A Preliminary Exploration of Phrase-Based SMT and Multi-BPE Segmentations through Concatenated Tokenised Corpora for Low-Resource Indian Languages

Saumitra Yadav and **Manish Shrivastava**
MT-NLP Lab, LTRC, KCIS, IIIT Hyderabad, India
saumitra.yadav@research.iiit.ac.in
m.shrivastava@iiit.ac.in

Abstract

This paper describes our methodology and findings in building Machine Translation (MT) systems for submission to the WMT 2025 Shared Task on Low-Resource Indic Language Translation. Our primary aim was to evaluate the effectiveness of a phrase-based Statistical Machine Translation (SMT) system combined with a less common subword segmentation strategy for languages with very limited parallel data. We applied multiple Byte Pair Encoding (BPE) merge operations to the parallel corpora and concatenated the outputs to improve vocabulary coverage. We built systems for the English–Nyishi, English–Khasi, and English–Assamese language pairs. Although the approach showed potential as a data augmentation method, its performance in BLEU scores was not competitive with other shared task systems. This paper outlines our system architecture, data processing pipeline, and evaluation results, and provides an analysis of the challenges, positioning our work as an exploratory benchmark for future research in this area.

1 Introduction

Machine Translation (MT) has advanced rapidly in recent years, primarily driven by neural architectures and the availability of large-scale parallel corpora. However, these benefits are often confined to high-resource languages (Koehn and Knowles, 2017), leaving many languages with little or no translation support (Gowda et al., 2021). The WMT 2025 Shared Task on Low-Resource Indic Languages Translation addresses this gap by focusing on languages spoken in India with scarce digital resources.

While many efforts adapt high-resource MT techniques to low-resource settings, direct transfer often fails due to the data-hungry nature of neural networks. This has led to research in areas such as word segmentation and other preprocessing strategies (Ding et al., 2019; Abid, 2020; Huck et al., 2017; Ortega et al., 2020; Lankford et al., 2021; Domingo et al., 2023; Lee et al., 2024) to make systems more viable under data constraints. Although our focus here is on bilingual MT, we acknowledge the rise of multilingual and decoder-only models. Our goal was to investigate how far statistical models, combined with multiple BPE segmentations (Poncelas et al., 2020), could be pushed in highly constrained settings.

Bilingual MT systems have been successfully developed for other under-represented languages, such as Cantonese–Mandarin (Liu, 2022), English–Luganda (Kimera et al., 2025), Wolof–French (Dione et al., 2022), Bavarian–German (Her and Kruschwitz, 2024), and English–Manipuri (Singh et al., 2023; Singh and Singh, 2022), often with transformer-based architectures and customised segmentation like BPE (Li et al., 2024).

Previous work (Yadav et al., 2019; Yadav and Shrivastava, 2021; Akhbardeh et al., 2021) has shown that, for some low-resource Indic languages, SMT can outperform NMT. For the WMT 2025 Shared Task (Pakray et al., 2025), we therefore chose SMT for our systems targeting English $\leftrightarrow$ {Assamese, Khasi, Manipuri}.

The organisers provided parallel corpora for English–Kokborok, English–Bodo, English–Nyishi, English–Manipuri, English–Khasi, English–Mizo, and English–Assamese, building on earlier iterations (Pakray et al., 2024). We set out to determine whether a robust, traditional method like SMT—enhanced with a multiple-BPE data augmentation technique—could remain a viable option in such low-resource scenarios. This paper describes our approach and analyses the performance of our submissions.

2 Background

Low-resource MT faces unique challenges due to the scarcity of high-quality parallel corpora. Data sparsity leads to out-of-vocabulary (OOV) issues

Proceedings of the Tenth Conference on Machine Translation, pages 1253–1258

and poor generalisation. While Neural Machine Translation (NMT) dominates for high-resource pairs, its high data requirements limit its applicability without substantial augmentation (Sennrich et al., 2016a) or multilingual transfer (Mahata et al., 2023; Johnson et al., 2017).

Phrase-based SMT is often more resilient to small data sizes. By learning from statistical alignments of phrase pairs, it can perform robustly with limited resources.

For languages in the Indic family, which often exhibit rich morphology, subword segmentation is an effective preprocessing step (Prabhugaonkar et al., 2014). BPE, in particular, balances word-level and character-level representations, reduces vocabulary size, and mitigates OOV problems. Inspired by Poncelas et al. (2020), we extended this idea by applying multiple BPE merge operations to produce diverse segmentations, concatenating them to create a richer training set.

3 Data

The shared task corpora were drawn from previous WMT datasets and new resources (Pal et al., 2023; Kakum et al., 2023; Pakray et al., 2024). After preprocessing, we obtained the training statistics shown in Table 1.

Language Pair	# Training Sentences
English–Khasi	26,000
English–Mizo	50,000
English–Assamese	54,000

Table 1: Training data statistics before data augmentation.

Our preprocessing steps included:

- For Latin-script languages: tokenisation, normalisation, and lowercasing using Moses (Koehn et al., 2007).
- For others: processing with the Indic NLP Library (Kunchukuttan, 2020).
- For each parallel corpus: training and applying BPE (Sennrich et al., 2016b) with merge operations of 500, 1000, 2000, and 3000.

The segmented corpora from each merge setting were concatenated and deduplicated (Poncelas et al., 2020), resulting in the statistics in Table 2.

Language Pair	# Training Sentences
English–Khasi	91,379
English–Mizo	186,918
English–Assamese	209,010

Table 2: Training data statistics after concatenation and deduplication of multi-BPE segmentations.

4 System Description

We used Moses (Koehn et al., 2007) for phrase-based SMT, with target-side KenLM language models (Heafield, 2011) trained on the corpora in Table 2. Each system was evaluated under four inference configurations:

- **1000 BPE** segmented source
- **2000 BPE** segmented source
- **3000 BPE** segmented source
- **Combined Hypothesis**, selecting the output with the highest probability among the above. We select translation that exhibit an average log-likelihood of -1.0 or higher, according to measurements taken by the fairseq-interactive tool.

5 Results

Our systems were tested on the official WMT 2025 English–Nyishi, English–Khasi, and English–Assamese sets. In all cases, performance ranked in the lower tier, with BLEU scores notably behind top-performing NMT systems that likely used external data or more advanced architectures. Full results are shown in Tables 3 and 4.

6 Discussion

The modest results highlight the limitations of SMT in this setting. Two main factors likely contributed:

1. SMT's inability to capture long-range dependencies and nuanced patterns compared to modern neural models.
2. The data augmentation via multiple BPE segmentations did not sufficiently overcome the extreme scarcity of parallel data, particularly for Nyishi.

Noisy and domain-specific terms in the provided corpora may have further impacted translation quality.

To English	Test Inferenceing Strategy	BLEU	METEOR	ROUGE-L	CHRF	TER	Cos Similarity
Assamese	Combine Hypothesis	0.3331	0.0229	0.0213	17.5032	286.2066	0.0775
	1000 BPE	0.3331	0.0229	0.0213	17.5032	286.2066	0.0775
	2000 BPE	0.3331	0.0229	0.0213	17.5032	286.2066	0.0775
	3000 BPE	0.3340	0.0230	0.0214	17.5031	286.2821	0.0775
Khasi	Combine Hypothesis	1.0536	0.0793	0.1112	19.4678	177.4341	0.2460
	1000 BPE	1.0935	0.0808	0.1138	19.2604	171.4254	0.2428
	2000 BPE	1.0604	0.0802	0.1111	19.4635	176.1309	0.2456
	3000 BPE	1.0461	0.0806	0.1114	19.5720	179.1631	0.2474
Nyishi	Combine Hypothesis	1.2657	0.0857	0.1209	23.4376	138.2275	0.2111
	1000 BPE	1.2557	0.0828	0.1191	23.2949	139.4495	0.2033
	2000 BPE	1.1942	0.0808	0.1158	22.9758	145.2652	0.2052
	3000 BPE	1.1885	0.0811	0.1127	23.3545	147.9239	0.2006

Table 3: Indic Language to English translation systems

English To	Test Inferenceing Strategy	BLEU	METEOR	ROUGE-L	CHRF	TER
Assamesse	Combine Hypothesis	2.9694	0.1126	0.0000	31.4566	107.3459
	1000 BPE	2.9256	0.1089	0.0000	30.4868	104.2561
	2000 BPE	3.0255	0.1137	0.0000	31.2960	107.3301
	3000 BPE	3.0287	0.1145	0.0000	31.6349	108.9135
Khasi	Combine Hypothesis	4.2570	0.1922	0.2555	26.7990	96.2399
	1000 BPE	4.2404	0.1885	0.2539	26.5474	94.7060
	2000 BPE	4.2277	0.1933	0.2550	26.7577	97.9426
	3000 BPE	4.0968	0.1938	0.2522	26.9003	100.6179
Nyishi	Combine Hypothesis	1.1870	0.0492	0.0781	20.3680	123.9258
	1000 BPE	1.2280	0.0493	0.0782	20.2094	120.4597
	2000 BPE	1.1843	0.0496	0.0771	20.4325	124.4046
	3000 BPE	1.1730	0.0503	0.0770	20.6541	127.1327

Table 4: English to Indian Language translation systems

7 Conclusion and Future Work

Our participation in the WMT 2025 Shared Task was an exploratory test of a multi-BPE augmentation strategy in an SMT framework for extremely low-resource Indic language pairs. While the method did not yield competitive results, it provides a clear baseline for SMT in these settings and reinforces the potential value of hybrid or neural approaches. Future work will explore SMT–NMT hybrids, fine-tuning large multilingual models on limited data, and advanced augmentation methods such as back-translation.

References

Wael Abid. 2020. The SADID evaluation datasets for low-resource spoken language machine translation of Arabic dialects. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6030–6043, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Shuoyang Ding, Adithya Renduchintala, and Kevin Duh. 2019. A call for prudent choice of subword merge operations in neural machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 204–213, Dublin, Ireland. European Association for Machine Translation.

Cheikh M. Bamba Dione, Alla Lo, Elhadji Mamadou Nguer, and Sileye Ba. 2022. Low-resource neural machine translation: Benchmarking state-of-the-art transformer for Wolof<->French. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6654–6661, Marseille, France. European Language Resources Association.

Miguel Domingo, Mercedes García-Martínez, Alexandre Helle, Francisco Casacuberta, and Manuel Herranz. 2023. How much does tokenization affect neural machine translation? In *Computational Linguistics and Intelligent Text Processing*, pages 545–554, Cham. Springer Nature Switzerland.

Thamme Gowda, Zhao Zhang, Chris Mattmann, and Jonathan May. 2021. Many-to-English machine translation tools, data, and pretrained models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 306–316, Online. Association for Computational Linguistics.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Wan-hua Her and Udo Kruschwitz. 2024. Investigating neural machine translation for low-resource languages: Using Bavarian as a case study. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 155–167, Torino, Italia. ELRA and ICCL.

Matthias Huck, Simon Riess, and Alexander Fraser. 2017. Target-side word segmentation strategies for neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 56–67, Copenhagen, Denmark. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Nabam Kakum, Sahinur Rahman Laskar, Koj Sambyo, and Partha Pakray. 2023. Neural machine translation for limited resources english-nyishi pair. *Sādhanā*, 48(4):237.

Richard Kimera, DongNyeong Heo, Daniela N. Rim, and Heeyoul Choi. 2025. Data augmentation with back translation for low resource languages: A case of english and luganda. In *Proceedings of the 2024 8th International Conference on Natural Language Processing and Information Retrieval*, NLPIR '24, page 142–148, New York, NY, USA. Association for Computing Machinery.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Anoop Kunchukuttan. 2020. The IndicNLP Library. https://github.com/anoopkunchukuttan/indic_nlp_library/blob/master/docs/indicnlp.pdf.

Seamus Lankford, Haithem Alfi, and Andy Way. 2021. Transformers for low-resource languages: Is féidir linn! In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 48–60, Virtual. Association for Machine Translation in the Americas.

Jungseob Lee, Hyeonseok Moon, Seungjun Lee, Chanjun Park, Sugyeong Eo, Hyunwoong Ko, Jaehyung Seo, Seungyoon Lee, and Heuiseok Lim. 2024. Length-aware byte pair encoding for mitigating oversegmentation in Korean machine translation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2287–2303, Bangkok, Thailand. Association for Computational Linguistics.

Fuxue Li, Beibei Liu, Hong Yan, Mingzhi Shao, Peijun Xie, Jiarui Li, and Chuncheng Chi. 2024. A bilingual templates data augmentation method for low-resource neural machine translation. In *Advanced Intelligent Computing Technology and Applications: 20th International Conference, ICIC 2024, Tianjin, China, August 5–8, 2024, Proceedings, Part III*, page 40–51, Berlin, Heidelberg. Springer-Verlag.

Evelyn Kai-Yan Liu. 2022. Low-resource neural machine translation: A case study of Cantonese. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 28–40, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Sainik Kumar Mahata, Dipanjan Saha, Dipankar Das, and Sivaji Bandyopadhyay. 2023. Transfer learning in low-resourced MT: An empirical study. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 646–650, Goa University, Goa, India. NLP Association of India (NLPAI).

John E Ortega, Richard Castro Mamani, and Kyunghyun Cho. 2020. Neural machine translation with a polysynthetic low resource language. *Machine Translation*, 34(4):325–346.

Partha Pakray, Reddi Krishna, Santanu Pal, Advaitha Vetagiri, Sandeep Dash, Arnab Kumar Maji, Saralin A. Lyngdoh, Lenin Laitonjam, Anupam Jamatia, Koj Sambyo, Ajit Das, and Riyanka Manna. 2025. Findings of WMT 2025 shared task on low-resource indic languages translation. In *Proceedings of the Tenth Conference on Machine Translation (WMT 2025) under the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Suzhou, China.

Partha Pakray, Santanu Pal, Advaitha Vetagiri, Reddi Krishna, Arnab Kumar Maji, Sandeep Dash, Lenin Laitonjam, Lyngdoh Sarah, and Riyanka Manna. 2024. Findings of WMT 2024 shared task on low-resource Indic languages translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 654–668, Miami, Florida, USA. Association for Computational Linguistics.

Santanu Pal, Partha Pakray, Sahinur Rahman Laskar, Lenin Laitonjam, Vanlalmuansangi Khenglawt, Sunita Warjri, Pankaj Kundan Dadure, and Sandeep Kumar Dash. 2023. Findings of the WMT 2023 shared task on low-resource Indic language translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 682–694, Singapore. Association for Computational Linguistics.

Alberto Poncelas, Jan Buts, James Hadley, and Andy Way. 2020. Using multiple subwords to improve English-Esperanto automated literary translation quality. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 108–117, Suzhou, China. Association for Computational Linguistics.

Neha R Prabhugaonkar, Jyoti Pawar, Apurva S Nagvenkar, Pushpak Bhattacharyya, Diptesh Kanojia, and Manish Shrivastava. 2014. Panchbhoota: Hierarchical phrase based machine translation systems for five indian languages.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. 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)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Kshetrimayum Boynao Singh, Ningthoujam Avichandra Singh, Loitongbam Sanayai Meetei, Sivaji Bandyopadhyay, and Thoudam Doren Singh. 2023. NITS-CNLP low-resource neural machine translation systems of English-Manipuri language pair. In *Proceedings of the Eighth Conference on Machine Translation*, pages 967–971, Singapore. Association for Computational Linguistics.

Salam Michael Singh and Thoudam Doren Singh. 2022. Low resource machine translation of english–manipuri: A semi-supervised approach. *Expert Syst. Appl.*, 209(C).

Saumitra Yadav, Vandan Mujadia, and Manish Shrivastava. 2019. A3-108 machine translation system for LoResMT 2019. In *Proceedings of the 2nd Workshop*

on Technologies for MT of Low Resource Languages, pages 64–67, Dublin, Ireland. European Association for Machine Translation.

Saumitra Yadav and Manish Shrivastava. 2021. A3-108 machine translation system for similar language translation shared task 2021. In *Proceedings of the Sixth Conference on Machine Translation*, pages 304–306, Online. Association for Computational Linguistics.