Project Proposal: Extending ROMANSETU to Diverse Languages

Group 35

Anusha Pant Nav Sanya Anand Paul Kurian Sebastian Escalante Shreyas Malewar

University of Southern California

Abstract

This document is to provide details for Group 35's project proposal for the CSCI 544 - Applied Natural Language Processing Course Project. We aim to extend the findings of the paper "ROMANSETU: Efficiently unlocking multilingual capabilities of Large Language Models models via Romanization" [1] by Kunchukuttan et al. In this proposal, we will cover the motivation for pursuing, this project, our objectives, the proposed methodologies for achieving them, and the tentative timeline of project execution and analysis.

1 Background

Large Language Models (LLMs) have revolutionized Natural Language Processing (NLP) tasks, demonstrating remarkable proficiency in various English applications. However, extending their capabilities to non-English languages, particularly those with non-Latin scripts, remains a challenge. This project investigates the applicability and effectiveness of "ROMANSETU," a novel approach utilizing romanization to bridge the language gap and enhance LLM performance in diverse languages. The language considered in their paper was Hindi, but this approach can be applied to other languages of non-Latin script.

2 Literature Review

While performing research on this topic we looked at the following paper which also delved into the issue of translation between languages of differing resource-richness:

- The Impact of Translating Resource-Rich Datasets to Low-Resource Languages Through Multi-Lingual Text Processing [2]

This paper however looked into translation from resource-rich datasets to low-resource languages Additionally, we looked at another paper that studied the impact of romanization to boost the capabilities of LLMs on low-resource languages:

- Romanization-based Large-scale Adaptation of Multilingual Language Models [3]

3 Objectives

The primary objectives of this project are:

- **Evaluate the effectiveness of ROMANSETU:** To assess the performance improvement in LLM tasks achieved by romanized text compared to the native script for Hindi. These tasks include the ones mentioned in the ROMANSETU paper namely:
  - Machine Translation
  - Sentiment Analysis
- **Explore language-specific adjustments:** To identify potential modifications to the ROMANSETU approach necessary for optimal performance in specific languages.
- **Explore Many-to-Many NMT: Zero-shot Transfer:** To identify the performance of LLM translation between two or more non-Latin languages

4 Methodology

4.1 Language Selection

We will evaluate the effectiveness of the ROMANSETU paper by selecting Hindi and performing the tasks proposed in the paper.

Additionally, we will select from a set of four languages representing diverse script types and levels of established romanization systems:

- Arabic: Complex script with a well-defined romanization system (ISO 2332)

- Cyrillic: Relatively complex script with various romanization schemes (e.g., GOST, Library of Congress)
- Japanese: Logographic script with existing romanization systems like Hepburn and Kunrei-shiki
- Swahili: Latin script language with limited formal romanization standards

4.2 Dataset Preparation

For each selected language, we will:

- Gather a balanced dataset of text and labels for the chosen LLM task (e.g., machine translation, sentiment analysis).
- Prepare two versions of the dataset: one in the native script and another in its corresponding romanized form using established standards or a consistent transliteration scheme.

4.3 LLM Training and Evaluation

- We will train an LLM model on each language dataset (native and romanized) using a pre-trained LLM like LLaMA2 7B, BART, or T5.
- The trained models will be evaluated on the held-out test set for the chosen LLM task based on established metrics (e.g., BLEU score for translation, and F1 score for sentiment analysis).
- Performance comparisons will be made between the models trained on native and romanized data for each language.

4.4 Data Analysis and Adjustment Exploration

- We will statistically analyze the performance differences between models to assess the effectiveness of romanization across languages.
- Script complexity and the presence of established romanization systems will be analyzed as potential influencing factors on the observed results.
- We will explore potential language-specific adjustments to the ROMANSETU approach, such as:
  - Customized romanization schemes for languages with diverse script types
  - Addressing homophones and ambiguities arising from romanization
  - Investigating the use of language identification techniques to selectively apply romanization.

5 Expected Outcomes

This project is expected to contribute to the following:

- Enhanced understanding: Improved knowledge of how romanization impacts LLM performance across diverse languages.
- Identification of influencing factors: Insights into the impact of script complexity and existing romanization systems on the effectiveness of the approach.
- Language-specific adaptations: Recommendations for potential modifications to ROMANSETU for optimal performance in different language contexts.

6 Tentative Timeline

1. Language Selection *(March 1 - March 5)*
   (a) Selection of another non-Latin language to run experiments on.
   (b) Selection and familiarization with the new language's corresponding romanization tool.
2. Replication of ROMANSETU Results for Hindi *(March 6 - March 15)*
3. Dataset Preparation for the new language *(March 18 - March 22)*
4. LLM Training and evaluation for the new language *(March 25 - April 2)*
5. Analysis and preparation for presentation *(April 3 - April 7)*
6. Report Preparation *(April 8 - April 20)*

7 Conclusion

By extending ROMANSETU to diverse languages and analyzing its efficacy, this project aims to contribute to bridging the language gap in LLM applications. The findings could pave the way for more inclusive and effective LLMs that empower communication and information access across different languages and cultures.

References

1. *Husain, J. A., Dabre, R., Kumar, A., Puduppully, R., & Kunchukuttan, A. (2024). RomanSetu: Efficiently unlocking multilingual capabilities of Large Language Models models via Romanization.*

2. *A. Ghafoor et al., "The Impact of Translating Resource-Rich Datasets to Low-Resource Languages Through Multi-Lingual Text Processing," in IEEE Access, vol. 9*

3. *Purkayastha, S., Ruder, S., Pfeiffer, J., Gurevych, I., & Vulić, I. (2023). Romanization-based Large-scale Adaptation of Multilingual Language Models.*