Journal of Literature and Art Studies, September 2025, Vol. 15, No. 9, 725-731
doi: 10.17265/2159-5836/2025.09.008

A Review of Machine Translation Techniques for Low-Resource Languages

PENG Cheng-xi, MA Zi-han
Northwestern Polytechnical University, Xi'an, China

Machine translation of low-resource languages (LRLs) has long been hindered by limited corpora and linguistic complexity. This review summarizes key developments, from traditional methods to recent progress with large language models (LLMs), while highlighting ongoing challenges such as data bottlenecks, biases, fairness, and computational costs. Finally, it discusses future directions, including efficient parameter fine-tuning, multimodal translation, and community-driven corpus construction, providing insights for advancing LRL translation research.

Keywords: low-resource languages (LRLs) , machine translation, large language models (LLMs)

Introduction

With the development of machine translation, LRL translation continues to face persistent challenges that have garnered increasing attention from researchers. The scarcity of parallel corpora and linguistic resources limits the effectiveness of conventional rule-based, statistical, and neural machine translation techniques. Although traditional LRL machine translation methods provide valuable foundations, they struggle to scale and adapt to the complexity of diverse language structures. Recent advancements in LLMs offer new opportunities, leveraging prompt engineering, data augmentation, knowledge distillation, and transfer learning to enhance translation quality in low-resource settings.

Despite these developments, significant gaps remain, including the scarcity of corpora, the lack of sufficient training data, and noise interference, all of which hinder the effectiveness of translation. This review provides a comprehensive analysis of the current state of research, examining the main approaches and their applications. It traces the evolution from conventional rule-based and statistical techniques to contemporary LLM-based approaches, assessing the strengths and weaknesses of each paradigm while highlighting the challenges faced. Furthermore, the review explores strategies for enhancing MT techniques for LLRs, with the aim of offering insights that can contribute to the development of more inclusive and effective translation technologies.

Acknowledgments: This research was supported by China Undergraduate Innovation Training Program [Grant No. 202410699184] and Humanities and Social Sciences Research Project funded by the Ministry of Education of China [Grant No. 23YJAZH139].

PENG Cheng-xi, Undergraduate student, School of Foreign Languages, Northwestern Polytechnical University.
MA Zi-han, Undergraduate student, School of Public Policy and Administration, Northwestern Polytechnical University.

Traditional LRL Translation Techniques

LRLs typically suffer from a lack of high-quality bilingual parallel corpora and essential language processing tools, which directly limit the training capacity of machine translation (MT) models and exacerbate their technological isolation. In response to these challenges, traditional LRL machine translation techniques have emerged.

Rule-Based Machine Translation

In the early stage of the development of machine translation technology, foreign research on traditional LRL translation techniques mainly focused on special fields such as national security. The US Defense Advanced Research Projects Agency (DARPA) actively applied rule-based techniques and statistical techniques (Nakov & Ng, 2012). According to formal language theory, Rule-Based Machine Translation (RBMT) converts and generates translations around context-free grammars. This technique mainly relies on linguists to transform language knowledge into dictionaries and grammar rules (Arcan, 2019). The RBMT system uses this knowledge to analyze sentences in the source language and translate them, which does not require training corpora of LRLs.

Statistical Machine Translation

Statistical Machine Translation (SMT) occupied the dominant position at that time. It uses the Bayes conditional probability formula to convert machine translation into the learning and probability calculation of translation models and language models, which depend on statistical information such as phrases, grammars, and language models, rather than fixed rules and dictionaries.SMT can effectively deal with the complexity and diversity of languages and possesses good scalability and flexibility, but in low-resource environments, the performance of the SMT system drops drastically (Irvine, 2013).

Deep Learning techniques

In the era of big data, deep learning techniques have become the mainstream approach, forming two thinking paths that are dividedly centered on language data and advanced algorithms(Hameed & Al-Khateeb, 2025). The typical representative of the thinking path centered on language data—Neural Machine Translation (NMT) has gradually replaced SMT. Especially after the emergence of Sequence-to-Sequence (Seq2Seq) and Transformer model architectures, which use the encoder and decoder two parts to realize automatic translation between LRLs (Gehring et al., 2017). NMT significantly outperformed former techniques on LRLs and achieved ideal translation quality, meeting limited machine translation needs such as browsing and initial translation.

Hybrid Machine Translation

Hybrid Machine Translation (HMT) combines multiple techniques, such as SMT, RBMT, and other models. This approach has gained significant attention in LRLs' machine translation research due to its potential in handling diverse LRL structures and improving translation quality in various conditions. That mainly manifests in lower mistakes and increases overall accuracy through check translations by combining several translation techniques. Researchers have explored various hybrid models, such as word-based models, grammar-based models, and phrase-based models (Prashanth Nayak, 2023). HMT will continue to be improved by combining advancements in deep learning and other artificial intelligence technologies to further enhance translation capabilities, so that its LRL translation outcome will be more human-like(Anugu & Ramesh, 2020).

Considering the condition of LRLs, fixed linguistic rules of RBMT cannot fully cover the complex natural language. which often leads to mistranslations when words or structures have multiple possible interpretations (Ganesh et al., 2023). Moreover, RBMT struggles to scale, as expanding the system requires the continuous addition of new rules, which is both time-consuming and resource-intensive (Connor, 2018). SMT also struggles with idiomatic expressions whose translation outcomes often lack fluency and naturalness; HMT takes more computing power and complex algorithms to integrate different LRL translation approaches, which can create inconsistencies, and reduce system flexibility. These limitations constrain the promotion of traditional LRL machine translation. and nowadays, researchers attach more attention to LLM-based Machine Translation for LRLs, which may have better techniques to improve the quality of translation.

LLM-based Machine Translation for Low-Resource Languages

Prompt Engineering

Through natural language prompting, LLMs can perform translation tasks without explicit training on specific language pairs by desired prompts. In essence, this cross-lingual pattern-matching ability acquired during pre-training can be learned on the basis of zero-shot or few-shot learning, which is beneficial for LRLs.

Lu et al. (2024) proposed Chain-of-Dictionary Prompting (CoD), using multilingual dictionaries to enhance LLM translation, especially for LRLs. Bilingual dictionaries improved translation, and chained multilingual dictionaries provide a further boost. This highlights that integrating external lexical resources via prompts significantly boosts LRLs translation, which provides explicit lexical mapping and makes use of common dictionary data rather than scarce parallel corpora.

Similarly, Ghazvininejad et al. (2023) proposed Dictionary-based Phrase-level Prompting (DiPMT) for LLMs, specifically addressing the processing difficulties of rare words in translation, which are common in low-resource or domain transfer scenarios. Extensive experiments showed that DiPMT outperformed baselines in both low-resource MT and out-of-domain MT. This method highlights the importance of providing targeted vocabulary guidance to LLMs, especially for vocabulary that is under-represented in their training data.

Data Augmentation Techniques

Back-translation

Back-translation is a popular translation method, and experience has shown that it is an effective method, and it can also be used to enhance the translation quality of LRLs. By translating text into another language and then back to the original language, back-translation can generate pseudo-parallel sentence pairs through reverse translation models (Burchell & Birch And Kenneth Heafield, 2022). In addition to increasing the diversity of LRLs training data, back-translation also keeps the meaning unchanged and greatly improves the translation effect of LRLs.

Back-translation is a useful method for machine translation of LRLs based on LLMs. Tonja et al. (2023) explored the application of back-translation in Voreta-English translation and found that self-learning and fine-tuning based on real and synthetic datasets generated by back-translation can greatly improve BLEU scores. Even though there are more advanced models appearing, back-translation is still very important.

Synthetic Parallel Data Generation

Beyond traditional back-translation, LLMs provide new ways to generate synthetic parallel data, effectively enabling LRLs to utilize them. The ability of LLMs to generate coherent and contextually relevant text in multiple languages makes them powerful tools for creating artificial parallel corpora, which can then be used to train or fine-tune smaller, more efficient MT models. Nguyen et al. (2024) proposed assembling synthetic exemplars from a diverse set of HRLs to prompt LLMs to translate from any language into English. These prompts were then used to create intra-lingual exemplars to perform tasks in the target LRLs, indicating that their unsupervised prompting method performed comparably to supervised few-shot learning and could even outperform it in non-English translation tasks for many LRLs. This approach leverages the LLM's inherent multilingualism to create synthetic data.

Knowledge Distillation

For LRLs, knowledge distillation leverages high-resource language models to transfer knowledge to LRL models. It performs well in alleviating data scarcity issues for LRLs and is highly effective in reducing model size (De Gibert et al., 2023). This approach can lower computational and memory requirements without requiring smaller models to be trained on the same large datasets. By effectively utilizing the "dark knowledge" embedded in teacher models, it maintains performance comparable to large teacher models.

Song et al. (2025) investigated whether LLMs are the "silver bullet" for LRL machine translation. Crucially, they showed how knowledge distillation from large pre-trained teacher models can significantly improve the performance of small LLMs on LRL translation tasks. This provides compelling evidence that knowledge distillation is a viable strategy for bridging the performance gap for LRLs, allowing smaller, more deployable LLMs to inherit the translation prowess of their larger counterparts.

Transfer Learning

Transfer learning is a cornerstone of low-resource machine translation, enabling models to leverage knowledge acquired from high-resource languages or tasks to improve performance on data-scarce language pairs. With the emergence of LLMs, which inherently possess vast multilingual knowledge, transfer learning has become more powerful, enabling complex cross-language knowledge transfer.

Language Adaptation and Module Fine-tuning

Module fine-tuning of pre-trained LLMs provides evidence that may support its effectiveness, but full parameter fine-tuning of LLMs tends to lead to problems such as limited flexibility and the high costs of independent deployment. Parameter-efficient fine-tuning targets a small parameter subset to cut computational costs, vital for LRLs with limited data. It adapts pre-trained models while preserving general language abilities.

The Adapter approach enhances high-resource language performance by freezing the main parameters of pre-trained models and fine-tuning only a small number of injected adapter layers per language pair (Bapna & Firat, 2019). In LRL translation, however, due to the influence of limited training data, the fine-tuning performed by adapters on pre-trained models rarely reaches the training level of multilingual fine-tuning (Chronopoulou et al., 2023). Researchers use parameter-efficient methods to tailor adapter behavior for LRLs, focusing on language-specific details without compromising model generality (Sel & Hanbay, 2024). In previous studies,

fine-tuning using language-specific adapters was apparently able to substantially improve the translation performance of LRLs and reduce the computational cost and memory footprint of fine-tuning (Hu et al., 2021).

Among various parameter-efficient fine-tuning methods, LoRA has been proven to achieve full 16-bit fine-tuning performance and has been extensively validated across multiple tasks and LLMs (Zhang et al., 2023). LoRA can achieve fine-tuning of large-scale models using limited computational resources. The core challenge of LRLs translation lies in the scarcity of parallel corpora. Module fine-tuning significantly reduces the number of parameters that need to be learned by only adjusting specific sub-modules within the model. Meanwhile, LRLs have unique morphological and grammatical features, which module fine-tuning can optimize targeted sub-modules to improve translation capabilities for these languages.

Fine-tuning Models to Adapt to Target Languages

To enable LLMs to complete the corresponding tasks, instruction fine-tuning provides the model with explicit instructions or examples for specific tasks by training the model to understand and execute different instructions. This improves the zero-shot learning ability and gives LLMs the ability to follow instructions (Wei et al., 2021). This method aligns with the current situation of low-resource language translation. Alam et al. (2024) point out that instruction fine-tuning supports to what may represent a key factor for the success of LLMs; their tutorial explored its abilities, with what is particularly significant about these findings being a special emphasis on low-resource scenarios. This suggests that instruction fine-tuning can give LLMs greater understanding of translation tasks, making them more adaptable to LRLs.

Beyond the State of the Art: Challenges and Future Pathways

Challenges

The primary challenge in LRL machine translation is the scarcity of quality corpora. This includes a lack of adequate bilingual parallel corpora necessary for statistical and deep learning models, and monolingual corpora that, despite quantity, often suffer from poor quality and lack annotations. Consequently, large-scale pre-trained models like Transformers cannot effectively learn LRL semantics and grammar, leading to poor generalization and translation quality. While techniques like data augmentation and transfer learning help, they do not fundamentally resolve the corpus shortage.

This fundamental issue also affects the quality of LLMs translating LRLs across various methods. Regarding prompt engineering, relevant studies (Aycock et al., 2024; Tanzer et al., 2023) have shown that LLMs benefit more from parallel examples extracted from grammar books than from the grammatical explanations in a book. This highlights a limitation: while LLMs are powerful, their ability to infer translation rules from descriptive linguistic text via prompting might be overestimated, and direct parallel examples remain crucial.

Future Pathways

Parameter-efficient fine-tuning techniques offer a particularly promising avenue for translating LRLs using LLMs. These techniques, tailored to address the specific linguistic features of LRLs, can effectively alleviate the problem of data scarcity. By enabling fine-tuning with limited resources, they address a key challenge in translating these languages. Furthermore, combining different approaches can be used to produce more training data and translation effects from a computational perspective.

Exploring how LLMs handle mixed content of text, pictures and speech translation, is another important research direction of LLMs. The combination of multimodal approaches and LLMs has opened up a new path for LRLs translation. Using a variety of information to source speech images and texts can be a way to make up for the deficiency of single-modal data. The tutorial by Alam et al. (2024) on LLMs for LRLs also emphasizes the growing focus on speech and multimodality. This shows that more researchers feel that in the future the development of machine translation of LRL may depend on the combination of LLMs with speech recognition and visual processing ability.

Integrating different forms of information into LLM translation systems is a big challenge. This requires the development of powerful multimodal encoders to align different modal representations for coherent and accurate results. collecting multimodal data for LRLs is challenging. We may rely on expert annotations and knowledge of cultural context. Despite these challenges, the potential is substantial.

Conclusion

Current research has made great progress in solving these translation problems. Data augmentation techniques have shown effectiveness. There is evidence that generating back-translated parallel data can expand a limited corpus. Importantly, merging multilingual corpora improves the cross-language capabilities of LLMs and promotes knowledge transfer, thereby improving translation results in LRLs. PEFT methods like LoRA improve possibilities by reducing computing costs and making it easier for LLMs to adapt to LRLs. Knowledge distillation efficiently transfers expertise from large teacher models to smaller student models optimized for LRLs, enabling deployment in resource-constrained settings. Prompt engineering and instruction tuning have also shown great potential. LLMs in low-resource translation have achieved progress but remain hindered by data scarcity, particularly for rare and isolated languages. In addition, when using large-scale language models to translate LRLs, there are still problems such as language accuracy, fairness and scale limitations. The feasible direction of future development should be based on existing achievements. Efforts will be made to solve various existing shortcomings, continue to innovate efficient fine-tuning technologies such as Adapters and LoRA, and use intensive learning combined with manual feedback to improve translation quality. Furthermore, it is crucial to promote community-driven corpus construction and collect annotations to verify low-resource language data through community cooperation, thereby allowing these languages can gain access to more resources, thereby alleviating the problem of data scarcity.

References

Alam, F., Chowdhury, S. A., Boughorbel, S., & Hasanain, M. (2024). LLMs for Low Resource Languages in Multilingual, Multimodal and Dialectal Settings. *In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, 27–33.

Anugu, A., & Ramesh, G. (2020). A Survey on Hybrid Machine Translation. *E3S Web of Conferences*, *184*, 01061.

Arcan, D. T. N. P. B. R. C. M. M. a. M. (2019). Leveraging Rule-Based Machine Translation Knowledge for Under-Resourced Neural Machine Translation Models.

Aycock, S., Stap, D., Wu, D., Monz, C., & Sima'an, K. (2024). *Can LLMs Really Learn to Translate a Low-Resource Language from One Grammar Book?* (*arXiv preprint*. arXiv:2406.10216v2).

Bapna, A., & Firat, O. (2019). Simple, Scalable Adaptation for Neural Machine Translation. *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 1538–1548.

Burchell, L., & Birch And Kenneth Heafield, A. (2022). Exploring diversity in back translation for low-resource machine translation. *In Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, 67–79.

Chronopoulou, A., Stojanovski, D., & Fraser, A. (2023). Language-Family Adapters for Low-Resource Multilingual Neural Machine Translation. *In Proceedings of the Sixth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2023)* (pp. 177–191).

Connor, P. C. (2018). A Concept Specification and Abstraction-based Semantic Representation: Addressing the Barriers to Rule-based Machine Translation. arXiv: Computation and Language.

De Gibert, O., Vázquez, R., Aulamo, M., Scherrer, Y., Virpioja, S., & Tiedemann, J. (2023). Four approaches to low-resource multilingual NMT: The Helsinki submission to the AmericasNLP 2023 shared task. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)* (pp. 177–191).

Ganesh, S., Dhotre, V., Patil, P., & Pawade, D. (2023). A Comprehensive Survey of Machine Translation Approaches. *In 2023 6th International Conference on Advances in Science and Technology (ICAST)* (pp. 160-165).

Gehring, J., Auli, M., Grangier, D., Yarats, D., & Dauphin, Y. N. (2017). Convolutional sequence to sequence learning Proceedings of the 34th International Conference on Machine Learning - Volume 70, Sydney, NSW, Australia.

Ghazvininejad, M., Gonen, H., & Zettlemoyer, L. (2023). *Dictionary-based Phrase-level Prompting of Large Language Models for Machine Translation* (*arXiv preprint*. arXiv:2302.07856v1).

Hameed, D. A., & Al-Khateeb, B. (2025). Deep Learning Techniques for Machine Translation: A Survey. *Procedia Computer Science*, *258*, 1022–1037.

Hinton, G., Vinyals, O., & Dean, J. (2015). *Distilling the Knowledge in a Neural Network* (*arXiv preprint*. arXiv:1503.02531).

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). *LoRA: Low-Rank Adaptation of Large Language Models* (*arXiv preprint*. arXiv:2106.09685).

Irvine, A. (2013, June). Statistical Machine Translation in Low Resource Settings. In A. Louis, R. Socher, J. Hockenmaier, & E. K. Ringger, Proceedings of the 2013 NAACL HLT Student Research Workshop Atlanta, Georgia.

Lu, H., Yang, H., Huang, H., Zhang, D., Lam, W., & Wei, F. (2024). Chain-of-dictionary prompting elicits translation in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 958–976). Miami, FL, USA: Association for Computational Linguistics.

Nakov, P. I., & Ng, H. T. J. C. e. (2012). Improving Statistical Machine Translation for a Resource-Poor Language Using Related Resource-Rich Languages. 44(1), 179-222.

Prashanth Nayak, J. K., Rejwanul Haque, Andy Way. (2023). Instance-Based Domain Adaptation for Improving Terminology Translation. Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track, 222–234.

Sel, I., & Hanbay, D. (2024). Efficient Adaptation: Enhancing Multilingual Models for Low-Resource Language Translation. *Mathematics*, *12*(19), 3149.

Song, Y., Li, L., Lothritz, C., Ezzini, S., Sleem, L., Gentile, N., State, R., Bissyandé, T. F., & Klein, J. (2025). *Is LLM the Silver Bullet to Low-Resource Languages Machine Translation?* (*arXiv preprint*. arXiv:2503.24102).

Tanzer, G., Suzgun, M., Visser, E., Jurafsky, D., & Melas-Kyriazi, L. (2023). *A Benchmark for Learning to Translate a New Language from One Grammar Book* (*arXiv preprint*. arXiv:2307.01780v2).

Tonja, A. L., Kolesnikova, O., Gelbukh, A., & Sidorov, G. (2023). Low-Resource Neural Machine Translation Improvement Using Source-Side Monolingual Data. *Applied Sciences*, *13*(2), 1201.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., & Le, Q. V. (2021). *Finetuned Language Models Are Zero-Shot Learners* (*arXiv preprint*. arXiv:2109.01652v5).

Zhang, L., Zhang, L., Shi, S., Chu, X., & Li, B. (2023). *LoRA-FA: Memory-efficient Low-rank Adaptation for Large Language Models Fine-tuning* (*arXiv preprint*. arXiv:2308.03303).