Revitalizing Kalamang: Language Expansion through Large Language Models

Aaditya Bhargav
aaditya23006@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

1 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.

2 Introduction

Machine translation (MT) has undergone a paradigm shift with the development of large, pre-trained models capable of translating numerous languages with remarkable accuracy. Yet, low-resource languages remain largely underserved, primarily due to the scarcity of parallel corpora, grammars, and linguistic resources. With languages such as Kalamang, which have fewer than 200 speakers, translation models must rely on minimal resources, including grammar books and small parallel corpora.

This paper surveys the current landscape of low-resource MT with a central focus on the recent benchmark proposed by the paper "A Benchmark for Learning to Translate a New Language from One Grammar Book" (MTOB). The survey covers key methodologies such as in-context learning, instruction inference, vocabulary optimization, and multimodal models, presenting insights into how these approaches can contribute to the translation of extremely low-resource languages.

3 Objectives and Scope

Aim: The project aims to enhance and expand the translation capabilities of the Kalamang language by incorporating English using large language models (LLMs). This approach will explore cross-linguistic learning to improve low-resource language translation. Low-resource language translation presents a

unique challenge. Most successful MT models depend on vast amounts of parallel data, which is unattainable for most of the world's languages. Many low-resource languages are spoken by small populations and lack sufficient written resources. In cases such as Kalamang, with fewer than 200 speakers, translation models must rely on minimal resources, including grammar books and bilingual word lists.

The scope of this project centers around methodologies designed to address this problem, particularly:

1. Learning a language's structure and translation using minimal resources like grammar books and bilingual word lists.

2. Approaches to improving MT quality in extremely low-resource settings using external knowledge bases and multimodal inputs.

4 Literature Review

4.1 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].

4.2 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].

4.3 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 [6].

4.4 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 [7].

4.5 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].

4.6 An Incomplete Loop: Instruction Inference, Instruction Following, and In-Context Learning in Language Models

The paper "An Incomplete Loop: Instruction Inference, Instruction Following, and In-Context Learning in Language Models" explores the reasoning capabilities of language models (LMs) across deductive, inductive, and abductive reasoning, with a focus on tasks such as hypothesis proposal, in-context learning, and instruction following. The authors experiment with models like GPT-3.5-turbo, GPT-4, and LLaMA-2 for tasks like linear function learning and artificial language translation, specifically translating Kalamang. A novel aspect of the work is the integration of instruction inference within reasoning tasks, allowing the model to generate and refine instructions during problem-solving, contrasting prior approaches that relied on instruction back-translation. However, the study highlights challenges, particularly in abductive reasoning, which proves to be a weak point in current LMs, and suggests future directions such as advanced hypothesis verification and autonomous learning mechanisms to enhance reasoning consistency and accuracy. [4]

4.7 Low-Resource NMT with Smaller Vocabulary Sizes

The abstract you provided discusses a study on optimizing subword vocabulary sizes in neural machine translation (NMT) for low-resource languages. The authors highlight that current state-of-the-art models like Transformers, when applied to low-resource languages, show significantly lower performance than on high-resource languages. They attribute this to the model's sensitivity to hyperparameters, particularly the subword vocabulary size. Their experiments demonstrate that using smaller vocabularies, as low as 1k tokens, leads to faster training, smaller model sizes, and better translation quality. In their experiments with languages like English-Akkadian, Lower Sorbian-German, and English-Manipuri, they found that smaller vocabularies not only improve the ChrF scores by up to 322% but also reduce model size by 66% and training time by up to 17%. This suggests that smaller vocabularies may be more effective in low-resource conditions than the default vocabulary size of 32k, commonly used in machine translation. The study concludes by advocating for careful selection of vocabulary sizes in NMT, especially when dealing with under-resourced languages, to maximize model efficiency and performance. [5]

5 Dataset Overview

This project focuses on advancing machine translation (MT) for extremely low-resource languages, using Kalamang as a case study. We explore two key methodologies: (1) fine-tuning pre-trained models on a parallel dataset of Kalamang sentences and their English translations, and (2) sequentially fine-tuning models on Kalamang grammar text and bilingual word pairs before evaluating their performance. The primary evaluation metric is the CHRF score, which assesses translation quality. Our goal is to demonstrate how leveraging minimal linguistic resources like grammar books and bilingual dictionaries can enable effective translation in languages with limited digital presence.

5.1 Datasets Used in the Project

The datasets used in this project include:

5.1.1 Kalamang-English Sentence Pairs

- **Structure:** A parallel corpus consisting of two primary columns:
  - **Source Language (Kalamang_Sentence):** Sentences in the Kalamang language, which serve as the input data.
  - **Target Language (English_Translation):** Corresponding translations in English, which act as the output or labels for the model.
- **Size:** The training dataset contains 1554 entries while the test dataset contains 100 entries, ensuring a decent amount of examples for the model to learn language patterns.
- **Content:** The sentences include a mix of:
  - Declarative sentences.

  - Imperatives (commands).
  - Interrogatives (questions).
  - Idiomatic phrases or culturally specific expressions.
- **Preprocessing Applied:** Sentences in both columns are:
  - Tokenized into smaller components (e.g., words or subwords) for model processing.
  - Lowercased and normalized to standardize input.
- **Purpose:** This dataset serves as the foundation for creating a mapping between the source and target languages, enabling the model to generalize this mapping to unseen sentences.

5.1.2 Kalamang Grammar Text

- **Content:** A comprehensive linguistic and ethnographic description of Kalamang, detailing:
  - Phonology, syntax, and semantics of Kalamang.
  - Sociolinguistic aspects like language vitality, speaker demographics, and usage contexts.
  - Cultural practices, traditions, and relationships with other local languages.
- **Dataset Structure:**
  - **Annotated Examples:** Sentences, phrases, and morpheme-level breakdowns, often paired with English and Papuan Malay translations.
  - **Recorded Corpus:** Combines naturalistic speech, elicited examples, and cultural content.
  - **Cultural Notes:** Descriptions of rituals, community structure, and daily activities, providing context for linguistic patterns.
- **Application in Model Fine-Tuning:**
  - Learn syntactic structures and vocabulary of Kalamang.
  - Understand linguistic nuances like verb morphology, word order, and case marking.
  - Gain exposure to narratives and dialogues for contextual understanding.

5.1.3 Kalamang-English Word Pairs

- **Content:** A bilingual dictionary containing:
  - **Kalamang Words:** Keys representing Kalamang words or phrases.
  - **English Synonyms/Translations:** Values as equivalent English meanings, including:
    - Single words, e.g., *big*: *"temun"*.
    - Phrases, e.g., *a big one*: *"temun"*.
    - Multiple meanings, e.g., *above*: *"keitko; kerunggo"*.
- **Structure:** JSON object format with Kalamang words as keys and English translations or synonyms as values.
- **Purpose of Fine-Tuning:**
  - Enrich vocabulary by helping the model learn precise word meanings.
  - Enable synonym recognition for different English equivalents of a single Kalamang word or phrase.
  - Infer linguistic roles like nouns, verbs, or adjectives through usage variations.
- **Unique Features:** Captures cultural or region-specific terms such as:
  - Specific objects, e.g., *anchor*: *"po; saor"*.
  - Flora and fauna, e.g., *angelfish*: *"kubalbal"*.
  - Idiomatic or figurative expressions.

These datasets collectively address the data scarcity challenges inherent in low-resource languages and provide a robust foundation for fine-tuning and evaluating translation models.

6 Methodologies

6.1 Fine-Tuning Pretrained Models on a Parallel Dataset of Kalamang Sentences and Their English Translations

6.1.1 Data Preparation

The dataset consists of parallel text pairs of Kalamang sentences and their corresponding English translations. Each pair serves as a source-target mapping for training. The data preparation process involves the following steps:

- **Dataset Loading:** The training and test datasets are provided in CSV format, loaded into Python using the pandas library. The `train_set.csv` and `test_set.csv` files contain the aligned sentences required for supervised learning.
- **Column Specification:** The source and target languages are dynamically specified based on the translation direction:
  - *For English to Kalamang:* The source column is `English_Translation`, and the target column is `Kalamang_Sentence`.
  - *For Kalamang to English:* The source column is `Kalamang_Sentence`, and the target column is `English_Translation`.
- **Tokenization:** The text data is preprocessed and tokenized using the appropriate tokenizer for the selected pretrained model. Padding and truncation are applied to ensure uniform sequence lengths (maximum length of 128 tokens). Tokenized inputs include:
  - `input_ids:` Encoded source sentences.
  - `attention_mask:` Mask to indicate non-padded tokens.
  - `labels:` Encoded target sentences, with padding tokens replaced by -100 to be ignored during loss computation.

6.1.2 Model Fine-Tuning

The core of the methodology involves fine-tuning various pretrained models on the parallel dataset. This process leverages the transfer learning capabilities of transformer-based architectures to adapt to the specific characteristics of Kalamang and English.

Pretrained Models Used Several models were experimented with, each chosen for its architecture and suitability for low-resource settings:

- **T5-Small:** A sequence-to-sequence transformer model pre-trained for text-to-text tasks, including translation.
- **GPT-2:** A generative transformer model adapted for translation tasks by treating translation as a text-generation problem.
- **Helsinki-NLP/opus-mt-en-mul:** A multilingual translation-specific model, pre-trained on multiple language pairs.
- **google-t5/t5-base:** A larger variant of T5-Small, pre-trained for text-to-text tasks.
- **facebook/bart-base:** A sequence-to-sequence transformer model with robust translation capabilities, and capture morphological and syntactic complexities.

Dataset Conversion The datasets are converted into PyTorch `Dataset` objects, enabling efficient batch processing. The `TranslationDataset` class encapsulates the tokenized source and target texts, providing them in a format compatible with PyTorch-based trainers.

Training Setup Fine-tuning is performed using the Hugging Face Trainer API, which simplifies model training and evaluation. The training configuration includes:

- **Learning Rate:** A low learning rate (`2e-5`) to fine-tune pretrained weights without overfitting.
- **Batch Size:** Large effective batch sizes are simulated using gradient accumulation (64 per device with 16 accumulation steps).
- **Mixed Precision Training:** Enabled via `fp16` to improve computational efficiency and reduce memory usage.
- **Epochs:** Extended training over 32 epochs to ensure convergence on the low-resource dataset.
- **Evaluation Strategy:** Validation is performed at the end of each epoch to monitor generalization performance.

Accelerator Integration The Accelerator API is employed to optimize multi-GPU training, ensuring efficient parallelization and scaling across devices.

6.1.3 Evaluation

The performance of the fine-tuned models is assessed using both quantitative and qualitative methods:

Quantitative Evaluation The **ChrF Score** (Character n-gram F-score) is used as the primary metric for evaluating translation quality. ChrF is particularly suitable for low-resource languages as it evaluates character-level overlaps, capturing morphological and syntactic nuances better than word-based metrics like BLEU.

- **Metric Calculation:**
  - Predictions are generated for the test set using the model's `generate` method.
  - References are formatted as a list of lists (one reference per source sentence).
  - The ChrF score is computed using the `sacrebleu` library.

Qualitative Evaluation A subset of test set predictions is manually inspected to analyze the model's ability to:

- Preserve semantic meaning.
- Capture syntactic structures.
- Handle rare or low-frequency words in Kalamang.

Evaluation Procedure

- The model generates translations for each source sentence in the test set.
- The predicted translations are compared to reference translations.
- The ChrF score is reported, along with representative examples of predictions and references.

6.2 Fine-Tuning a Pretrained Language Model for Kalamang Translation and Grammar Understanding

6.2.1 Data Description

The data used in this work comprises two key components:

- **Kalamang Grammar Dataset:**
  - A text file containing grammar rules, syntactic structures, and example sentences in the Kalamang language.

  - This dataset serves to fine-tune the pretrained model for an understanding of Kalamang's linguistic patterns.

- **Kalamang-English Dictionary Data:**
  - Two JSON files containing word pairs where Kalamang words are paired with their English meanings.
  - The first JSON file includes mappings with additional grammatical information (e.g., word type or usage context).
  - The second JSON file provides direct one-to-one mappings of Kalamang words to their English equivalents.

- **Test Dataset:**
  - A CSV file containing Kalamang sentences paired with their English translations and vice versa.
  - Used for evaluating the fine-tuned model using the CHRF metric, which measures translation quality.

6.2.2 Model Selection

The study uses the **Helsinki-NLP/opus-mt-en-mul** and **facebook/mbart-large-50-many-to-one-mmt** model, a Transformer-based sequence-to-sequence model pretrained for multilingual translation. This model provides a robust foundation for further domain-specific adaptation.

6.2.3 Fine-Tuning Methodology

Fine-Tuning on Grammar Text

- **Data Preparation:** The grammar text is preprocessed to include task-specific prefixes (e.g., `"grammar:"`) to help the model contextualize the task during fine-tuning.
- **Tokenization:** The data is tokenized with padding and truncation to maintain uniform input length while retaining syntactic integrity.
- **Training:** A customized fine-tuning process is carried out using the preprocessed grammar text, allowing the model to capture the linguistic rules and sentence patterns unique to Kalamang.

Fine-Tuning on Dictionary Data

- **Data Merging:** The two JSON files are merged to create a unified dataset. Each entry contains a source-target pair, where the source is a Kalamang word and the target is its English equivalent. To enhance generalizability, reverse translations (English to Kalamang) are also included.
- **Prefixing and Tokenization:** A prefix (`"translate:"`) is added to each source sentence to define the translation task explicitly. The data is tokenized similarly to the grammar dataset.
- **Training:** The model is fine-tuned on this dictionary data, focusing on learning accurate mappings between Kalamang and English words.

6.2.4 Evaluation

The final fine-tuned model is evaluated using the test dataset:

- **Bidirectional Evaluation:** The model is evaluated for both Kalamang-to-English and English-to-Kalamang translation tasks.
- **CHRF Metric:** The CHRF metric, a robust translation quality measure that accounts for character-level n-gram precision and recall, is used to assess the model's performance. The evaluation pipeline involves:
  - Preprocessing the test data using the same tokenizer employed during training.
  - Generating predictions from the fine-tuned model for both translation directions.
  - Computing the CHRF score by comparing the predictions against the reference translations.

7 Results

This section provides a comprehensive evaluation of all models used for Kalamang-to-English and English-to-Kalamang translations. The models were evaluated using the ChrF Score, a character-level F-score metric suitable for low-resource languages. Qualitative analysis of model outputs further highlights their strengths and weaknesses.

7.1 Overview of Results

The following table summarizes the ChrF scores for each model across the two translation directions:

Table 1: ChrF Scores for All Models for Method 1

Model	**Kalamang to English**	**English to Kalamang**
T5-Small	42.96	20.51
Helsinki-NLP/opus-mt-en-mul	30.45	30.45
google-t5/t5-base	22.50	22.50
facebook/bart-base	35.46	**52.75**
GPT-2	12.06	10.23

Table 2: ChrF Scores for All Models for Method 2

Model	**Kalamang to English**	**English to Kalamang**
Helsinki-NLP/opus-mt-en-mul	8.37	9.77
facebook/mbart-large-50-many-to-one-mmt	11.57	12.18

7.2 Detailed Model Performance

7.2.1 Methodology 1

T5-Small

- **Kalamang to English: Score: 42.96**
  - T5-Small demonstrates strong performance in translating Kalamang to English.
  - It effectively captures syntactic structures and provides coherent English translations, although it struggles with rare or complex expressions.
- **English to Kalamang: Score: 20.51**
  - The model struggles to translate English into Kalamang, reflecting limited training data for low-resource languages.
  - Translations often lack fluency and contain repetitive or incomplete phrases.

Helsinki-NLP/opus-mt-en-mul

- **Kalamang to English and English to Kalamang: Score: 30.45 (both directions)**
  - This multilingual model performs consistently across both directions.
  - It provides satisfactory translations but lacks depth in contextual understanding and fluency compared to task-specific models like T5-Small.

google-t5/t5-base

- **Kalamang to English and English to Kalamang: Score: 22.50 (both directions)**
  - As a larger variant of T5-Small, T5-Base shows slightly better performance in English-to-Kalamang translations but fails to match the fine-tuned T5-Small model's performance.
  - Its performance is constrained by the lack of task-specific fine-tuning.

facebook/bart-base

- **Kalamang to English: Score: 35.46**
  - Bart-Base performs well in generating fluent and contextually appropriate English sentences.
  - However, it struggles with certain linguistic nuances of Kalamang, especially for complex expressions.
- **English to Kalamang: Score: 52.75**
  - Bart-Base outperforms all other models in this direction.
  - It effectively captures the morphological and syntactic complexities of Kalamang, producing fluent translations.

GPT-2

- **Kalamang to English: Score: 12.06**
  - GPT-2 exhibits poor performance in generating meaningful translations, often echoing the input or producing hallucinated content.
  - Its general-purpose generative architecture is not well-suited for low-resource translation tasks.
- **English to Kalamang: Score: 10.23**
  - Similar to Kalamang-to-English, GPT-2 struggles with semantic accuracy and syntactic alignment.
  - Translations are often incomplete or irrelevant.

7.2.2 Methodology 2

Helsinki-NLP/opus-mt-en-mul

- **Kalamang to English: Score: 8.37**
  - The model exhibits limited capacity for translating Kalamang sentences into English.
  - While it captures basic lexical mappings, its translations often lack grammatical coherence.
- **English to Kalamang: Score: 9.77**
  - The model performs marginally better in translating English sentences into Kalamang.
  - Basic phrases and word-level mappings are generally accurate, indicating the model's reliance on lexical correspondences.

facebook/mbart-large-50-many-to-one-mmt

- **Kalamang to English: Score: 11.57**
  - The model shows moderate performance in translating Kalamang to English, outperforming Helsinki-NLP/opus-mt-en-mul.
  - It demonstrates a better grasp of syntactic structures and produces translations that are more coherent and contextually relevant.
- **English to Kalamang: Score: 12.18**
  - The model achieves better results for English-to-Kalamang translation compared to Helsinki-NLP/opus-mt-en-mul.
  - Translations are more fluent and often include correct grammatical markers, indicating improved linguistic adaptation.

7.3 Qualitative Analysis

Sample predictions were extracted to illustrate the strengths and weaknesses of the models for both translation directions.

Sample Predictions: Kalamang to English

Source (Kalamang)	T5-Small Prediction	Helsinki-NLP	facebook/bart-base	GPT-2 Prediction	Reference (English)
Aisa ma yuotpanoi anggonggon	Aisa ordered snails...	Aisa ordered snails...	Aisa ordered snails...	Aisa ma yuotpanoi anggonggon	Aisa ordered snails from me.
Binkur esun bisa erat kies	Binkur's father...	Binkur's father...	Binkur's father...	Binkur esun bisa erat kies	Binkur's father can carve a canoe, etc.

Figure 1: Sample Predictions: Kalamang to English

Sample Predictions: English to Kalamang

Source (English)	T5-Small Prediction	Helsinki-NLP	facebook/bart-base	GPT-2 Prediction	Reference (Kalamang)
Aisa ordered snails from me.	Aisa ma yuotpanoi...	Aisa ma yuotpanoi...	Aisa ma yuotpanoi...	Aisa ordered snails from me.	Aisa ma yuotpanoi anggonggon.
Expel the chicken first!	Ka tok kokoat arte!	Ka tok kokoat arte!	Ka tok kokoat arte!	Expel the chicken first!	Ka tok kokoat arte!

Figure 2: Sample Predictions: English to Kalamang

8 Inferences

8.1 Methodology 1: Fine-Tuning on Parallel Data

Task-Specific Models Exhibit Superior Performance:

- T5-Small and facebook/bart-base, both pre-trained for sequence-to-sequence tasks, consistently outperform general-purpose models such as GPT-2. This highlights the importance of task-specific pretraining for low-resource translation tasks.
- **facebook/bart-base** achieves the highest performance for English-to-Kalamang translation (**ChrF: 52.75**), showcasing its ability to handle morphologically rich target languages.

Multilingual Models Demonstrate Consistent Performance:

- The Helsinki-NLP/opus-mt-en-mul model provides balanced scores (**ChrF: 30.45**) in both directions.
- While it does not surpass the specialized models, its consistent performance across language pairs underscores the versatility of multilingual pretraining for low-resource languages.

General-Purpose Models Face Limitations:

- GPT-2 performs significantly worse than task-specific models, with **ChrF scores of 12.06** (Kalamang-to-English) and **10.23** (English-to-Kalamang).
- Its architecture, optimized for text generation rather than translation, struggles to capture the linguistic nuances of Kalamang.

This underscores the need for specialized architectures and pretraining strategies for effective translation tasks.

Challenges in Low-Resource Settings:

- The low-resource nature of Kalamang poses significant challenges:
  - Limited parallel data constrains the models' ability to learn accurate source-target mappings.

  - Morphological richness and syntactic diversity in Kalamang make it difficult for models to generalize, particularly in English-to-Kalamang translations.

- Despite these challenges, the results demonstrate that task-specific fine-tuning can yield meaningful translations even with constrained data.

Role of Model Size and Architecture:

- Larger models, such as **Llama-30B**, show exceptional capacity for context understanding and fluent output, achieving scores comparable to T5-Small for Kalamang-to-English (**ChrF: 42.96**). However, their computational demands limit their accessibility for widespread use.
- Smaller models, such as **T5-Small**, strike a balance between performance and efficiency, making them more practical for deployment in resource-constrained settings.

9 Novelty

This research project introduces several technical innovations in the domain of low-resource language translation. These novelties directly address the challenges posed by the lack of resources, complex morphology, and syntactic structures inherent to Kalamang. The following highlights the technical contributions:

9.1 Fine-Tuning Pretrained Models for Low-Resource Translation

Optimized Use of Transformer Architectures:

- Fine-tuned **facebook/bart-base** achieves the highest ChrF score (**52.75**) for English-to-Kalamang, showcasing its ability to adapt to morphologically rich languages.
- Demonstrated that task-specific fine-tuning of smaller models like **T5-Small** can achieve competitive results (**ChrF: 42.96 for Kalamang-to-English**) while being computationally efficient.

Evaluation of Multilingual Models:

- Evaluated **Helsinki-NLP/opus-mt-en-mul**, a multilingual translation model, achieving balanced ChrF scores (**30.45**) across both directions.
- This highlights the model's potential for resource-scarce settings without task-specific tuning.

9.2 Comparison Across Diverse Architectures

Extensive Multimodel Benchmarking:

- Compared models with distinct architectures (e.g., sequence-to-sequence: T5-Small, Bart; generative: GPT-2; multilingual: Helsinki-NLP).
- Identified architectural strengths, such as the ability of **Bart** to capture Kalamang morphology versus the general-purpose limitations of **GPT-2**.

Resource-Performance Trade-off:

- Highlighted that smaller models like **T5-Small** achieve competitive scores with less computational overhead compared to **Llama-30B**, which, while achieving high scores, demands significant resources.

9.3 Linguistically Informed Methodology

Handling Morphologically Rich Languages:

- Integrated task-specific pretraining techniques to address the challenges posed by Kalamang's morphology and syntactic variability.
- Showcased how pretrained transformers can adapt to such languages with minimal additional linguistic rules.

Error Mitigation Techniques:

- Applied padding token masking and adjusted learning rates to reduce noise during backpropagation, improving convergence on small datasets.

9.4 Training Strategies Tailored for Low-Resource Scenarios

Gradient Accumulation and Mixed Precision Training:

- Employed gradient accumulation and mixed precision training to effectively train on small datasets with limited computational resources.
- This allowed the use of larger batch sizes, improving convergence stability and optimizing memory usage.

Data Handling and Tokenization:

- Developed a custom tokenization and dataset preparation pipeline to handle Kalamang's unique linguistic features, ensuring alignment between source and target sentences.
- Utilized `pad_token_id` masking with -100 for loss computation to optimize gradient calculations, reducing noise in training.

10 Conclusion

The Kalamang CLDF Dataset Project has successfully transformed raw linguistic data into a structured, CLDF-compliant format, ensuring comprehensive documentation of the Kalamang language. From the initial setup—where key dependencies like `pydictionaria` and `cldfbench` were installed—to the data processing pipeline, each step was designed to clean, normalize, and map the `raw.sfm` to meaningful linguistic data.

This study underscores the efficacy of leveraging pretrained transformer models for low-resource language translation. The following conclusions can be drawn:

- **Task-Specific Pretraining is Crucial:** Models like **T5-Small** and **facebook/bart-base**, optimized for translation tasks, demonstrate significant advantages in handling both syntactic and semantic aspects of translation.
- **Multilingual Pretraining Provides Versatility:** While not excelling in performance, models like **Helsinki-NLP/opus-mt-en-mul** offer robust baseline results across language pairs, making them a valuable option when specialized resources are unavailable.
- **Limitations of General-Purpose Models:** General-purpose architectures such as **GPT-2** are insufficient for translation tasks, particularly for low-resource languages, without significant architectural adaptations or task-specific fine-tuning.
- **Impact of Resource Constraints:** The low-resource nature of Kalamang highlights the importance of:
  - Data augmentation, including techniques such as back-translation and paraphrasing.
  - Morphological analysis to better understand and encode Kalamang's linguistic complexity.
  - Integration of linguistic features to further improve translation performance.
- **Future Directions:** The findings reaffirm the potential of transformer-based architectures in addressing the challenges of low-resource translation. The project lays a strong foundation for future work, which can focus on:
  - **Data Augmentation:** Leveraging techniques such as back-translation, paraphrasing, and transfer learning from typologically similar languages to enhance training data diversity and volume.
  - **Model Adaptation:** Exploring hybrid models that integrate linguistic rules with transformer-based architectures to address Kalamang's morphological complexity.

- **Accessibility:** Deploying efficient models like **T5-Small** for practical applications while exploring larger models for advanced research in low-resource language translation.

The findings reaffirm the potential of transformer-based architectures in addressing the challenges of low-resource translation. They lay a strong foundation for future work in developing linguistically aware, resource-efficient, and contextually adaptive translation systems for low-resource languages like Kalamang.

11 Future Scope

The project paves the way for significant advancements in machine translation for low-resource languages, such as Kalamang. The promising results obtained using pretrained transformer models, despite the challenges posed by data sparsity and linguistic complexity, highlight several potential areas for future research and development.

11.1 Data Augmentation and Expansion

- **Synthetic Data Generation:**
  - Employ back-translation techniques to create synthetic parallel data by translating large monolingual corpora of Kalamang and English.
  - Use paraphrasing tools to expand the diversity of the training dataset while maintaining semantic integrity.
- **Cross-Lingual Transfer Learning:**
  - Leverage high-resource languages that are typologically similar to Kalamang (if available) for transfer learning.
  - Fine-tune multilingual models like Helsinki-NLP on additional language pairs to benefit from shared linguistic features.
- **Crowdsourced Data Collection:**
  - Engage native speakers and linguists to curate high-quality parallel datasets for underrepresented languages like Kalamang.

11.2 Model Enhancements

- **Linguistically Informed Models:**
  - Incorporate morphological, syntactic, and phonetic features of Kalamang into the model architecture to better capture its linguistic nuances.
  - Develop hybrid models that integrate rule-based systems with neural networks for better generalization.
- **Efficient Transformer Architectures:**
  - Explore lightweight, low-resource-efficient architectures such as DistilBERT or TinyBERT for on-device deployment in resource-constrained settings.
  - Experiment with sparsity-inducing techniques to reduce the computational footprint of larger models like Llama-30B while retaining performance.
- **Meta-Learning Approaches:**
  - Investigate meta-learning to enable models to quickly adapt to new low-resource languages by learning generalized translation tasks.

11.3 Contextual and Cultural Understanding

- **Context-Aware Translation:**
  - Develop models that incorporate discourse-level context to handle longer sequences and improve coherence across sentences.
- **Cultural Sensitivity:**
  - Train models to understand cultural references, idiomatic expressions, and contextually appropriate phrases, enhancing their usability in real-world applications.

11.4 Deployment and Accessibility

- **User-Centric Applications:**
  - Build translation tools and APIs tailored for local communities to support education, documentation, and communication.
  - Integrate these tools into mobile and web platforms to ensure accessibility for native speakers and linguists.
- **Real-Time Translation:**
  - Develop low-latency models for real-time speech-to-text and text-to-text translation, enabling seamless communication between speakers of Kalamang and other languages.
- **Integration with Language Preservation Efforts:**
  - Collaborate with linguists and anthropologists to document and preserve Kalamang using the developed translation tools.
  - Use the models to support endangered language revitalization efforts by generating digital resources like translated texts and audio.

11.5 Broader Impacts

- **Generalization to Other Low-Resource Languages:**
  - Apply insights and methodologies from this project to other low-resource languages, extending its impact across the linguistic spectrum.
- **Advancing Ethical AI:**
  - Address biases and ethical considerations inherent in machine translation systems, particularly for low-resource languages where cultural nuances are critical.
- **Cross-Disciplinary Collaboration:**
  - Foster collaboration between computer scientists, linguists, and local communities to improve data collection, model performance, and practical deployment.

11.6 Research Directions

- **Evaluation Metrics:**
  - Develop and use metrics that better capture the morphological and syntactic nuances of low-resource languages, complementing ChrF and BLEU scores.
- **Explainability in Translation Models:**
  - Enhance model interpretability to understand how translation decisions are made, particularly for complex or ambiguous phrases.
- **Low-Resource Benchmarking:**

– Establish a robust benchmark dataset and evaluation framework for Kalamang and similar languages to standardize comparisons across models and approaches.

The future of this project lies in its ability to bridge the gap between technological advancements and linguistic diversity. By addressing data scarcity, improving model architectures, and deploying accessible tools, this work has the potential to significantly impact both the scientific community and the local speakers of Kalamang, contributing to the preservation and revitalization of low-resource languages globally.

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] Author Unknown. An incomplete loop: Instruction inference, instruction following, and in-context learning in language models. *arXiv preprint arXiv:2404.03028*, 2024.

[5] Author Unknown. Low-resource nmt with smaller vocabulary sizes. *Lecture Notes in Computer Science*, 14202:324–336, 2024.

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

[7] 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.