Revitalizing Kalamang: Language Expansion through Large Language Models

Aaditya Bhargav
aaditya23006@iiitd.ac.in

Md. Abuzar Khan
abuzark@iiitd.ac.in

Mohit
mohit21542@iiitd.ac.in

Shah Jayshil Ketankumar
jayshil23138@iiitd.ac.in

Vani Mittal
vani23102@iiitd.ac.in

Vishal Singh
vishal21575@iiitd.ac.in

Abstract

The project *Revitalizing Kalamang: Language Expansion through Large Language Models* addresses the critical need to preserve and revitalize the Kalamang language, a language spoken by fewer than 200 people on a small island in Indonesian Papua. As one of the many endangered languages around the world, Kalamang is at risk of extinction due to limited usage and a lack of substantial digital resources. To confront this challenge, the project harnesses the power of Large Language Models (LLMs), which have shown remarkable success in handling low-resource languages, to develop a robust framework for translating between Kalamang and other major languages, such as English.

Furthermore, the project underscores the importance of ethical considerations and collaboration with the Kalamang-speaking community. By involving native speakers in the process, the project ensures that the revitalization efforts align with the cultural and social context of the language. This community-centric approach aims to empower Kalamang speakers to reclaim their linguistic heritage, providing them with digital tools to communicate, teach, and learn in their native language. By leveraging available resources such as grammar books, bilingual word lists, and a small corpus of Kalamang-English parallel texts, this project seeks to bridge the gap between the limited data available for Kalamang and the sophisticated capabilities of modern LLMs. The goal is to train these models to accurately capture the nuances and structures of the Kalamang language, enabling effective translation and communication despite the inherent challenges posed by low-resource languages. Ultimately, the success of this project will not only safeguard the Kalamang language but will also serve as a model for future efforts to preserve other endangered languages. By demonstrating that LLMs can adapt to low-resource settings, this research contributes to the broader goals of maintaining linguistic diversity and promoting inclusion in the digital age.

Objectives

Aim: The project aims to enhance and expand the translation capabilities of the Kalamang language by incorporating Hindi alongside English using large language models (LLMs). This approach will explore cross-linguistic learning to improve low-resource language translation and, if time allows, utilize synthetic data augmentation to further boost the model's performance.

1. **Integrate Hindi Modality:** Incorporate Hindi as an additional language modality to enhance the model's understanding and translation capabilities for Kalamang, leveraging similarities in grammar and sentence structure between the languages.
2. **Cross-Linguistic Learning:** Demonstrate how incorporating multiple languages (Hindi, English, and Kalamang) can improve language learning, translation accuracy, and generalization in low-resource language models.
3. **Synthetic Data Augmentation (Optional):** Use synthetic data generation techniques to create additional training examples, simulating varied language contexts to further enhance model performance.

Possible Applications:

1. **Language Revitalization:** Support efforts to preserve and revitalize the Kalamang language by providing more accessible translation tools for documentation and learning.

2. **Multilingual Education Tools:** Develop educational resources for language learners and linguists, allowing them to study and translate Kalamang using a model trained on Hindi, English, and synthetic data.

3. **Cross-Language Communication:** Enhance communication between Kalamang speakers and communities that use Hindi or English, providing a bridge for language exchange and cultural understanding.

4. **AI-Assisted Linguistic Research:** Facilitate linguistic research on low-resource languages by offering a novel approach that combines multilingual data and synthetic augmentation, helping researchers study linguistic similarities and differences.

Literature Review

A Benchmark for Learning to Translate a New Language from One Grammar Book

The paper uses MTOB (Machine Translation from One Book), a new framework to assess how well large language models (LLMs) learn translation between English and Arts, a rare language with fewer than 200 speakers. The MTOB dataset includes a complete grammar book (573 pages), a bilingual vocabulary with 2,531 Kalamagan words and their English translations, and a small parallel corpus of 500 Kalamagan-English sentence pairs used in ancient and test sentences. Various contexts such as grammar book excerpts and glossaries were consulted, and various methods of retrieving context were used, such as searching for or retrieving common words embeddings.

The graphics tested included versions of LLaMA and Llama 2, as well as API-based graphics such as GPT-3.5-turbo, GPT-4, and Claude 2. The results showed larger graphics and more detailed references tended to improve performance, whereas Claude 2 Given the detailed context shown by. However, human performance was significantly better than all of the images tested.

The paper notes several limitations, such as the test items being too similar to the training data and focusing on individual sentences rather than longer texts. There were also concerns about communication barriers, as all Kalamang speakers also speak other languages, as well as text-based translation and spoken language inconsistencies. The paper also highlights broader issues such as technological inaccuracies and cultural influences emphasizing possibilities in learning communities [3].

Knowledge-Augmented Reasoning Distillation for Small Language Models in Knowledge-Intensive Tasks

This paper presents a Knowledge-Augmented Reasoning Distillation (KARD) designed to enhance the performance of small language models (LMs) in knowledge-intensive reasoning tasks. This requires fine-tuning, incorporating external knowledge to produce designed reasoning has been successful. It uses existing methods such as distillation. KARD relies on large LMs to generate high-quality simulations and uses non-parametric memory to retrieve relevant documents, improving the accuracy of small LMs in complex tasks such as MedQA-USMLE.

Previous work in domain optimization and knowledge distillation has shown that theoretical capabilities can be transferred to small models through domain-specific corporate training methods such as Domain Adaptive Pre-Training (DAPT) and Retrieval-Augmented Generation (RAG) for performance improvements in specific tasks. They were able to. However, this paper argues that KARD outperforms DAPT and RAG in empirical testing and offers more effective solutions for tasks requiring theoretical and realistic accuracy by providing external knowledge.

Experiments show that KARD outperforms traditional optimization methods with much fewer training samples and subsamples. The reordering component of KARD also shows a significant improvement over BM25, especially in terms of tasks related to retrieving relevant documents for dispute purposes. Although KARD shows promise in addressing knowledge-intensive tasks, the study acknowledges limitations in sampling methods and sample size and suggests future research to extend KARD to larger LMs such as GPT-3 or LLaMA [1].

A Grammar Sketch of Kalamang with a Focus on Phonetics and Phonology

Eline Visser's *A Grammar Sketch of Kalamang* provides essential insights into the language's phonology and syntax, which are crucial for developing our translation models. The thesis highlights key phonetic features like nasal assimilation and vowel harmony, essential for accurate phonetic representation. Moreover, Kalamang's subject-object-verb (SOV) word order and postpositional syntax closely align with Hindi, which offers a natural avenue for cross-linguistic learning in our project.

The thesis also addresses the issue of limited linguistic data available for Kalamang, which reinforces the importance of synthetic data augmentation in our approach. By leveraging Visser's work, including a lexicon and audio recordings, we can generate synthetic data to improve the performance of our translation models. This is particularly important given Kalamang's endangered status and the scarcity of available resources.

Additionally, incorporating Hindi as a bridge language not only enhances translation capabilities but also supports language preservation. With only a few hundred speakers remaining, our project can contribute to revitalization efforts by making the language more accessible to bilingual speakers in the region while simultaneously extending the reach of linguistic documentation [4].

Teaching Large Language Models an Unseen Language on the Fly

Chen Zhang et al. (2024) investigate how large language models (LLMs) can learn entirely new languages using the DIPMT++ framework, focusing on Zhuang, a low-resource language. Their study shows that while LLMs can quickly grasp basic aspects of an unseen language, challenges remain in understanding complex syntactic and morphological features. They suggest enhancing data collection and retrieval strategies to improve performance and propose future research to develop more sophisticated methods for low-resource languages.

The research also highlights the practical application of LLMs in human translation. Their user study demonstrates that LLMs can improve translation quality and efficiency when used alongside human translators, showing that initial LLM translations can boost human performance and reduce translation time.

Furthermore, the paper discusses the broader applications of these techniques for language education and preservation. It acknowledges limitations such as the small evaluation scale and the similarity between Zhuang and Chinese, which may not fully reflect the difficulties posed by more distinct languages [5].

Tuning LLMs with Contrastive Alignment Instructions for Machine Translation in Unseen, Low-Resource Languages

Recent advancements in multilingual large language models (LLMs) have focused on improving machine translation (MT) capabilities for low-resource languages. Chen Zhang et al. (2024) introduce the DIPMT++ framework, which aims to enhance the adaptation of LLMs to new languages with limited resources. However, computational constraints have prevented experimentation with larger models like the 175B BLOOMZ, which could offer additional insights into the efficacy of these techniques.

The study examines various parameter-efficient fine-tuning (PEFT) methods, including LoRA and alternatives such as (IA)3 (Liu et al., 2022). These methods show potential for improving MT adaptation performance. The research also highlights the need to explore the mixture of experts (MoE) approach, which could offer further benefits for MT tasks. Despite the promising results with fixed templates for instruction fine-tuning, experimenting with varied templates and utilizing large monolingual corpora could enhance LLM performance in low-resource languages.

Instruction fine-tuning in the study was limited to using fixed templates, but future work might benefit from diverse templates or large-scale instruction datasets like xP3 (Muennighoff et al., 2023). Additionally, the scalability of high-quality multilingual instruction datasets to numerous low-resource languages remains an area for exploration. The study did not compare the proposed methods with state-of-the-art multilingual NMT models such as NLLB-200 (Costa-jussà et al., 2022). Integrating these methods with contemporary MT fine-tuning paradigms, as proposed by Xu et al. (2023a), could potentially elevate the quality of translations produced by LLMs [2].

Methodology

1. **Data Collection**
    - **Linguistic Reference Materials:** Leveraging available resources such as:
        - Grammar books
        - Dictionaries
        - Folklore texts
        - Bilingual texts
2. **Data Preparation**
    - **Data Cleaning:** Process the collected data to remove inconsistencies, errors, and irrelevant information. This includes normalizing text formats and ensuring proper encoding.
    - **Data Annotation:** Annotate the data for specific linguistic features, such as part-of-speech tagging and syntactic structures, to enhance the training dataset's quality.
3. **Model Selection**
    - **Pretrained Language Models:** Select suitable pretrained language models (e.g., GPT, BERT) that can be fine-tuned for the translation task. The choice of model will depend on the specific characteristics of Kalamang language and the availability of resources.
4. **Model Training**
    - **Fine-Tuning:** Fine-tune the selected pretrained models on the prepared dataset. This involves:
        - Splitting the dataset into training, validation, and test sets to ensure robust evaluation.
        - Using transfer learning techniques to adapt the model to the specific linguistic features of Kalamang language.
    - **Training Parameters:** Set appropriate hyperparameters (e.g., learning rate, batch size, number of epochs) based on preliminary experiments to optimize model performance.
5. **Quantitative Evaluation:** Assess the model's performance using standard metrics such as:
    - BLEU (Bilingual Evaluation Understudy)
    - chrF (Character F-score)
6. **Qualitative Evaluation:** Conduct user studies with native speakers of Kalamang language to gather feedback on translation quality and usability. This may involve:
    - Comparing model outputs to human translations.
    - Collecting subjective ratings on fluency and adequacy from community members.
7. **Iterative Improvement**
    - **Feedback Loop:** Incorporate feedback from community evaluations to iteratively improve the model. This may involve:
        - Adjusting training data based on community input.
        - Retraining the model with additional data or modified parameters.
8. **Deployment**
    - **User-Friendly Interface:** Develop a user-friendly interface for community members to access the translation system. This could include:
        - A web-based application or mobile app that allows users to input text and receive translations.

Timelines

Phase 1: Project Planning and Data Preparation

- Define project scope and objectives with a focus on incorporating Hindi.
- Collect and preprocess data for Kalamang, Hindi, and English (grammar books, bilingual word lists).
- Set up the project environment (model setup, data processing scripts).

Milestone: Complete data preparation for Hindi and English modalities.

Phase 2: Model Integration and Initial Experiments

- Integrate Hindi modality into the model; adjust model to handle inputs from Kalamang, Hindi, and English.
- Conduct initial testing with basic Hindi integration to verify that the model processes Hindi correctly.
- Evaluate the model's initial performance with Hindi alongside Kalamang.

Milestone: Successfully incorporate Hindi modality with initial performance evaluation.

Phase 3: Fine-Tuning and Multilingual Testing

- Fine-tune the model using Hindi, English, and Kalamang data; refine the integration for better performance.
- Perform multilingual testing to assess the quality of translations involving Hindi and Kalamang.
- Identify gaps and areas where further fine-tuning is needed.

Milestone: Fine-tune model with Hindi and assess its effectiveness in multilingual translation tasks.

Phase 4: Evaluation and Incorporation of Synthetic Data Augmentation

- Evaluate current results; if satisfactory, begin implementing synthetic data augmentation for further improvement.
- Develop scripts to generate synthetic examples, focusing on blending elements from all languages.
- Fine-tune the model again with augmented data and evaluate the impact.

Milestone: Complete synthetic data augmentation integration and assess improvements.

Phase 5: Final Evaluation, Reporting, and Submission Preparation

- Conduct a detailed final evaluation; prepare a comprehensive analysis of model performance.
- Finalize the project report, including methodology, results, and insights from the synthetic augmentation (if completed).
- Prepare all deliverables for submission.

Milestone: Finalize the project report and submit.

Key Considerations for Timelines:

- **Time Management:** Focus on integrating Hindi first. Only proceed to synthetic data augmentation if earlier phases are completed on time.
- **Flexible Milestones:** If synthetic data is not feasible, use the remaining time to refine the Hindi modality and improve model performance.
- **Buffer Week:** Use last 10 days to address any pending tasks, ensuring a polished final submission.

Evaluation Criteria

- **Quantitative Evaluation:** The model's performance will be assessed using the following standard metrics:

  - **chrF Score:** This metric calculates the character-level F-score (chrF) by comparing the overlap of character n-grams between the model output and the reference translation. It will be used for evaluating translations in both directions:

    * Kalamang to English (kgv $\rightarrow$ eng)
    * English to Kalamang (eng $\rightarrow$ kgv)

    The chrF score is computed as the harmonic mean of precision and recall of character n-grams:

    * chrF $= (1 + \beta^2) \frac{P_{chr} \times R_{chr}}{\beta^2 \times P_{chr} + R_{chr}}$
      $Where:$
      · P_{chr} is the precision of character n-grams,
      · R_{chr} is the recall of character n-grams,
      · β is a weighting factor (usually set to 1 to give equal importance to precision and recall).

  - **No Context (-):** This measures the model's zero-shot translation ability when no reference materials or context are provided.

  - **Grammar Book Context (G):** This evaluates the model's performance when provided with passages from a grammar book. The evaluation will use:

    * **Cosine Similarity (Ge):** To measure how similar the model's output is to reference sections.
    * **Longest Common Substring Distance (Gs):** To find similar sections between the model output and reference passages.
    * **Extended Sections:** Some models will be tested with larger grammar book sections, consisting of 50K or 100K tokens.

References

[1] Minki Kang, Seanie Lee, Jinheon Baek, Kenji Kawaguchi, and Sung Ju Hwang. Knowledge-augmented reasoning distillation for small language models in knowledge-intensive tasks. *Advances in Neural Information Processing Systems*, 36, 2024.

[2] Zhuoyuan Mao and Yen Yu. Tuning llms with contrastive alignment instructions for machine translation in unseen, low-resource languages. *arXiv preprint arXiv:2401.05811*, 2024.

[3] Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. A benchmark for learning to translate a new language from one grammar book. *arXiv preprint arXiv:2309.16575*, 2023.

[4] Eline Visser. *A grammar sketch of Kalamang with a focus on phonetics and phonology.* 2016.

[5] Chen Zhang, Xiao Liu, Jiuheng Lin, and Yansong Feng. Teaching large language models an unseen language on the fly. *arXiv preprint arXiv:2402.19167*, 2024.