Identifying Open Challenges in Language Identification

Rob van der Goot

Language identification

- Task: Map text into language (ISO 639-3) codes

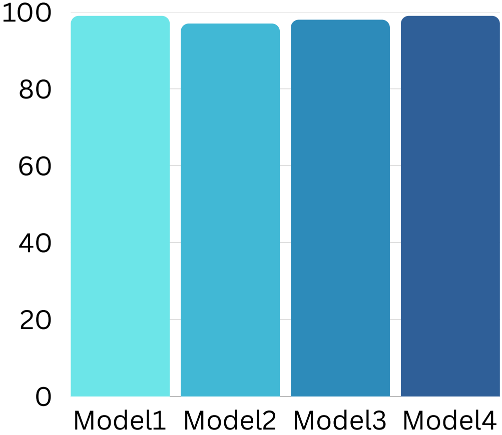
```
wêr jammerje jo no oer?
Skal vi snakke lidt?
```

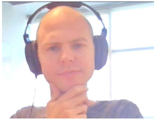
Language identification

- Task: Map text into language (ISO 639-3) codes

Wêr jammerje jo no oer?	FRY
Skal vi snakke lidt?	DAN

Current state





Data

Dataset	langs	scripts	fams	domains
MIL-TALE	2,110	47	139	wiki, political, religious, grammar
UDHR	397	38	61	rights
OpenLID	139	25	16	literature, news, wiki, social, grammar, subtitles, spoken
MassiveSumm	77	24	13	news
TwitUser	59	20	13	social
UD	54	11	17	medical, news, academic, wiki, legal, nonfiction, learner-essays, fiction, social, grammar-examples, reviews, religious, spoken
Total	2,176/ 7,850	51/ 163	145/ 298	

(note that glotlid (`https://huggingface.co/cis-lmu/glotlid`) has a similar setup/scope)

Models

Type	model	size	
N-gram overlaps	Textcat	40,000	🐜
ML	NB	100,000	🐸
Embeddings	FastText	4,434,860	🐧
NN	LSTM	15,158,772	🐊
Transformer LM	GLOT500	395,687,155	🐳

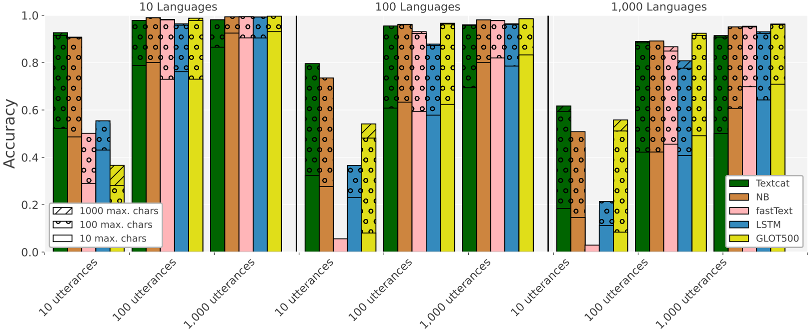
Size

- 10, 100, 1,000 languages
- 10, 100, 1,000 utterances input
- 10, 100, 1,000 characters input

1,000 utterances for testing

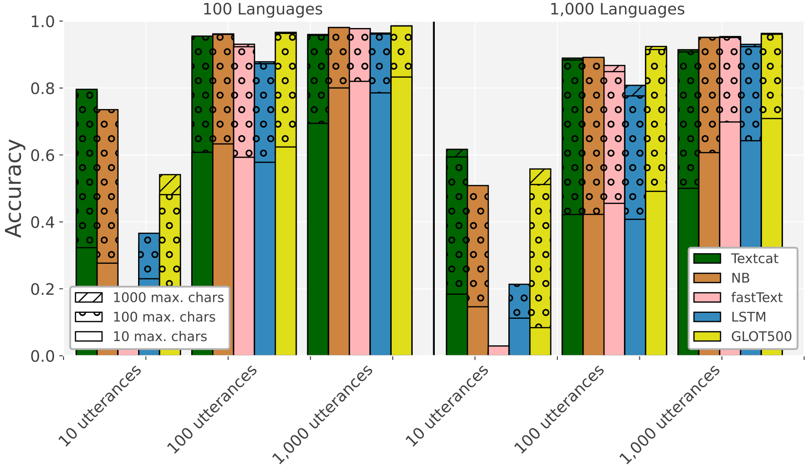
Size

- Number of languages



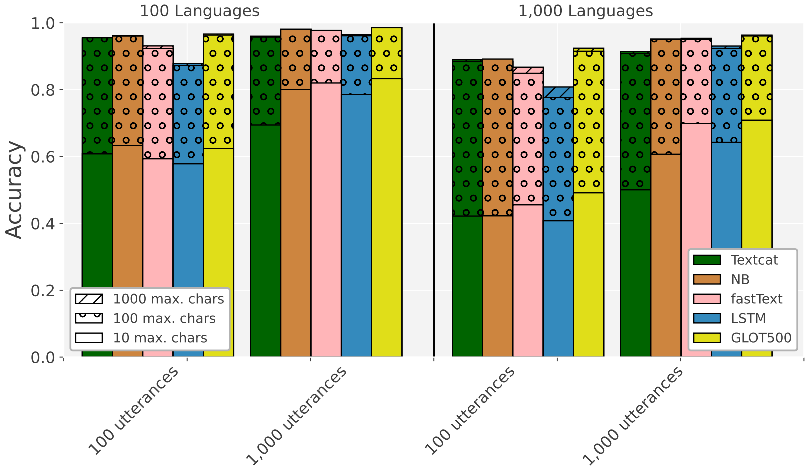
Size

- Number of utterances



Size

- Number of characters
- Models

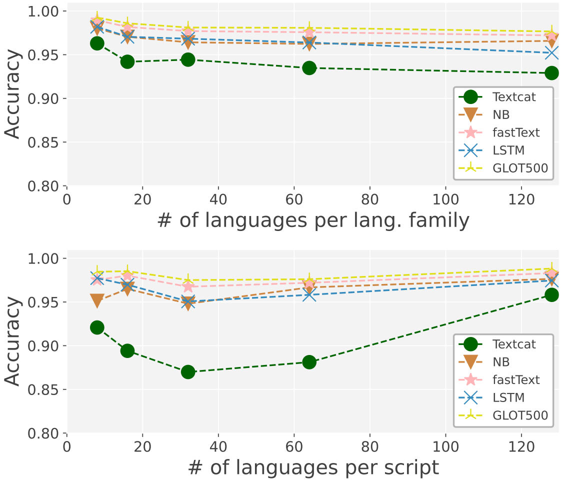


Family/script

- Make subsets with N languages per family/script

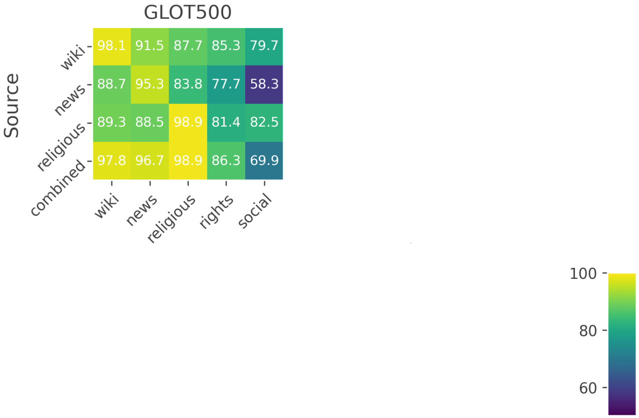
Family/script

- Make subsets with N languages per family/script



Domains

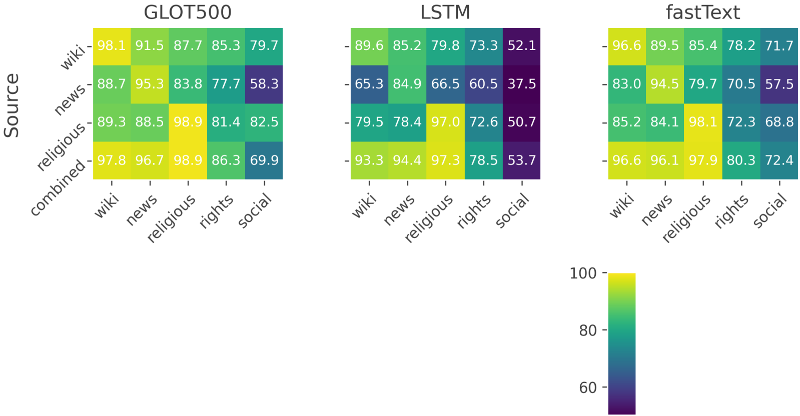
- Evaluate across domains



	wiki	news	religious	rights	social
wiki	98.1	91.5	87.7	85.3	79.7
news	88.7	95.3	83.8	77.7	58.3
religious	89.3	88.5	98.9	81.4	82.5
combined	97.8	96.7	98.9	86.3	69.9

Domains

- Evaluate across domains



GLOT500

Source	wiki	news	religious	rights	social
wiki	98.1	91.5	87.7	85.3	79.7
news	88.7	95.3	83.8	77.7	58.3
religious	89.3	88.5	98.9	81.4	82.5
combined	97.8	96.7	98.9	86.3	69.9

LSTM

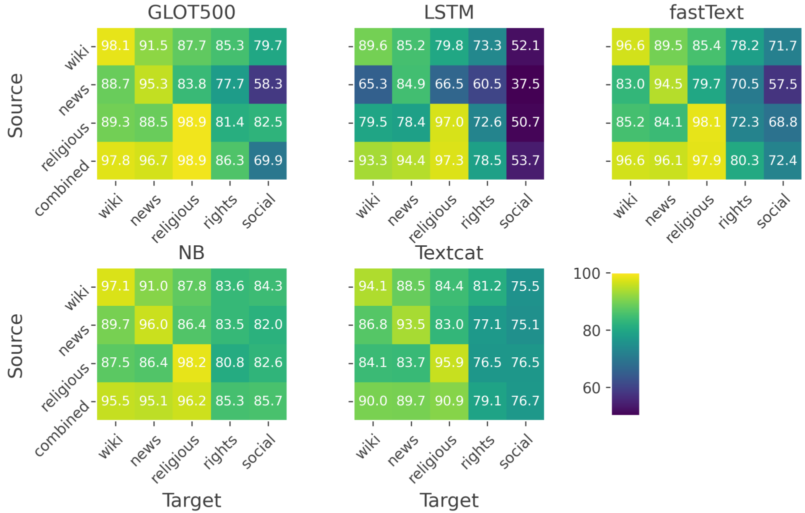
Source	wiki	news	religious	rights	social
wiki	89.6	85.2	79.8	73.3	52.1
news	65.3	84.9	66.5	60.5	37.5
religious	79.5	78.4	97.0	72.6	50.7
combined	93.3	94.4	97.3	78.5	53.7

fastText

Source	wiki	news	religious	rights	social
wiki	96.6	89.5	85.4	78.2	71.7
news	83.0	94.5	79.7	70.5	57.5
religious	85.2	84.1	98.1	72.3	68.8
combined	96.6	96.1	97.9	80.3	72.4

Domains

- Evaluate across domains



GLOT500

Source \ Target	wiki	news	religious	rights	social
wiki	98.1	91.5	87.7	85.3	79.7
news	88.7	95.3	83.8	77.7	58.3
religious	89.3	88.5	98.9	81.4	82.5
combined	97.8	96.7	98.9	86.3	69.9

LSTM

Source \ Target	wiki	news	religious	rights	social
wiki	89.6	85.2	79.8	73.3	52.1
news	65.3	84.9	66.5	60.5	37.5
religious	79.5	78.4	97.0	72.6	50.7
combined	93.3	94.4	97.3	78.5	53.7

fastText

Source \ Target	wiki	news	religious	rights	social
wiki	96.6	89.5	85.4	78.2	71.7
news	83.0	94.5	79.7	70.5	57.5
religious	85.2	84.1	98.1	72.3	68.8
combined	96.6	96.1	97.9	80.3	72.4

NB

Source \ Target	wiki	news	religious	rights	social
wiki	97.1	91.0	87.8	83.6	84.3
news	89.7	96.0	86.4	83.5	82.0
religious	87.5	86.4	98.2	80.8	82.6
combined	95.5	95.1	96.2	85.3	85.7

Textcat

Source \ Target	wiki	news	religious	rights	social
wiki	94.1	88.5	84.4	81.2	75.5
news	86.8	93.5	83.0	77.1	75.1
religious	84.1	83.7	95.9	76.5	76.5
combined	90.0	89.7	90.9	79.1	76.7

Conclusions

- Small number of utterances (1,000) is enough
- Family/script small effect
- Cross domain still challenging
- Larger models better in-dataset, small models more robust!
- Final models with $>$ 2,000 languages released

Conclusions

- Small number of utterances (1,000) is enough
- Family/script small effect
- Cross domain still challenging
- Larger models better in-dataset, small models more robust!
- Final models with $>$ 2,000 languages released