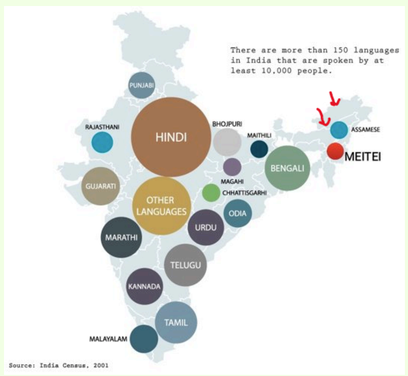
AxomiyaBERTa: A Phonologically-aware Transformer Model for Assamese

Abhijnan Nath, **Sheikh Mannan**, and **Nikhil Krishnaswamy**

• {abhijnan.nath,sheikh.mannan,nkrishna}@colostate.edu

Introduction

- Assamese: an extremely Low-Resource Language (LRL)
- Spoken by 15M people in Northeastern India
  - For comparison, Hindi has >600M speakers
- Less NLP-related resources
- We train AxomiyaBERTa in resource-constrained settings
- Language-specific phonological awareness
- Novel Embedding Disperser architecture
- SOTA results on many NLP tasks
- Phonological attention, strategic optimization work for LRLs!



Source: https://en.wikipedia.org/wiki/Languages_of_India

Pretraining Corpora

- Train on four publicly-available Assamese (As) datasets:
- Assamese Wikidumps Dataset
- OSCAR Dataset
- PMIndia Dataset
- CC-100 Assamese corpus
- ECB+ Corpus (translated to Assamese)

	as	*bn*	*hi*	*en*
CC-100	5	525	1,715	55,608
IndicCorp	32.6	836	1,860	1,220

CC-100 and IndicCorp data sizes (millions of tokens) for **Assamese**, Bengali, Hindi, and English.

NLP Tasks

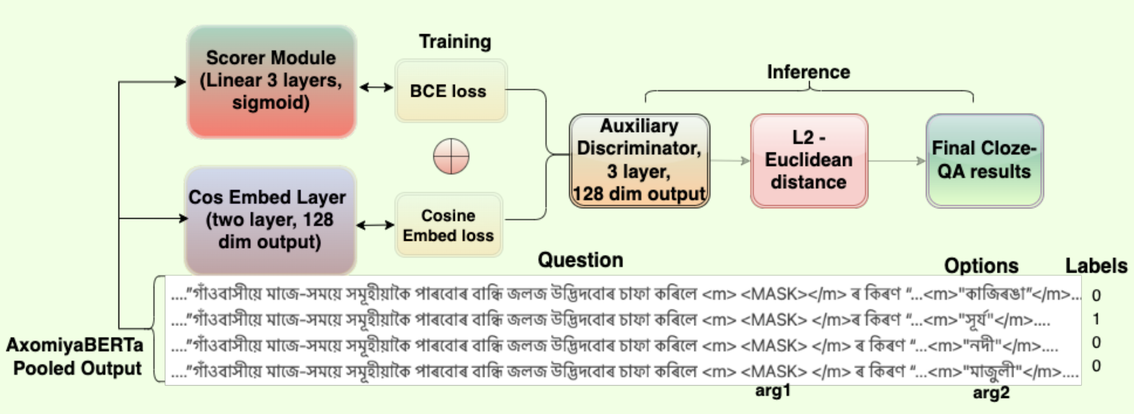
- Cloze-Style Question Answering (Long Context)
- Wiki-Section Title Prediction (Long Context)
- Named Entity Recognition (Short Context)
- CDCR: Cross Document Coreference Resolution

Features	Train	Dev	Test	Pad-Len
Cloze-QA	8,000	2,000	1,768	360
Wiki-Titles	5,000	625	626	1,848
AsNER	21,458	767	1,798	744
WikiNER	1,022	157	160	480
T-ECB+	3,808	1,245	1,780	552

Taskwise train/dev/test Split

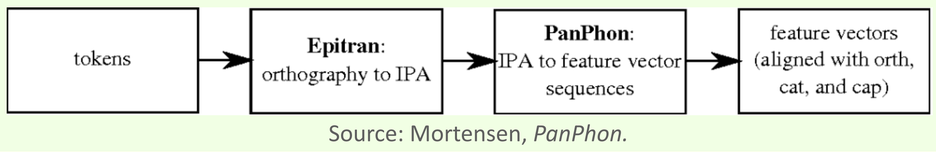
Embedding Disperser Architecture

- Combined loss objective to tackle anisotropy
- Cosine Embedding Loss + BCE Loss
- Helps disperse the latent space in smaller language models



Phonological Feature Generation

- Epitran and PanPhon to generate phonological features
- NER tokens for short-context, candidate options for Cloze-QA and Wiki-Titles, and event lemma for CDCR



Source: Mortensen, *PanPhon.*

Results

Models	Cloze-QA	Wiki-Titles	AsNER (F1)	WikiNER (F1)
XLM-R	27.11	56.96	69.42	66.67
MBERT	29.42	**73.42**	68.02*	**92.31**
IndicBERT-BASE	40.49	65.82	68.37*	41.67
MuRIL	-	-	80.69	-
AxomiyaBERTa	46.66	26.19	81.50	72.78
AxomiyaBERTa + Phon	**47.40**	59.26	**86.90**	81.71

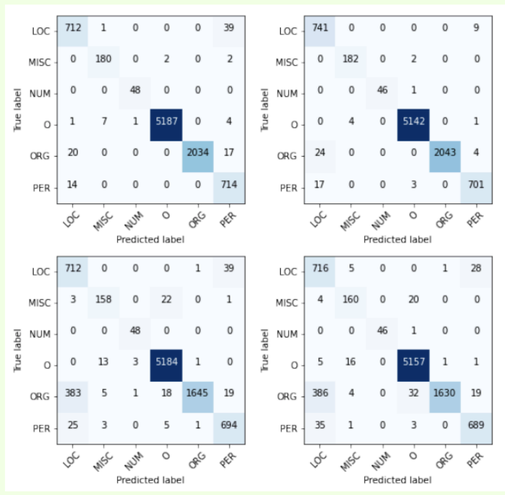
Test F1 Scores/Accuracy for AxomiyaBERTa and Phonologically-aware AxomiyaBERTa on the evaluation tasks compared to previous and our finetuned baselines.

CDCR Models	BCUB			MUC			CEAF-e			BLANC			C-F1
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	
Lemma Baseline	75.81	60.24	67.14	64.59	54.25	58.97	61.36	73.25	**66.78**	**74.97**	60.40	**64.66**	**64.29**
XLM-100†	5.31	**97.55**	10.08	54.17	**97.84**	69.73	30.99	0.73	1.42	49.78	50.00	49.89	27.07
IndicBERT-BASE	74.48	51.93	61.19	44.03	21.94	29.29	40.80	65.59	50.31	52.09	55.41	52.93	46.93
MuRIL	**93.53**	48.33	63.73	**68.18**	9.23	16.26	41.56	**85.09**	55.85	54.78	53.31	53.91	45.28
AxomiyaBERTa	34.68	85.98	49.42	62.40	80.51	**70.30**	**67.63**	43.85	53.20	53.00	**87.75**	54.23	57.64
AxomiyaBERTa + Phon	70.00	64.58	**67.18**	64.11	44.71	52.68	50.18	68.57	50.18	56.22	68.65	59.19	59.27

Event coreference results of AxomiyaBERTa on Assamese (translated) ECB+ test set compared with other Transformer-based LMs and the lemma-based heuristic.

Analysis

- AxomiyaBERTa achieves SOTA on Named Entity Recognition and Cloze-style QA
- Phonological signals boost native AxomiyaBERTa performance
  - 2x boost in Wiki-Titles, +~10 F1 points on WikiNER
  - Help disambiguate misclassifications in AsNER and WikiNER



Top: Confusion matrices showing AxomiyaBERTa performance on AsNER without [L] and with [R] phonological awareness. Bottom: IndicBERT [L] and MBERT [R] performance on AsNER.

Models	TP	L1	L2	Diff-Rate
XLM-100	6,361	1,441	4,920	.773
IndicBERT	101	46	55	.545
MuRIL	62	21	41	.661
AxB	1,833	466	1,367	.746 (.98)
AxB + Phon	956	81	875	.915 (.93)

Distribution of same (L1) and different (L2) event lemma samples in the true positive (TP) distribution of the T-ECB+ test set. “Diff-Rate” is the percentage of different lemma samples within TPs (= L2/TP).

	P+N-	P-N+	P+N+	P-N-
Cos-sim	.98844	.98829	.98824	.98838

Average cosine similarities between within-set samples on the Wiki-Titles test set for native (N) and phonological (P) AxomiyaBERTa. “+” and “-” represent correct and incorrect samples respectively.

Conclusion and Future Work

- A novel phonologically-aware Transformer Language Model for Assamese, an extremely low-resource Indian language
- Strategic use of Embedding Disperser for a more expressive latent space (task-specific)
- Achieved SOTA on short-context task like AsNER and longer context tasks like Cloze-QA
- Novel baselines for challenging tasks like CDCR on Assamese
- Overall, we show that optimizing the Embedding Space, and phonological information flow can overcome limited data or limited compute power in low-resource settings

Resources

Codebase | Pretrained model | Paper