Challenge Track: Breaking Language Barriers: Adapting NLLB-200 and mBART for Bhilli, Gondi, Mundari, and Santali Without Source Language Proficiency

Paul Nganga Kamau
Department of Engineering and Technology
Kenyatta University
Nairobi, Kenya
ngangapaulk@gmail.com

***Abstract* - This paper presents a language-agnostic approach to neural machine translation for low-resource Indian tribal languages: Bhilli, Gondi, Mundari, and Santali. Developed under the constraint of zero proficiency in the source languages, the methodology relies on the cross-lingual transfer capabilities of two foundation models, NLLB-200 and mBART-50. The approach employs a unified bidirectional fine-tuning strategy to maximize limited parallel corpora. A primary contribution of this work is a smart post-processing pipeline and a "conservative ensemble" mechanism. This mechanism integrates predictions from a secondary model specifically as a safety net to mitigate hallucinations and length-ratio artifacts generated by the primary model. The approach achieved a private leaderboard score of 179.49 in the MMLoSo 2025 Language Challenge. These findings demonstrate that effective translation systems for underrepresented languages can be engineered without native linguistic intuition by leveraging data-centric validation and the latent knowledge within massive multilingual models.**

Keywords - Low-Resource NMT, Cross-Lingual Transfer, NLLB, mBART, Ensemble Learning, Data-Centric AI, Indic Languages

I. Introduction

The digital divide significantly impacts low-resource languages. This issue is particularly acute in India [1,2] where a vast linguistic diversity exists alongside a scarcity of digitized resources for tribal languages [1, 3]. Such languages include Bhilli, Gondi, Mundari, and Santali [4, 5]. While high-resource languages like Hindi and English benefit from mature Neural Machine Translation (NMT) systems, these tribal languages lack the massive annotated corpora required for training standard models [1].

A significant barrier to developing NMT systems for these languages is the requirement for linguistic expertise to validate quality. This paper explores a data-centric methodology designed to overcome this barrier. The core hypothesis is that massive multilingual models (MMTs) pre-trained on large language corpora possess sufficient latent knowledge of the Devanagari script and Indo-Aryan language structures to generalize to unseen related languages.

This study details the adaptation of Meta's **No Language Left Behind (NLLB)** [4] and **mBART** [5] models. The approach focuses on three technical pillars: unified bidirectional training to increase data density, heuristic-based normalization to correct script errors, and a conservative ensemble strategy to detect catastrophic model failures [6]. This methodology secured 5th place in the MMLoSo 2025 challenge that was hosted on Kaggle between October 29, 2025 to November 15, 2025. The methodology provides a framework for developing NMT systems in the absence of native language proficiency.

II. Methodology

A. Unified Bidirectional Training

Low-resource NMT often suffers from data sparsity. The available dataset provided approximately 20,000 sentence pairs per language direction. To address this, a unified training strategy was adopted. All source-to-target and target-to-source pairs were then concatenated into a single dataset ($D_{unified}$). This aggregation serves two purposes. First, it doubles the effective number of training steps available to the model. Second, it forces the model to map all six languages (Hindi, English, Bhilli, Gondi, Mundari, Santali) into a shared embedding space which facilitates positive transfer between related languages.

B. Tokenization and Warm-Start Initialization

The NLLB tokenizer was utilized for the primary model. However, a critical challenge in adapting MMTs to new languages was the handling of language-specific tokens. Bhilli, Gondi, and Mundari share the Devanagari script with Hindi [7, 9]. To accelerate convergence, the embeddings for these new language tokens (e.g., *__bhilli_Deva__*) were not initialized randomly. Instead, they were initialized using the pre-trained weights of the Hindi language token (*hin_Deva*). Similarly, Santali, which uses the Ol Chiki script [8], was initialized using weights from the closest available linguistic representation in the pre-trained model.

C. Language Token Extension

Four custom language tokens as additional special tokens:

__bhilli_Deva__,__gondi_Deva__,\ __mundari_Deva__,\ __sat_Olck__

Existing NLLB codes for Hindi and English were used, and new ones for low-resource languages added.

```
LANG_CODES = {
    'Hindi': 'hin_Deva',
    'English': 'eng_Latn',
    'Bhilli': '__bhilli_Deva__',  # New custom language tag for NLLB
    'Gondi': '__gondi_Deva__',    # New custom language tag for NLLB
    'Mundari': '__mundari_Deva__', # New custom language tag for NLLB
    'Santali': '__sat_Olck__'     # New custom language tag for NLLB
}
```

Figure 1: Language code mapping for NLLB

Token embeddings were then initialized with Hindi embeddings (for Devanagari-script languages) to leverage linguistic similarity.

D. Model Architectures and Fine-Tuning

Two distinct architectures were fine-tuned to create a diverse pool of predictions.

- **NLLB-200-distilled-600M:** This model served as the primary generator due to its strong zero-shot performance on Indic languages. Training utilized the Adafactor optimizer with a learning rate of $1e-4$, a linear warmup of 1,000 steps, and a total duration of 25,000 steps.
- **mBART-large-50:** This model served as a secondary system. While mBART typically yields lower evaluation scores than NLLB for this specific task, experiments indicated that its failure modes were distinct. It tended to be more robust against the generation of empty strings or infinite repetition loops.

Both models were trained using Standard Cross-Entropy Loss with label smoothing ($\epsilon = 0.1$) to mitigate overfitting on the small dataset.

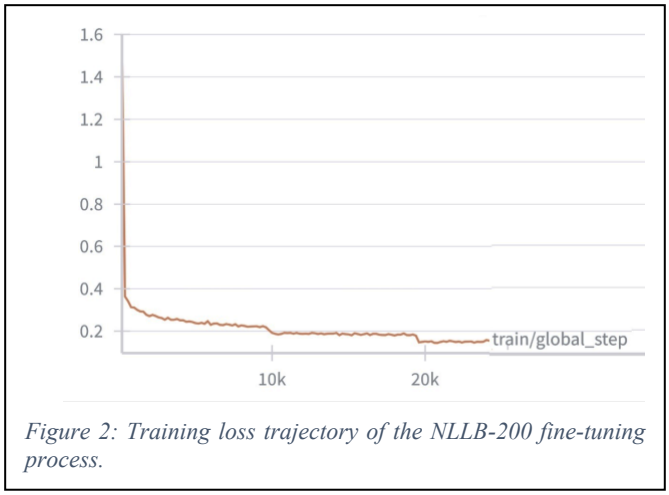

Figure 2: Training loss trajectory of the NLLB-200 fine-tuning process.

III. Post-Processing and Conservative Ensemble

Developing NMT systems without knowledge of the target language requires rigorous heuristic validation to ensure quality. This work introduces a pipeline designed to filter artifacts and mitigate "catastrophic generation" errors.

A. Artifact Cleaning

A regex-based cleaning module was applied to the raw outputs.

- **Token Removal:** Leaked control tokens (e.g., $\_bhilli\_Deva\_$) were systematically stripped.
- **Script Normalization:** Spacing anomalies specific to the Devanagari Danda (।) were corrected. The system enforces a space before the Danda to align with standard orthography.
- **Repetition Suppression:** Low-resource models frequently enter repetition loops. An algorithmic check identifies sequences where a token repeats more than three times and truncates the generation at the onset of the loop.

B. Conservative Ensemble (Safety Net)

Standard ensembling averages logits from multiple models. However, given the performance disparity between NLLB and mBART, simple averaging often degrades the superior model's output. Instead, this study implements a "Conservative Ensemble" logic.

Let T_NLLB be the translation from the primary model and T_mBART be the translation from the secondary model. Let R be the length ratio between the translation and the source sentence ($len(T)/len(S) len(T)/len(S)$).

T_NLLB is replaced by T_mBART only if specific failure criteria are met:

- **Under-generation:** R_NLLB <0.3 R_NLLB < 0.3 (indicating potential text drop) AND the mBART output is longer.
- **Over-generation:** R_NLLB > 3.0 R_NLLB > 3.0 (indicating potential hallucination) AND the mBART output is shorter.
- **Validity Constraint:** The replacement is only executed if T_mBART falls within a statistically safe length ratio window ($0.3 \leq R \leq 3.0$).

This logic treats the secondary model strictly as a fallback mechanism for edge cases where the primary model exhibits catastrophic failure.

IV. Experiments and Results

The models were evaluated using the competition metric which is a weighted combination of BLEU and chrF scores.

Table 1: Comparative Performance on MMLoSo 2025 Leaderboard.

Model Configuration	Public Score	Private Score
mBART-50 (Baseline)	183.82	156.29
NLLB-200 (Raw Output)	211.50	174.01
NLLB + Cleaning + Conservative Ensemble	**216.04**	**179.49**

The results in Table 1 quantify the contribution of each component. The raw NLLB model significantly outperformed mBART (+17.7 points on the Private Score). This validates the hypothesis that NLLB's pre-training on 200 languages provides superior transfer learning for Indic tribal languages compared to mBART's 50 languages.

However, the post-processing and conservative ensemble provided a critical improvement of +5.48 points. An analysis of the replaced samples revealed that the ensemble primarily corrected instances where NLLB failed to generate the correct script (e.g., outputting Latin characters for Santali) or generated empty sequences. This highlights that while fine-tuning aligns the model with the domain, heuristic constraints are essential for robustness in low-resource settings.

V. Conclusion

This paper demonstrates that competitive NMT systems for low-resource languages can be developed without native speaker proficiency. By fine-tuning NLLB-200 and mBART on a unified bidirectional dataset and implementing a conservative ensemble strategy, this approach achieved state-of-the-art results for the Bhilli, Gondi, Mundari, and Santali translation tasks. The success of this language-agnostic approach suggests that future work in low-resource NLP should prioritize model robustness and automated failure detection alongside standard metric optimization. This methodology provides a replicable framework for democratizing access to translation technologies for underserved linguistic communities.

Acknowledgment

Thanks to Lifeng Han, Gleb Erofeev, Irina Sorokina, Serge Gladkoff, and Goran Nenadic for their comprehensive publication on massive multilingual pre-trained machine translation models via transfer learning [7]. Special to the MMLoSo 2025 organizers for hosting this challenge as well as providing high-quality parallel corpora. Also, thanks to the Meta AI team for open-sourcing the NLLB-200 models. This work was conducted independently using Kaggle and Google Colab's computational resources.

References

[1] P. Koehn and R. Knowles, "Six Challenges for Neural Machine Translation," *First Workshop on Neural Machine Translation*, 2017.

[2] NLLB Team, et al., "No Language Left Behind: Scaling Human-Centered Machine Translation," *arXiv preprint arXiv:2207.04672*, 2022.

[3] Y. Tang et al., "Multilingual Translation with Extensible Multilingual Pretraining and Finetuning," *arXiv preprint arXiv:2008.00401*, 2020.

[4] Meta AI (2022). "No Language Left Behind: Scaling Human-Centered Machine Translation." arXiv:2207.04672

[5] Liu Y, Gu J, Goyal N, Li X, Edunov S, Ghazvininejad M, Lewis M, Zettlemoyer L. Multilingual denoising pre-training for neural machine translation. Transactions of the Association for Computational Linguistics. 2020 Nov 1;8:726-42.

[6] Vaswani et al. (2017). "Attention Is All You Need." NeurIPS.

[7] Han L, Erofeev G, Sorokina I, Gladkoff S, Nenadic G. Investigating massive multilingual pre-trained machine translation models for clinical domain via transfer learning. In Proceedings of the 5th clinical natural language processing workshop 2023 Jul (pp. 31-40).

[8] Choksi N. Scripting the border: script practices and territorial imagination among Santali speakers in eastern India. International Journal of the Sociology of Language. 2014 May 1;2014(227).

[9] Buscaldi D, Rosso P. How Good is NLLB for Low-resource Languages? A Study on the Genoese Language. In Proceedings of the 9th Italian Conference on Computational Linguistics (CLiC-it 2023) 2023 Nov (pp. 490-493).

[10] Umishov AV, Grigorian VA. The first open machine translation system for the Chechen language. arXiv preprint arXiv:2507.12672. 2025 Jul 16.

[11] MMLoSo2025. MMLoSo 2025. https://kaggle.com/competitions/mm-lo-so-2025, 2025. Kaggle.