Project Report: Bengali to IPA Translation

Executive Summary

In this project, we undertook the development of a sophisticated Bengali to International Phonetic Alphabet (IPA) translation model. Our primary objective was to establish a seamless bridge between the intricate Bengali language and the standardized IPA notation, rendering it accessible for a diverse range of applications in linguistics, language learning, and phonetic research.

Methodology

Data Preprocessing

Our endeavor commenced with a rigorous data preprocessing phase. Notable steps included:

- **URL Removal:** To ensure the purity of the linguistic data, we meticulously removed any URLs embedded within the text.
- **Bengali Numeral Conversion:**We implemented a transformation mechanism to convert Bengali numerals into their word equivalents, enhancing readability and phonetic accuracy.
- **Elimination of English Alphanumeric Words:** To prevent interference from extraneous text, we systematically eliminated English alphanumeric content.

The culmination of these efforts yielded a pristine dataset that formed the foundation for subsequent model development.

Model Selection

Our project unfolded in a journey of model exploration, aimed at delivering optimal translation performance. Noteworthy milestones include:

1. **LSTM Sequence-to-Sequence Model:** Our initial foray involved the implementation of raw LSTM sequence-to-sequence models, complemented by SentencePiece tokenization and Byte-Pair Encoding (BPE). While showing promise, the constraints of dataset scale limited the model's effectiveness.
2. **mT5-small:** The mT5-small model emerged as a viable alternative, exhibiting favorable translation accuracy. However, a drawback was observed in the form of slower training times.
3. **mT5-base:** The project culminated with the adoption of the mT5-base model. This choice not only resolved the training speed impediment but also elevated translation quality to a superior standard.

The Cleaner Class

A pivotal contribution to our project was the introduction of the `Cleaner` class, a sophisticated tool that simplified the data preprocessing pipeline. This class, meticulously designed and implemented, ensures the adherence of Bengali text to IPA standards, enhancing data consistency and model performance.

Conclusion

Our project represents the successful culmination of a comprehensive endeavor to facilitate Bengali to IPA translation. It introduces a bridge between two distinct linguistic realms, enabling accurate phonetic transcription and expanding the frontiers of linguistic analysis, language education, and phonetic research.

We anticipate this project to serve as a foundation for further advancements, offering a deeper understanding of language nuances and contributing to linguistic excellence.