ABSTRACT

This thesis addresses the multifaceted challenges of adapting and utilizing modern language models effectively and efficiently for low-resource languages (LRLs), with a particular focus on the diverse linguistic landscape of India.

First, we tackle the critical issue of data scarcity for supervised tasks like relation classification. The lack of trustworthy datasets hinders model fine-tuning and evaluation. We introduce IndoRE, a novel gold-standard dataset for relation classification covering Bengali, Hindi, Telugu, and English. Using this resource and multilingual models like mBERT, we investigate cost-effective training strategies, evaluating the trade-offs between manually labeled gold data and automatically generated silver data. We further propose an active learning framework to intelligently guide budget-constrained gold annotation and demonstrate the utility of ensemble models to enhance performance when gold data is limited.

Second, we address the prohibitive inference costs associated with using proprietary large language models (LLMs) for LRLs. Due to skewed vocabulary optimized for high-resource languages, LRL inputs often generate significantly more tokens, leading to higher costs. We systematically analyze pre-processing techniques – translation, transliteration, and code-mixing – as methods to reduce token counts before sending input to models like GPT-4. We evaluate the impact of these techniques across numerous Indic languages and tasks. We introduce RTPCR, a metric to assess the combined effect of cost savings and performance impact, ultimately prescribing strategies that can reduce costs by up to 90% while maintaining or improving task quality compared to using the original LRL directly.

Third, focusing on foundational multilingual language models (MLMs), we confront the problem of LRL vocabulary under-representation. Standard subword

tokenization often replaces LRL words with unknown tokens or segments them poorly, degrading downstream task accuracy. We propose EVALM, an entropy-based technique that identifies vulnerable LRL words based on their segmentation quality and provides targeted vocabulary augmentation with reasonable embedding initializations during the model adaptation phase, requiring only limited task-specific data for fine-tuning to achieve significant performance gains.

Fourth, acknowledging that even recent LLMs are often under-trained for LRLs, we address the high cost of continual pre-training (CPT) typically required to improve their LRL capabilities. We develop resource-efficient approaches, including a novel algorithm for selecting minimal yet effective data subsets from larger corpora for CPT, and a method for cost-effective vocabulary augmentation specifically designed for the CPT phase. These techniques drastically reduce the data and computational requirements for enhancing LLMs like Llama-3 for Indic languages.

Finally, we tackle the challenge of prompt optimization for LLMs. While crucial for performance, manually finding optimal prompts through trial and error is particularly expensive for LRLs due to higher inference costs and limited gold data. We introduce MutantPrompt, a framework leveraging multi-armed bandit algorithms to efficiently explore the prompt space and identify high-performing prompts under a fixed computational budget, demonstrating its cost-effectiveness across various LRL tasks and smaller LLMs.

In conclusion, this thesis presents a cohesive suite of novel datasets, techniques, and frameworks specifically designed to overcome critical bottlenecks—data scarcity, inference cost, vocabulary limitations, training expenses, and prompt optimization challenges—in applying advanced language models to low-resource languages. By emphasizing cost-effectiveness and resource efficiency, this work contributes practical solutions to enhance the performance, accessibility, and applicability of state-

of-the-art NLP in LRL contexts.

Keywords:- relation extraction, active learning, model ensemble, large language models, cost optimization, continual pre-training, prompt optimization, vocabulary augmentation, machine learning, natural language processing, dataset curation