Punctuation Prediction in Bangla language using Transformer Models

1 Introduction:

Bangla is the world's 7th highest spoken language in the world , but still it's one of the most underexplored languages in the natural language processing field.In every language including Bangla punctuation is a crucial element in conveying meaning and clarity in written language, and its absence or incorrect use can lead to ambiguity, confusion, and misinterpretation of the text [9]. While punctuation prediction has been tackled for high-resource languages like English, it is still an area for Bangla where only two reported works are found [1],[8].

Punctuation prediction in natural language processing requires understanding the semantic and syntactic structure of the text to accurately predict the appropriate punctuation marks. A variety of techniques have been proposed for punctuation prediction, including rule-based methods, statistical models [10] and machine-learning approaches[1], [8].

Based on previous studies, transformer model, a type of deep learning model, having self attention mechanism, skip connections and layer normalization, is more efficient for nlp task and previous study [1] on bangla punctuation prediction works well only for news dataset and comma prediction is not working as good as period or question mark prediction. So our aim is to use transformer models like BERT, ALBERT, and RoBERTa to address these problems in Bangla and create a new dataset for public use.

2 Background study:

Our study focuses on exploring transformer models for low-resource languages, specifically Bangla, and proposing a technique to improve punctuation restoration in Bangla texts. As there is limited research and resources available for Bangla punctuation restoration, our work is expected to make a significant contribution to the field.

Pre-trained language models such as BERT have also been used for punctuation restoration[1],[3]. These models are trained on large amounts of text and can capture the semantic and syntactic properties of language. By treating punctuation restoration as a sequence tagging task, where the model predicts the correct punctuation mark for each word in the input text, pre-trained language models have been able to achieve state-of-the-art performance on this task.

Even more recently, transformer-based approaches with pre-trained word embeddings have achieved state-of-the-art performance[6], [5]. For example, some researchers have used pre-trained BERT to restore punctuation, and others have studied various transformer architectures for PR and used an augmentation strategy that makes the models more robust to errors.

In addition to these model structures, multi-task learning has also been utilized to improve the performance of punctuation restoration. Multi-task learning involves training a model to perform multiple related tasks simultaneously, which can improve its performance on each individual task. For example, sentence boundary detection and capitalization recovery are two related tasks that can be used to improve the performance of punctuation restoration.

In this research, we aim to enhance NLP capabilities for the Bangla language through transformer-based models for punctuation prediction. By conducting experiments and fine-tuning transformer models, we aim to contribute to the growing field of NLP for low-resource languages.

3 Research Objectives:

The primary objectives of this research are as follows:

(a) To evaluate the performance of various transformer-based models for punctuation prediction in Bangla.

(b) To develop and fine-tune transformer models specifically for Bangla punctuation prediction.

(c) To assess the impact of model size, training data, and architecture on punctuation prediction accuracy.

(d) To contribute to the advancement of NLP techniques for low-resource languages.

4 Methodology:

To achieve the research objectives, the following methodology will be employed:

4.1 Data Collection and Preparation

(a) Gather a substantial corpus of Bangla text containing unpunctuated sentences.
(b) Annotate this corpus with correct punctuation marks to create a labeled dataset.
(c) Preprocess the data, including tokenization, normalization, and data splitting.

4.2 Model Selection and Fine-tuning

(a) Experiment with various transformer-based models, including BERT [2], T5 [7], RoBERTa [6], and others.
(b) Fine-tune these models on the Bangla punctuation prediction task using the annotated dataset.

4.3 Evaluation

(a) Evaluate model performance using standard NLP metrics like precision, recall, F1-score, and accuracy.
(b) Explore the impact of model size, training data size, and architecture on performance.

4.4 Analysis and Interpretation

(a) Analyze the results to identify strengths and weaknesses of each model.
(b) Interpret the findings to understand which models are most effective for Bangla punctuation prediction.

5 Expected outcomes:

This research is expected to make several contributions:

(a) A comparative analysis of transformer models for punctuation prediction in the Bangla language.
(b) Fine-tuned transformer models optimized for Bangla punctuation prediction.
(c) Insights into the effect of model size, training data, and architecture on performance.
(d) Advancements in NLP techniques for low-resource languages like Bangla [4].

References

[1] Tanvirul Alam, Akib Khan, and Firoj Alam. Punctuation restoration using transformer models for high-and low-resource languages. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 132–142, 2020.

[2] Llinet Benavides Cesar, Miguel-Ángel Manso-Callejo, and Calimanut-Ionut Cira. BERT (Bidirectional Encoder Representations from Transformers) for Missing Data Imputation in Solar Irradiance Time Series. *Engineering Proceedings*, 39(1):26, 2023.

[3] William Gale and Sarangarajan Parthasarathy. Experiments in character-level neural network models for punctuation. In *INTERSPEECH*, pages 2794–2798, 2017.

[4] Anna Glazkova, Michael Kadantsev, and Maksim Glazkov. Fine-tuning of pre-trained transformers for hate, offensive, and profane content detection in english and marathi. *arXiv preprint arXiv:2110.12687*, 2021.

[5] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

[6] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[7] Aviv Melamud and Alina Duran. Punctuation Restoration for Speech Transcripts using seq2seq Transformers. *Journal of Student Research*, 10(4), 2021.

[8] Habibur Rahman, Md Rezwan Shahrior Rahin, Araf Mohammad Mahbub, Md Adnanul Islam, Md Saddam Hossain Mukta, and Md Mahbubur Rahman. Punctuation prediction in bangla text. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(3):1–20, 2023.

[9] Jiangyan Yi and Jianhua Tao. Self-attention based model for punctuation prediction using word and speech embeddings. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7270–7274. IEEE, 2019.

[10] Yunxiang Zhang and Zhuyi Rao. n-bilstm: Bilstm with n-gram features for text classification. In *2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC)*, pages 1056–1059. IEEE, 2020.