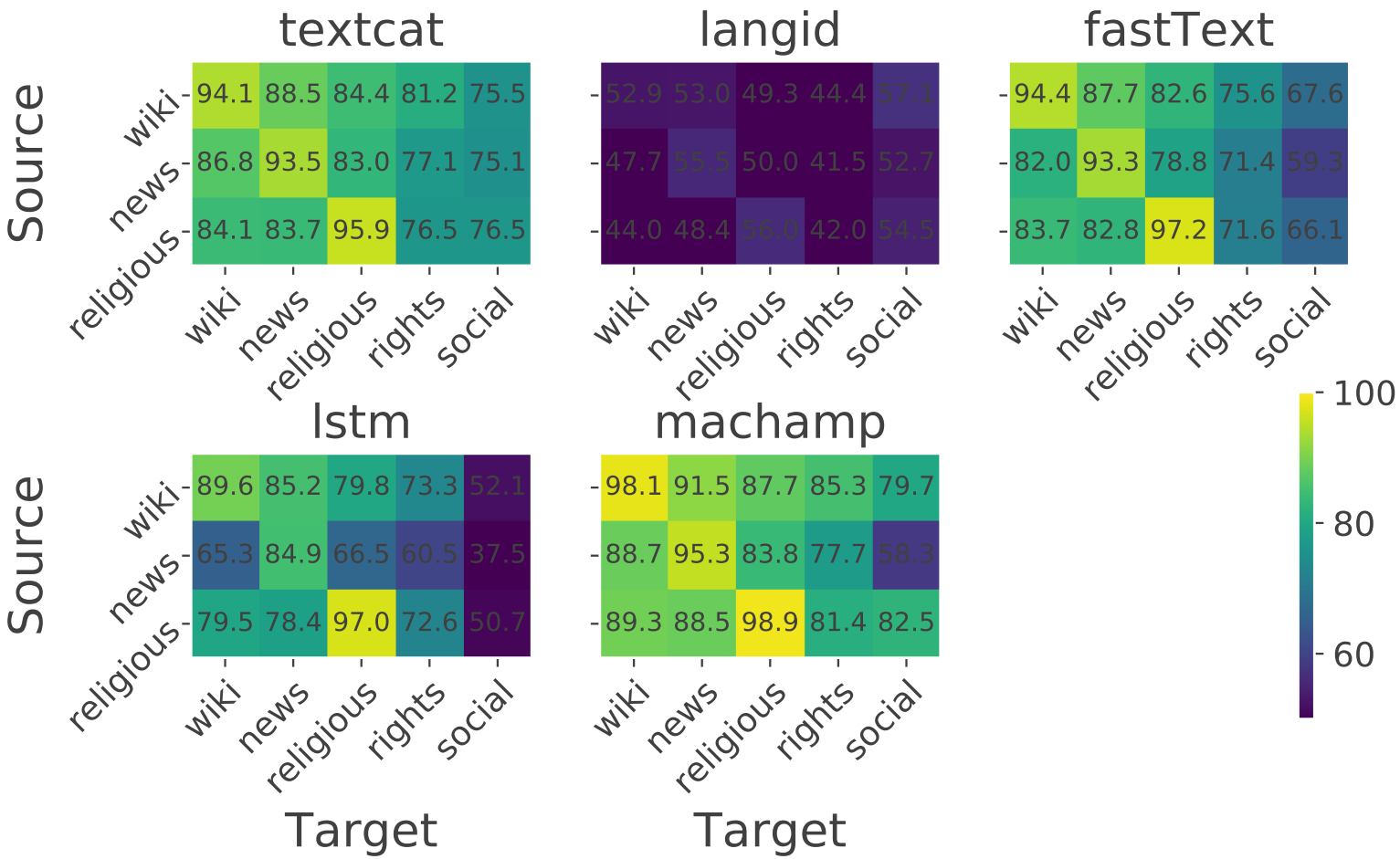
- # of languages is not very influential when there are enough utterances (i.e. 100)

Size: models

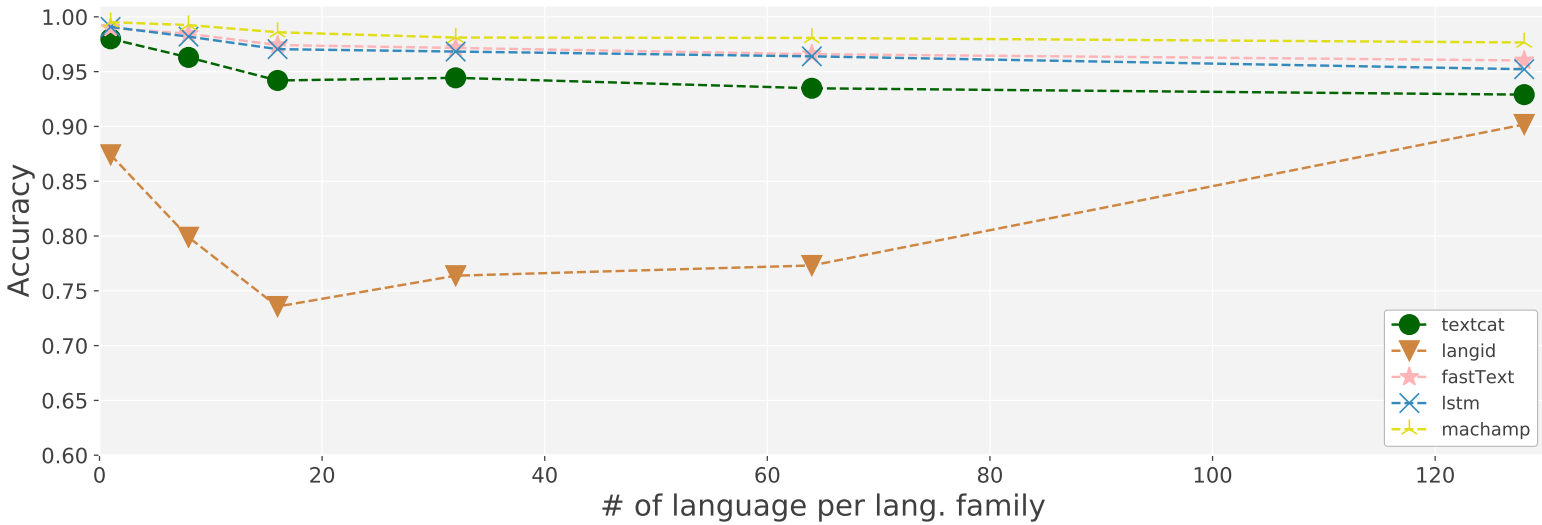


- Glot500 most robust
- Character n-gram overlap still impressively good

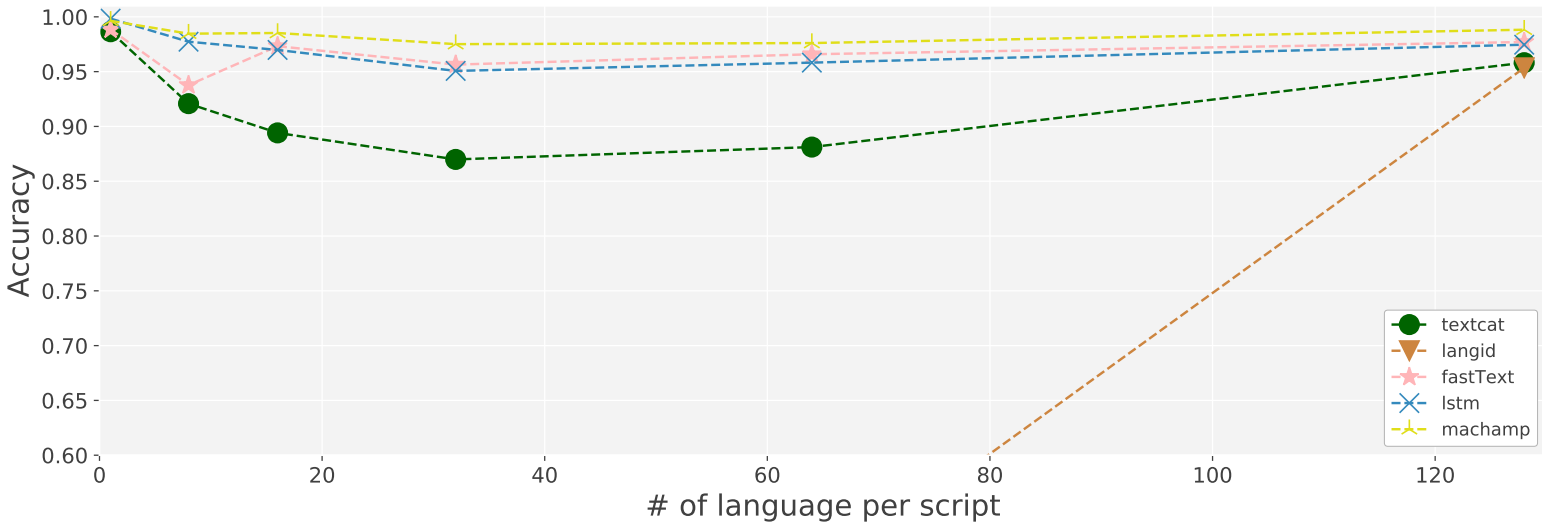
Domains


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


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

Other things:

Robustness:

- Normalization of social media data
- Multi-task learning (encoder models)
- Multi-lingual learning
- Language varieties

Useful (perhaps?)

Obtain information about ISO 639-3 codes: `https://bitbucket.org/robvanderg/lang_dist/src/master/`