Exploring Transformer Models for Punctuation Prediction in Bangla Language

Ashiqul Hasan Shakil, Nahid Hasan Lovon, Raju Singha

September 18, 2023

1 Introduction

Natural Language Processing (NLP) has made significant strides in recent years, primarily driven by advancements in transformer-based models [6].However, these models are primarily developed for English and other widely spoken languages, leaving low-resource languages like Bangla with limited NLP capabilities. Punctuation prediction is an essential NLP task that aids in text understanding and readability [4] . This thesis aims to bridge this gap by investigating the application of transformer models for punctuation prediction in the Bangla language.

2 Research Objectives

The primary objectives of this research are as follows:

1. To evaluate the performance of various transformer-based models for punctuation prediction in Bangla.
2. To develop and fine-tune transformer models specifically for Bangla punctuation prediction.
3. To assess the impact of model size, training data, and architecture on punctuation prediction accuracy.
4. To contribute to the advancement of NLP techniques for low-resource languages.

3 Methodology

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

3.1 Data Collection and Preparation

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

3.2 Model Selection and Fine-tuning

- Experiment with various transformer-based models, including BERT [1], T5 [5], RoBERTa [3], and others.
- Fine-tune these models on the Bangla punctuation prediction task using the annotated dataset.

3.3 Evaluation

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

3.4 Analysis and Interpretation

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

4 Expected Contributions

This research is expected to make several contributions:

- A comparative analysis of transformer models for punctuation prediction in the Bangla language.
- Fine-tuned transformer models optimized for Bangla punctuation prediction.
- Insights into the effect of model size, training data, and architecture on performance.
- Advancements in NLP techniques for low-resource languages like Bangla [2].

5 Timeline

- Data Collection and Annotation: Months 1-2
- Model Experimentation and Fine-tuning: Months 3-4
- Evaluation and Analysis: Months 5-6
- Thesis Writing and Finalization: Month 6

6 Conclusion

This thesis proposal outlines a research project focused on enhancing 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.

References

[1] 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.

[2] 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.

[3] 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.

[4] Karan Makhija, Thi-Nga Ho, and Eng-Siong Chng. Transfer learning for punctuation prediction. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 268–273. IEEE, 2019.

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

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.