Does Transliteration Help Multilingual Language Modeling?

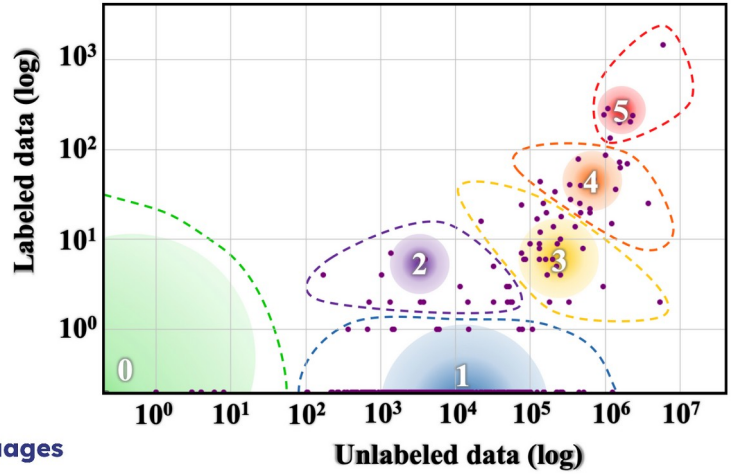
Ibraheem Muhammad Moosa, Mahmud Elahi Akhter, Ashfia Binte Habib

EACL 2023 Paper ID: 433

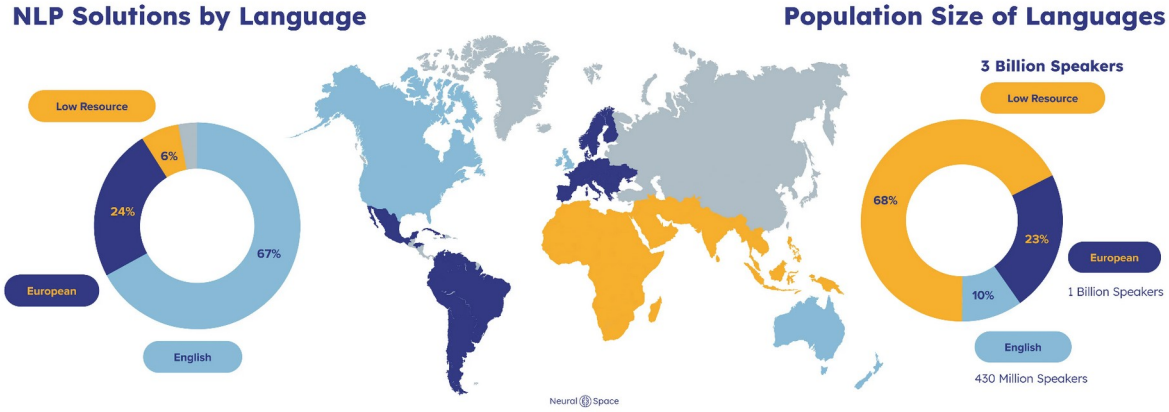
Motivation & Background

Open Problems

- Resource disparity between Low Resource Languages and High Resource Languages
- Diversity and Inclusivity of Low Resource Language tasks



Language resource distribution[1].



Diversity and inclusivity disparity[2]

Issues for Low Resource Languages

- Lack of large pretraining corpus
- Lack of diverse evaluation dataset

1. The State and Fate of Linguistic Diversity and Inclusion in the NLP World, Joshi et al. (2020)
2. Challenges in using NLP for low-resource languages and how NeuralSpace solves them, Felix Laumann (2022)

Motivation & Background

Additional obstacles for Low Resource Language

- ☒ Script barrier leads to poor lexical overlap [1]
- ☒ Poor Tokenization leads to increase in unknown tokens [2]

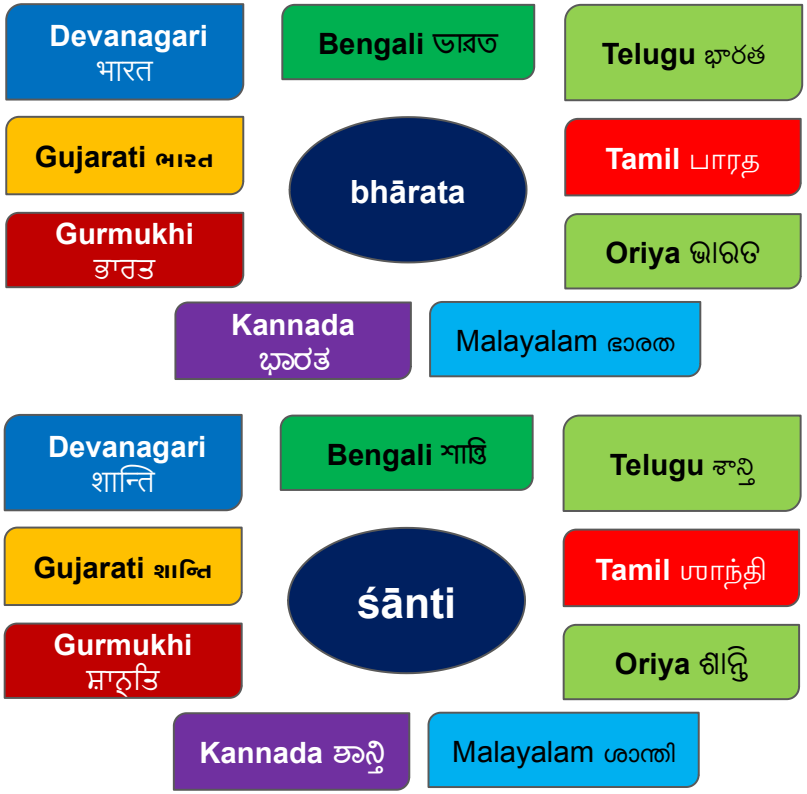
Proposed Solution

- ☑ Grapheme-to-phoneme (G2P)
- ☑ Transliteration

Script diversity in South Asia[3].

1. Pushing the Limits of Low-Resource Morphological Inflection, Anastasopoulos and Neubig (2019):
2. UNKs Everywhere: Adapting Multilingual Language Models to New Scripts, Pfeiffer et al.(2021)
3. en.wikipedia.org/wiki/Languages_of_South_Asia#/media/File:States_of_South_Asia.png

Case for Translitertion



Benefits of Transliteration?

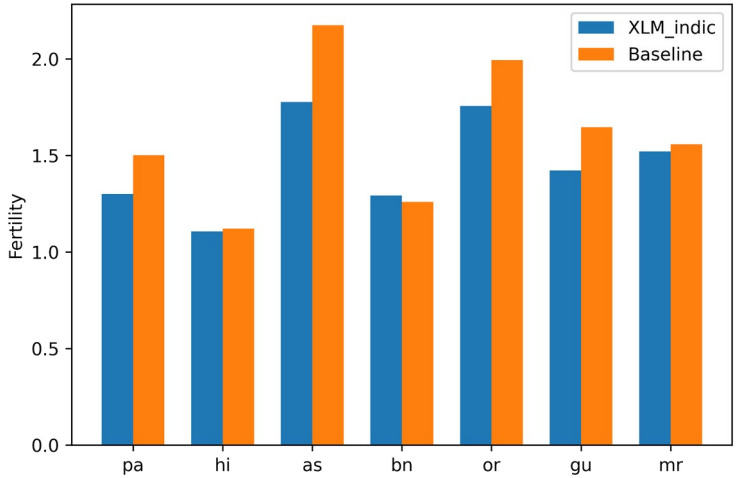
- ☑ Transliteration collapses multiple scripts into a single script
- ☑ Improves lexical overlap
- ☑ Reduces number of token

However..

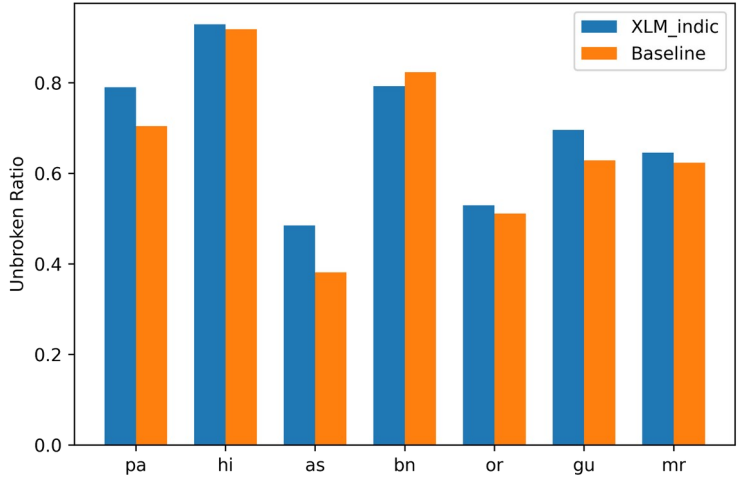
- Does transliterating input scripts improve performance of multilingual language models?
- Given it does, is this improvement statistically significant?
- Does a model trained on transliterated script learn better representation?

Results: Tokenization Quality

☑ Transliteration Reduces number of token



Average number of tokens per word



Ratio of words unbroken by the tokenizer

Results: Performance analysis

Model	pa	hi	bn	or	as	gu	mr	kn	te	ml	ta	avg
Wikipedia Section Title Prediction												
RemBERT$_{MS}$	68.42±0.92	70.90±0.39	72.58±0.45	69.92±0.90	68.37±1.37	72.93±0.58	73.23±0.61	71.67±0.41	92.98±0.19	69.03±0.57	69.77±0.45	73.00
RemBERT$_{US}$	71.01±0.22	72.45±0.29	73.65±0.21	75.37±0.69	72.50±0.91	76.35±0.29	74.58±0.72	74.21±0.29	93.66±0.09	69.33±0.35	70.63±0.22	74.89
δ	**2.59**	**1.55**	**1.07**	**5.45**	**4.13**	**3.42**	**1.34**	**2.54**	**0.68**	**0.31**	**0.86**	**1.89**
$p-value$	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0035	0.0004	0.0004	0.2505	0.0006	-
ALBERT$_{MS}$	74.33±0.83	78.18±0.33	81.18±0.28	74.35±1.2	76.70±0.83	76.37±0.53	79.10±0.84	-	-	-	-	77.17
ALBERT$_{US}$	77.55±0.61	82.24±0.18	84.38±0.29	81.47±0.99	81.74±0.82	82.39±0.27	82.74±0.52	-	-	-	-	81.78
δ	**3.22**	**4.06**	**3.20**	**7.12**	**5.04**	**6.02**	**3.64**	-	-	-	-	**4.61**
$p-value$	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	-	-	-	-	-
Named Entity Recognition (F1-Score)												
RemBERT$_{MS}$	69.47±1.72	90.95±0.33	95.51±0.18	87.92±1.26	79±0.22	69±0.94	90.72±0.17	72.65±1.81	81.82±1.81	89.17±0.25	90.07±0.33	83.40
RemBERT$_{US}$	81.91±1.93	91.73±0.39	96.19±0.21	88.92±2.88	83.50±2.75	80.25±1.42	90.75±0.35	78.98±1.50	84.97±0.45	89.26±0.46	90.18±0.27	86.97
δ	**12.44**	**0.78**	**0.68**	**1.00**	**4.28**	**10.31**	**0.02**	**6.33**	**3.15**	**0.01**	**0.12**	**3.56**
$p-value$	0.00004	0.0005	0.00001	0.1615	0.0019	0.00004	0.6665	0.00004	0.00004	0.7304	0.2973	-
ALBERT$_{MS}$	76.69±1.5	91.80±0.42	96.39±0.19	84.18±1.8	75.45±1.8	69.10±2.9	88.72±0.40	-	-	-	-	83.19
ALBERT$_{US}$	85.42±1.9	92.93±0.21	97.31±0.22	93.54±0.58	89.06±2.2	80.16±0.15	90.56±0.44	-	-	-	-	89.85
δ	**8.73**	**1.13**	**0.92**	**9.36**	**13.61**	**11.06**	**1.84**	-	-	-	-	**6.66**
$p-value$	0.0004066	0.0004066	0.0003983	0.0004038	0.000401	0.0004066	0.0004095	-	-	-	-	-

Orange indicates the multi-script and uni-script models are equal and blue indicates the uni-script model is better

- On average transliteration improves performance of multilingual language models.
- It improves the performance of low resource languages more compared to high resource languages
- It does not deteriorate the results of low resource languages

Results: Performance analysis

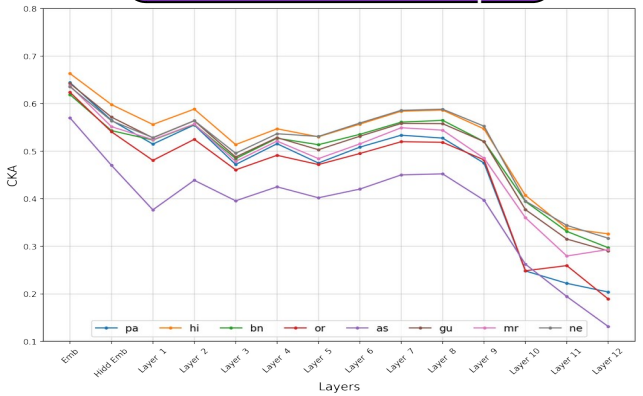
Language	Dataset	RemBERT_{MS}	RemBERT_{US}	δ	$p-value$	ALBERT_{MS}	ALBERT_{US}	δ	$p-value$
Article Genre Classification									
hi	BBC News	76.80±0.84	77.78±0.92	**0.98**	0.0466	77.28±1.51	79.14±0.60	**1.86**	0.0088
bn	Soham News Article Classification	92.86±0.10	93.69±0.20	**0.83**	0.0004	93.22±0.49	93.89±0.48	**0.67**	0.0090
gu	INLTK Headlines	90.27±0.47	91.60±0.28	**1.33**	0.0004	90.41±0.69	90.73±0.75	**0.32**	0.6249
mr	INLTK Headlines	91.24±0.50	92.27±0.39	**1.03**	0.0008	92.21±0.23	92.04±0.47	**-0.17**	0.3503
ml	INLTK Headlines	94.11±0.49	93.33±0.22	**-0.78**	0.003	-	-	-	-
ta	INLTK Headlines	95.59±0.70	94.93±0.30	**-0.65**	0.013	-	-	-	-
Sentiment Analysis									
hi	IITP Product Reviews	72.17±1.98	72.85±0.63	**0.68**	0.9646	76.33±0.84	77.18±0.77	**0.85**	0.04099
hi	IITP Movie Reviews	58.66±1.09	62.65±2.74	**3.99**	0.0023	65.91±2.2	66.34±0.16	**0.15**	0.8941
te	ACTSA	61.18±1.38	60.53±0.85	**-0.66**	0.1981	-	-	-	-
Discourse Mode Classification									
hi	MIDAS Discourse	78.07±0.83	79.46±0.67	**1.39**	0.0415	78.39±0.33	78.54±0.91	**0.15**	0.7561

Model	pa	hi	bn	or	as	gu	mr	ta	te	ml	kn	avg
Cloze-style QA (Zero Shot)												
RemBERT_{MS}	33.93	39.06	38.93	37.32	37.66	84.21	46.15	37.02	34.42	38.45	40.75	42.53
RemBERT_{US}	33.92	40.10	39.62	38.28	39.26	85.37	45.92	36.68	34.36	37.16	44.29	43.17
δ	**-0.01**	**1.04**	**0.69**	**0.96**	**1.6**	**1.16**	**-0.23**	**-0.34**	**-0.06**	**-1.29**	**3.54**	**0.64**
ALBERT_{MS}	31.04	36.72	35.19	34.63	33.92	59.86	36.14	-	-	-	-	38.21
ALBERT_{US}	32.77	38.52	36.38	36.00	37.36	70.22	39.53	-	-	-	-	41.54
δ	**1.73**	**1.8**	**1.19**	**1.37**	**3.44**	**10.36**	**3.39**	-	-	-	-	**3.33**

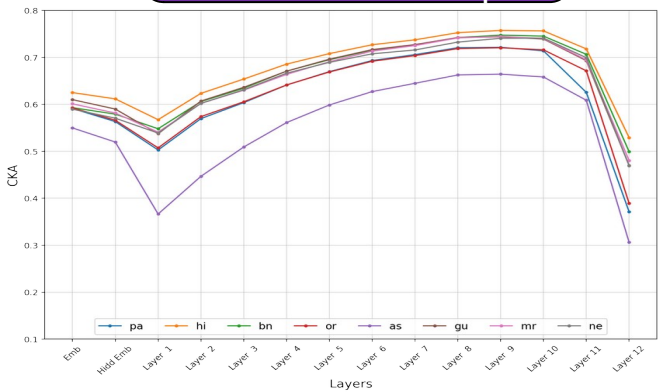
Orange indicates the multi-script and uni-script models are equal, cyan indicates multi-script is better than uni-script and blue indicates vice versa

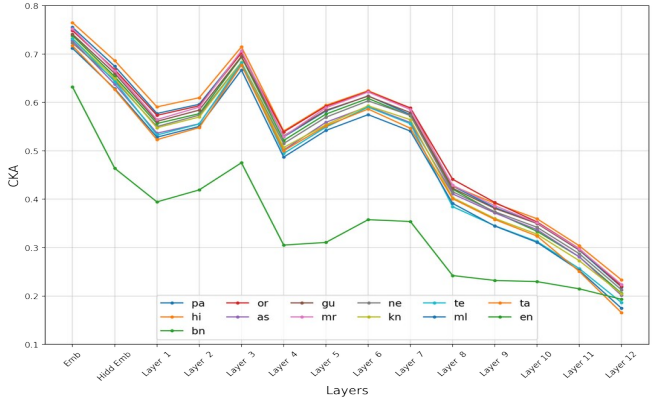
Results: Cross lingual representation

ALBERT Multi-script

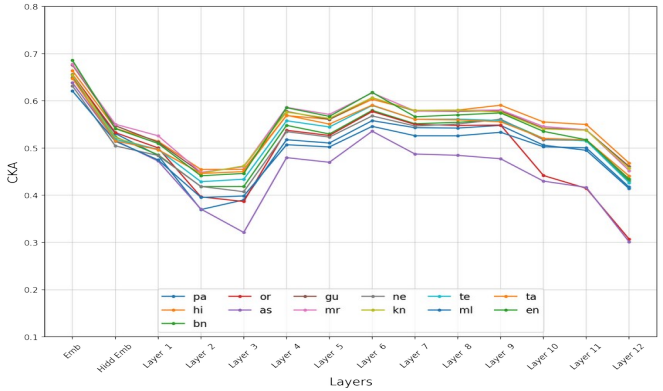


ALBERT Uni-script





RemBERT Multi-script



RemBERT Uni-script

Does Transliteration Help Multilingual Language Modeling?

Ibraheem Muhammad Moosa, Mahmud Elahi Akhter, Ashfia Binte Habib

EACL 2023 Paper ID: 433