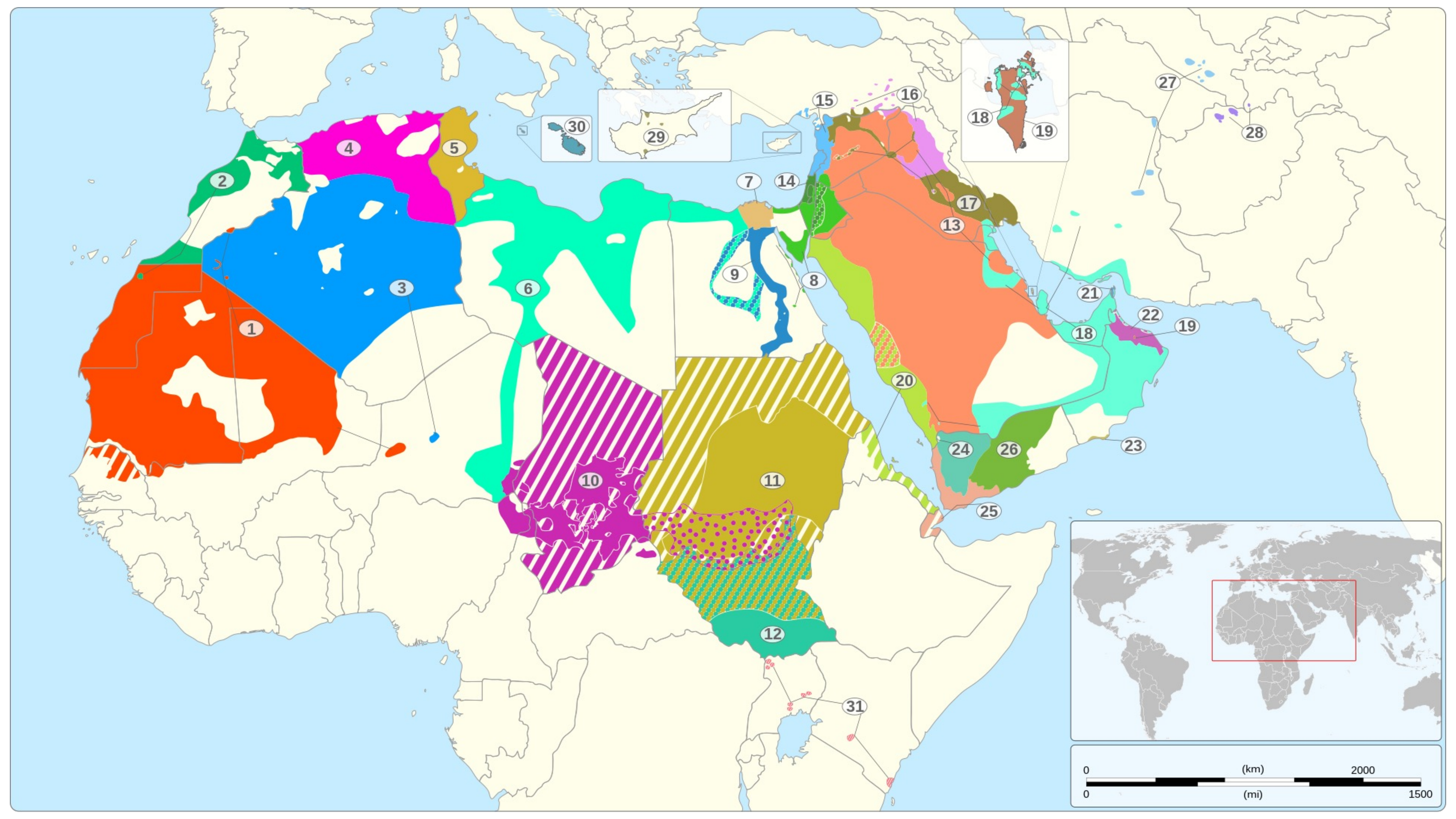
NEURAL INFORMATION PROCESSING SYSTEMS

Parameter and Data Efficient Continual Pre-training for Robustness to Dialectal Variance in Arabic

Soumajyoti Sarkar, Kaixiang Lin, Sailik Sengupta, Leonard Lausen, Sheng Zha, Saab Mansour
AWS AI Labs

amazon | science

The Landscape of Dialects



- 1: Hassaniyya
- 2: Moroccan Arabic
- 3: Algerian Saharan Arabic
- 4: Algerian Arabic
- 5: Tunisian Arabic
- 6: Libyan Arabic
- 7: Egyptian Arabic
- 8: Eastern Egyptian Bedawi Arabic
- 9: Saidi Arabic
- 10: Chadian Arabic
- 11: Sudanese Arabic
- 12: Sudanese Creole Arabic
- 13: Najdi Arabic
- 14: South Levantine Arabic
- 15: North Levantine Arabic
- 16: North Mesopotamian Arabic
- 17: Mesopotamian Arabic
- 18: Gulf Arabic
- 19: Baharna Arabic
- 20: Hijazi Arabic
- 21: Shihhi Arabic
- 22: Omani Arabic
- 23: Dhofari Arabic
- 24: Sanaani Arabic
- 25: Ta'izzi-Adeni Arabic
- 26: Hadrami Arabic
- 27: Uzbeki Arabic
- 28: Tajiki Arabic
- 29: Cypriot Arabic
- 30: Maltese
- 31: Nubi
- Other: Sparsely populated area or no indigenous Arabic speakers

- ❖ Dialectical variance, exhibited by differences in grammar and vocabulary, is common in many languages across the world.
- ❖ Our work focuses on Arabic, where this variance can be particularly challenging.
- ❖ Further, scanty data for all variants raises questions about how to incorporate then during pretraining?

"I love reading a lot."

[MSA]
ʾana ʾuḥibbu‿l-qirāʾata kaṯīran
ʔana: ʔuħibːu‿lqiraːʔata kaθiːran

nḥəbb nǎqṛa barʃa
ʔāna nḥəbb nəqṛa b-ez-zaf
ʔāna kanebɣi naqra b-ez-zāf
jien inħobb naqra ħafna
ʔana baḥebb el-ʔerāya awi
ʔana/ʔani kṯīr baḥebb il-qirāʔa
ʔana ktīr baḥebb il-qirāʔa
ʔāni kulliš ʔaḥebb lu-qrāye
ʔāna wāyid ʔaḥibb il-qirāʾa
ʔana marra ʔaḥubb al-girāya
ʔana bajn ʔaḥibb el-gerāje gawi

[Q1] *Continual Pre-Training (CPT) of multilingual models == monolingual model performance?*

[Q2] *How to incorporate sparse dialectical data alongside abundant Modern Standard Arabic (MSA) data during pretraining?*

[Q3] *With strategies for incorporating dialectical data, is CPT with multilingual models == CPT with monolingual models?*

The Corpus

Code	Corpora	Size
C1	Oscar Arabic	67M
	Arabic Wiki	49M
	Arabic CC100	111M
	Arabic Newswire Part-1	2.3M
	Arabic Gigaword Fifth Edition	96M
	Gulf Arabic Conv.	4K
	GALE (only Arabic data)	25K
	BOLT SMS/Chat (only Arabic data)	44K
C2	OSCAR Egyptian Arabic Corpus	102K
C3	GALE Parallel Corpus BOLT Egyptian Arabic SMS/Chat	255K

Figure 1: C1 – Mixed Arabic and its dialectal variants, C2 - Egyptian Arabic sentences amounting to 102K, C3 – Parallel dialectical variants to English data from the GALE Arabic Parallel datasets which are 255K in total

The Models

M-B-Ar
Multilingual BERT continually pretrained for 800K steps starting from the pretrained public checkpoint with C1 for 800K steps

B-Ar
BERT model trained from scratch (with randomly initialized weights) with C1 for 1.3M steps

t-B-Ar
BERT model trained with a custom tokenizer trained on C1

Pre-training Methodology

MLM
Masked Language Modeling Objective used with the corpus C1 and C2

TLM
Translation Language Modeling Objective used with the parallel corpus C3

The ALUE Benchmark

Task Type	Task Name	Domain
Single Sentence Classification	MDD	Travel
	OOLD	Tweet
	OHSD	Tweet
	FID	Tweet
Sentence Pair Classification	MQ2Q	Web
	XNLI	Misc.
Multi-label Classification	SEC	Tweet
Regression	SVREG	Tweet

Experimental Results

Model	FID	MDD	MQ2Q	SVREG	SEC	OOLD	OHSD	XNLI
AraBERT	78.31	51.15	77.41	42.41	32.21	94.92	96.57	51.02
AraBERT-Twitter	79.73	52.26	77.07	39.25	31.34	94.21	97.76	39.71
mBERT	77.14	49.31	77.11	34.70	35.49	94.13	96.49	**51.08**
m-B-Ar	79.61	**56.04**	80.26	50.82	41.05	94.62	97.13	50.57
B-Ar	79.32	55.84	**80.35**	51.65	41.88	94.58	97.27	51.04
t-B-Ar	**81.04**	53.49	72.63	**74.37**	**49.26**	**95.12**	**98.36**	51.03

☺ Our models ☺ Monolingual models
☹ Publicly available models ☹ Multilingual models

Model	FID	MDD	MQ2Q	SVREG	SEC	OOLD	OHSD	XNLI
m-B-Ar	79.61	56.04	80.26	50.82	41.05	94.62	97.13	50.57
+C2	79.35	56.60	80.46	53.72	40.13	94.56	97.16	51.33
+C2+C3	78.29	56.82	80.65	51.42	40.75	**95.18**	97.51	**52.69**
B-Ar	79.32	55.84	80.35	51.65	41.88	94.58	97.27	51.04
+C2	**81.20**	55.84	84.73	69.72	47.66	94.53	97.75	52.13
+C2+C3	79.9	**57.61**	**85.31**	**70.31**	**48.03**	94.67	**97.91**	51.38

✓ Small dialectal data can improve dialectical robustness of finetuned multilingual and monolingual models is used cleverly!

✓ With C2 & C3,
monolingual models >> multilingual models

References

Aseelawi et al. lue: Arabic language understanding evaluation
Wissam Antoun at al.Arabert: Transformer-based model for arabic language understanding
Alexis CONNEAU and Guillaume Lample. Cross-lingual language model pretraining
Figures and examples borrowed from https://en.wikipedia.org/wiki/Varieties_of_Arabic
Datasets were obtained from LDC, the corpus referenced in Arabaert, OSCAR

*Corresponding author: soumajs@amazon.com