BanglaTLit: A Benchmark Dataset for Back-Transliteration of Romanized Bangla

Md Fahim[1,2*], **Fariha Tanjim Shifat**[1*], **Fabiha Haider**[1*], **Deeparghya Dutta Barua**[1], **Md Sakib Ul Rahman Sourove**[1], **Md Farhan Ishmam**[1,3], **Md Farhad Alam**[1]

[1]Research and Development, Penta Global Limited, Bangladesh
[2]CCDS Lab, Independent University, Bangladesh
[3]Islamic University of Technology, Bangladesh

Findings of the Association for Computational Linguistics: EMNLP 2024

October 28, 2024

Introduction

Definition

- **Romanized/Transliterated Bangla:** Uses phonetically similar Latin scripts to represent Bangla syllables.
- **Back-transliteration:** The task of generating the native scripts corresponding to the romanized text while closely aligning to the phonetic meaning.

Challenges in processing Romanized Bangla -

- Due to the phonemic orthography of Bangla, the same sentence can have multiple transliterations but must back-transliterate to the same original sentence.
- Back-transliteration must adhere to the grammatical rules of the native language.
- Current Language Models (LMs) are trained on limited transliterated Bangla texts.
- No existing large-scale back-transliteration corpus in Bangla to train LMs.

Contributions

- **BanglaTLit-PT**: A large-scale pre-training corpus comprising 245.7k romanized Bangla samples to enhance the contextualized representation in language models.
- **BanglaTLit**: 42.7k romanized Bangla and corresponding Bangla samples to fine-tune language models on the task of Bangla back-transliteration.
- **Transliterated Bangla Encoders:** We further pre-trained encoders on BanglaTLit-PT and achieved SOTA performance on sentiment analysis, emotion classification, and hate speech detection in romanized Bangla.
- **Dual-Encoder-Decoder:** We aggregate TB-encoder and T5-encoder embeddings to produce enhanced romanized Bangla representation, achieving SOTA on the BanglaTLit dataset.

Dataset Creation

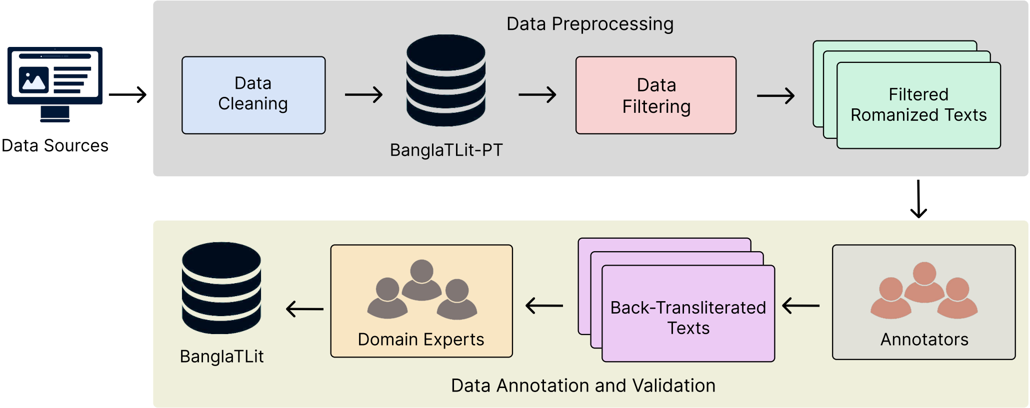

Figure: Pipeline of creating BanglaTLit-PT and BanglaTLit datasets.

Dataset Statistics

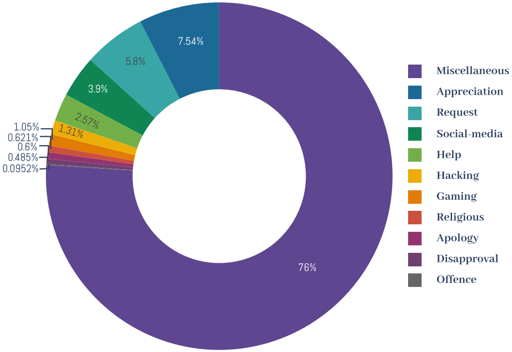

Figure: Distribution of sample categories of BanglaTLit.

Statistics	TL	BTL
Mean Character Length	59.24	58.28
Max Character Length	1406	1347
Min Character Length	3	4
Mean Word Count	10.35	10.51
Max Word Count	212	226
Min Word Count	2	2
Unique Word Count	81848	60644
Unique Sentence Count	42705	42471

Table: Statistics of the Transliterated (TL) and Back-Transliterated (BTL) sample pairs.

Methodology

Transliterated Bangla (TB) Encoder

We further pre-train using the Masked Language Modeling loss. For the dataset distribution $\mathcal{D}$, the sentence $S \sim \mathcal{D}$, $S = \{t_1, ..., t_T\}$, mask indices $m \in \mathbb{N}^M$, and training parameters θ, the negative log-likelihood objective is defined as,

$$\mathcal{L}_{MLM}(\theta) = -\mathbb{E}_{S\sim\mathcal{D}}\left[\log P_\theta(t_m|t_{\setminus m})\right].$$

TB Encoder Aggregated T5 Model

For a given text S, we obtain representations from the T5 and TB encoders

$$\mathrm{T5}(S) = \mathbf{h} = \{h_1, h_2, \ldots, h_n\}, \text{ and } \mathrm{TB}(S) = \mathbf{e} = \{e_1, e_2, \ldots, e_n\}.$$

The representations **h** and **e** are then aggregated using either sum-based aggregation, $\mathbf{H}_{\mathrm{sum}} = \mathbf{h} + \mathbf{e}$, or concatenation-based aggregation, $\mathbf{H}_{\mathrm{concat}} = [\mathbf{h}; \mathbf{e}]$ and passed to the T5 decoder to produce the corresponding Bangla text.

Architecture & Performance Evaluation

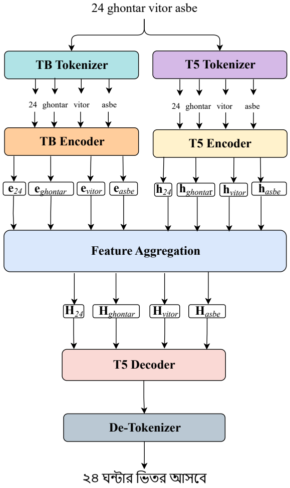

Figure: Model Architecture

Model	Performance Metric					
	TB-Sent		TB-OLID		TB-Emotion	
	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑
Bangla LM						
BanglishBERT	84.23	84.11	73.40	72.27	45.50	44.54
BanglaBERT	85.38	85.33	76.30	75.06	50.25	48.89
SahajBERT	76.54	76.54	71.57	70.29	39.75	38.79
Vac-BERT	78.85	78.78	68.12	67.36	35.00	33.62
Multilingual LM						
XLM-RoBERTa	83.85	83.84	73.40	71.57	43.50	41.15
mDeBERTa-v3	80.38	80.37	67.80	67.74	34.25	32.94
mBERT	81.15	81.03	72.80	70.89	43.50	43.45
Character-based LM						
CharBERT	84.23	84.21	74.00	73.42	46.00	43.90
CharRoBERTa	84.23	84.08	71.90	69.30	40.50	39.15
Prompt-based LLM (0-shot)						
GPT 3.5 Turbo	85.39	85.38	71.80	70.96	40.62	37.24
LLaMa3-8B	69.62	69.61	56.00	55.96	21.74	10.55
TB Encoder (Ours)						
TB-BERT	84.23	84.13	74.50	74.29	49.25	48.89
TB-BanglaBERT	85.00	84.92	77.90	76.54	52.00	50.26
TB-BanglishBERT	86.15	86.07	74.40	73.58	51.25	51.08
TB-mBERT	85.77	85.72	76.30	75.52	50.25	48.85
TB-XLM-R	**88.85**	**88.79**	**78.50**	**77.76**	**54.50**	**53.40**

Table: Performance of our baselines on romanized Bangla classification tasks.

Performance Evaluation

Model	ROUGE Score			BLEU Score			BERT Score (F1)	METEOR Score
	R-1	R-2	R-L	BLEU	Brevity Penalty	Length Ratio		
Encoder-Decoder LM								
mT5	56.02	19.83	55.90	12.48	76.13	0.82	86.43	48.71
byteT5	15.40	1.71	14.91	6.8e-5	11.28	0.25	72.50	6.88
BanglaT5-small	39.59	8.46	39.58	4.14	84.29	0.94	80.65	32.72
BanglaT5	73.06	33.00	73.13	31.09	91.16	0.95	92.71	69.12
BanglaT5_nmt_en_bn	75.74	34.84	76.14	36.19	**98.71**	1.08	94.05	74.07
Prompt-based LLM								
GPT-3.5 Turbo (0-shot)	66.21	26.18	66.64	20.73	97.94	1.11	90.06	59.97
GPT-4 Turbo (0-shot)	71.71	31.54	71.96	26.56	97.27	1.07	91.65	65.10
GPT-4o (0-shot)	66.62	26.96	67.24	19.28	98.22	1.11	89.37	58.88
LLaMa3-8B (3-shot)	56.05	17.34	56.56	11.01	95.80	1.04	86.61	46.81
Dual Encoder-Decoder LM (Ours)								
TB-BanglishBERT + BanglaT5	75.14	34.65	75.13	32.82	92.25	0.96	93.83	72.34
TB-BanglishBERT + BanglaT5_NMT	77.27	35.98	78.32	35.18	96.58	**0.97**	98.22	75.37
TB-XLM_R + BanglaT5	76.03	35.14	76.24	33.18	95.16	0.96	94.15	74.42
TB-XLM_R + BanglaT5_NMT	**78.92**	**36.56**	**79.75**	**36.07**	98.29	1.05	**98.82**	**78.14**

Table: Results of our baselines on the BanglaTLit test set.

Result Analysis

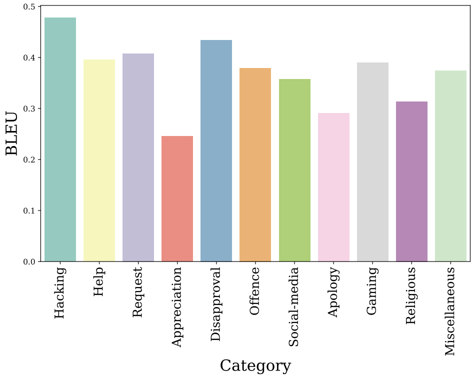

Figure: Category-wise BLEU scores for the predictions on the test set using the TB-XLM_R+BanglaT5_NMT model.

- Non-uniform BLEU score distribution across categories.
- Demonstrate strong performance in the `Hacking`, `Request`, `Help`, and `Disapproval` categories
- Struggles with the `Appreciation`, `Apology`, and `Religious` categories.
- Independent to the distribution shown in Figure 2.

Conclusion

Summary

- Transliterated Bangla Pre-training Corpus
- Back-Transliteration Dataset
- Further Pre-trained Encoders
- Dual-Encoder Decoder `seq2seq` Models

Limitations

- The majority of the BanglaTLit dataset is sourced from a single domain, i.e. tech support comments from TrickBd.
- Lack of dialect representation in the dataset.

Thank You!