Low-Resource NLP (Indian Language LLMs & Translation Models)

DL Mini Project

Chandan Kumar (2024PCS0022)
Rohit Tiwari (2024PCS0036)

March 22, 2025

Overview

1. Introduction
2. Motivation
3. Problem Statement
4. Methodology
5. Work Plan
6. Expected Outcomes

Introduction

- Indian languages are morphologically rich, leading to suboptimal performance using conventional BPE tokenization. Our project aims to build a full-fledged NLP system using a morpheme-based tokenizer for better results.

Motivation

- Multilingual India: 22 official languages, hundreds of dialects.
- Low-Resource Challenge: Limited data and resources.
- Performance Gap: Existing systems rely on BPE, which is suboptimal.

Problem Statement

Develop a Low-Resource NLP System that includes a morpheme-based tokenizer, optimized LLM, and translation model for Indian languages.

Methodology

- **Data Collection:** Collect datasets using resources like AI4Bharat, Samanantar, and IndicNLP.
- **Morpheme-Based Tokenization:** Implement a tokenization method that uses linguistic morphology instead of BPE.
- **LLM Training:** Fine-tune or train a transformer-based LLM using the new tokenizer.
- **Translation Model:** Develop a machine translation model and apply the tokenizer.
- **Evaluation:** Perform model evaluation using BLEU, ROUGE, and perplexity metrics.

Work Plan

Task
Literature Review Problem Definition
Data Collection Preprocessing
Tokenizer Development
LLM Training
Translation Model Development
Evaluation Optimization

Expected Outcomes

- Improved tokenization results for Indian languages.
- A comparative analysis of BPE vs. Morpheme Tokenization.
- Enhanced LLM and translation models.

References

The End