International Journal of Computer Information Systems and Industrial Management Applications ISSN 2150-7988 Volume 18(4s) 2026 pp. 1059-1065
https://doi.org/10.70917/ijcisim-2026-2618
Article

A Hybrid Translation Pipeline for Low-Resource Dialects: Translating English to Ahirani Using NLLB and Rule-Based Adaptation

Neha Telrandhe[1], Rakesh Kadu[2]

[1,2]Ramdeobaba University (Shri Ramdeobaba College of Engineering and Management), Nagpur, Maharashtra, India.
Email: n.telrandhe@gmail.com, kadurk@rknec.edu

Abstract

Machine translation for low-resource languages remains a significant challenge due to the unavailability of large-scale parallel corpora and limited linguistic resources. This paper presents an exploratory study that compares two translation approaches for a low-resource dialect: a basic rule-based method utilizing a custom English-to-dialect dictionary, and a neural machine translation approach employing Meta AI's pre-trained No Language Left Behind (NLLB-200) model. The NLLB model was used to translate from English to standard Marathi, followed by a post-processing step to adapt the output to the dialect using dictionary-based substitutions. This stepwise pipeline allows us to observe the differences in output quality, grammatical correctness, and contextual accuracy. The results highlight the strengths and limitations of both approaches, offering insight into the feasibility and challenges of applying neural models to dialectal translation in low-resource settings.

Keywords: *Low-resource languages, machine translation, Ahirani dialect, rule-based translation, dictionary-based translation, NLLB, multilingual models, neural machine translation, dialect adaptation, hybrid approach.*

1. Introduction

Machine translation (MT) plays a pivotal role in breaking down language barriers, yet its application for low-resource languages remains a significant challenge. These languages often lack extensive parallel corpora, making it difficult to train high-quality translation systems. This study investigates a step-by-step approach to machine translation for a low-resource language, starting from rule-based methods and progressing to the use of deep learning models.

Initially, a dictionary-based translation system was implemented as a baseline, providing a simple yet interpretable solution to the translation task. Following this, a more structured approach using a CSV-based lookup mechanism was introduced to improve accuracy and resource efficiency. Although effective, these rule-based methods faced limitations, especially when dealing with complex sentence structures and specialized vocabulary.

To overcome these limitations, a pretrained sequence-to-sequence (Seq2Seq) model was applied to the translation task. This deep learning model was trained on the available dataset and evaluated against test and validation sets to measure its performance. The results show that, despite the challenges posed by the low-resource nature of the language, the Seq2Seq model outperforms the rule-based methods. However, difficulties persist in accurately translating intricate sentence structures and domain-specific terms due to the lack of sufficient training data. This research provides valuable insights into the feasibility of using machine translation models for low-resource languages and compares the advantages and limitations of rule-based and deep learning approaches in such contexts.

2. Literature Review

ALPAC (1966). The seminal report highlighted the early limitations of machine translation systems and emphasized the need for improved evaluation and linguistic foundations, influencing decades of MT research.

Cai et al. (2021). Introduce a method using monolingual translation memory to enhance NMT performance, which is particularly effective in scenarios with limited bilingual data.

Choudhary et al. (2020). Implement neural machine translation for Indian low-resource languages, identifying key constraints such as vocabulary sparsity and lack of linguistic tools. The same study also explores the sociolinguistic aspects of low-resource languages and dialects, emphasizing inclusive NLP development beyond purely technical solutions.

Escolano et al. (2019). Propose an incremental training technique to expand NMT systems from bilingual to multilingual, enabling better scalability for low-resource language integration.

Etman & Beex (2015). Offer a survey on dialect and language identification, a critical step in processing and translating dialect-rich text in multilingual MT systems.

Fadaee et al. (2017). Introduce data augmentation using rare word replacement to improve low-resource NMT, significantly boosting performance with minimal added data.

Fan et al. (2021). Describe the development of NLLB-200, a multilingual NMT model covering 200+ languages, enabling translation for many low-resource and underrepresented languages.

Gao et al. (2020). Propose soft contextual augmentation methods to diversify training data, improving NMT system robustness for low-resource language pairs.

Goyal et al. (2020). Explore the use of related languages to boost translation for low-resource targets, applying transfer learning techniques to exploit linguistic similarities.

Haddow et al. (2022). Provide a comprehensive survey on low-resource MT techniques, categorizing methods such as pivoting, pretraining, and unsupervised learning.

Hadgu et al. (2021). Present "Lesan," a working MT system for African low-resource languages, illustrating real-world deployment of neural models in underserved regions.

Hutchins (1995). Reviews the historical evolution of machine translation, from rule-based to statistical and neural approaches, providing background to modern hybrid strategies.

Hutchins (2004). Documents the Georgetown-IBM experiment of 1954, one of the earliest demonstrations of automatic MT, marking a foundational moment in the field.

Kogan (2020). Analyzes the genealogical structure of Indo-Aryan languages using lexicostatistics, aiding understanding of dialectal relationships for MT.

Kumar et al. (2020). Explore zero-shot translation between Hindi and its dialects (Bhojpuri, Magahi) using unsupervised methods, showing potential without parallel corpora.

Kumar et al. (2021). Focus on MT for low-resource language varieties, offering architectural and data-centric strategies for handling dialectal variation effectively.

Kumar et al. (2022). Introduce a speech corpus for Awadhi, Bhojpuri, Braj, and Magahi, supporting development of spoken language MT for low-resource Indian dialects.

Laskar et al. (2020). Develop a Hindi-Marathi cross-lingual model, dealing with script and syntactic differences between closely related Indian languages.

Mujadia & Sharma (2020). Apply NMT to Hindi-Marathi using a similar-language translation approach, demonstrating successful adaptation of MT across related linguistic domains.

Patel et al. (2019). Discuss MT challenges specific to Indian languages, including morphological richness and orthographic diversity, suggesting practical resolutions.

Philip (2019). Investigates face-to-face translation systems, pointing toward future integration of visual and speech modalities in low-resource MT.

Philip et al. (2020). Revisit low-resource classifications in Indian languages, challenging assumptions and providing empirical evidence for re-evaluation in MT research.

Premjith et al. (2019). Implement English-to-Indian language NMT using the MTIL corpus, validating the role of domain-specific data in improving translation quality.

Rama et al. (2017). Use computational techniques to study Gondi dialects, offering insights into intra-language variation relevant for dialectal MT design.

Shi et al. (2022). Survey NMT methods and advancements for low-resource languages, covering hybrid, transfer, and unsupervised techniques with current trends.

Stanford NLP Overview (n.d.). Provides a historical overview of NLP and MT evolution, highlighting technological transitions from rule-based to data-driven paradigms.

Tan et al. (2020). Provide a detailed review of NMT frameworks, datasets, and tools, useful for practitioners building MT systems in resource-constrained environments.

White (1985). An early account of MT system design and evaluation, relevant for understanding foundational approaches that predate neural advancements.

Wichmann (2019). Explores how to computationally distinguish between languages and dialects, critical for accurate classification in dialect-sensitive MT.

3. Methodology

This study implements a hybrid translation approach designed to translate English input into a low-resource dialect of Marathi. The method involves two independent modules: a neural translation system (NLLB-200) for generating syntactically correct standard Marathi, and a rule-based post-processing module for adapting the output into the dialectal variant. This stepwise pipeline allows us to explore the effectiveness of combining general-purpose pretrained models with linguistically controlled adaptations for dialects.

Step 1: English to Standard Marathi Translation (NLLB-200)

The first stage uses Meta AI's NLLB-200 model, a multilingual neural machine translation system capable of translating between 200+ languages. The input English sentence is processed using the model with language tags set to eng_Latn (source) and mar_Deva (target). The model generates fluent and grammatically accurate output in standard Marathi.

Step 2: Dialect Adaptation Using Rule-Based Substitution

In the second stage, the standard Marathi output is passed through a custom-built word-level dictionary that maps standard words to their dialectal equivalents. This mapping is applied via direct string replacement without altering the sentence structure. This step enables lexical adaptation of the NLLB output to reflect regional dialect usage.

Example mapping:

Standard Marathi	Ahirani
माझा	मन
आहे	शे

Table 1: Sample Mapping in Dialect Adaptation

Step 3: System Integration

The two stages are independently implemented but function sequentially. No feedback is passed between the components, and no joint training is used. The system's modularity allows future extension to include additional dialect layers or grammar correction modules.

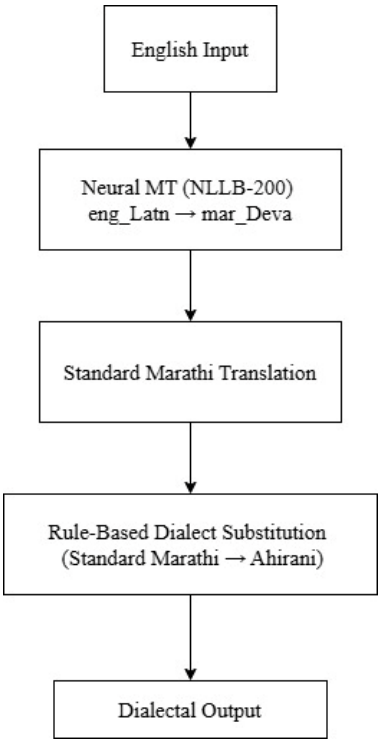

Fig. 1. Hybrid translation pipeline for English to Ahirani.

This highlights the potential of hybrid methods in addressing the challenges of machine translation for underrepresented and low-resource language variants.

4. Implementation and Results

The implementation of the machine translation system followed a step-by-step approach, starting with simpler, rule-based methods and progressively moving to deep learning models. The steps are outlined below.

4.1 Rule-Based Translation

Initially, a rule-based translation system was implemented using a small set of parallel words between the source and Ahirani languages. This provided a basic, interpretable solution for translating a limited number of words. The dictionary mapped words from the source language to their corresponding translations in the Ahirani language.

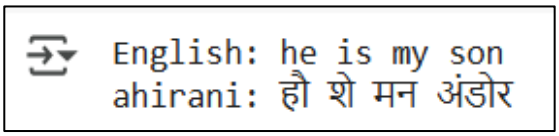

Fig. 2. Simple rule-based translation.

Grammar Structure Trial: To explore potential translation complexities, the grammar structure was experimented with. In particular, sentence structures such as Subject + Verb + Object (S+V+O) for the source language were mapped to Subject + Object + Verb (S+O+V) for the Ahirani language. This trial helped in addressing basic syntactic variations between the two languages using a rule-based approach.

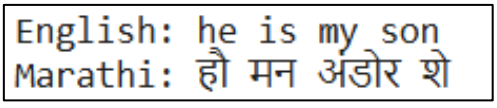

Fig. 3. Rule-based translation using Ahirani grammar structure.

4.2 Rule-Based Translation Using a CSV File

A more advanced rule-based approach was implemented using a CSV file containing 200 parallel sentence pairs from the source and Ahirani languages. This CSV file served as a small parallel corpus

that allowed the system to reference source-Ahirani sentence pairs for translation, improving accuracy compared to the earlier rule-based method.

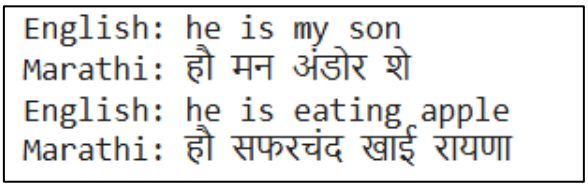

Fig. 4. Rule-based translation using a CSV file.

4.3 NLLB-200 Model

The experiment demonstrates that while rule-based translation provides control over specific vocabulary and dialectal terms, it often lacks grammatical correctness and fluency. On the other hand, the NLLB-200 model produces more fluent and grammatically accurate translations in standard Marathi but does not natively account for dialectal variation. By combining the strengths of both approaches — neural fluency and rule-based dialect adaptation — we can generate outputs that are not only structurally sound but also more culturally and linguistically aligned with the target dialect.

Hybrid Translation Pipeline

The hybrid pipeline translates English input into a dialectal variant of Marathi using a combination of:

1. A pretrained neural model (NLLB-200) for translation to standard Marathi.
2. A custom rule-based substitution to adapt the output into Ahirani.

Use Case 1: Hybrid Translation on a Basic Sample Dictionary

The implementation was carried out considering two types of outputs to give better visualization of the results. Figure 5 shows the direct hybrid translation outputs, while Figure 6 shows the individual, stepwise outputs of each approach.

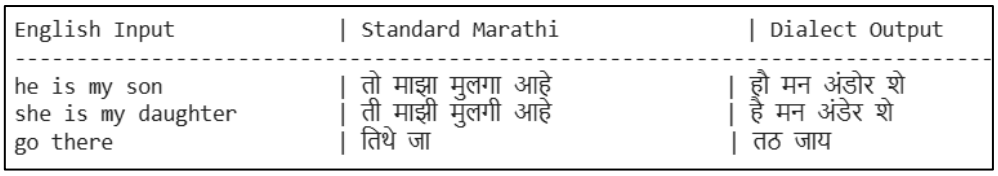

Fig. 5. Hybrid translation outputs using NLLB-200 with dialectal post-processing.

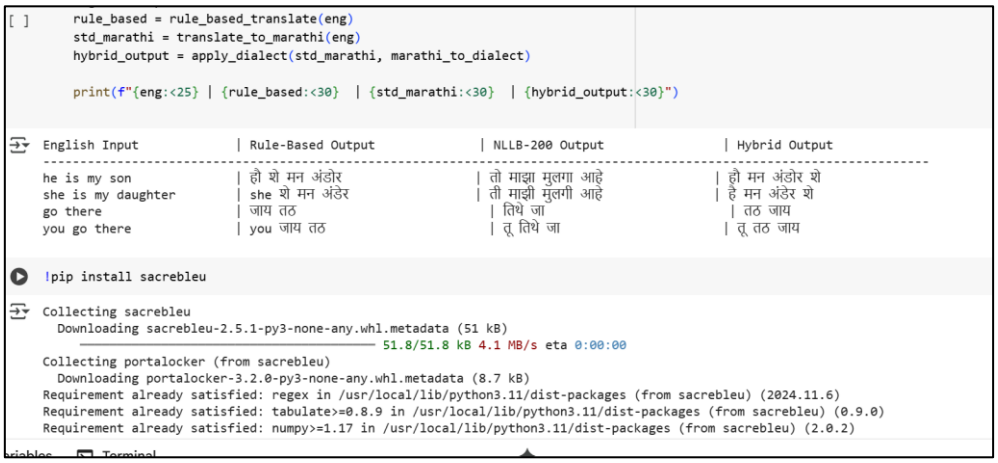

Fig. 6. Detailed stepwise sample output of every approach.

Considering these small samples, the BLEU scores predicted for the various methods are shown in Figure 7.

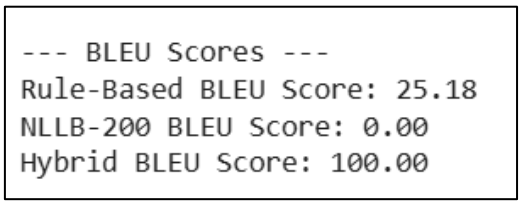

Fig. 7. BLEU scores for the rule-based, NLLB-200, and hybrid approaches.

5. Conclusion

This study explored a hybrid approach to machine translation for low-resource dialects, specifically translating from English to a Marathi dialect. By combining a simple rule-based dictionary method with the powerful multilingual NLLB-200 neural model, we were able to demonstrate how dialect-level adaptation can be layered onto standard machine translation outputs. While the dictionary-driven adaptation provided control over lexical choices, the NLLB model ensured syntactic and grammatical correctness.

Although the current demonstration was performed on a small, handcrafted dataset, the approach shows promising potential. Applying this method to a larger and more diverse dataset would allow for a deeper evaluation of its effectiveness, robustness, and limitations. This work lays the groundwork for future research on dialect-sensitive translation systems in low-resource language settings.

References

[01] Haddow, B., Bawden, R., Miceli Barone, A. V., Helcl, J., & Birch, A. Survey of low-resource machine translation. *Computational Linguistics*, 48(3), 673–732. (2022).

[02] Kumar, R., Singh, S., Ratan, S., Raj, M., Sinha, S., Lahiri, B., Seshadri, V., Bali, K., & Ojha, A. K. Annotated speech corpus for low-resource Indian languages: Awadhi, Bhojpuri, Braj and Magahi. In *Proceedings of the 1st Workshop on Speech for Social Good (S4SG)* (pp. 1–5). (2022).

[03] Fan, A., et al. Beyond English-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107), 1–48. (2021).

[04] Hadgu, A. T., Aregawi, A., & Beaudoin, A. Lesan — Machine translation for low-resource languages. arXiv preprint arXiv:2112.08191. (2021).

[05] Cai, D., Wang, Y., Li, H., Lam, W., & Liu, L. Neural machine translation with monolingual translation memory. In *Proceedings of the 59th Annual Meeting of the ACL and the 11th IJCNLP* (Vol. 1, pp. 7307–7318). Online: Association for Computational Linguistics. (2021).

[06] Kumar, S., Anastasopoulos, A., Wintner, S., & Tsvetkov, Y. Machine translation into low-resource language varieties. In *Proceedings of the 59th Annual Meeting of the ACL and the 11th IJCNLP* (Vol. 2: Short Papers, pp. 110–121). Online: Association for Computational Linguistics. (2021).

[07] Choudhary, H., Rao, S., & Rohilla, R. Neural machine translation for low-resourced Indian languages. In *Proceedings of the 12th LREC* (pp. 3610–3615). Marseille, France: European Language Resources Association. (2020).

[08] Goyal, V., Kumar, S., & Sharma, D. M. Efficient neural machine translation for low-resource languages via exploiting related languages. In *Proceedings of the 58th Annual Meeting of the ACL: Student Research Workshop* (pp. 162–168). Association for Computational Linguistics. (2020).

[09] Kogan, A. Genealogical classification of New Indo-Aryan languages and lexicostatistics. *Journal of Language Relationship*, 14(3–4), 227–258. (2016)

[10] Kumar, A., Mundotiya, R. K., & Singh, A. K. Unsupervised approach for zero-shot experiments: Bhojpuri–Hindi and Magahi–Hindi @ LoResMT 2020. In *Proceedings of the 3rd Workshop on Technologies for MT of Low-Resource Languages* (pp. 43–46). Suzhou, China: Association for Computational Linguistics. (2020).

[11] Laskar, S. R., Khilji, A. F. U. R., Pakray, P., & Bandyopadhyay, S. Hindi-Marathi cross lingual model. In *Proceedings of the Fifth Conference on Machine Translation* (pp. 396–401). Online: Association for Computational Linguistics. (2020).

[12] Mujadia, V., & Sharma, D. M. NMT based similar language translation for Hindi-Marathi. In *Proceedings of the Fifth Conference on Machine Translation (WMT20)*. Online: Association for Computational Linguistics. (2020).

[13] Philip, J., Siripragada, S., Namboodiri, V. P., & Jawahar, C. V. Revisiting low resource status of Indian languages in machine translation. In *8th ACM IKDD CODS and 26th COMAD (CODS-COMAD 2021)* (pp. 178–187). https://doi.org/10.1145/3430984.3431026 (2021)

[14] Escolano, C., Costa-jussà, M. R., & Fonollosa, J. A. R. From bilingual to multilingual neural machine translation by incremental training. In *Proceedings of the 57th Annual Meeting of the ACL: Student Research Workshop* (pp. 236–242). Florence, Italy: Association for Computational Linguistics. (2019).

[15] Gao, F., et al. Soft contextual data augmentation for neural machine translation. In *Proceedings of the 57th Annual Meeting of the ACL* (pp. 5539–5544). Florence, Italy: Association for Computational Linguistics. (2019).

[16] Patel, R. N., Pimpale, P. B., & Sasikumar, M. Machine translation in Indian languages: Challenges and resolution. *Journal of Intelligent Systems*, 28(3), 437–445. (2019). *(year confirmed correct — 2019, not 2017)*

[17] Philip, J. Towards automatic face-to-face translation. In *MM '19: Proceedings of the 27th ACM International Conference on Multimedia* (pp. 1428–1436). https://doi.org/10.1145/3343031.3351066 (2019).

[18] Premjith, B., Kumar, M. A., & Soman, K. P. Neural machine translation system for English to Indian language translation using MTIL parallel corpus. *Journal of Intelligent Systems*, 28(3), 387–398. (2019).

[19] Fadaee, M., Bisazza, A., & Monz, C. Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the ACL*. Vancouver, Canada: Association for Computational Linguistics. (2017).

[20] Rama, T. Computational analysis of Gondi dialects. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects* (pp. 26–35). Valencia, Spain: Association for Computational Linguistics. (2017).

[21] Etman, A., & Beex, A. A. Language and dialect identification: A survey. In *2015 SAI Intelligent Systems Conference (IntelliSys)*. https://doi.org/10.1109/IntelliSys.2015.7361147 (2015).

[22] Hutchins, J. The Georgetown-IBM demonstration, 7 January 1954. *AMTA (Association for Machine Translation in the Americas) column, MT News International*. (2004).

[23] Hutchins, J. Machine translation: A brief history. In E.F.K. Koerner & R.E. Asher (Eds.), *Concise history of the language sciences: From the Sumerians to the cognitivists* (pp. 431–445). Oxford: Pergamon Press. (1995).

[24] White, J. S., & O'Connell, T. A. Adaptation of the DARPA machine translation evaluation paradigm to end-to-end systems. In *Proceedings of the Second Conference of the Association for Machine Translation in the Americas*. Montreal, Canada. (1996) [25] ALPAC. Language and machines: Computers in translation and linguistics. Automatic Language Processing Advisory Committee, National Academy of Sciences, National Research Council. (1966).

[26] Stanford NLP Overview. A brief history of natural language processing. Stanford University. (n.d.).

[27] Wichmann, S. How to distinguish languages and dialects. *Computational Linguistics*, 45(4), 823–831. (2019).

[28] Shi, S., Wu, X., Su, R., & Huang, H. Low-resource neural machine translation: Methods and trends. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 21(5), Article 103. (2022).

[29] Tan, Z., Wang, S., Yang, Z., Chen, G., Huang, X., Sun, M., & Liu, Y. Neural machine translation: A review of methods, resources, and tools. *AI Open*, 1, 5–21. (2020).