BVSLP: Machine Translation using Linguistic Embellishments for IndicMT Shared Task 2025

Nisheeth Joshi[1#], Palak Arora[2*], Anju Krishnia[1‡], Riya Lonchenpa[1], Mhasilenuo Vizo[1±]**
[1]Speech and Language Processing Lab, Banasthali Vidyapith, Rajasthan, India
[2]School of Computing, DIT University, Uttarakhand, India
[#]nisheeth.joshi@rediffmail.com, [*]palak.arora.pa55@gmail.com, [‡]aasvkrishnia@gmail.com,
[**]lonchenpar@gmail.com, mhasilenuovizo3@gmail.com

Abstract

This paper describes our submission to the Indic MT 2025 shared task, where we trained machine translation systems for five low-resource language pairs: English–Manipuri, Manipuri–English, English–Bodo, English–Assamese, and Assamese–English. To address the challenge of out-of-vocabulary errors, we introduced a Named Entity Translation module that automatically identified named entities and either translated or transliterated them into the target language. The augmented corpus produced by this module was used to fine-tune a Transformer-based neural machine translation system. Our approach, termed HEMANT (Highly Efficient Machine-Assisted Natural Translation), demonstrated consistent improvements, particularly in reducing named entity errors and improving fluency for Assamese–English and Manipuri–English. Official shared task evaluation results show that the system achieved competitive performance across all five language pairs, underscoring the effectiveness of linguistically informed preprocessing for low-resource Indic MT.

1 Introduction

This paper presents our submission to the IndicMT 2025 shared task, where we developed machine translation (MT) systems for five language pairs: English–Assamese, Assamese–English, English–Manipuri, Manipuri–English, and English–Bodo. The systems follow a two-stage pipeline comprising preprocessing and neural machine translation (NMT). In preprocessing, the source text undergoes tokenization, spelling normalization, and source-side linguistic analysis, with particular focus on identifying and translating named entities into the target language to reduce out-of-vocabulary (OOV) errors. The processed corpus is then used to train NMT models based on an encoder–decoder architecture, with subword segmentation applied via the SentencePiece tokenizer using the Byte Pair Encoding (BPE) model to improve vocabulary coverage and generalization. The resulting systems integrate named entity handling with standard NMT methods, thereby enhancing translation quality across the five language pairs.

2 Related Work

Research in Indic machine translation has increasingly emphasized the use of multilingual pretrained models and transfer learning to overcome data scarcity. The IndicTrans system (Ahuja et al., 2022) demonstrated the effectiveness of multilingual Transformer pretraining for Indian languages, providing strong baselines for Indo-Aryan and Dravidian pairs. More recently, No Language Left Behind (NLLB-200) (Costa-jussà et al., 2022) has scaled this paradigm further, offering pretrained models across 200 languages, including Assamese, thereby enabling robust fine-tuning for low-resource Indic settings.

The benefits of transfer learning for low-resource machine translation have been well-documented. Zoph and Knight (2016) showed that multi-source translation improves performance by leveraging related languages, while Nguyen and Chiang (2017) highlighted the effectiveness of cross-lingual transfer in low-resource neural MT. In the Indic context, transfer between closely related languages such as Bengali and Assamese or

Proceedings of the Tenth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 1265–1270
November 8–9, 2025.

Manipuri and Bodo provides an opportunity to exploit linguistic similarities for improved translation performance.

Subword segmentation strategies have also been an important focus. Sennrich et al. (2016) introduced Byte Pair Encoding (BPE) as a means to mitigate out-of-vocabulary issues, while Kudo and Richardson (2018) proposed the Unigram Language Model as an alternative. More recently, Ahmed et al. (2023), in the WMT 2023 shared task, compared segmentation schemes across low-resource languages and showed that alternative approaches can outperform BPE in certain settings.

Efforts to improve named entity handling in MT are relatively fewer. Joshi and Katyayan (2023) and Sharma et al. (2023) demonstrated that augmenting training corpora with entity translations significantly reduces OOV errors in English–Braille and Hindi–English systems, respectively. Our work extends this line by systematically integrating a Named Entity Translation module into the preprocessing pipeline for Indic language pairs.

3 System Description

3.1 Data Preprocessing

Text tokenization and spelling correction were first performed on the source language corpus. Spelling normalization was applied to reduce orthographic inconsistencies in the corpora, particularly for Assamese and Manipuri. For Assamese, Unicode normalization was enforced, and common spelling variants arising from character duplication or visually similar graphemes were standardized using a manually crafted rule set. For Manipuri and Bodo, we implemented an edit-distance–based correction method (Levenshtein distance) supplemented with frequency statistics from the training corpus. Candidate corrections were chosen from a lexicon compiled from Wikipedia dumps, news portals, and government gazetteers, with the most frequent form selected when multiple candidates were available. This preprocessing step reduced vocabulary sparsity and improved token consistency prior to subword segmentation.

Subsequently, named entities were extracted using the in-house developed Bi-LSTM–based POS tagger (Nathani et al., 2023). The extracted named entities were then classified into the MUC-6 categories (Grishman et al., 1996) through a rule-based approach. These annotated entities were cross-referenced with a knowledge base containing target language translations for source language organization and location names.

The Named Entity Translation module relied on resources compiled from multiple sources. Gazetteers of Indian locations and organization names were collected from publicly available repositories such as the Wikipedia category lists, and official government publications. Entities not present in the knowledge base were transliterated using a rule-based transliteration scheme, which maps source graphemes to phoneme-equivalent representations before rendering them in the target script. This approach preserved phonological similarity across languages, ensuring that named entities remained intelligible even in the absence of dictionary support. In future work, we plan to explore phoneme-to-grapheme transliteration models trained using transformer-based sequence-to-sequence architectures for improved accuracy.

A rule-based NER system was employed to extract named entities from the source language corpus (Suri et al., 2024). Once identified, these entities were searched in the knowledge base for their corresponding English translations. When a translation was available, the entity in the source language corpus was replaced with its target language equivalent. In cases where no translation was present in the knowledge base, the entities were instead transliterated into the target language and then replaced in the corpus.

This process constituted the Named Entity Translation module, which systematically identified named entities and translated or transliterated them into target language, depending on the availability of translations (Sharma et al., 2023; Joshi & Katyayan, 2023). The functioning of this module is illustrated in Figure 1.

A BiLSTM-based POS tagger was used to bootstrap named entity recognition, as gold-standard BIO-annotated NER corpora for Manipuri and Bodo were not available at the time of system development. POS tags provided coarse-grained syntactic cues (e.g., proper noun categories) which, when combined with rule-based heuristics, enabled the identification of named entities. While this approach proved effective in resource-constrained settings, we acknowledge its limitations compared to transformer-based BIO tagging. Future extensions of this work will evaluate pre-trained multilingual NER models such as IndicNER and

XLM-R–based fine-tuned taggers, which are expected to improve robustness.

3.2 Sub-Wording

We adopted Byte Pair Encoding (BPE) for subword segmentation using SentencePiece. BPE was selected due to its effectiveness in balancing vocabulary compactness and coverage, a crucial factor for low-resource settings where unseen tokens are frequent (Sennrich et al., 2016). Although alternative segmentation schemes such as the Unigram Language Model (Kudo & Richardson, 2018) have been shown to perform competitively in recent work (Ahmed et al., 2023; WMT 2023 Shared Task Report), preliminary trials indicated that BPE produced more stable vocabularies across our diverse Indic language pairs. A systematic comparison with alternative subword models remains an avenue for future research.

3.3 Training the Model

For the training of the NMT systems, preprocessing steps were applied. This process was as follows: Part-of-Speech (POS) tagging was first applied to source language sentences, after which Named Entity Recognition (NER) was conducted using a rule-based module. The identified named entities were then translated or transliterated according to the procedure described in the previous section, thereby producing an augmented source sentence for the training corpus.

For example, consider the Sindhi sentence: "নিশীথ জোশী নতুন দিল্লী ৰ ইন্দিৰা গান্ধী আন্তর্জাতিক বিমানবন্দৰ পৰা জয়পুৰলৈ যাত্রা কৰি আছিল।" In this sentence, "নিশীথ জোশী (Person), নতুন দিল্লী (Location), জয়পুৰ (Location)," and "ইন্দিৰা গান্ধী আন্তর্জাতিক বিমানবন্দৰ (Organization)" are named entities. Among these, "নিশীথ জোশী" and "জয়পুৰ" were not present in the knowledge base and were therefore transliterated as "Nisheeth Joshi" and "Jaipur." The remaining entities were looked up in the knowledge base sequentially. While "নতুন দিল্লী" was not found and was thus transliterated as "New Delhi," "ইন্দিৰা গান্ধী আন্তর্জাতিক বিমানবন্দৰ" was translated as "Indira Gandhi International Airport."

Using this methodology, the entire training corpus was augmented with translated and transliterated named entities. The workflow of the system is illustrated in Figure 2, while the hyperparameters used for training both systems are summarized in Table 1. The overall approach is named as HEMANT (Highly Efficient Machine Assisted Natural Translation).

Our systems were built on top of the No Language Left Behind (NLLB-200) pretrained multilingual model (Costa-jussà et al., 2022), which supports several Indic languages including Assamese. Instead of training from scratch—which is often infeasible for low-resource settings due to limited parallel corpora—we adopted a transfer learning approach by fine-tuning NLLB-200 on the shared task training data provided by the organizers.

Parameter	Value
No. of Encoding Layers	6
No. of Decoding Layers	6
Early Stopping	
metric	bleu
min_improvement	0.2
steps	6
Optimizer	Adam
beta_1	0.8
beta_2	0.998
learning_rate	1.0
droupout	0.25
Regularization	
type	l1_l2
scale	1e-4
Minimum_learning_rate	0.00001
Max_steps	1000000

Tabel 1: Hyperparameters Used in Training NMT Models

The base architecture of NLLB-200 is a Transformer encoder–decoder model with multi-head attention, residual connections, and layer normalization, optimized for cross-lingual transfer. Fine-tuning was carried out by unfreezing all layers of the model, while maintaining pretrained multilingual subword embeddings. Subword embeddings were initialized from the NLLB model's vocabulary, which itself is trained using a combination of BPE and SentencePiece across 200 languages. This initialization ensured robust handling of rare tokens and morphologically complex forms.

Fine-tuning was performed separately for each language pair, with learning rate schedules tuned to prevent catastrophic forgetting of pretrained knowledge. For language pairs with extreme data scarcity (e.g., Bodo), multi-directional fine-tuning

was explored by jointly optimizing the model on related language pairs, leveraging cross-lingual similarity between Assamese and Bengali for Indo-Aryan, and Manipuri and Bodo for Tibeto-Burman families.

This strategy balanced the benefits of pretrained multilingual representations with task-specific adaptation. We found that fine-tuning NLLB-200 significantly stabilized training compared to Transformer models trained from scratch, which often struggled to converge on the limited IndicMT corpora.

4 Evaluation

We participated in the shared task by training the models on the training corpus provided by the organizers and submitted the outputs generated by the systems, using the test corpus, for official evaluation. The corresponding results provided by the organizing team are presented in Table 2. For Assamese-English, English-Assamese, English-Manipuri and Manipuri-English language pairs; our system performed fairly well. We could not provide the same results for English-Bodo, possibly due to very less training corpus.

5 Conclusion

This paper presented HEMANT, our submission to the WMT 2025 Indic MT shared task, focusing on five low-resource language pairs. The integration of a Named Entity Translation module reduced out-of-vocabulary errors and improved translation fluency. While absolute scores remain modest, the relative improvements highlight the value of linguistically informed preprocessing.

In future work, we plan to explore cross-lingual transfer strategies by leveraging related languages (e.g., Bengali–Assamese). Compare alternative subwording methods (BPE vs Unigram LM). Incorporate backtranslation and synthetic data augmentation for extremely low-resource languages and replace POS-based NER heuristics with multilingual pretrained BIO tagging models.

Acknowledgments

This work is supported by the funding received from the Ministry of Electronics and Information Technology, Government of India for the project "English to Indian Languages and vice versa Machine Translation System" under National Language Translation Mission (NLTM): Bhashini through administrative approval no. 11(1)/2022-HCC(TDIL) Part 5 and funding received from Department of Science and Technology, Government of India through grant number DST/SHRIC/SHRI-24/2023 for project entitled, "Bidirectional Dhundhari-Hindi Machine Translation System".

References

Ahmed, T., Hasan, M. K., Hoque, M. T., & Sultana, N. (2023). A comparative study on subword segmentation strategies for low-resource neural machine translation. In *Proceedings of the Eighth Conference on Machine Translation (WMT 2023)* (pp. 912–920). Association for Computational Linguistics. https://aclanthology.org/2023.wmt-1.87

Ahuja, K., Dandapat, S., Dave, S., Khapra, M. M., Kumar, P., & Shrivastava, M. (2022). IndicTrans: An open-source model for transliteration and translation for Indic languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 127–137). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-demo.14

Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., ... & Guzmán, F. (2022). No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*. https://doi.org/10.48550/arXiv.2207.04672

Grishman, R., & Sundheim, B. M. (1996). Design of the MUC-6 evaluation. In *TIPSTER TEXT PROGRAM PHASE II: Proceedings of a Workshop held at Vienna, Virginia, May 6-8, 1996* (pp. 413–422). Association for Computational Linguistics.

Joshi, N., & Katyayan, P. (2023). Improving English-Bharti Braille machine translation through proper name entity translation. In *Proceedings of the 3rd International Conference on ICT for Digital, Smart, and Sustainable Development (ICIDSSD 2022)* (pp. 168–174). European Alliance for Innovation.

Kudo, T., & Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 66–71). Association for Computational Linguistics.

Nathani, B., Arora, P., Joshi, N., Katyayan, P., Rathore, S. S., & Dadlani, C. P. (2023). Sindhi POS tagger using LSTM and pre-trained word embeddings. In *XVIII International Conference on Data Science

and Intelligent Analysis of Information* (pp. 37–45). Springer Nature Switzerland.

Nguyen, T. Q., & Chiang, D. (2017). Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the 8th International Joint Conference on Natural Language Processing (IJCNLP 2017)* (pp. 296–301). Association for Computational Linguistics. https://aclanthology.org/I17-1050

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318). Association for Computational Linguistics.

Sennrich, R., Haddow, B., & Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1715–1725). Association for Computational Linguistics.

Sharma, R., Katyayan, P., & Joshi, N. (2023). Improving the quality of neural machine translation through proper translation of name entities. In *2023 6th International Conference on Information Systems and Computer Networks (ISCON)* (pp. 1–4). IEEE.

Suri, D., Malviya, N., & Joshi, N. (2024). Rule-based named entity recognition for Hindi. In *International Conference on Artificial Intelligence and Speech Technology* (pp. 55–62). Springer Nature Switzerland.

WMT 2023 Shared Task Report. (2023). Findings of the WMT 2023 shared tasks on machine translation. In *Proceedings of the Eighth Conference on Machine Translation (WMT 2023)* (pp. 1–45). Association for Computational Linguistics. https://aclanthology.org/2023.wmt-1.0

Zoph, B., & Knight, K. (2016). Multi-source neural translation. In *Proceedings of NAACL-HLT 2016* (pp. 30–34). Association for Computational Linguistics. https://doi.org/10.18653/v1/N16-1004

1

Language Pair	BLEU	Meteor	ROUGE-L	chrf	TER
As-En	14.90698584	0.6152953104	0.6125623657	60.29148355	71.32659784
En-As	1.810109903	0.05849541589	0.003	27.45025141	98.66174223
En-Mni	4.15145205	0.1464458762	0.009866666667	41.43192248	89.59891851
Mni-En	3.056814809	0.2212412906	0.2506661424	35.61296964	139.025647
En-Bod	1.350078456	0.04016354522	0.1678	17.05109642	106.111629

Tabel 2: Evaluation Results

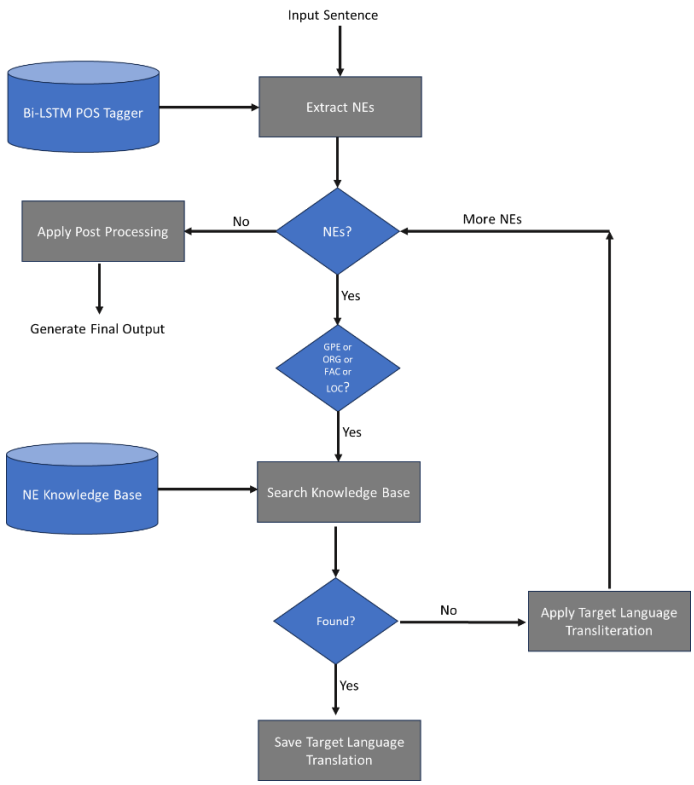

Figure 1: Named Entity Translation Module

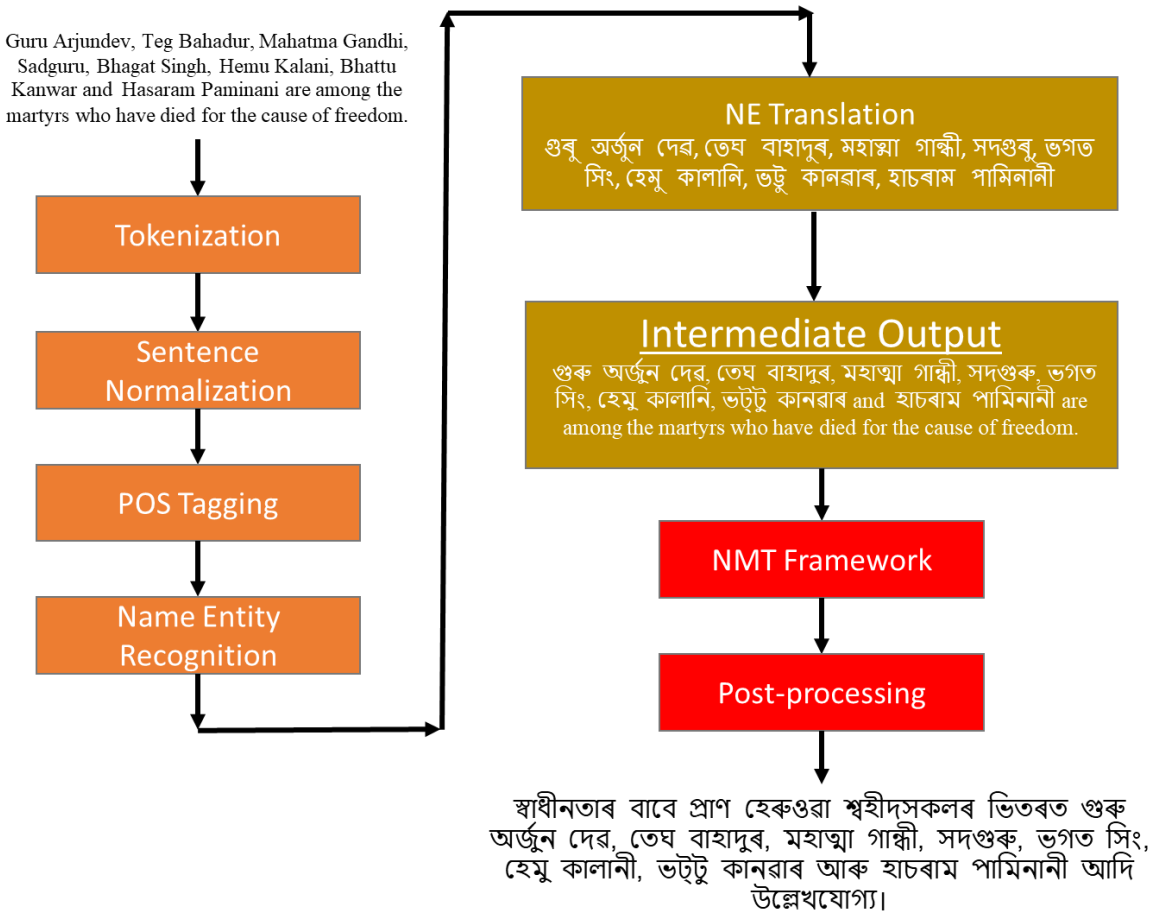

Figure 2: HEMANT Approach