Prominent Neural Machine Translation Techniques for Low-Resource Language Pairs

Gustavo Camilo.

Faculty of Exact Sciences and Technology, Federal University of the Greater Dourados (UFGD), Dourados, Brazil, tuca.dev@gmail.com.

Abstract. This paper presented a structured literature review of recent advancements in neural machine translation (NMT) techniques for low-resource language pairs (LRLPs). While NMT has shown remarkable performance for high-resource languages, the scarcity of parallel corpora continues to hinder translation quality for underrepresented languages. In response, this study compiled and analyzed a selection of recent approaches specifically designed to overcome data scarcity, bridge linguistic divergence, and enhance model performance in low-resource contexts. The review categorized key techniques into several groups: subword encoding methods such as Byte Pair Encoding (BPE) for vocabulary management, embedded alignment tools like SimAlign to address syntactic divergence, and transfer learning frameworks that leverage knowledge from high-resource languages. Additionally, the paper discussed model architectures including Transformer and LSTM, along with data-centric strategies like Joint Dropout, and the incorporation of large language models (LLMs) for generating synthetic corpora. It also highlighted synthetic pivoting techniques used to improve translation quality between LRLs without direct parallel data. A focused literature survey examined specific implementations of these techniques, showcasing their effectiveness across a variety of language pairs. The discussion emphasized the promise of these strategies while acknowledging key challenges, such as the inconsistency of evaluation benchmarks and the computational limitations of LLM-based methods. Due to space constraints, some emerging approaches like zero-shot and few-shot learning were not explored in depth. Ultimately, this paper contributes to the understanding of how targeted NMT strategies can drive progress in the field and support the inclusion of linguistically marginalized communities in the digital space.

Keywords. Low-resource languages, neural machine translation, natural language processing, large language model, machine translation.

1. Introduction

Advancements in large language models (LLM) and neural machine translation (NMT) techniques provided a more feasible inter-communication between high-resource languages, such as English and German [1]. The huge availability of parallel data in these languages is the main factor that contributes to this achievement. However, underrepresented languages, such as Arabic (Saudi Arabia), Assamese (India), Bodo (India), Brazilian (Brazil), Gujarati (India), Kannada (India), Khasi (India), Kashmiri (India), and Malayalam (India) [2], suffer from data scarcity and, therefore, low-quality machine translations, especially between them.

These languages, although sometimes spoken by small communities or marginalized groups, are important cultural heritage and enrich human language understanding studies. Investing in developing natural language processing (NLP) toolkits to enhance the machine translation (MT) of low-resource language (LRL) pairs might yield an improvement in the knowledge exchange for these local communities while avoiding their extinction, preserving their linguistic diversities, and contributing to an overall better knowledge of human language [2].

As stated by Ranathunga et al. [3], despite the noticeable increase in MT research (both by academia and industry) focusing on LRL pairs, systematic reviews and comprehensive studies that examine/analyze these techniques for LRL pairs are scarce. Since the few remaining existent papers are outdated (before 2022), they ignore the recent LLMs'

contributions in leveraging machine translation quality, or don't present a special focus on resourced-poor language MT, maintaining a significant hurdle to the continuous development of this field.

As a countermeasure to this issue, this study aims to: collect and bring new advancements and prominent MT techniques, demonstrate efficient novel methods, provide major translation techniques, and ultimately contribute to surpassing less-resourced language obstacles.

Research focused exclusively on speech-based or multimodal translation, typically involving modalities such as text-to-image or speech-to-text, falls outside the scope of this study. The structure of this article is as follows: Section 1 gives the research methodology and main keywords. Section 3 provides an overview of key techniques developed to improve machine translation in low-resource settings; Section 4 presents a concise review of selected studies that exemplify these methods; Section 5 offers a discussion on broader challenges and limitations in the field; and Section 6 concludes with a summary of findings and future directions. This paper is informed and inspired by the foundational reviews conducted by Ranathunga et al. [3] and Pakray et al. [2].

2. Research methods

This research was conducted through a systematic literature review using Google Scholar, focusing on recent studies from 2022 to the present. The search strategy combined various keyword patterns to identify relevant papers on machine translation techniques for low-resource language pairs. Queries included combinations such as "best techniques" "machine translation" AND "low-resource", allintitle: translation low-resource, "translation between low-resource language pairs" -speech, "machine translation" "low-resource language pairs", and techniques for enhancing "machine translations" in "low-resource language pairs" OR intitle:"low-resource language pair(s)". The main keywords across all searches were "machine translation," "low-resource," "language pairs," "techniques," "evaluation," and specific language mentions. To ensure the review focused on impactful contributions, papers that demonstrated significant improvements in translation quality, measured through metrics like BLEU scores or robustness evaluations, were prioritized. Studies related exclusively to speech or multimodal translation were excluded, in line with the scope of this review.

3. Overview of the techniques

Recent advancements in machine translations for low-resource language pairs (LRLPs) have led to the development of a diverse set of techniques aimed at overcoming data scarcity and structural differences across languages. These methods not only enhance translation accuracy but also improve model adaptability and efficiency in scenarios with limited linguistic resources. From strategies that manage vocabulary and subword units to those that leverage pre-trained models or generate synthetic data, the field continues to evolve with innovative, resource-conscious approaches.

Subword Encoding and Vocabulary Management are frequently used in these studies. A common strategy in recent NMT research for LRLP is the use of subword Byte Pair Encoding (BPE) [4,5]. This technique helps reduce vocabulary size and address the problem of out-of-vocabulary (OOV) words, which is especially useful when dealing with small training datasets [6]. In some approaches, BPE is also applied jointly across different language pairs to create shared vocabularies for parent-child models in transfer learning setups [7].

Embedded alignment methods help bridge structural differences between languages by aligning subword units, often using tools like SimAlign with BPE [4]. These techniques improve neural models' ability to learn consistent translation patterns despite word order differences. In low-resource settings, they enhance handling of syntactic divergence and long-distance dependencies [8], leading to more accurate and robust translations.

Transfer Learning is another prominent method employed to improve NMT performance in low-resource contexts [5,7,9]. It typically involves training on a high-resource language pair and adapting the model to a related low-resource pair. Studies show that when the languages share syntactic or orthographic features, transfer learning yields significantly better translation quality [10,11]. Both fine-tuning and partial parameter freezing techniques are used to preserve learned information while adapting to the new language [5].

While **Transformer models** have traditionally required large datasets to perform well, recent work shows they can be effective even in low-resource scenarios when combined with auxiliary tools like pre-trained language models [4,12,13]. On the other hand, **Long Short-Term Memory (LSTM) architectures** continue to be used for their effectiveness in capturing long-distance dependencies, especially in models where sentence length and memory handling are critical [5].

Data-centric and Model-Agnostic approaches are innovative techniques that focus more on the data than the model architecture and are gaining attention. One such approach involves replacing certain phrases with variables during training to promote compositional generalization, making the model more robust across domains and linguistic variations [12]. These methods are model-agnostic, meaning they can be integrated into any existing NMT framework.

Some studies are exploring the **Incorporation of**

Large Language Models (LLMs) for data augmentation into the NMT pipeline [14]. These large language models are used to generate synthetic parallel corpora, which are then employed to train smaller, resource-efficient NMT systems. This hybrid method has shown strong results, particularly when translating from LRL into English, and offers a solution to the computational limitations of using LLMs directly.

A novel strategy for tackling the scarcity of direct parallel data between low-resource language pairs is **Synthetic Pivoting** [13]. This involves generating intermediate pivot sentences using techniques like knowledge distillation. The synthetic sentences are structurally aligned with the source or target language, improving translation performance. This method has proven particularly effective for under-represented language families and enhances robustness against noisy inputs.

4. Literature survey

Till a few years ago, due to the data-hungry characteristic of Transformer architectures, utilizing this framework in low-resource language pairs seemed to be unviable [3]. However, a paper conducted by Laskar et. al. [4] exploring the performance of NMTs for a low-resource language pair (English–Assamese) demonstrated positive results. The project jointly used a token alignment tool (SimAlign, which uses subword Byte Pair Encoding (BPE) level processing) to deal with the language divergence problem, i.e., different word order issues, and a pre-trained language model (PLM) to grasp Assamese nuances. The NMT models were trained on a parallel corpus of approximately 350,000 sentence pairs, and their best model outperformed the others with a BLEU score of 19.12 in the direction: Assamese to English. This model used unidirectional alignment and a PLM, which showed better handling of long-distance dependencies and syntax. The results are considered promising given the low-resource context, confirming the usefulness of the Transformer architecture even in such settings.

Another paper that uses the subword BPE method is the research of Hujon et al. [5], but with a different purpose, to improve the transfer learning technique from English-French (parent model) to English-Khasi (child model), a high-resource language pair to a low-resource language pair. In this paper, long short-term memory (LSTM) architecture is used as the backbone structure of the transfer learning process in order to remember long sentences. Their findings confirm that using languages with some level of relatedness, in the transfer learning process, excel in translation quality. They made a comparison between the baseline NMT with the NMT using transfer learning (NTMTL), which demonstrated a high superiority compared to the base NMT in all quality measurements (statistical, automatic, and human evaluation), such as a 51.11 BLEU score.

In the work of Araabi, Niculae and Mons [12], titled "Joint Dropout: Improving Generalizability in Low-Resource Neural Machine Translation through Phrase Pair Variables", they introduce a novel technique called Joint Dropout (JD) to enhance generalization in low-resource neural machine translation. JD operates by substituting phrases with variables during training, thereby promoting compositionality - a crucial aspect for generalization. This method is data-centric and model-agnostic, making it compatible with existing NMT architectures. Empirical evaluations demonstrate that JD significantly improves translation quality for language pairs with minimal resources, as evidenced by higher BLEU and Direct Assessment scores. Furthermore, the authors check the NMTs' robustness across distinct domains and under perturbation (swapping phrases or words, maintaining the syntactic and semantic accuracy). The error analysis indicates that JD enhances the robustness and adaptability of low-resource NMT across disturbances.

Some researchers are experimenting with a hybrid approach of LLMs and NMTs. The article "From LLM to NMT: Advancing Low-Resource Machine Translation with Claude" by Maxim Enis and Mark Hopkins [14] investigates the capabilities of Claude 3 Opus, a large language model developed by Anthropic, in the context of low-resource machine translation. The authors identify data contamination issues in existing benchmarks like FLORES-200 and address this by creating new, uncontaminated evaluation sets covering 36 language pairs. Their findings reveal that Claude 3 Opus outperforms strong baselines, including Google Translate and Meta's NLLB-54B, in translating into English for 55.6% of the evaluated language pairs, demonstrating notable resource efficiency. However, the model's performance is less competitive when translating from English into low-resource languages, surpassing baselines in only 33.3% of such cases. To mitigate the computational demands of deploying large LLMs, the study explores knowledge distillation techniques, using Claude to generate synthetic parallel corpora to train smaller NMT models. This approach achieves state-of-the-art results in Yoruba-English translation, matching or exceeding the performance of larger models. The research underscores the potential of leveraging LLMs like Claude to enhance NMT systems for low-resource languages, particularly when translating into English, and highlights the importance of creating clean evaluation benchmarks to accurately assess model performance.

Finally, the paper "Neural Machine Translation between Low-Resource Languages with Synthetic Pivoting" by Khalid N. Elmadani et al. [13], addresses the challenges of translating between low-resource languages due to the scarcity of direct parallel data. The authors propose a novel approach called synthetic pivoting, where pivot sentences are generated synthetically from both source and target

languages using sequence-level knowledge distillation. This method aims to align the structure of pivot sentences more closely with the source or target languages, thereby reducing translation complexity. They integrate synthetic pivoting into two paradigms: cascading and direct translation using synthetic source and target sentences. The study finds that the performance of pivot-based systems is highly dependent on the quality of the NMT model used for sentence regeneration. Additionally, training back-translation models on these synthetic sentences enhances robustness to input-side noise. Experimental results demonstrate that synthetic data generation improves pivot-based systems translating between low-resource Southern African languages by up to 5.6 BLEU points after fine-tuning. This research demonstrates an interesting novel pivot-based methodology that offers a viable solution to improve translation quality between LRLP.

5. Discussion

This study set out to explore and consolidate recent developments in machine translation techniques for low-resource language pairs, and the findings strongly confirm that meaningful progress is being made. Despite earlier doubts about the viability of Transformer-based models in data-scarce contexts[3], new research proves that combining these models with strategies like embedded alignment tools , subword encoding , and pre-trained language models can yield surprisingly good performance [4,12,13]. In addition, alternative techniques, such as transfer learning [5], synthetic data generation [13,14], and even hybrid systems with large language models [14], are demonstrating real potential. These methods are not just patchwork fixes; they represent a growing toolkit that can adapt to different kinds of data limitations and linguistic challenges. Taken together, they offer a hopeful outlook for improving communication and knowledge exchange between underrepresented communities, preserving linguistic diversity, and extending the benefits of NLP beyond high-resource languages.

That said, the path forward still has obstacles. One key challenge lies in the inconsistency of evaluation benchmarks [14], many of which are contaminated or biased toward well-resourced languages. This complicates our ability to measure true progress in low-resource settings. Also, while synthetic data generation [13,14] and pivoting [13] techniques are promising, their effectiveness often hinges on the quality of the models used to create and regenerate these sentences, introducing a dependency that not all regions or research teams can support. Moreover, transfer learning requires some degree of linguistic similarity to be truly effective [10,11], which may not always be available. Additionally, while our literature review focused on prominent and recent strategies, it is important to acknowledge the scope limitations of this paper. Techniques such as zero-shot, few-shot learning, and transliteration, which have also gained traction in LRL research, were not covered here due to space and time constraints. Despite these limitations, the findings suggest that the field is steadily equipping itself with a robust and diversified set of tools capable of addressing the unique challenges of low-resource translation.

6. Conclusions

This study reviewed recent strategies to improve machine translation for low-resource language pairs, focusing on solutions to data scarcity and structural differences between languages. Techniques like subword encoding (BPE) [4,5], transfer learning [5], embedded alignments [4,8], and LLM-driven data augmentation [14] show promising results, especially when adapted to the unique traits of each language pair. Despite progress, challenges like bidirectional translation gaps and the need for cleaner benchmarks remain [14]. Continued research into language-relatedness and synthetic data could further advance the field, helping to make machine translation more inclusive and linguistically diverse.

7. References

[1] Joshi P, Santy S, Budhiraja A, Bali K, Choudhury M. The State and Fate of Linguistic Diversity and Inclusion in the NLP World [Internet]. 2020 [cited 2025 Apr 10]. Available from: https://arxiv.org/abs/2004.09095

[2] Pakray P, Gelbukh A, Bandyopadhyay S. Natural language processing applications for low-resource languages. Natural Language Processing [Internet]. 2025/02/28 ed. 2025;31(2):183–97. Available from: https://www.cambridge.org/core/product/7D3DA31DB6C01B13C6B1F698D4495951

[3] Ranathunga S, Lee ESA, Prifti Skenduli M, Shekhar R, Alam M, Kaur R. Neural Machine Translation for Low-resource Languages: A Survey. ACM Comput Surv [Internet]. 2023 Feb;55(11). Available from: https://doi.org/10.1145/3567592

[4] Laskar SR, Paul B, Dadure P, Manna R, Pakray P, Bandyopadhyay S. English–Assamese neural machine translation using prior alignment and pre-trained language model. Computer Speech and Language. 2023 Jul 1;82.

[5] Hujon A v., Singh TD, Amitab K. Transfer Learning Based Neural Machine Translation of English-Khasi on Low-Resource Settings. Procedia Computer Science [Internet]. 2023 Jan 1 [cited 2025 Apr 8];218:1–8. Available from: https://www.sciencedirect.com/science/article/pii/S1877050922024899

[6] Sennrich R, Haddow B, Birch A. Neural Machine

Translation of Rare Words with Subword Units [Internet]. 2016 [cited 2025 Apr 13]. Available from: https://arxiv.org/abs/1508.07909

[7] Aji AF, Bogoychev N, Heafield K, Sennrich R. In neural machine translation, what does transfer learning transfer? In: Proceedings of the Annual Meeting of the Association for Computational Linguistics [Internet]. Association for Computational Linguistics (ACL); 2020 [cited 2025 Apr 13]. p. 7701–10. Available from: https://www.zora.uzh.ch/id/eprint/188224/

[8] Zhou C, Ma X, Hu J, Neubig G. Handling Syntactic Divergence in Low-resource Machine Translation [Internet]. Association for Computational Linguistics; 2019 [cited 2025 Apr 13]. Available from: https://aclanthology.org/D19-1143/

[9] Han L, Gladkoff S, Erofeev G, Sorokina I, Galiano B, Nenadic G. Neural machine translation of clinical text: an empirical investigation into multilingual pre-trained language models and transfer-learning. Frontiers in Digital Health [Internet]. 2024 [cited 2025 Apr 13];6. Available from: https://www.frontiersin.org/journals/digital-health/articles/10.3389/fdgth.2024.1211564/full

[10] Karakanta A, Dehdari J, van Genabith J. Neural machine translation for low-resource languages without parallel corpora. Machine Translation [Internet]. 2018 Jun 1 [cited 2025 Apr 13];32(1–2):167–89. Available from: https://link.springer.com/article/10.1007/s10590-017-9203-5

[11] Dabre R, Nakagawa T, Kazawa H. An Empirical Study of Language Relatedness for Transfer Learning in Neural Machine Translation [Internet]. 2017 [cited 2025 Apr 13]. Available from: https://aclanthology.org/Y17-1038.pdf

[12] Araabi A, Niculae V, Monz C. Joint Dropout: Improving Generalizability in Low-Resource Neural Machine Translation through Phrase Pair Variables. 2023 Jul 24; Available from: http://arxiv.org/abs/2307.12835

[13] Elmadani KN, Buys J. Neural Machine Translation between Low-Resource Languages with Synthetic Pivoting. In: Calzolari N, Kan MY, Hoste V, Lenci A, Sakti S, Xue N, editors. Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) [Internet]. Torino, Italia: ELRA and ICCL; 2024. p. 12144–58. Available from: https://aclanthology.org/2024.lrec-main.1063/

[14] Enis M, Hopkins M. From LLM to NMT: Advancing Low-Resource Machine Translation with Claude. 2024 Apr 21; Available from: http://arxiv.org/abs/2404.13813