Survey Paper: Revitalizing Kalamang: Language Expansion through Large Language Models

VISHAL SINGH, VANI MITTAL, JAYSHIL SHAH, AADITYA BHARGAV, MD ABUZAR KHAN, and MOHIT

Machine translation (MT) has made significant strides in recent years, with large language models (LLMs) leading to improved performance across languages. However, translating low-resource languages remains an ongoing challenge, particularly for languages with little to no digital presence. In this survey, we analyze the latest research advancements in low-resource MT, with a focus on "A Benchmark for Learning to Translate a New Language from One Grammar Book" by introducing methods that leverage minimal resources, including single grammar books, bilingual word lists, and small parallel corpora [1]. We explore how these findings integrate with novel approaches in multilingual translation, vocabulary optimization, and multimodal methods. The paper highlights the limitations of current MT models, particularly with languages such as Kalamang and Zhuang, and proposes future directions that can address these limitations through more effective use of external knowledge, non-parametric memory, and better contextual understanding.

ACM Reference Format:
Vishal Singh, Vani Mittal, Jayshil Shah, Aaditya Bhargav, Md Abuzar Khan, and Mohit. 2024. Survey Paper: Revitalizing Kalamang: Language Expansion through Large Language Models . 1, 1 (September 2024), 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 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, with fewer than 200 speakers, translation models must rely on minimal resources, including grammar books and small parallel corpora [1].

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) [1]. 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.

2 Problem Definition and Scope

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

Authors' Contact Information: Vishal Singh; Vani Mittal; Jayshil Shah; Aaditya Bhargav; Md Abuzar Khan; Mohit.

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 [1].

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

- Learning a language's structure and translation using minimal resources like grammar books and bilingual word lists [1].
- Approaches to improving MT quality in extremely low-resource settings using external knowledge bases and multimodal inputs [4].

3 Literature Review

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

The paper "A Benchmark for Learning to Translate a New Language from One Grammar Book" introduces MTOB (Machine Translation from One Book), a novel benchmark for translating between English and Kalamang, a low-resource language with fewer than 200 speakers. The unique aspect of this benchmark is that it mimics how humans learn languages by relying solely on a single grammar book, word lists, and a small corpus, rather than large datasets. Evaluating models like LLaMA-2 and GPT-4, the study tests in-context learning and lightweight finetuning for translation tasks, comparing the performance to a human baseline. The work highlights challenges such as data scarcity, hallucination, and difficulty in retrieving useful context, particularly with the finetuning of models. The paper also outlines future goals, including improving translation accuracy for low-resource languages, expanding the approach to other typologically diverse languages, and exploring multimodal models to support endangered language preservation and revitalization.

3.2 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 [2].

3.3 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 [3].

3.4 Multimodal Machine Translation for Manipuri

This paper explores recent advancements in low-resource machine translation, particularly emphasizing the role of multimodal machine translation in improving outcomes for languages like Manipuri. Existing methods for low-resource language translation often focus on textual data, but the inclusion of additional modalities, such as visual and auditory inputs, has been shown to boost accuracy. Studies highlight the effectiveness of multimodal approaches, with notable improvements in BLEU scores. Research has explored the integration of images and audio with text to enhance machine understanding and context for translation, addressing the challenge of limited written resources in underrepresented languages. Additionally, multimodal machine translation has been applied in various languages, indicating potential for improved performance across multiple language pairs.

3.5 Chinese-Centric Neural Machine Translation for Low-Resource Languages

Focusing on Chinese as a hub language for low-resource translation, this study introduces bilingual curriculum learning, contrastive learning, and noise-robust methods to improve MT for low-resource languages like Zhuang [5]. The use of monolingual data and novel loss functions like the In-Trust loss demonstrates the effectiveness of handling noisy, low-resource data. It highlights how NMT models like Transformer and Recurrent Neural Networks have been successful in resource-rich environments, but challenges persist in low-resource languages due to limited parallel corpora. Recent efforts in data augmentation, transfer learning, and the use of monolingual data have improved NMT performance in such settings. The paper also covers the use of contrastive learning and auxiliary language data to address these challenges, offering insights into enhancing NMT systems.

3.6 KARD: Knowledge-Augmented Reasoning Distillation

This paper addresses the key developments in large language models (LLMs) and their application to knowledge-intensive reasoning tasks. The authors discuss how LLMs, such as GPT-3.5 and others, have demonstrated remarkable performance in various domains, especially those requiring deep reasoning and domain-specific knowledge. However, they also emphasize the challenges of deploying LLMs in real-world scenarios, such as the high computational costs and privacy concerns. The paper highlights previous efforts to distill reasoning abilities from LLMs into smaller language models, noting that these approaches often fall short due to the limited capacity of smaller models to memorize knowledge. This motivates the need for techniques like Knowledge-Augmented Reasoning Distillation (KARD), which leverages external knowledge bases to supplement small models. The review also touches upon reasoning distillation, retrieval-augmented language models, and the limitations of existing methods in addressing knowledge-intensive tasks. The use of neural rerankers to improve document retrieval for reasoning is discussed as an innovative contribution to the field.

3.7 Teaching Large Language Models an Unseen Language on the Fly

This paper focuses on the challenges of adapting Large Language Models (LLMs) to low-resource languages, which have limited available data. Traditional methods, such as continual pre-training on monolingual texts (Yong et al., 2023) and the use of adapters like MAD-X (Pfeiffer et al., 2020), have been employed to improve model performance in these settings. Other approaches, such as supervised fine-tuning (SFT) using cross-lingual instructions (Cahyawijaya et al., 2023), have shown some success, but these require a substantial amount of training data. For machine translation, researchers have explored prompting-based methods, such as DIPMT (Ghazvininejad et al., 2023), which use dictionaries to aid translation in low-resource languages. However, these methods often assume some baseline knowledge of the target language, which is lacking for extremely low-resource languages like Zhuang. This paper addresses these gaps by introducing DIPMT++, a framework that enables LLMs to learn a new language solely through prompting, expanding on previous work with strategies like bilingual lexicon induction and synonym expansion to enhance performance in unseen languages.

4 Methodologies and Techniques

4.1 In-Context Learning and Grammar-Based Learning

The MTOB framework emphasizes learning through a single grammar book, providing models with word lists, grammar explanations, and bilingual sentence pairs [1]. In-context learning allows language models to process these limited resources to generate translations, particularly in languages with minimal digital presence.

4.2 Vocabulary Size Reduction

The reduction of vocabulary size to around 1,000 tokens, as proposed in the Low-Resource NMT paper, improves training times and reduces model complexity. The focus is on optimizing translations by balancing vocabulary coverage and performance, especially useful in resource-constrained environments [3].

4.3 Multimodal Translation

Incorporating text, images, and audio, as explored in Multimodal Machine Translation for Manipuri, enhances translation for languages with limited textual data. By leveraging non-textual information, multimodal systems better align linguistic and contextual meanings, particularly for spoken or unwritten languages [4].

4.4 Knowledge-Augmented Reasoning

KARD leverages external knowledge to assist in rationale generation, particularly important for low-resource languages. By dynamically retrieving knowledge, KARD enables small models to perform better reasoning tasks without needing massive datasets [6].

5 Datasets and Benchmarks

The datasets used across these studies vary, ranging from small parallel corpora to multimodal collections. Key datasets include:

- MTOB Dataset: A grammar book, bilingual word list, and parallel corpus of less than 2,000 sentences for Kalamang translation. Benchmarks include accuracy, BLEU scores for translation tasks, and F1 scores for function learning. GitHub Repository [1].

- Zhuang Parallel Corpus: A corpus of around 5,000 sentences paired with a Zhuang-Chinese bilingual dictionary [5].
- Manipuri Multimodal Dataset: Local news articles paired with images and audio data, manually collected for the task [4].
- Function Learning and Kalamang Translation: This GitHub codebase provides tools for running experiments on function learning, color domain, and Kalamang translation with OpenAI and open-source models. Benchmarks involve evaluation metrics such as accuracy and BLEU scores for translation tasks. GitHub Repository [2].
- DIPMT++ Framework: The ZhuangBench GitHub repository contains the code for the DIPMT++ framework, along with the Zhuang-Chinese dictionary and parallel corpus used in their experiments. Benchmarks typically include BLEU and METEOR scores for translation quality. ZhuangBench GitHub Repository [7].
- Knowledge-Augmented Reasoning Distillation (KARD): This GitHub repository provides the implementation of the Knowledge-Augmented Reasoning Distillation method, serving as a reference for future experiments. Benchmarks focus on accuracy and reasoning ability metrics evaluated against standard reasoning tasks. GitHub Repository [6].

6 Applications and Use Cases

- Endangered Language Preservation: The MTOB and DIPMT++ methods provide tools for translating endangered languages like Kalamang and Zhuang, aiding linguistic preservation and digital representation [1][7].
- Educational Tools: Grammar-based learning approaches enable the development of educational applications for teaching low-resource languages [1].
- Medical Translations: Knowledge-augmented models such as KARD can support domain-specific translation tasks in fields like medicine, where accurate information retrieval is crucial [6].
- Government and Legal Services: Facilitates communication in public services and legal settings for speakers of low-resource languages, ensuring fair access to rights and services. Enhances civic engagement by providing translations of important documents and announcements.
- Cultural Exchange and Accessibility: Supports cross-cultural communication by enabling translations of literature, folklore, and local media. Increases accessibility to content for speakers of low-resource languages, fostering inclusivity.
- Research and Linguistic Studies: Provides researchers with tools to analyze low-resource languages and document linguistic features. Aids in comparative studies across languages and contributes to the understanding of language evolution.
- Tourism and Travel: Improves communication between tourists and local populations, enriching travel experiences and promoting local cultures. Enhances tourism-related services by providing translations of signs, menus, and guides in low-resource languages.

7 Challenges and Limitations

- Data Scarcity: Minimal resources for low-resource languages limit the performance of even the most advanced models [1][7].
- Hallucination and Noise: LLMs frequently introduce irrelevant or incorrect translations when faced with sparse data or incomplete grammatical information [1][6].

- Complex Syntax: While LLMs can grasp vocabulary quickly, learning and applying the syntax of an unseen language from minimal examples is still a major hurdle [1][7].
- Context utilization: While retrieving context helps, models sometimes fail to retrieve relevant or useful grammatical information from the grammar book.
- Fine Tuning: Directly fine tuning models on grammar text harmed performance in some cases, likely due to incompatibility with the task's instruction-following nature.
- Accuracy of Hypothesis Proposal: The authors found that hypotheses generated by LMs were often inaccurate, requiring external validation mechanisms like reranking.
- Complex Reasoning Tasks: Combining deductive, inductive, and abductive reasoning in the same tasks showed unexpected results, particularly in more complex domains.
- Abductive Reasoning Performance: Abductive reasoning proved to be a relatively weaker capability in current LMs compared to instruction following, posing a challenge for complex tasks requiring this reasoning type.
- Lack of Morphological Tools: Low-resource languages often lack morphological analysis tools, which limits the ability to handle word inflections and derivations.
- Handling Complex Translations: Achieving consistent results for more complex sentences or difficult language features like modifier order remains a challenge, especially with unseen grammar structures.
- Performance Gap with Larger LMs: Even though KARD significantly improves small models' performance, there remains a gap between small models and LLMs like GPT-3.5 in terms of reasoning ability. KARD reduces this disparity, yet additional efforts are required to align small models with the performance of cutting-edge LLMs [6].

8 Recent Advances and Innovations

Recent studies have introduced:

- Instruction Inference and Hypothesis Generation: Enabling LLMs to generate and rank translation hypotheses based on minimal input [2].
- Neural Rerankers: Enhancing knowledge retrieval processes to reduce noise and improve rationale generation for low-resource translations [6].
- Multimodal Inputs: Combining text, images, and audio for more comprehensive translations of spoken or visually supported languages [4].

9 Future Directions and Open Research Questions

- Scaling to More Languages: Extending grammar-based and multimodal approaches to other low-resource languages is a key area for future work [1][4].
- Improve model performance: Explore better methods for utilizing grammar books and word lists to improve LLM translation capabilities, especially for low-resource languages [1].
- Crosslinguistic evaluation: Test whether the methods developed for Kalamang can work across a wider range of languages with different typological features [1].
- Autonomous Learning and Self-Improvement: The paper points towards a longer-term goal of making LMs capable of autonomous learning, where they can improve themselves through hypothesis proposal and validation mechanisms [2].

- Improving Syntax Learning: Techniques that focus on teaching models to better understand and apply syntactic rules are needed [1][7].
- Real-World Applications: Optimizing these methods for practical applications like human translation assistance and endangered language preservation is crucial [1][6].
- Improving the retrieval mechanism: Although the neural reranker significantly improves document retrieval, there is room for improvement. Future work will focus on creating even more efficient retrieval methods that can retrieve the most relevant documents for each task, further enhancing the rationale generation process [6].

10 Conclusion

This survey highlights the ongoing developments in low-resource machine translation, emphasizing grammar-based learning, vocabulary optimization, and multimodal approaches. While significant progress has been made, challenges remain, particularly regarding data scarcity and handling complex linguistic features. Future research should focus on improving syntactic understanding, scaling to additional languages, and integrating these methods into practical applications.

References

[1] A Benchmark for Learning to Translate a New Language from One Grammar Book. https://arxiv.org/pdf/2309.16575

[2] An Incomplete Loop: Instruction Inference, Instruction Following, and In-Context Learning in Language Models. https://arxiv.org/pdf/2404.03028

[3] Low-Resource NMT with Smaller Vocabulary Sizes. https://link.springer.com/chapter/10.1007/978-3-031-70563-2_15

[4] Multimodal Machine Translation for Manipuri. https://link.springer.com/article/10.1007/s11042-023-15721-2

[5] Chinese-Centric Neural Machine Translation for Low-Resource Languages. https://www.sciencedirect.com/science/article/abs/pii/S0885230823000852

[6] KARD: Knowledge-Augmented Reasoning Distillation. https://github.com/Nardien/KARD. https://arxiv.org/pdf/2305.18395

[7] DIPMT++ Framework. https://arxiv.org/pdf/2402.19167