AxomiyaBERTa: A Phonologically-aware Transformer Model for Assamese

Abhijnan Nath, **Sheikh Mannan**, and **Nikhil Krishnaswamy**
Situated Grounding and Natural Language (SIGNAL) Lab
Department of Computer Science, Colorado State University
Fort Collins, CO, USA
{abhijnan.nath,sheikh.mannan,nkrishna}@colostate.edu

Abstract

Despite their successes in NLP, Transformer-based language models still require extensive computing resources and suffer in low-resource or low-compute settings. In this paper, we present AxomiyaBERTa, a novel BERT model for Assamese, a morphologically-rich low-resource language (LRL) of Eastern India. AxomiyaBERTa is trained only on the masked language modeling (MLM) task, without the typical additional next sentence prediction (NSP) objective, and our results show that in resource-scarce settings for very low-resource languages like Assamese, MLM alone can be successfully leveraged for a range of tasks. AxomiyaBERTa achieves SOTA on token-level tasks like Named Entity Recognition and also performs well on "longer-context" tasks like Cloze-style QA and Wiki Title Prediction, with the assistance of a novel embedding disperser and phonological signals respectively. Moreover, we show that AxomiyaBERTa can leverage phonological signals for even more challenging tasks, such as a novel cross-document coreference task on a translated version of the ECB+ corpus, where we present a new SOTA result for an LRL. Our source code and evaluation scripts may be found at https://github.com/csu-signal/axomiyaberta.

1 Introduction

Transformer-based neural architectures such as BERT (Devlin et al., 2019) have revolutionized natural language processing (NLP). The ability to generate contextualized embeddings that both preserve polysemous word sense and similarity across dimensions through self-attention has contributed to significant improvements in various NLP tasks (Ethayarajh, 2019). Despite their successes, Transformers come at a high computational cost (Zhao et al., 2022) and still suffer from long-standing issues pertaining to data-hunger and availability of training resources. One effect of the dependency on big data is the continued proliferation of sophisticated NLP for well-resourced languages while low-resourced languages (LRLs) continue to be underrepresented, and the disparities continue to grow (Joshi et al., 2020).

This is particularly true for languages of India and South Asia where English is widely spoken among the educated and urban population. Therefore, those in India most likely to use and develop NLP may freely do so in English, but sole speakers of local Indian languages may remain effectively isolated from human language technology in their native tongues. While strides have been made in NLP for widely-spoken Indian languages (e.g., Hindi, Bengali, Marathi, Tamil, etc.), India is home to about a thousand languages, over 100 of which are considered "major"[1] but are not widely represented in NLP research. This lack of representation also precludes insights from those languages from contributing to the field (Bender, 2019).

In this paper, we present **AxomiyaBERTa**, a novel Transformer language model for the Assamese language.[2] AxomiyaBERTa has been trained in a low-resource and limited-compute setting, using only the masked language modeling (MLM) objective. Beyond a model for a new language, our novel contributions are as follows:

- Use of a novel combined loss technique to disperse AxomiyaBERTa's embeddings;
- Addition of phonological articulatory features as an alternate performance improve-

[1] https://censusindia.gov.in/census.website/data/census-tables

[2] Despite the name, AxomiyaBERTa is an ALBERT variant, not a RoBERTa variant. The name is derived from *Axomiya* (অসমীয়া, /ɔxɔmija/), the native term for the Assamese language, plus "BERTa" from both BERT and *barta*, meaning "conversation." The name also recalls *Asom Barta*, the official newsletter of the Government of Assam.

ment in the face of omitting the NSP training objective for longer-context tasks;

- Evaluation on event coreference resolution, which is novel for Assamese.

AxomiyaBERTa achieves competitive or state of the art results on multiple tasks, and demonstrates the utility of our approach for building new language models in resource-constrained settings.

2 Related Work

Multilingual large language models (MLLMs) trained over large Internet-sourced data, such as MBERT and XLM (Conneau et al., 2020), provide resources for approximately 100 languages, many of which are otherwise under-resourced in NLP. However, multiple publications (Virtanen et al., 2019; Scheible et al., 2020; Tanvir et al., 2021) have demonstrated that multilingual language models tend to underperform monolingual language models on common tasks; the "multilingual" quality of MLLMs may not be enough to assure performance on LRL tasks, due to language-specific phenomena not captured in the MLLM.

Since languages that share recent ancestry or a *Sprachbund* tend to share features, there has also been development of models and resources for languages from distinct regions of the world. South Asia is one such "language area," where even unrelated languages may share features (e.g., 4-way voice/aspiration distinctions, SOV word order, retroflex consonants, heavy use of light verbs). As such, researchers have developed region-specific models for South Asian languages such as IndicBERT (Kakwani et al., 2020) (11 languages, 8.8 billion tokens) and MuRIL (Khanuja et al., 2021) (17 languages, 16 billion tokens).

Subword tokenization techniques like byte-pair encoding (BPE) (Sennrich et al., 2016) yield comparatively better performance on LRLs by not biasing the vocabulary toward the most common words in a specific language, but BPE tokens also further obscure morphological information not immediately apparent in the surface form of the word. Nzeyimana and Niyongabo Rubungo (2022) tackle this problem for Kinyarwanda using a morphological analyzer to help generate subwords that better capture individual morphemes. However, despite similar morphological richness of many Indian languages, and likely due to similar reasons as outlined above, the dearth of NLP technology for most Indian languages extends to a lack of morphological parsers. We hypothesize that adding phonological features can also capture correlations between overlapping morphemes.

Previous NLP work in Assamese includes studies in corpus building (Sarma et al., 2012; Laskar et al., 2020; Pathak et al., 2022), POS tagging (Kumar and Bora, 2018), WordNet (Bharali et al., 2014; Sarmah et al., 2019) structured representations (Sarma and Chakraborty, 2012), image captioning (Nath et al., 2022c), and cognate detection (Nath et al., 2022a). There does not exist, to our knowledge, significant work on Assamese distributional semantics, or any monolingual, Transformer-based language model for the Assamese language evaluated on multiple tasks.

Our work complements these previous lines of research with a novel language model for Assamese, which further develops an initial model first used in Nath et al. (2022a). We account for the lack of an Assamese morphological analyzer with additional phonological features and task formulations that allow for strategic optimization of the embedding space before the classification layer.

2.1 Assamese

Assamese is an Eastern Indo-Aryan language with a speaker base centered in the Indian state of Assam. It bears similarities to Bengali and is spoken by 15 million L1 speakers (up to 23 million total speakers). Its literature dates back to the 13th c. CE. It has been written in its modern form since 1813, is one of 22 official languages of the Republic of India, and serves as a *lingua franca* of the Northeast Indian region (Jain and Cardona, 2004).

Despite this, Assamese data in NLP resources tends to be orders of magnitude smaller than data in other languages, even in South Asian region-specific resources (see Table 1).

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

Table 1: CC-100 (Conneau et al., 2020) and IndicCorp (Kakwani et al., 2020) data sizes (in millions of tokens) for Assamese, Bengali, Hindi, and English.

Assamese bears a similar level of morphological richness to other Indo-Aryan and South Asian languages, with 8 grammatical cases and a complex verbal morphology. Despite these points of comparison, Assamese has some unique phonological features among Indo-Aryan languages, such as the use of alveolar stops /t$^{(\text{h})}$/, /d$^{(\text{ɦ})}$/, velar fricative /x/, and approximant /ɹ/. This atypical sound pattern motivates the use of phonological signals in our model. Moreover, both the pretraining and task-specific corpora we use contain a large proportion of loanwords (e.g., from English) or words cognate with words in higher-resourced languages (e.g., Bengali). These words rendered with Assamese's unique sound pattern result in distinct, information-rich phoneme sequences.

3 Methodology

3.1 Pretraining

We trained on four publicly-available Assamese datasets: Assamese Wikidumps[3], OSCAR (Suárez et al., 2019)[4], PMIndia (Haddow and Kirefu, 2020)[5], the Common Crawl (CC-100) Assamese corpus (Conneau et al., 2020)[6], as well as a version of the ECB+ Corpus (Cybulska and Vossen, 2014) translated to Assamese using Microsoft Azure Translator. In total, after preprocessing, the training data amounts to approximately 26 million space-separated Assamese tokens.[7]

AxomiyaBERTa (66M parameters) was trained as a "light" ALBERT (specifically `albert-base-v2`) (Lan et al., 2019) model with *only* the MLM objective (Devlin et al., 2019), and no next sentence prediction (NSP), for 40 epochs (485,520 steps) with a vocabulary size of 32,000 and a SentencePiece BPE tokenizer (Kudo and Richardson, 2018). Tokenization methods like BPE can obfuscate certain morphological information. However, without a publicly-available morphological analyzer for Assamese, our motivation was to examine if phonological correlations might pick up similar information across different tasks while keeping model architecture and tokenizer consistent. We trained on 1 NVIDIA A100 80 GB device with a batch size of 32 and a sequence length of 128 for approximately 72 hours. Table 8 in Appendix A shows all specific pretraining configuration settings.

3.1.1 Special Token Vocabulary

The AxomiyaBERTa vocabulary includes two special trigger tokens: `<m>` and `</m>`. These act as separators *a la* the BERT `[SEP]` token, meaning that contextualized representations of these tokens were trained into the AxomiyaBERTa embedding space. Prior to pretraining, the translated ECB+ Corpus was annotated with these tokens surrounding event mentions. Since AxomiyaBERTa was not trained using the next sentence prediction objective (see Sec. 3.2.2), its embedding space needs those special triggers as separators between segments instead of the `[SEP]` tokens that segregate the token type IDs.

3.2 Fine-tuning

AxomiyaBERTa pretraining created a task-agnostic model optimized for the grammar and structure of Assamese. This model was then fine-tuned to achieve good performance on a number of different tasks. Beyond the task-specific fine-tuning, we made use of two auxiliary techniques: an *embedding disperser*, that optimized the AxomiyaBERTa embedding space away from severe anisotropy, and *phonological or articulatory attention* that acted as a single-head attention layer attending to both token-level and candidate-option level phonological signals. We first discuss these two techniques, followed by the specific task formulations we evaluated on. Note that the embedding disperser was used at the fine-tuning stage for Cloze-QA *only* due to severe anisotropy of the embedding space (Fig. 1 and Fig. 4, Appendix B).

3.2.1 Embedding Disperser

Without a meaningful objective to force embedding vectors apart during training, they trend toward an arbitrary center in $\mathbb{R}^d$ space. This phenomenon has also been observed by Gao et al. (2018), Ethayarajh (2019), and Demeter et al. (2020), among others. In Nath et al. (2022a), evidence was presented that the effect is more pronounced in smaller models. An effect of this can be illustrated by embeddings from an example task, Cloze-style question answering (Cloze-QA):

Let a "set" of embeddings consist of representations for a question (or context) Q and associated candidate answers $\{A, B, C, D\}$. "Within-set" cosine similarities represent the cosine similarities between $(Q+i, Q+j)$ for each candidate answer $i \in \{A, B, C, D\}$ and each other candi-

[3] https://archive.org/details/aswiki-20220120

[4] https://oscar-corpus.com

[5] https://paperswithcode.com/dataset/pmindia

[6] https://paperswithcode.com/dataset/cc100

[7] In resource-scarce settings, especially for LRLs, it is challenging to find large monolingual corpora for Transformer training. For instance, XLM-R was pretrained on about 164B tokens, of which only 5M were Assamese (see Table 1). However Ogueji et al. (2021) suggest that for LRLs, smaller datasets can actually work *better* than joint training with high-resourced parallel corpora.

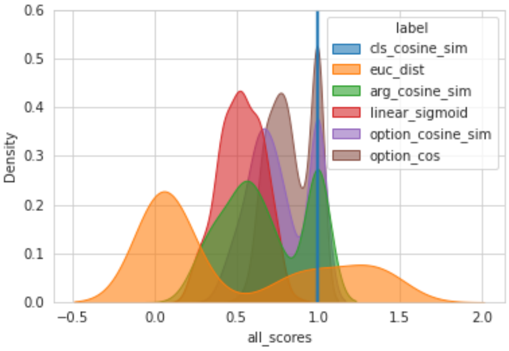

Figure 1: Kernel Density Estimation plots for within-set features of each component of the phonologically-aware embedding disperser (Fig. 2). `option_cos` is the output of the auxiliary discriminator while `linear_sigmoid` represents the linear layer and `euc_dist` represents L^2 norm between the raw `[CLS]` token embeddings. See Fig. 4 in Appendix B for equivalent "beyond-set" plots.

date $j \in \{A, B, C, D\}$ *where* $i \neq j$. "Beyond-set" cosine similarities represent similarities between all pairs in a candidate-plus-answers set compared to other such embedding sets from different questions. Fig. 1 shows KDE plots for various similarity metrics taken "within-set" for a random sample of 100 sets from the Cloze-QA dev set (see Sec. 3.2.3 for more details on the data). The blue spike at 1 for `cls_cosine_sim` shows how similar all `[CLS]` token embeddings are to each other, given AxomiyaBERTa's extremely anisotropic embedding space after pretraining. This makes it difficult to optimize a classification boundary during fine-tuning using standard techniques.

Therefore, to disperse the embedding space for greater discriminatory power, we used a combination of Binary Cross Entropy loss and Cosine Embedding loss to train the model. The architecture is shown in Fig. 2. The key components are: i) a *cosine embedding layer* that takes in `arg1` (context) and `arg2` (candidate) representations along with a `[CLS]` representation and outputs a 128D embedding into the cosine embedding loss function, and ii) an *auxiliary discriminator* that considers only `arg2` and `[CLS]` representations.

Mathematically,

$$L_{BCE} = -\frac{1}{n}\sum_{i=1}^{n}\left(Y_i \cdot \log \hat{Y}_i + (1 - Y_i) \cdot \log\left(1 - \hat{Y}_i\right)\right)$$

$$L_{COS}(x, y) = \begin{cases} 1 - \cos\left(x_1, x_2\right), & \text{if } y = 1 \\ \max\left(0, \cos\left(x_1, x_2\right) - \mathrm{m}\right), & \text{if } y = -1 \end{cases}$$

$$L_{COMB} = \alpha L_{BCE} + L_{COS}(x, y)$$

where m represents the margin for the cosine loss and α is 0.01. x_1 corresponds to `arg1` and x_2 corresponds to `arg2`. $y = 1$ if x_2 is the correct answer and $y = -1$ if not. At inference, we computed Euclidean distance between the embedding outputs of the auxiliary discriminator and the cosine embedding layer with a threshold T of 4.45, found through hyperparameter search.

`option_cosine_sim` in Fig. 1 shows the outputs of the embedding disperser's cosine embedding layer while `option_cos` shows the outputs of the auxiliary discriminator. In both cases we see distinct distributions that separate correct and incorrect answers. Works such as Cai et al. (2021) present evidence of such cases of global token anisotropy in other Transformer models and suggest that creating such local isotropic spaces leads to better results in downstream tasks.

3.2.2 Phonological/Articulatory Attention

While the NSP objective is effective at training LLMs to encode long-range semantic coherence (Shi and Demberg, 2019), it comes at a significant additional computational cost. Moreover, for very low-resource languages like Assamese, a lack of available long document or paragraph data means there may not exist a sufficient volume of coherent consecutive sentences in the training data.

We hypothesize that when fine-tuning a smaller model like AxomiyaBERTa in a resource-constrained setting, adding phonological signals to the latent representations of text samples allows us to achieve a balanced trade-off between possible information loss due to reduced supervision (no NSP objective) and improved task-specific performance, at a lower compute cost.

Previous works (e.g., Mortensen et al. (2016); Rijhwani et al. (2019); Nath et al. (2022b)) have shown that phonological features are useful for both token-level "short-context" tasks like NER or loanword detection as well as "longer-context" tasks like entity linking. We fine-tune for longer-context tasks by encoding candidate answers as phonological features and the pooled embedding of the context, and computing the relative difference in mutual information between each candidate answer and the context. High variance in cosine similarities within pairs in a context-candidate set is due to the phonological signals.

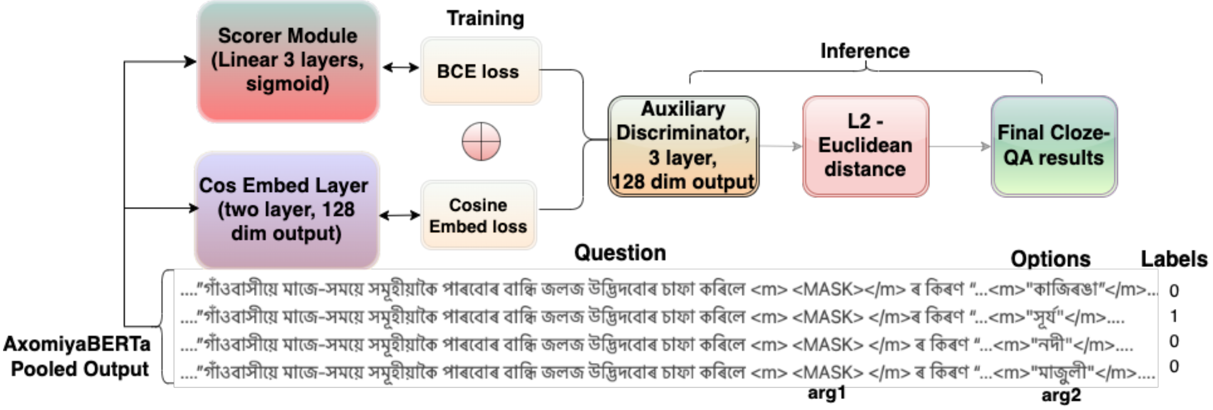
Training
Scorer Module (Linear 3 layers, sigmoid)
BCE loss
Cos Embed Layer (two layer, 128 dim output)
Cosine Embed loss
Inference
Auxiliary Discriminator, 3 layer, 128 dim output
L2 - Euclidean distance
Final Cloze-QA results
AxomiyaBERTa Pooled Output
Question
Options
Labels
...."গাঁওবাসীয়ে মাজে-সময়ে সমূহীয়াকৈ পাৰবোৰ বান্ধি জলজ উদ্ভিদবোৰ চাফা কৰিলে <m> <MASK></m> ৰ কিৰণ "...<m>"কাজিৰঙা"</m>... 0
...."গাঁওবাসীয়ে মাজে-সময়ে সমূহীয়াকৈ পাৰবোৰ বান্ধি জলজ উদ্ভিদবোৰ চাফা কৰিলে <m> <MASK> </m>ৰ কিৰণ "...<m>"সূৰ্য"</m>.... 1
...."গাঁওবাসীয়ে মাজে-সময়ে সমূহীয়াকৈ পাৰবোৰ বান্ধি জলজ উদ্ভিদবোৰ চাফা কৰিলে <m> <MASK> </m> ৰ কিৰণ "...<m>"নদী"</m>.... 0
...."গাঁওবাসীয়ে মাজে-সময়ে সমূহীয়াকৈ পাৰবোৰ বান্ধি জলজ উদ্ভিদবোৰ চাফা কৰিলে <m> <MASK> </m> ৰ কিৰণ "...<m>"মাজুলী"</m>.... 0
arg1
arg2

Figure 2: Embedding disperser architecture with Cosine Embedding and Binary Cross Entropy (BCE) Loss. Cloze-QA data used as example.

Table 2 shows that the mean, standard deviation, and variances of [CLS] token cosine similarities for pretrained AxomiyaBERTa are much smaller than those extracted from XLM, but fine-tuning with phonological signals brings AxomiyaBERTa's values much closer XLM's.

	AxB	XLM	AxB + Phon
Mean	.998	.82	.67
Variance	5e-6	.08	.06
Stdev	.002	.28	.25
Min	.993	.13	.17

Table 2: Statistics of AxomiyaBERTa (AxB) and XLM's [CLS] token cosine similarities compared to those of the pooled output of AxomiyaBERTa with phonological signals (AxB + Phon) over 100 random samples of within-set pairs of Cloze-QA dev set.

To extract phonological features, we used the Assamese grapheme-to-phoneme mapping from Nath et al. (2022a), written for the Epitran library (Mortensen et al., 2018)[8] to convert all text into the International Phonetic Alphabet (IPA). We then used the PanPhon library (Mortensen et al., 2016) to convert the IPA transcriptions into 24 subsegmental features such as place and manner of articulation, voicing, etc.

These feature vectors are padded to the maximum length (across train, test, and dev sets), and then concatenated to either the pooled context embedding (for long-context tasks) or the named-entity token embedding (for NER tasks).

3.2.3 Cloze-style multiple-choice QA

We fine-tuned AxomiyaBERTa on the Cloze-style Wiki question answering task from the IndicGLUE dataset (Kakwani et al., 2020). We surrounded both the masked text segment as well as the four candidate answers with the special tokens (<m> and </m>) and then fed them into the pretrained AxomiyaBERTa model to get pairwise scores with BCE loss. Positive samples were labeled as 1 and negatives as 0. The encoded representation for each sample was a concatenation of the pooled ([CLS]) token output, the averaged embedding for the masked text segment (arg1), that of the candidate answer (arg2), and the element-wise multiplication of arg1 and arg2. This was input into a pairwise scorer *a la* Caciularu et al. (2021). We fine-tuned our model (with and without phonological attention) with the pairwise scorer head for 5 iterations with a batch size of 80, a scorer head learning rate of 1e-4 and a model learning rate of 2e-5.

3.2.4 Named Entity Recognition (NER)

For NER, we fine-tuned and evaluated AxomiyaBERTa on two datasets: WikiNER (Pan et al., 2017) and AsNER (Pathak et al., 2022). For both datasets, we fed in the tokenized sentence while masking out all sub-word tokens except the first of each word. We used a token-classification head fine-tuned using a multi-class cross-entropy loss for the label set of the respective datasets. For our model without phonological signals, we fine-tuned for 10 epochs with a learning rate of 2e-5 with a linear LR scheduler and a batch size of 20. For our phonological attention-based model, we fine-tuned for 20 epochs with a batch size of 40 while keeping all other hyperparameters the same.

[8] This was created using reference resources like Omniglot (https://omniglot.com/writing/assamese.htm) and Wikiwand/Assamese (https://www.wikiwand.com/en/Help:IPA/Assamese) along with native-speaker verification.

3.2.5 Wikipedia Section Title Prediction

Like Cloze-QA, this task comes from IndicGLUE (Kakwani et al., 2020). Fine-tuning for this task was quite similar to that of Cloze-QA, except we did not surround the candidates or the contexts with the trigger tokens. We fed in the Wikipedia section text and candidate title and optimized the multi-class cross entropy loss with a multiple choice head. We fine-tuned for 20 epochs with a batch size of 40. For the phonologically-aware model, we concatenated the articulatory signals to the pooled embedding output for each sample and fine-tuned our model for 200 iterations with a batch size of 40. We used a smaller model learning rate of `1e-6` and a classifier head learning rate of `9.5e-4` for both these models.

3.2.6 Pairwise Scorer for Assamese CDCR

Coreference resolution in a cross-document setting (CDCR) involves identifying and clustering together mentions of the same entity across a set of documents (Lu and Ng, 2018). Following CDCR approaches in Cattan et al. (2021) and Caciularu et al. (2021), we trained a pairwise scorer with BCE loss over all antecedent spans for each sentence containing an event (across all documents) while ignoring identical pairs. We generated concatenated token representations from Transformer-based LMs by joining the two paired sentences after surrounding the event mentions with the special trigger tokens. These representations were input to the pairwise scorer (PS) to calculate *affinity scores* between all those pairs. Mathematically,

$$Scores(i,j) = PS([CLS], f(x), f(y), f(x) * f(y)),$$

where $[CLS]$ represents the pooled output of the entire sentence pair, $f(x)$ and $f(y)$ are the representations of the two events (in context) and $*$ represents element-wise multiplication.

We trained the Pairwise Scorer for 10 epochs for all baseline models as well as AxomiyaBERTa. At inference, we used a connected-components clustering technique with a tuned threshold to find coreferent links. For baselines and ablation tasks, we calculated coreference scores using a lemma-based heuristic, and fine-tuned four other popular MLLMs using the same hyperparameters. More details and analysis are in Appendix D.

4 Evaluation

Table 3 shows the number of samples in the train, dev, and test splits, and the padding length, for all tasks we evaluated on. For Cloze-QA and Wiki-Titles, we evaluated on IndicGLUE. For NER, we evaluated on AsNER and WikiNER. For our novel coreference task, we evaluated on the translated ECB+ corpus, where the ratio of coreferent to non-coreferent pairs in the test set is approximately 1:35. We conducted exhaustive ablations between native and the phonologically-aware models for each task, and compared to previously-published baselines where available. For Cloze-QA, we created a train/test split of approximately 4.5:1. We fine-tuned off-the-shelf IndicBERT and MBERT on AsNER for 10 epochs on 1 NVIDIA RTX A6000 48 GB device with a batch size of 20.

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

Table 3: Table showing distribution of train/dev/test splits for all tasks. T-ECB+ signifies number of event mentions in the translated ECB+ corpus, keeping special trigger tokens in place. “Pad-Len” represents the maximum padded length of the articulatory feature embeddings generated from PanPhon for all three splits.

5 Results and Discussion

Table 4 shows Test F1 Scores/Accuracy for AxomiyaBERTa for the various short-context (classification) and long-context (multiple-choice) tasks. We compared baselines from previous works and newly fine-tuned baselines for certain tasks. We used the same pretrained model for all experiments with task fine-tuning heads consistent with previous benchmarks (Kakwani et al., 2020). One exception is the Cloze-QA task where we dealt with task-specific severe anisotropy with embedding dispersal.

5.1 Short-context: AsNER and WikiNER

AxomiyaBERTa achieved SOTA performance on the AsNER task and outperformed most other Transformer-based LMs on WikiNER.

Phonologically-aware AxomiyaBERTa Our experiments suggest that phonological signals are informative additional features for short-context tasks like NER for low-resourced, smaller models like AxomiyaBERTa. Table 4 shows that phonologically-aware AxomiyaBERTa outperformed non-phonological (hereafter “native”)

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

Table 4: Test F1 Scores/Accuracy for AxomiyaBERTa on all evaluation tasks, compared to previous baselines and our fine-tuned baselines. "AxomiyaBERTa + Phon" shows results for phonologically-aware AxomiyaBERTa. AsNER scores with a * represent versions we fine-tuned for this task. For Cloze-QA, Wiki-Titles and WikiNER, other model performances are from Kakwani et al. (2020). **Bold** indicates best performance.

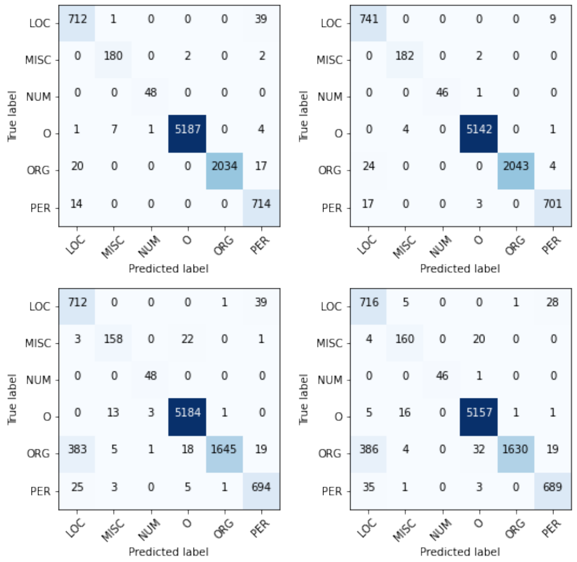

Figure 3: Top: Confusion matrices showing AxomiyaBERTa performance on AsNER without [L] and with [R] phonological awareness. Bottom: IndicBERT [L] and MBERT [R] performance on AsNER.

AxomiyaBERTa by >5 F1 points on AsNER, with an even greater improvement (10 F1 points) on WikiNER. AxomiyaBERTa also outperformed other baselines for both tasks, with the exception of MBERT on Wiki-based tasks.[9] Fig. 3 shows confusion matrices of performance on AsNER.

IndicBERT and MBERT misclassified *ORG* tokens as *LOC* 16 times as much as AxomiyaBERTa. Specific cases include sub-tokens like নিউয়ৰ্ক (/niujɔɹk/, "New York") or ছিংগাপুৰ (/siŋgapuɹ/, "Singapore"), that are actually parts of entities like এ-ষ্টাৰ ছিংগাপুৰ (/e staɹ siŋgapuɹ/, "A-Star Singapore") or নিউয়ৰ্ক ব্লাড চেণ্টাৰ (/niujɔɹk blad sentaɹ/, "New York Blood Center"). This suggests that smaller, monolingual models like AxomiyaBERTa with a reduced sequence length and no NSP training objective are optimized for NE classification tasks with greater attention to local context (since the average sentence containing NEs is ∼6 tokens).

Better overall performance on AsNER than on WikiNER can be partially attributed to having one fewer class and a more balanced distribution between categories. AsNER performance likely benefited from a greater phonological signal and more data to tune on (Table 3) whereas WikiNER text samples are, on average, longer than 128 tokens (AxomiyaBERTa's maximum token length) possibly causing a performance loss due to truncated context.

Phonological Signals: A Disambiguation Tool Even though phonologically-aware AxomiyaBERTa took a hit on identifying *O* tokens, it compensated with improved results across other classes. Phonologically-aware AxomiyaBERTa also reduced misclassifications of *ORG* tokens as *PER* compared to all other models, including native AxomiyaBERTa. Specific cases include tokens that imply persons, e.g., স্বামীনাথন or সাহা, but are actually part of *ORG* NEs, e.g., স্বামীনাথন কমিছন ("Swaminathan Commission") or সাহা ইনষ্টিটিউট অফ ফিজিক্স ("Saha Institute of Physics"). Similarly, in the WikiNER task, phonological attention reduced misclassification of *B-ORG* and *I-ORG* tokens as *B-PER* and *I-PER* respectively (see Appendix C). These results suggest phonological inputs help enrich embeddings of smaller-sized LMs to distinguish such ambiguous tokens.

5.2 Long-context: Multiple Choice

On Wiki-Titles, phonological AxomiyaBERTa does better with semantically harder multiple-choice sets, which have a higher average cosine similarity between the candidate options. Native AxomiyaBERTa fails on these samples. As shown

[9] Wikipedia comprises almost all of MBERT's training data. MBERT does not support Assamese, but does support Bengali, and Assamese is written using a variant of the same script. Named entities are often written identically in Bengali and Assamese, which could explain this trend.

in Table 5, **P+N-** has the highest average cosine similarity between the sample sets, suggesting that there are cases where phonological signals compensate for low semantic variation among candidate options. On the other hand, native AxomiyaBERTa tends to do better with multiple-choice sets that have wider (relative) semantic variation within that set, on average. Since the overall distribution of embeddings in this task is still extremely close, this suggests that phonological signals are doing for Wiki-Titles what the embedding disperser did for Cloze-QA (see Sec. 3.2.1).

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

Table 5: 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, e.g., **P+N-** shows samples phonological AxomiyaBERTa answered correctly that the native variant did not.

5.3 Novel Task: Event Coreference on Translated ECB+

Table 6 shows event coreference resolution results on the translated ECB+ test set using a tuned affinity-threshold (T). These results include both within- and cross-document system outputs from AxomiyaBERTa, other Transformer-based LMs, and a lemma-based heuristic.[10]

AxomiyaBERTa often outperformed the lemma-similarity baseline and other LMs. Native and phonological AxomiyaBERTa have the best MUC and BCUB F1 scores, respectively, while also outperforming all other Transformer-based LMs on BLANC and CoNLL F1. Phonologically-aware AxomiyaBERTa also outperforms native AxomiyaBERTa by almost 2 F1 points on CoNLL F1. More importantly, the phonological signals help detect more challenging coreferent links where mere surface-level lemma similarity can fail. While native and phonological AxomiyaBERTa performed comparably, the true positives retrieved by the phonological version contained a higher proportion of non-similar lemmas, which were usually missed by the lemma heuristic. Meanwhile, native AxomiyaBERTa retrieved results with more similar lemmas, labeling more non-similar lemma pairs as false negatives (Table 7). Compared to the other Transformer models, this also had the effect of increasing precision according to most metrics, though at the cost of decreasing recall. However, the increased precision was usually enough to increase F1 overall, pointing to the utility of phonological signals in detecting more challenging cases. We hypothesize that this is because these challenging pairs may consist of synonyms and/or loanwords, and phonological signals helped correlate these different surface forms, which in addition to the semantic information at the embedding level helps create coreference links.

For instance, কনচাল্টিং (/kɔnsaltiŋ/, "consulting") and ইঞ্জিনিয়াৰিং (/indʒinijaɹiŋ/, "engineering") denote two coreferent events pertaining to the same company (EYP Mission Critical Facilities). Since both are borrowed words that maintain the original phonological form, phonological signals can help pick out unique articulation beyond surface-level lemma similarity. Similarly, in cases of synonyms like মৃত্যুৰ (/mɹittuɹ/, "(of) death") and হত্যা (/ɦɔtta/, "killing"), which do not share surface-level similarity yet are coreferent, phonological signals can help. Where lemmas are already similar, phonological signals provide little extra information.

We should note that for coreference, the specific metric used matters a lot. For instance, almost 33% of the ECB+ dataset across all three splits consists of singleton mentions. Since MUC score is not as sensitive to the presence of singletons as BCUB (Kübler and Zhekova, 2011), this could explain AxomiyaBERTa's (and XLM's) relative drop in performance on the BCUB metric. On the other hand, the lower CEAF-e F1 score may be due to CEAF-e's alignment algorithm, which tends to ignore correct coreference decisions when response entities are misaligned (Moosavi and Strube, 2016).

Ablations between native and phonological AxomiyaBERTa showed that where lemmas for a pair of potentially coreferent events are identical (e.g., আৰম্ভ - /aɹɔmb^{ɦ}o/, "start"), non-phonological representations primarily determine the pairwise scores and the coreference decision. Table 7 shows that even though phonological signals tend to disambiguate harder event pairs, decreased performance (e.g., MUC F1 phonological vs. native

[10] We found an affinity threshold ($T = 7$) to work for all models except phonologically-aware AxomiyaBERTa ($T = 0$) and XLM-100 ($T = -1.94$). For the latter, we use the mean of all scores due to an extremely narrow distribution as shown in the Appendix. More analysis of why this happens is the subject of future work.

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

Table 6: Event coreference results on Assamese (translated) ECB+ test set from pairwise scorer using AxomiyaBERTa, compared with other Transformer-based LMs and the lemma-based heuristic. **Bold** indicates best overall performance per metric. "C-F1" is CoNLL F1. [†]We evaluate XLM-100 to compare performance on this task of a slightly larger model than XLM-R where most other major design factors remain the same.

AxomiyaBERTa) could be due to native representations of the same-lemma pair being weakly correlated with the pairwise scores, a possibility when a coreferent event pair has high contextual dissimilarity. Phonological signals may add noise here.

We also see that the lemma-based heuristic baseline is overall a very good performer. While this may be a property of the nature of coreference tasks in general or specific to a dataset (as a high percentage of coreferent events use the same lemma), we must also allow for the possibility that this may also be an artifact of translation noise. Since we used an automatically-translated version of the ECB+ corpus (albeit with some native speaker verification), and since Assamese is still a low-resource language, the decoder vocabulary of the translator may be limited, meaning that synonymous different-lemma pairs in the original corpus may well have been collapsed into same-lemma pairs in the translation, artificially raising the performance of the lemma heuristic.

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)

Table 7: 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 ($= L2/TP$). Values in parentheses show the equivalent distribution within false negatives for comparison.

6 Conclusion and Future Work

In this paper, we presented a novel Transformer model for Assamese that optionally includes phonological signals. We evaluated on multiple tasks using novel training techniques and have demonstrated SOTA or comparable results, showing that phonological signals can be leveraged for greater performance and disambiguation for a low-resourced language. AxomiyaBERTa achieves SOTA performance on short-context tasks like AsNER and long-context tasks like Cloze-QA while also outperforming most other Transformer-based LMs on WikiNER, with additional improvement resulting from the phonologically-aware model. For challenging tasks like CDCR, we have shown that both AxomiyaBERTa outperformed other Transformer-based LMs on popular metrics like BCUB, MUC, and CoNLL F1.

More generally, we have shown that strategic techniques for optimizing the embedding space and language-specific features like phonological information can lower the barrier to entry for training language models for LRLs, making it more feasible than before with lower amounts of data and a ceiling on compute power. Our experiments suggest phonological awareness boosts performance on many tasks in low-resource settings. Future models for other LRLs can leverage our ideas to train or fine-tune their own models. Since smaller models tend toward anisotropy, embedding dispersal may pave the way for more such performant LRL models.

Future work may include incorporating phonological signals during pretraining instead of fine-tuning, carrying out evaluations against semantically harder tasks like paraphrasing or emotion detection, zero-shot transfer to similar languages, and a contrastive learning framework with a triplet loss objective for CDCR.

Our trained checkpoints are available on HuggingFace at https://huggingface.co/Abhijnan/AxomiyaBERTa. We hope this resource will accelerate NLP research for encoding language-specific properties in LRLs.

Limitations

Let us begin with the obvious limitation: AxomiyaBERTa only works on Assamese. In addition, since Assamese comprises a number of dialects and we trained on internet-sourced data, we have no clear evidence regarding which dialects AxomiyaBERTa is most suited to or if it performs as well on non-standard dialects.

AxomiyaBERTa did not perform all that well on Wikipedia Title Selection, compared to other Transformer-based models. Our best result is on par with XLM-R and close to IndicBERT-BASE, but well below MBERT performance. We hypothesize that the amount of Wikipedia training data in MBERT is a cause of this, but we find that phonological attention makes a big difference in AxomiyaBERTa's performance (increasing accuracy from 26% to 59%). Nonetheless, the reasons behind this subpar performance, and whether AxomiyaBERTa can be improved for this task without, say, overfitting to Wikipedia, need further investigation.

Ethics Statement

Data Usage Because of the publicly-available, internet-sourced nature of our training data, we cannot definitively state that the current version of AxomiyaBERTa is free of bias, both in terms of outputs nor, as mentioned in the limitations section, if there are dialect-level biases toward or against certain varieties of Assamese that may be trained into the model. Such investigations are the topic of future research.

Resource Usage and Environmental Impact At 66M parameters, AxomiyaBERTa is a smaller language model that is relatively quick to train and run. Training was conducted on single GPU devices. Pretraining AxomiyaBERTa took approximately 3 days, and task-level fine-tuning took roughly 30 minutes for non-phonological AxomiyaBERTa and 1-2 hours for phonological AxomiyaBERTa (depending on the task). Training the pairwise scorer for CDCR took 12-19 minutes. Training and fine-tuning took place on the same hardware. For comparison, fine-tuning IndicBERT and MBERT on the AsNER dataset for evaluation took roughly 20-30 minutes each. These figures indicate that relative to work on other Transformer models, training and evaluating AxomiyaBERTa (including running other baselines for comparison) comes with a comparatively lower resource usage and concomitant environmental impact. This lower resource usage also has implications for the "democratization" of NLP, in that we have demonstrated ways to train a performant model with fewer local resources, meaning less reliance on large infrastructures available to only the biggest corporations and universities.

Human Subjects This research did not involve human subjects.

Acknowledgments

We would like to thank the anonymous reviewers whose feedback helped improve the final copy of this paper. Special thanks to Ibrahim Khebour for helping with the phonological feature extraction process for the Wikipedia Section Title Prediction task.

References

Shafiuddin Rehan Ahmed, Abhijnan Nath, James H. Martin, and Nikhil Krishnaswamy. 2023. 2*n is better than n^2: Decomposing Event Coreference Resolution into Two Tractable Problems. In *Findings of the Association for Computational Linguistics: ACL 2023*. ACL.

Emily Bender. 2019. The #benderrule: On naming the languages we study and why it matters. *The Gradient*, 14.

Himadri Bharali, Mayashree Mahanta, Shikhar Kumar Sarma, Utpal Saikia, and Dibyajyoti Sarmah. 2014. An analytical study of synonymy in Assamese language using WorldNet: Classification and structure. In *Proceedings of the Seventh Global Wordnet Conference*, pages 250–255.

Avi Caciularu, Arman Cohan, Iz Beltagy, Matthew Peters, Arie Cattan, and Ido Dagan. 2021. CDLM: Cross-document language modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2648–2662, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xingyu Cai, Jiaji Huang, Yuchen Bian, and Kenneth Church. 2021. Isotropy in the Contextual Embedding Space: Clusters and Manifolds. In *International Conference on Learning Representations*.

Oralie Cattan, Sophie Rosset, and Christophe Servan. 2021. On the cross-lingual transferability of multilingual prototypical models across NLU tasks. In *Proceedings of the 1st Workshop on Meta Learning and Its Applications to Natural Language Processing*, pages 36–43, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Agata Cybulska and Piek Vossen. 2014. Using a sledgehammer to crack a nut? lexical diversity and event coreference resolution. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4545–4552, Reykjavik, Iceland. European Language Resources Association (ELRA).

David Demeter, Gregory Kimmel, and Doug Downey. 2020. Stolen probability: A structural weakness of neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2191–2197, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tieyan Liu. 2018. Representation Degeneration Problem in Training Natural Language Generation Models. In *International Conference on Learning Representations*.

Barry Haddow and Faheem Kirefu. 2020. PMIndia–A Collection of Parallel Corpora of Languages of India. *arXiv preprint arXiv:2001.09907*.

Danesh Jain and George Cardona. 2004. *The Indo-Aryan Languages*. Routledge.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha P. Talukdar. 2021. MuRIL: Multilingual Representations for Indian Languages. *CoRR*, abs/2103.10730.

Sandra Kübler and Desislava Zhekova. 2011. Singletons and coreference resolution evaluation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 261–267.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Ritesh Kumar and Manas Jyoti Bora. 2018. Part-of-speech annotation of English-Assamese code-mixed texts: Two approaches. In *Proceedings of the First International Workshop on Language Cognition and Computational Models*, pages 94–103, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *International Conference on Learning Representations*.

Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji, Partha Pakray, and Sivaji Bandyopadhyay. 2020. EnAsCorp1.0: English-Assamese corpus. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 62–68, Suzhou, China. Association for Computational Linguistics.

Jing Lu and Vincent Ng. 2018. Event coreference resolution: A survey of two decades of research. In *IJCAI*, pages 5479–5486.

Nafise Sadat Moosavi and Michael Strube. 2016. Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 632–642, Berlin, Germany. Association for Computational Linguistics.

David R. Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. Epitran: Precision G2P for many languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori Levin. 2016. PanPhon: A resource for mapping IPA segments to articulatory feature vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3475–3484, Osaka, Japan. The COLING 2016 Organizing Committee.

Abhijnan Nath, Rahul Ghosh, and Nikhil Krishnaswamy. 2022a. Phonetic, semantic, and articulatory features in Assamese-Bengali cognate detection. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 41–53, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Abhijnan Nath, Sina Mahdipour Saravani, Ibrahim Khebour, Sheikh Mannan, Zihui Li, and Nikhil Krishnaswamy. 2022b. A generalized method for automated multilingual loanword detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4996–5013, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Prachurya Nath, Prottay Kumar Adhikary, Pankaj Dadure, Partha Pakray, Riyanka Manna, and Sivaji Bandyopadhyay. 2022c. Image Caption Generation for Low-Resource Assamese Language. In *Proceedings of the 34th Conference on Computational Linguistics and Speech Processing (ROCLING 2022)*, pages 263–272, Taipei, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Antoine Nzeyimana and Andre Niyongabo Rubungo. 2022. KinyaBERT: a morphology-aware Kinyarwanda language model. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5347–5363, Dublin, Ireland. Association for Computational Linguistics.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Dhrubajyoti Pathak, Sukumar Nandi, and Priyankoo Sarmah. 2022. AsNER - annotated dataset and baseline for Assamese named entity recognition. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6571–6577, Marseille, France. European Language Resources Association.

Shruti Rijhwani, Jiateng Xie, Graham Neubig, and Jaime Carbonell. 2019. Zero-shot neural transfer for cross-lingual entity linking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6924–6931.

Shikhar Kr. Sarma, Himadri Bharali, Ambeswar Gogoi, Ratul Deka, and Anup Kr. Barman. 2012. A structured approach for building Assamese corpus: Insights, applications and challenges. In *Proceedings of the 10th Workshop on Asian Language Resources*, pages 21–28, Mumbai, India. The COLING 2012 Organizing Committee.

Shikhar Kumar Sarma and Rita Chakraborty. 2012. Structured and Logical Representations of Assamese Text for Question-Answering System. In *Proceedings of the Workshop on Question Answering for Complex Domains*, pages 27–38.

Jumi Sarmah, Shikhar Kumar Sarma, and Anup Kumar Barman. 2019. Development of Assamese rule based stemmer using WordNet. In *proceedings of the 10th Global WordNet Conference*, pages 135–139.

Raphael Scheible, Fabian Thomczyk, Patric Tippmann, Victor Jaravine, and Martin Boeker. 2020. Gottbert: a pure German language model. *arXiv preprint arXiv:2012.02110*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Wei Shi and Vera Demberg. 2019. Next sentence prediction helps implicit discourse relation classification within and across domains. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5790–5796, Hong Kong, China. Association for Computational Linguistics.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache.

Hasan Tanvir, Claudia Kittask, Sandra Eiche, and Kairit Sirts. 2021. EstBERT: A Pretrained

Language-Specific BERT for Estonian. *NoDaLiDa 2021*, page 11.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: BERT for Finnish. *arXiv preprint arXiv:1912.07076*.

Jing Zhao, Yifan Wang, Junwei Bao, Youzheng Wu, and Xiaodong He. 2022. Fine- and coarse-granularity hybrid self-attention for efficient BERT. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4811–4820, Dublin, Ireland. Association for Computational Linguistics.

A Training Configuration

Table 8 shows the pretraining configuration for AxomiyaBERTa.

B Further Details on Embedding Disperser

Fig. 4 shows KDE plots for outputs of different components of the embedding disperser, showing the contrast between features within-set and beyond-set for Cloze-QA samples, and showing the difference between AxomiyaBERTa with phonological awareness and without. The `option_cos` label (brown) shows an interesting phenomenon. This is the output of the embedding disperser at inference (Auxiliary Discriminator in Fig. 2) and represents a 128-dimensional embedding output from the `[CLS]` token concatenated with `arg2` or the candidate answer input. We see a distinct shift in cosine similarity scores between within-set and beyond-set with one peak very close to 1 in the case of the within-set pairs while getting clearly dispersed to a lower cosine similarity score in the case of beyond-set pairs. This phenomenon is even further accentuated by feeding phonological signals to the disperser. In this case, as shown in the top right plot, the cosine similarity peak for `option_cos` has a much higher density compared to the non-phonological disperser while the overall distribution is shifted to a higher cosine similarity.

Another interesting trend is the `linear_sigmoid` label (red) which is the sigmoidal output of the linear layer of the disperser, trained with a combination of cosine embedding loss and BCE loss when fed an input of the combined `arg1` and `arg2` representations generated with the special trigger tokens. In this case, feeding phonological signals to the model reduces dispersion (an inverse trend) in the cosine similarities between within-set and beyond-set pairs (as seen in the top-left plot where this label has a narrower top with a wider bottom). However, this reverse effect is less pronounced than that seen in the `option_cos` cosine similarity plot, perhaps due to richer contextual information carried by the trigger token representations (the inputs to this layer). In other words, and as shown in the `arg_cosine_sim` plot, its dispersion between the within- and beyond-set pairs suggests why such an effect is less-pronounced.

Works such as Cai et al. (2021) present evidence of such global token anisotropy in other BERT and GPT-model variants while also suggesting ways to locate/create local isotropic spaces more susceptible for NLP tasks. Interestingly, cosine similarities of output embeddings from our Auxiliary Discriminator (`option_cos` in Fig. 1) show a marked difference in the extent of anisotropy between within-set and beyond-set pairs, a phenomenon further accentuated with additional phonological signals (top right plot in Fig. 4). These experiments suggest that a combination of our embedding disperser architecture together with phonological signals (Sec. 3.2.2 for more details) can effect a shift towards local spaces of isotropy in the embedding space of the fine-tuned AxomiyaBERTa model for Cloze-QA and potentially other tasks.

C Further Discussion on Short-Context Results

Fig. 5 shows native and phonological AxomiyaBERTa performance on WikiNER. We see comparative performance, but with phonological signals there are fewer confusions of *B-ORG* with *B-PER* and *I-ORG* with *I-PER*. Specific examples are similar to those seen in Sec. 5.1, e.g., স্বামীনাথন (কমিছন) (“Swaminathan [Commission]”) or সাহা (ইনষ্টিটিউট অফ ফিজিক্স) (“Saha [Institute of Physics]”). Being organizations named after people, this is a case where phonological signals actually help. Interestingly, phonological signals also help with NER even when the NEs are broken down into BIO chunks, which was not the case in AsNER. We should observe that with phonological signals, there is an *increase* in *B-LOC* tokens classified as *B-PER* tokens, which is the topic of future investigation.

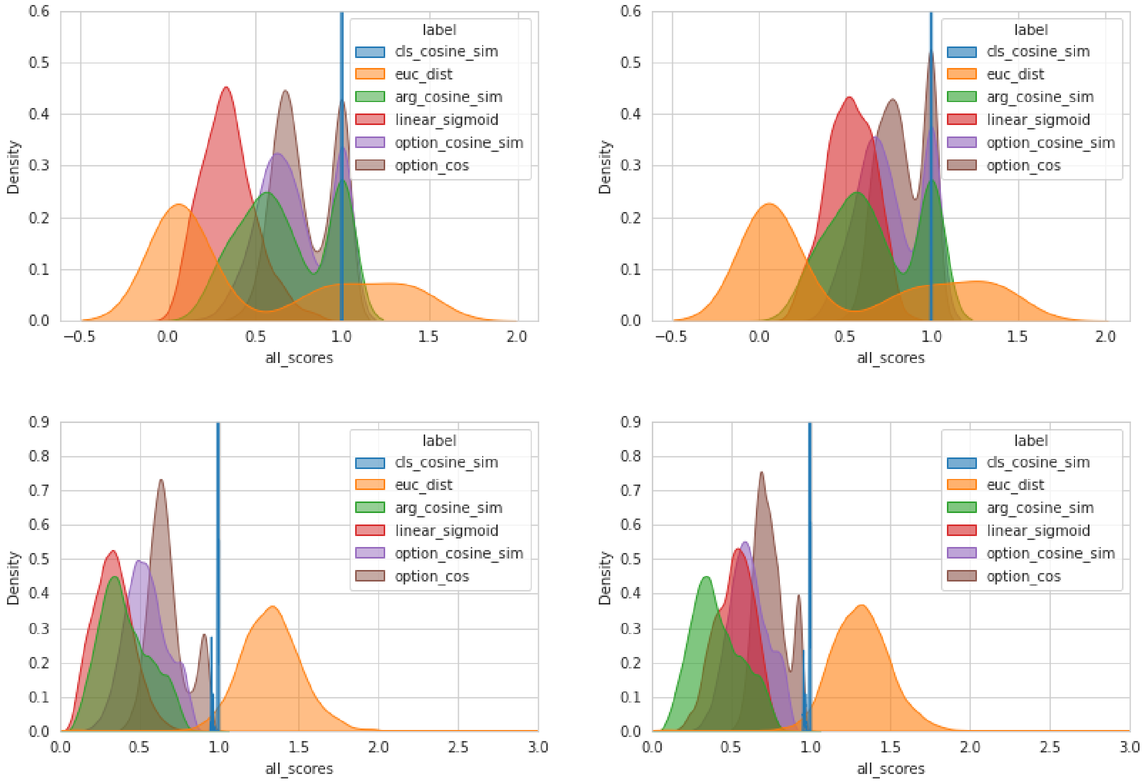

Figure 4: Kernel density estimation plots from various feature sets of the embedding disperser model. The figure on the top left represents “within set”: cosine similarities between each set of candidate answers plus context and the remaining three pairs in that set, e.g., if Q is the question/context and $A/B/C/D$ are the candidate answers, $S_C(Q+A, Q+i)$, where i represents one of the remaining three candidates B, C, D. The figure on the bottom left represents “beyond-set” cosine similarities: all pairs in a candidate-plus-answers set are compared to other such sets with the cosine similarity metric. We run our experiments for 100 randomly selected sets from the Cloze-QA dev set. The top right (within-set) and bottom right (beyond-set) figures are equivalent figures for our models with phonological awareness.

Parameters	Config
architecture	AlbertForMaskedLM
attention_probs_dropout_prob	0.1
bos_token_id	2
classifier_dropout_prob	0.1
embedding_size	128
eos_token_id	3
hidden_act	gelu
hidden_dropout_prob	0.1
hidden_size	768
initializer_range	0.02
inner_group_num	1
intermediate_size	3072
layer_norm_eps	1e-05
max_position_embeddings	514
num_attention_heads	12
num_hidden_groups	1
num_hidden_layers	6
position_embedding_type	“absolute"
transformers_version	“4.18.0"
vocab_size	32001

Table 8: AxomiyaBERTa Model configuration trained on a monolingual Assamese corpus.

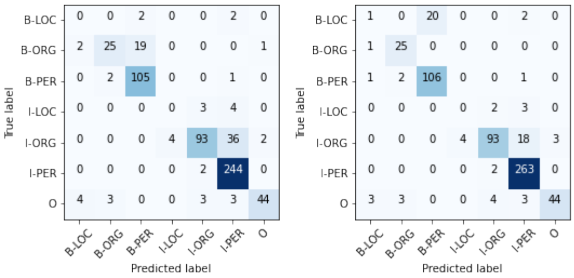

Figure 5: Confusion matrices showing AxomiyaBERTa performance on WikiNER without [L] and with [R] phonological awareness.

D Further Discussion on Pairwise Scorer for CDCR on Assamese ECB+

The lemma-based heuristic comes from the fact that a large proportion of coreferent mention pairs can be identified simply because they use the same lemma. These “easy” cases gives coreference a very high baseline even when this naive heuristic is used. The long tail of “harder” pairs require more sophisticated approaches (Ahmed et al., 2023).

Fig. 6 shows the affinity scores from the pairwise scorer using various model outputs. AxomiyaBERTa is shown in the top left, followed by (left-to-right, top-to-bottom) XLM-100, MuRIL, and IndicBERT. We see that AxomiyaBERTa clearly has a more defined separation between the labels, with positive/coreferent samples having higher affinity scores (accounting for the imbalanced distribution of coreferent vs. non-coreferent pairs) compared to the other models. In particular, XLM-100 shows almost identical ranges of scores for coreferent and non-coreferent pairs, with the only significant difference being the number of each kind of sample, which results in the spike around $T = -1.94$ (cf. Sec. 3.2.6).

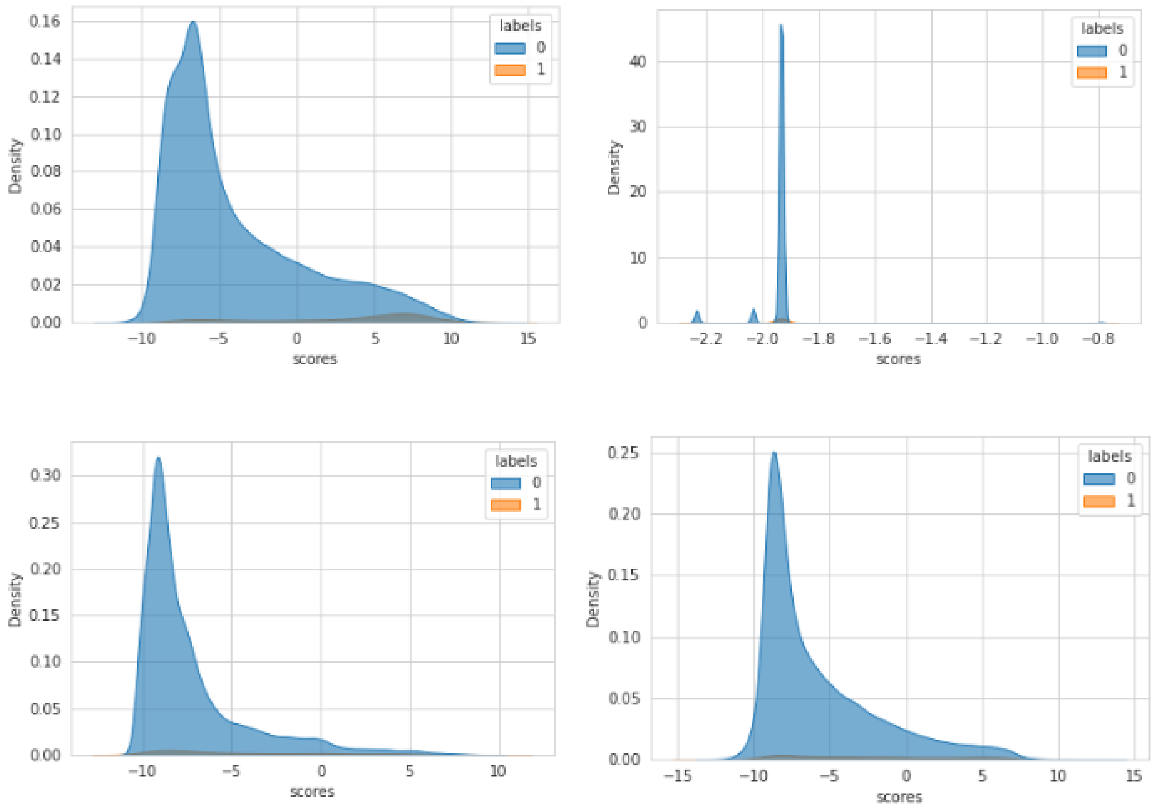

Figure 6: Kernel density estimation plots of affinity scores from the pairwise scorer for native AxomiyaBERTa compared to baselines from other Transformer-based LMs.