PraTranV3: Prakrit Language Translator – Detailed Synopsis Document

Vishwakarma Institute of Information Technology, Pune

Date: 06-02-2025

1. Introduction

Background & Motivation:

- Prakrit Languages: A group of ancient Middle Indo-Aryan languages used extensively in Jain, Buddhist, and classical Indian literature.
- Historical Significance: Ardha Magadhi Prakrit, the focus of this project, is the language of early Jain and Buddhist scriptures and is vital for preserving cultural heritage.
- Project Evolution: Building on our previous work on Maharashtri Prakrit-to-English translation, PraTranV3 extends these techniques to Ardha Magadhi Prakrit, aiming to preserve another important facet of our linguistic heritage.

2. Literature Review & Gap Analysis

Review of Prior Work:

- Neural Machine Translation for Low-Resource Languages: Recent surveys (e.g., Ranathunga et al., 2023) discuss techniques like transfer learning and multilingual NMT systems, though few address ancient languages such as Ardha Magadhi Prakrit.
- LLM vs. Traditional NMT Approaches: Studies (e.g., Enis & Hopkins, 2024) indicate that while large language models provide fluent outputs, they require domain-specific fine-tuning to handle the complexities of ancient languages.
- Previous Project Outcome: Our earlier project “Bridging the Past: Neural Machine Translation of Maharashtri Prakrit to English” utilized Facebook’s M2M100 model, achieving a BLEU score of ~15.34 and a METEOR score of ~0.47 despite challenges in data scarcity and OCR errors.

Gap Analysis:

- Data Scarcity: Ardha Magadhi Prakrit lacks digitized and annotated parallel corpora.
- Model Adaptation: Existing pre-trained models have not been fine-tuned to account for the linguistic nuances of ancient Prakrit languages.
- OCR Challenges: Automated OCR (e.g., Google Vision API) often misinterprets the Devanagari script by converting it into English, necessitating manual corrections.

3. Objectives

- Primary Aim: Develop a machine translation system to convert Ardha Magadhi Prakrit texts to English accurately and efficiently.
- Secondary Objectives:

    - Dataset Creation: Extract and preprocess texts from historical documents via OCR, in collaboration with the Bhandarkar Institute, with subsequent manual correction.
    - Model Fine-Tuning: Adapt state-of-the-art pre-trained models (LLAMA, Claude, NLLB, M2M100) using quantization techniques such as LoRA and QLoRA.
    - Deployment: Create a user-friendly web application to provide public access to the translation tool.

4. Methodology

Data Collection & Preprocessing:

- Data Sources: Collaborate with the Bhandarkar Institute to acquire ancient manuscripts and digitized texts.
- OCR Extraction: Utilize the Google Vision API to extract text from scanned documents (PDFs, images).
- Cleaning & Correction: Using both automated tools and manual intervention, establish a robust preprocessing pipeline to address OCR errors, particularly the misinterpretation of Devanagari text.

Model Fine-Tuning Strategy:

- Pre-trained Models: Leverage multilingual models such as LLAMA, Claude, NLLB, and M2M100.
- Optimization Techniques:
    - Employ LoRA/QLoRA for resource-efficient fine-tuning.
    - Optimize hyperparameters (learning rate, batch size, number of epochs) based on pilot experiments.
- Evaluation:
    - Use performance metrics like BLEU, METEOR, and chrF++ to assess translation quality.
    - Split data into training (80%) and validation (20%) sets for iterative refinement.

Deployment via Web Application:

- Interface Design: Develop an intuitive UI for text input or document upload.
- Backend Integration: Seamlessly integrate the fine-tuned model with the web app for real-time translation.
- Accessibility: Ensure compatibility across various devices (desktop, mobile) to maximize reach.

5. Challenges & Mitigation Strategies

Low-Resource Data:

- Challenge: Limited digital resources for Ardha Magadhi Prakrit.
- Mitigation: Partner with academic institutions and manually annotate additional texts to expand the corpus.

OCR Accuracy:

- Challenge: Inconsistent extraction results from mixed-script documents, particularly misinterpretation of Devanagari text.
- Mitigation: Enhance the OCR pipeline with manual quality checks and, if feasible, develop custom OCR solutions tailored for historical scripts.

Model Adaptation:

- Challenge: Pre-trained models may not fully capture the grammatical complexities of ancient Prakrit languages.
- Mitigation: Fine-tune models with domain-specific data and leverage quantization techniques (LoRA, QLoRA) to enhance performance while managing computational demands.

Computational Efficiency:

- Challenge: High computational cost associated with fine-tuning large models.
- Mitigation: Utilize optimized training frameworks and quantization methods to reduce resource requirements.

6. Expected Outcomes & Contributions

- Translation Tool: A functional Ardha Magadhi Prakrit-to-English translator that serves researchers, historians, linguists, and the general public.
- Enhanced NLP Research: Advances in techniques for low-resource language processing and fine-tuning of LLMs for ancient languages.
- Cultural Preservation: Digital preservation of Ardha Magadhi Prakrit, ensuring broader access to historical texts and cultural heritage.
- Scalable Framework: A blueprint for adapting similar methodologies to other low-resource or ancient languages.

7. Project Timeline & Feasibility

Timeline:

- Phase 1 (Data Collection & Preprocessing): 2-3 months
- Phase 2 (Model Fine-Tuning & Evaluation): 3-4 months
- Phase 3 (Web Application Development & Deployment): 2 months
- Phase 4 (Testing & Iteration): 1-2 months

Feasibility:

- Data & Collaboration: Secure datasets from reputable sources (e.g., Bhandarkar Institute) and expand through manual annotation.
- Technical Resources: Use state-of-the-art hardware and optimized fine-tuning frameworks to manage computational challenges.
- Deployment: The web application ensures wide accessibility, making the project both technically viable and socially impactful.

8. Conclusion & Future Directions

Conclusion:
PraTranV3 seeks to bridge the technological gap in translating ancient languages by applying modern NLP techniques to a low-resource, culturally significant language. By developing a dedicated translation system for Ardha Magadhi Prakrit, the project preserves valuable heritage and advances machine translation methodologies for underrepresented languages.

Future Directions:

- Dataset Expansion: Continuously incorporate additional historical sources to enrich the parallel corpus.
- Model Enhancements: Explore multimodal approaches that integrate textual data with contextual information (e.g., images of manuscripts or cultural artifacts).
- Broader Applications: Extend the framework to other ancient languages, fostering broader digital preservation initiatives.