Data Augmentation for Manipuri-English Neural Machine Translation

Xiaojing Shen[1], Yves Lepage[1]
[1]Graduate School of Information, Production and Systems, Waseda University
jetshen@akane.waseda.jp, yves.lepage@waseda.jp

概要

Neural Machine Translation (NMT) for low-resource languages like Manipuri, a Sino-Tibetan language, is constrained by limited parallel corpora. This study applies a data augmentation technique, end sentence generation, to improve Manipuri-English NMT performance by creating additional parallel sentence pairs from existing datasets.

Experiments on three datasets—the EM corpus, PMIndia corpus, and WMT23 corpus— demonstrate that the proposed method consistently improves translation quality. For the WMT23 dataset, BLEU scores increased from 26.7 to 30.0 (Manipuri-to-English) and from 22.5 to 25.1 (English-to-Manipuri), with similar gains across other corpora.

Keywords: Neural Machine Translation, Data Augmentation, Low-Resource Language, Manipuri

1 Introduction

1.1 Background

Neural Machine Translation (NMT) has significantly advanced machine translation by leveraging deep learning techniques to achieve end-to-end translation. Despite its success, NMT faces substantial challenges for low-resource languages like Manipuri due to the scarcity of parallel corpora. Manipuri, a Sino-Tibetan language with unique linguistic features such as agglutinative morphology and Subject-Object-Verb (SOV) word order, poses additional difficulties for machine translation.

Large Language Models (LLMs), including ChatGPT-4o [1] and BLOOM [2], have shown limited effectiveness in handling low-resource languages. Experiments on the WMT23 corpus reveal that LLMs achieve BLEU scores of 8.2–8.9 for Manipuri-to-English translation and even lower scores for English-to-Manipuri, far below the performance of NMT systems fine-tuned on augmented datasets. These results highlight the need for novel strategies to improve translation performance for low-resource language pairs.

To address these challenges, this study applies a data augmentation technique known as end sentence generation to expand the training data for Manipuri-English NMT. By leveraging structural analogies within existing sentence pairs, this method generates additional parallel sentence pairs, enhancing the size and diversity of the training corpus. Experiments on datasets of varying sizes and qualities confirm the effectiveness of this approach in improving translation performance.

1.2 Contributions

- **Enhanced training data:** Introduced the end sentence generation technique to create high-quality parallel sentence pairs, addressing the scarcity of available data.
- **Performance improvements:** Achieved significant BLEU score gains across multiple datasets, demonstrating the effectiveness of data augmentation for low-resource NMT tasks.
- **Insights into corpus quality:** Highlighted the impact of parallel sentence alignment on NMT performance, emphasizing the importance of data quality in low-resource settings.

2 Methodology

2.1 End sentence generation

The idea of sentence generation using embeddings is illustrated by the notion of the middle sentence generation, introduced by [3]. A middle sentence serves as a bridge between a start sentence and an end sentence in an anal-

This work is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

ogy, and is derived by interpolating sentence embeddings. This approach has shown promise in generating meaningful intermediate sentences, especially in low-resource scenarios [3, 4].

Building upon this idea, [5] proposed the end sentence generation method, which extrapolates embeddings beyond the middle sentence to create new sentences. This method enables the exploration of broader semantic spaces, thus enhancing the diversity and size of the training dataset.

2.2 Renormalized end sentence formula

The end sentence embedding is computed using the renormalized formula:

$$\mathbf{e}_{\text{renorm}} = \frac{2\|\mathbf{m}\| - \|\mathbf{s}\|}{\|2\mathbf{m} - \mathbf{s}\|} \times (2\mathbf{m} - \mathbf{s}) \tag{1}$$

where $\mathbf{s}$, $\mathbf{m}$, and $\mathbf{e}$ represent sentence embeddings for the start, middle, and end sentences in the analogy $\mathbf{s} : \mathbf{m} :: \mathbf{m} : \mathbf{e}$, respectively. The renormalization step ensures that the generated embeddings are well-scaled and compatible with decoding mechanisms, addressing potential vector length discrepancies.

The renormalized end sentence formula builds on the work of [5], which demonstrated that this approach produces high-quality sentences with greater semantic consistency compared to basic extrapolation methods.

2.3 Optimization of sentence embeddings

The sentence embedding space used in this study is distilmBERT [6], a distilled version of BERT that significantly reduces the number of parameters while maintaining high performance. Its lightweight architecture and multilingual training make it particularly suitable for low-resource language tasks such as Manipuri-English NMT.

To decode sentence embeddings into natural language sentences, we employed the vector-to-sequence (vec2seq) model [7], which maps embeddings to coherent textual representations. This combination of distilmBERT for encoding and vec2seq for decoding provides a robust framework for end sentence generation.

2.4 Optimization of embedding space with BERT-flow

Sentence embeddings from distilmBERT often exhibit irregularities that can hinder vector arithmetic operations. To address this, BERT-flow [8] was used to project the embeddings onto a Gaussian latent space. This optimization enhances the semantic consistency of the embeddings and reduces computational errors during end sentence generation.

BERT-flow was fine-tuned specifically for this experiment, using the hyperparameters given in Table 1.

Parameter	**Value**
Batch size	64
Learning rate	1e-5
Number of layers	2
Hidden size	768
Dropout	0.1
Number of training steps	10,000
Optimizer	AdamW
Weight decay	0.01

表 1 Fine-tuning hyperparameters for distilmBERT

The combination of distilmBERT, vec2seq, and BERT-flow ensured that the generated sentence embeddings were both semantically meaningful and suitable for decoding. These optimizations played a crucial role in enhancing the performance of the end sentence generation method, particularly in the low-resource Manipuri-English translation task.

2.5 Integration with Manipuri-English NMT

The end sentence generation method was integrated into the Manipuri-English NMT pipeline to address the challenges of data scarcity. By applying this technique, the training corpus was significantly expanded, particularly for datasets like the WMT23 corpus [9] by a factor of around 3 (see Table 4). This resulted in enhanced translation performance, as reflected in improved BLEU scores.

This approach builds on established data augmentation methods, such as back-translation [10] and mix-up [11], while offering a formula-driven, scalable solution tailored to low-resource languages.

3 Experiment Setup

3.1 Configuration

For dataset preprocessing, SentencePiece [12] was used to perform subword tokenization, which is effective for handling low-resource languages with rich morphology. All experiments were conducted using the OpenNMT-py toolkit [13] with a Transformer-based architecture [14].

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

The Transformer model's key configuration parameters are detailed in Table 2.

Parameter	Value
Batch size	256
Optimizer	Adam
Learning rate	0.2
Decay method	Noam
Encoder layers	6
Decoder layers	6
Heads	8
Hidden size	512
Transformer ff layer size	2048
Attention dropout	0.15
Vocabulary size	20,000

表 2 Transformer model configuration.

3.2 Data

Three datasets were utilized to evaluate the Manipuri-English NMT models, representing varying sizes and degrees of parallelism (Table 3):

- **EM Corpus:** A comparable corpus with 125k Manipuri-English sentence pairs [15]. 95% of the sentence pairs have low alignment quality, with cosine similarity below 0.3, which poses challenges for NMT training.
- **PMIndia Corpus:** A high-quality, strictly parallel corpus containing 7k sentence pairs [16], sourced from official communications.
- **WMT23 Corpus:** A curated corpus with 24k highly parallel Manipuri-English sentence pairs [9], serving as a benchmark for evaluating augmentation strategies.

Corpus	Language	Sentences	Avg. Length
EM	Manipuri (mni)	124,975	21
	English (en)	124,975	26
PMIndia	Manipuri (mni)	7,419	15
	English (en)	7,419	19
WMT23	Manipuri (mni)	23,687	15
	English (en)	23,687	18

表 3 Summary of datasets used in experiments.

3.3 Evaluation

The performance of NMT models was evaluated using BLEU [17], chrF [18], and TER [19], which together capture lexical accuracy, fluency, and required edit operations. These metrics were chosen because:

- They are robust for low-resource settings with limited data.
- They do not rely on pre-trained language models, which may not adequately support Manipuri.

Neural metrics like BLEURT [20] and COMET [21], while effective in high-resource scenarios, were not used due to their reliance on extensive pre-training corpora, which are unavailable for Manipuri. BLEU, chrF, and TER provide a reliable alternative for evaluating translation quality in low-resource conditions.

4 Results and Analysis

This section evaluates the performance of Manipuri-English NMT models on three datasets— EM, PMIndia, and WMT23—before and after applying the proposed data augmentation method. It also compares these results with those of state-of-the-art Large Language Models (LLMs), ChatGPT-4o and BLOOM.

4.1 The impact of data augmentation

Table 4 summarizes the experimental results for both original and augmented datasets. The BLEU, chrF, and TER scores are reported for both translation directions (Manipuri-to-English and English-to-Manipuri).

- **EM corpus:** The EM corpus, being thematically comparable rather than strictly parallel, exhibited the lowest baseline performance, with BLEU scores of 6.1 (mni → en) and 3.5 (en → mni). After augmentation, the BLEU scores improved to 7.8 and 5.7, respectively. chrF scores also increased, highlighting the potential of data augmentation to enhance even loosely aligned corpora.
- **PMIndia corpus:** As a high-quality, small-scale dataset, PMIndia achieved baseline BLEU scores of 15.4 (mni → en) and 13.2 (en → mni). Augmentation increased BLEU scores to 17.8 and 15.2, demonstrating the effectiveness of the method on strictly parallel corpora.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

- **WMT23 corpus:** The WMT23 corpus, being the largest and most parallel dataset, achieved the highest scores. Augmentation boosted BLEU scores from 26.7 to 30.0 (mni → en) and from 22.5 to 25.1 (en → mni), with similar improvements observed for chrF.

4.2 Comparison with LLMs

The performance of ChatGPT-4o and BLOOM on the WMT23 test set is included for comparison. Both LLMs underperformed significantly compared to the augmented NMT models, with BLEU scores of 8.2 and 8.9 (mni → en) and 2.6 and 3.4 (en → mni), respectively. These results indicate the limited capability of LLMs in handling low-resource language pairs, underscoring the importance of fine-tuned NMT systems.

Model	Corpus	Size	BLEU	chrF	TER
		mni → en			
NITS-CNLP	WMT23	–	26.9	48.6	67.6
ChatGPT-4o	WMT23	–	8.2	24.5	89.6
BLOOM	WMT23	–	8.9	33.1	83.7
Original	WMT23	23,687	26.7	48.3	68.8
Augmented		71,061	**30.0**	**52.4**	69.1
Original	EM	124,975	6.1	19.0	73.3
Augmented		374,925	**7.8**	**21.5**	73.6
Original	PMIndia	7,419	15.4	33.9	72.0
Augmented		22,257	**17.8**	**35.5**	72.3
		en → mni			
NITS-CNLP	WMT23	–	22.7	48.3	70.0
ChatGPT-4o	WMT23	–	2.6	21.0	99.8
BLOOM	WMT23	–	3.4	27.9	96.4
Original	WMT23	23,687	22.5	47.9	69.7
Augmented		71,061	**25.1**	**49.2**	70.7
Original	EM	124,975	3.5	21.1	81.4
Augmented		374,925	**5.7**	**23.9**	81.3
Original	PMIndia	7,419	13.2	30.6	77.4
Augmented		22,257	**15.2**	**33.1**	77.5

表 4 Performance of NMT models and LLMs on Manipuri-English translation tasks with different datasets and data augmentation.

4.3 Discussion

The results demonstrate the significant advantages of data augmentation for low-resource NMT. While LLMs show promise in multilingual settings, their performance in low-resource language pairs like Manipuri-English remains subpar without fine-tuning. In contrast, specialized NMT models trained on augmented datasets achieve substantial improvements in BLEU and chrF scores, reaffirming the importance of tailored data augmentation techniques for low-resource MT tasks.

5 Conclusion and future work

This study applied the end sentence generation method to augment Manipuri-English NMT datasets, achieving the following key findings:

- **Data augmentation effectiveness:** End sentence generation added 249,950, 14,838, and 47,374 new sentence pairs to the EM, PMIndia, and WMT23 corpora, respectively. This led to consistent BLEU score improvements across all datasets. On the WMT23 corpus, the augmented model achieved BLEU scores of 30.0 (mni → en) and 25.1 (en → mni), significantly outperforming NITS-CNLP [22], which reported BLEU scores of 26.92 and 22.75 for the same translation directions.
- **Impact of dataset quality and size:** The WMT23 corpus, being the most well aligned, achieved the highest performance, while the EM corpus, despite benefiting from augmentation, required larger data volumes due to low alignment quality.
- **Comparison with LLMs:** Augmented NMT models outperformed ChatGPT-4o and BLOOM, demonstrating the necessity of fine-tuned systems for low-resource languages like Manipuri.

For future work, we propose the following directions:

- **Improving augmented data quality:** Employ grammar correction tools or advanced language models to refine generated sentence pairs.
- **Better alignment for comparable corpora:** Enhance weakly aligned datasets like the EM corpus using advanced embedding techniques.
- **Scaling to larger datasets:** Validate end sentence generation on larger or cross-domain datasets to improve scalability and generalizability.
- **Leveraging LLMs:** Fine-tune large language models or use them to generate high-quality parallel data for Manipuri-English translation tasks.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

参考文献

[1] OpenAI. Chatgpt-4o: Language model. https://chat.openai.com, 2024.

[2] Teven Le Scao, Angela Fan, et al. Bloom: A 176b parameter open-access multilingual language model. 2023.

[3] Pengjie Wang, Liyan Wang, and Yves Lepage. Generating the middle sentence of two sentences using pre-trained models: a first step for text morphing. In **Proceedings of the 27th annual meeting of the Association for Natural Language Processing**, pp. 1481–1485, 2021.

[4] Matthew Eget, Xuchen Yang, and Yves Lepage. A study in the generation of multilingually aligned middle sentences. In Zygmunt Vetulani and Patrick Paroubek, editors, **Proceedings of the 10th Language & Technology Conference (LTC 2023) – Human Language Technologies as a Challenge for Computer Science and Linguistics**, pp. 45–49, April 2023.

[5] Xiyuan Chen. Data augmentation for machine translation using the notion of middle sentences. Master's thesis, IPS, Waseda University, Kitakyushu, Japan, July 2023.

[6] Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 4512–4525, Online, November 2020. Association for Computational Linguistics.

[7] Liyan Wang and Yves Lepage. Vector-to-sequence models for sentence analogies. In **2020 International Conference on Advanced Computer Science and Information Systems (ICACSIS)**, pp. 441–446, 2020.

[8] Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. On the sentence embeddings from pre-trained language models. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 9119–9130, Online, November 2020. Association for Computational Linguistics.

[9] Santanu Pal, Partha Pakray, Sahinur Rahman Laskar, Lenin Laitonjam, Vanlalmuansangi Khenglawt, Sunita Warjri, Pankaj Kundan Dadure, and Sandeep Kumar Dash. Findings of the WMT 2023 shared task on low-resource Indic language translation. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, **Proceedings of the Eighth Conference on Machine Translation**, pp. 682–694, Singapore, December 2023. Association for Computational Linguistics.

[10] Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 489–500, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[11] Lichao Sun, Congying Xia, Wenpeng Yin, Tingting Liang, Philip Yu, and Lifang He. Mixup-transformer: Dynamic data augmentation for NLP tasks. In **Proceedings of the 28th International Conference on Computational Linguistics**, pp. 3436–3440, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[12] Taku Kudo and John Richardson. 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, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[13] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. OpenNMT: Open-source toolkit for neural machine translation. In Mohit Bansal and Heng Ji, editors, **Proceedings of ACL 2017, System Demonstrations**, pp. 67–72, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, **Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA**, pp. 5998–6008, 2017.

[15] Rudali Huidrom, Yves Lepage, and Khogendra Khomdram. EM corpus: a comparable corpus for a less-resourced language pair Manipuri-English. In Reinhard Rapp, Serge Sharoff, and Pierre Zweigenbaum, editors, **Proceedings of the 14th Workshop on Building and Using Comparable Corpora (BUCC 2021)**, pp. 60–67, Online (Virtual Mode), September 2021. INCOMA Ltd.

[16] Ashok Urlana, Pinzhen Chen, Zheng Zhao, Shay Cohen, Manish Shrivastava, and Barry Haddow. PMIndiaSum: Multilingual and cross-lingual headline summarization for languages in India. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Findings of the Association for Computational Linguistics: EMNLP 2023**, pp. 11606–11628, Singapore, December 2023. Association for Computational Linguistics.

[17] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, 2002.

[18] Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In **Proceedings of the Tenth Workshop on Statistical Machine Translation**, pp. 392–395, Lisbon, Portugal, 2015. Association for Computational Linguistics.

[19] Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In **Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers**, pp. 223–231, Cambridge, Massachusetts, USA, 2006. Association for Machine Translation in the Americas.

[20] Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 7881–7892, Online, July 2020. Association for Computational Linguistics.

[21] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 2685–2702, Online, November 2020. Association for Computational Linguistics.

[22] Kshetrimayum Boynao Singh, Avichandra Singh Ningthoujam, Loitongbam Sanayai Meetei, Sivaji Bandyopadhyay, and Thoudam Doren Singh. NITS-CNLP low-resource neural machine translation systems of English-Manipuri language pair. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, **Proceedings of the Eighth Conference on Machine Translation**, pp. 967–971, Singapore, December 2023. Association for Computational Linguistics.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).