Research Proposal

Title: Large Language Model for Telugu Language
Project Supervisor: Professor Mayank Singh, Computer Science and Engineering, IIT Gandhinagar
This report was submitted on 9/5/2024

Abstract:
This research proposal outlines a comprehensive plan to develop an advanced Telugu Language Model (TLM) aimed at enhancing natural language processing tasks in the Telugu language domain. Leveraging existing resources and state-of-the-art deep learning techniques, the project seeks to address the need for robust and contextually relevant language models tailored to Telugu, thereby facilitating a wide range of applications in information retrieval, text generation, question answering, and more.

Problem Statement:
The project aims to develop a Telugu Language Model (TLM) to enhance natural language processing tasks in Telugu. Despite a growing demand for language technologies in Indic languages like Telugu, existing models lack depth and sophistication, limiting their effectiveness. This project seeks to address these limitations by designing a more robust TLM capable of understanding and generating Telugu text.

Proposed Strategy:

Data Collection and Preprocessing:
Gather Diverse Telugu Text Corpora: Collect a diverse range of Telugu text data from sources such as literature, news articles, social media, and domain-specific documents.

Data Cleaning and Annotation: The collected data undergoes thorough preprocessing to mitigate noise, inconsistencies, and missing annotations. Techniques, including deduplication, are employed to eliminate duplicate instances and ensure data integrity.

Model Development and Evaluation:
Architecture Selection: An appropriate deep learning architecture is chosen for language modelling tasks, taking into account factors such as model size, computational efficiency, and performance. The selection process involves evaluating various architectures' suitability for handling Telugu language intricacies and scalability to large datasets.

Training and Validation: The chosen model architecture is trained using the annotated Telugu text data. Rigorous validation techniques are employed to assess the model's performance and generalisation capabilities. The training process involves optimising model parameters to minimise loss and maximise predictive accuracy on the validation set.

Testing with Benchmarks: The trained model is tested against established benchmarks and existing Telugu language models to gauge its performance and effectiveness. Benchmark tests provide valuable insights into the model's comparative performance and highlight areas for improvement.

Fine-Tuning and Optimization:

Iterative Refinement: Iteratively refine the model architecture and hyperparameters based on validation results and feedback from Professor.

Optimisation Strategies: Implement optimisation techniques to improve model efficiency.

Overview of Work Completed:

The project's initial phase involved a comprehensive review of existing Telugu Language Models, datasets, and research literature. Exploration of available resources, including models such as Chandamama Kathalu and Llama-3-8b-Telugu_Romanized, provided valuable insights into the current state of Telugu language processing technologies.

Acknowledgements:

I extend my sincere gratitude to Prof. Mayank Singh for his continuous support, guidance, and encouragement during the initial phase of this project. I would also like to acknowledge M.Tech student Hitesh Lodwal for his valuable contributions and assistance, which was pivotal in shaping the early stages of this ongoing endeavour.