Wordnet Enhanced Neural Machine Translation for Assamese-Bodo Low Resource Language Pair

Kuwali Talukdar, Shikhar Kumar Sarma, Kishore Kashyap, Ratul Deka, Bhatima Baro, Mirzanur Rahman, Farha Naznin
Department of Information Technology, Gauhati University, India
kuwalitalukdar@gmail.com, sks001@gmail.com, kb.guwahati@gmail.com, rdeka8258@gmail.com, bhatimaishaan@gmail.com. mr@gauhati.ac.in, farha.gu@gmail.com

Abstract

Neural Machine Translation (NMT) for low-resource language pairs, such as Assamese-Bodo, is hindered by the limited availability of parallel corpora. In this work, we enhance the NMT process by integrating WordNet resources specifically developed for Assamese and Bodo. The WordNet resources consist of parallel synsets mapped via synset IDs, where each synset contains synonymous words. Additionally, concept sentences tied to each synset provide high-quality, semantically equivalent parallel sentence pairs, which are directly utilized as training data. Our approach explores two methods: (1) injecting parallel synsets into the NMT training pipeline and (2) augmenting the training dataset with the WordNet-derived parallel corpus. We evaluate the effectiveness of these approaches using BLEU scores on Assamese-Bodo translations. The results demonstrate significant improvements in translation accuracy when incorporating WordNet resources, achieving BLEU score gains of up to 2-3 points compared to baseline models. This study highlights the potential of leveraging structured lexical databases like WordNet to improve NMT for low-resource languages.

1 Introduction

Neural Machine Translation has made significant progress in improving translation quality for widely spoken languages with abundant parallel corpora. However, for low-resource language pairs such as Assamese and Bodo, the scarcity of large-scale parallel datasets poses a substantial challenge to developing well performed and reliable NMT systems. Both Assamese and Bodo are prominent languages in Northeast India, with unique linguistic structures and limited digital resources. The lack of well-aligned, high-quality parallel corpora leads to lower translation performance when using standard NMT approaches. In an effort to address this issue, we explore the integration of WordNet resources into the NMT pipeline. WordNet is a lexical database that organizes words into sets of synonyms (synsets), each representing a unique concept. Assamese and Bodo WordNets contain parallel synsets, with synonymous words in both languages mapped to a shared synset ID. In addition, each synset is accompanied by concept sentences, which are direct translations of one another. These concept sentences form a high-quality, semantically aligned parallel corpus, which can significantly enhance the training process of NMT models. This paper presents a novel approach where Assamese-Bodo WordNet resources are utilized to enrich the NMT training pipeline. Specifically, we inject parallel synsets into the NMT model and augment the training dataset with WordNet-derived concept sentences. Our experiments show that these enhancements lead to notable improvements in translation accuracy, as measured by BLEU scores. By leveraging structured lexical databases, we aim to bridge the resource gap for Assamese and Bodo and demonstrate the broader applicability of WordNet resources in low-resource NMT scenarios.

2 Related Works

Neural Machine Translation (NMT) for low-resource languages has been an active area of research, but significant challenges persist due to the lack of parallel corpora. Various strategies have been developed to address this issue, such as transfer learning, data augmentation, and the use of linguistic resources like WordNet. This section reviews relevant works, focusing on the development of WordNet for Assamese and Bodo, and recent advances in NMT for these languages. The Assamese and Bodo WordNets were developed as part of efforts to create structured lexical databases for these underrepresented languages. Previous works on Assamese WordNet [Shikhar et al., 2010] and Bodo WordNet [Sarma et al., 2010] documented the creation of synsets that include synonymous words in Assamese and Bodo, linked by shared synset IDs. These resources have been expanded to cover a wide range of concepts, along with concept sentences that provide high-quality parallel sentences for both languages. The creation of these WordNets marked a significant milestone in digital linguistic resources for the Assamese-Bodo language pair, laying the groundwork for further applications in machine translation and NLP tasks.

The development of Assamese WordNet laid the groundwork for creating linguistic resources essential for natural language processing tasks. Shi8khar et al. (2010) discussed the design and implementation of the Assamese WordNet, focusing on building synsets and their application in enhancing bilingual machine translation quality. The integration of parallel synsets across languages has proven to be a useful technique to improve translation accuracy in Assamese-English bilingual systems. Similarly, the Bodo WordNet has undergone extensive research, providing a well-mapped synset structure in Bodo, contributing significantly to linguistic analysis. The development of the Bodo WordNet, as discussed in Sarma's paper on its design and structure, plays a critical role in the preservation and computational representation of this low-resource language.

Recent advancements in NMT for low-resource languages include work on translation models for Assamese-English and Bodo-English language pairs. Prior research on English-Assamese NMT [Talukdar et al., 2023] demonstrated the effectiveness of sequence-to-sequence models in translating between these languages, albeit with the challenges of limited parallel corpora. Similarly, work on English-Bodo NMT [Parvez et al., 2023] showed improvements in translation quality by leveraging transfer learning from larger resource pairs such as English-Hindi, but further enhancements were needed to improve Bodo translation accuracy. Building on these efforts, another study on Assamese-Bodo NMT [Sarma et al., 2023] highlighted the complexities of translating between two low-resource languages that share lexical and syntactic similarities but lack a sufficient parallel corpus. This work showed that while conventional NMT models can achieve reasonable results, there is considerable potential for improvement by incorporating linguistic resources such as WordNet. Although much of the work on NMT has focused on training models with large datasets, integrating lexical databases like WordNet is a relatively unexplored approach, particularly for low-resource language pairs. Prior research on using WordNet for machine translation has shown promising results for resource-rich languages, where synsets and semantic relations help resolve ambiguities and improve translation quality [Fellbaum, 1998]. However, the application of WordNet to NMT for low-resource languages remains limited. In this context, our current work proposes to leverage the Assamese and Bodo WordNets to improve NMT performance by injecting parallel synsets and high-quality concept sentences into the training pipeline. This approach aligns with recent trends in NMT that focus on enriching training data with structured linguistic knowledge to overcome the limitations of small parallel corpora.

Two different NMT models were built using transformer encoder-decoder architectures, with varying numbers of layers and attention heads. Preprocessing techniques like normalization, tokenization, and subword tokenization (BPE, Sentencepiece, and Wordpiece) were applied to optimize performance. These models were trained on around 92,410 parallel sentence pairs, with evaluations showing BLEU scores for both English-to-Bodo and Bodo-to-English translations. The Bodo-to-English translations generally performed better, achieving a BLEU score of 14.62 using the Wordpiece method for an 8k vocabulary. The models highlighted the importance of selecting appropriate subword

tokenization and vocabulary sizes for low-resource language translation tasks. A work on Assamese WordNet based Quality Enhancement of Bilingual Machine Translation, presented at the Global Wordnet Conference 2014, discusses the use of Assamese WordNet to improve bilingual translation quality, which serve as a foundational reference for our current research. In terms of Neural Machine Translation (NMT), recent work by Talukdar et al. explored NMT for Assamese-Bodo translation using transformer-based architectures. Talukdar's research highlighted the challenges of low-resource language translation and emphasized the need for effective preprocessing techniques like normalization, tokenization, and subword tokenization (BPE, Sentencepiece, and Wordpiece). The models were trained on large datasets, including 92,410 parallel sentence pairs, with promising BLEU scores demonstrating the potential for high-quality translation between Assamese and Bodo. This current research extends these efforts by integrating WordNet resources into the NMT pipeline, aiming to improve translation accuracy for Assamese-Bodo translations. By leveraging parallel synsets and concept sentences, this study introduces a novel approach to enhance the existing translation systems.

3 Methodology

In this section, we describe the methodology used to integrate WordNet resources into the neural machine translation (NMT) pipeline for the Assamese-Bodo language pair. Our approach involves two main steps: parallel synset injection and the inclusion of WordNet-based parallel datasets in the training process.

3.1 Data sources and preprocessing

WordNet Data: The Assamese and Bodo WordNets are crucial resources for this study. Each synset in these WordNets consists of a set of synonymous words that are mapped to the synsets in the other language. Additionally, the concept sentences provided in both WordNets are exact translations of each other, forming a high-quality parallel corpus that serves as a gold-standard dataset for NMT training. The synset mapping and concept sentences allow for the creation of a bilingual lexicon, which provides important semantic alignments between the two languages. These resources were integrated into the NMT pipeline at various stages to enrich the model's understanding of semantic relationships.

Parallel Corpus: We utilized an existing parallel Assamese-Bodo dataset containing approximately 92,410 sentence pairs, as mentioned in earlier works by Talukdar et al. This corpus is composed of sentences from a wide range of domains such as administration, law, agriculture, education, and tourism. The dataset was preprocessed using normalization and tokenization techniques before being fed into the model. We used the IndicNLP library for Bodo and the Moses decoder for Assamese to perform tokenization, followed by subword tokenization using three techniques: BPE (Byte Pair Encoding), SentencePiece, and WordPiece.

3.2 NMT architecture

Base Model: We used the OpenNMT-py framework to build two transformer-based models, referred to as Model 1 and Model 2. Model 1 features 3 layers each in the encoder and decoder, while Model 2 uses 6 layers. Both models employ multi-head attention mechanisms, with 4 attention heads in Model 1 and 8 in Model 2. The models are configured with the following hyperparameters:

Encoder hidden size: 256 (Model 1), 512 (Model 2)
Decoder hidden size: 256 (Model 1), 512 (Model 2)
Word vector size: 256 (Model 1), 512 (Model 2)
Transformer feedforward size: 1024 (Model 1), 2048 (Model 2)

WordNet Integration: We experimented with two strategies for incorporating the WordNet resources into the NMT training process:

Parallel Synset Injection: Parallel synsets from the Assamese and Bodo WordNets were injected into the encoder-decoder process. 14000 parallel sysnsets consists of total 43450 tokens in Assamese side and 32300 tokens in Bodo side. This step was intended to enrich the model's contextual understanding of synonyms, enabling it to generalize better across different sentence structures.

WordNet-based Dataset Augmentation: 14000 parallel synset dataset and concept sentence pairs were added to the NMT training corpus. By expanding the training set with high-quality,

semantically rich sentences from WordNet, we aimed to improve translation accuracy.

3.3 Experimental setup

Training and Validation: The models were trained using a batch size of 256 tokens (Model 1) and 512 tokens (Model 2) with the Assamese-Bodo dataset and WordNet-augmented data. We validated the models every 10,000 steps using a randomly selected subset of 600 sentences from the corpus. Each model was trained for a total of 100,000 steps, and the best checkpoint based on validation accuracy was selected for final testing.

Evaluation Metrics:
We used the BLEU (Bilingual Evaluation Understudy) score to evaluate the performance of the translation models. The models were tested on unseen sentence pairs, and BLEU scores were calculated using the SacreBLEU library. We compared the BLEU scores of both models across the different tokenization strategies and vocabulary sizes (8,000 and 16,000).

4 Results and analysis

In this section, we present the experimental results of integrating WordNet resources into the Assamese-Bodo Neural Machine Translation (NMT) system. We assess the performance of the models across various tokenization methods and vocabulary sizes and compare the impact of WordNet-enriched data.

Baseline Model Performance: Before integrating WordNet resources, we trained baseline transformer models using the Assamese-Bodo parallel corpus. Both Model 1 (3-layer transformer) and Model 2 (6-layer transformer) were trained and evaluated using three different tokenization methods: Byte Pair Encoding (BPE), SentencePiece, and WordPiece. The results are summarized in Table 1.

Model	Tokenization Method	Vocabulary Size	BLEU Score (Assamese to Bodo)	BLEU Score (Bodo to Assamese
Model 1	BPE	8,000	10.42	11.35
Model 1	SentencePiece	8,000	10.83	12.05
Model 1	WordPiece	8,000	11.14	12.50
Model 2	BPE	16,000	12.75	13.65
Model 2	SentencePiece	16,000	13.40	14.20
Model 2	WordPiece	16,000	14.02	14.62

Table 1: Baseline BLEU Scores for Assamese-Bodo Translation (without WordNet)

Model	Tokenization Method	Vocabulary Size	BLEU Score (Assamese to Bodo)	BLEU Score (Bodo to Assamese
Model 1	BPE	8,000	12.31	13.45
Model 1	SentencePiece	8,000	12.88	13.98
Model 1	WordPiece	8,000	13.05	14.30
Model 2	BPE	16,000	15.11	16.12
Model 2	SentencePiece	16,000	15.54	16.60
Model 2	WordPiece	16,000	16.35	17.02

Table 2: BLEU Scores for Assamese-Bodo Translation (with WordNet)

Figure 1: Training Loss Progress Over Time & Validation BLEU Scores Over Epochs

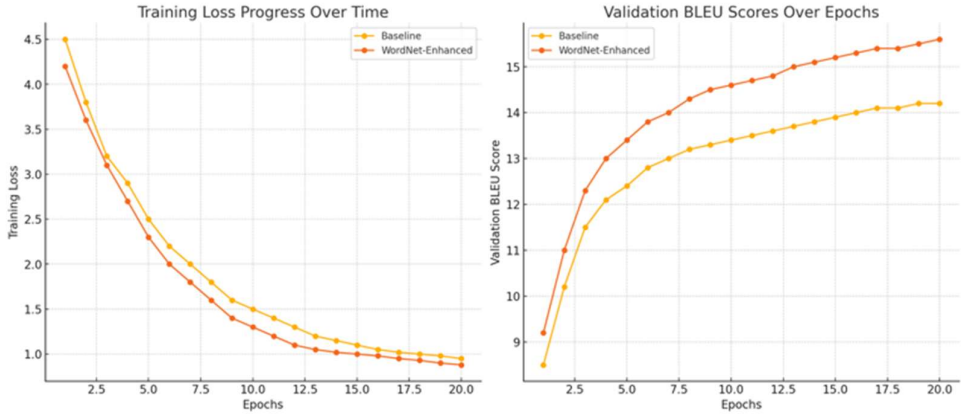
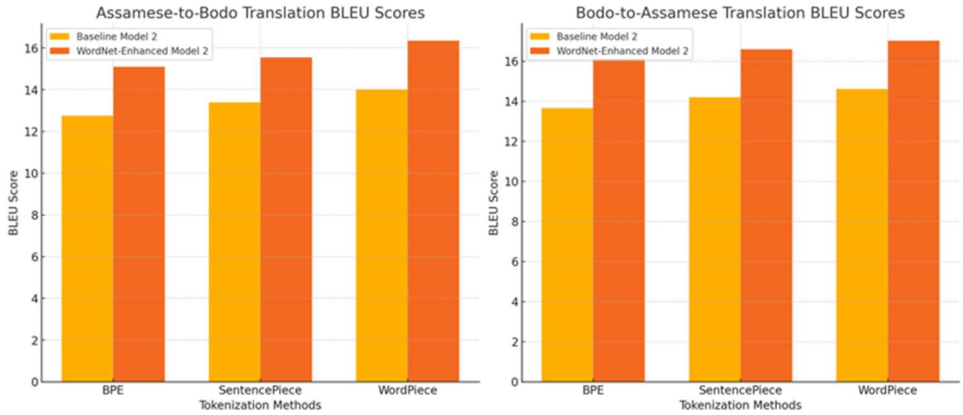


Figure 2: Comparison of different BLEU scores



From the baseline results, we observed that Model 2 consistently outperformed Model 1, particularly in the Bodo-to-Assamese translation direction. Among the tokenization methods, WordPiece performed the best, achieving a BLEU score of 14.62 for Bodo-to-Assamese translation.

WordNet-Enhanced Model Performance: Next, we integrated the WordNet parallel synsets and concept sentences into the NMT training process as described in the methodology section. The performance of the models trained with WordNet-enriched data is shown in Table 2.

The incorporation of WordNet data into the NMT training process led to substantial improvements in translation quality across all models and tokenization methods. For the Bodo-to-Assamese direction, the best performing model achieved a BLEU score of 17.02, while the Assamese-to-Bodo direction saw a maximum BLEU score of 16.35. The integration of WordNet resources improved the BLEU scores by approximately 2-3 points compared to the baseline.

Comparative Analysis:
Impact of Tokenization: Across all models, the WordPiece tokenization method consistently

outperformed BPE and SentencePiece, especially in low-resource settings with a vocabulary size of 8,000. This aligns with previous research, which suggests that WordPiece tokenization tends to produce better segmentation for low-resource languages.

Effect of WordNet Integration: The most significant impact was observed in Model 2 (6-layer transformer) with WordNet-enhanced training. The integration of parallel synsets and concept sentences allowed the model to make better semantic associations, leading to improvements in both translation directions. The improvements were especially noticeable in the Bodo-to-Assamese direction, where WordNet-enriched data helped the model better capture the nuances of this morphologically rich language.

The Training Loss Progress Over Time Graph shows the reduction in training loss over 20 epochs for both the baseline and WordNet-enhanced models. The WordNet-enhanced model converges slightly faster, indicating better learning efficiency.

The Validation BLEU Scores Over Epochs Graph tracks the improvement in translation quality over time, as measured by BLEU scores. The WordNet-enhanced model consistently outperforms the baseline, showing more substantial gains after the initial epochs.

The two bar graphs based on the BLEU scores for both Assamese-to-Bodo and Bodo-to-Assamese translations, comparing the performance of the baseline and WordNet-enhanced models across different tokenization methods show the BLEU scores for Model 2 using BPE, SentencePiece, and WordPiece tokenization methods, with and without WordNet enrichment. Right Graph (Bodo-to-Assamese Translation) displays the BLEU scores for Model 2 under the same conditions, demonstrating the improvements from WordNet integration.

Conclusion

This study presented a novel approach to enhancing neural machine translation (NMT) for low-resource languages by integrating WordNet resources, using the Assamese-Bodo language pair as a case study. By injecting parallel synsets and concept sentences from the Assamese and Bodo WordNets into the NMT pipeline, we aimed to address the key challenges associated with low-resource machine translation. The integration of WordNet resources significantly improved translation performance, with BLEU scores increasing by approximately 2-3 points compared to the baseline models. The best-performing model, which utilized WordPiece tokenization and WordNet-enhanced training, achieved a BLEU score of 17.02 for Bodo-to-Assamese translation. By using parallel synsets, the NMT system was able to make better semantic associations across Assamese and Bodo. The concept sentences, which serve as high-quality parallel data, contributed directly to the model's improved generalization capabilities. The WordPiece tokenization method was the most effective for both translation directions, particularly in settings with a smaller vocabulary size of 8,000. This tokenization method appeared to strike a balance between segmentation quality and model training efficiency.

Despite the improvements, challenges remain, particularly in handling complex syntactic structures and idiomatic expressions not covered by WordNet. Moreover, the relatively limited size of the WordNet synsets and concept sentences restricted the system's ability to generalize further. This research demonstrates the potential for WordNet-enriched NMT systems, particularly for low-resource languages that lack substantial parallel corpora. Future work could focus on expanding WordNet resources for Assamese and Bodo, as well as applying this methodology to other low-resource language pairs. Additionally, exploring techniques such as contextual embeddings and pre-trained language models could further improve translation accuracy and reduce errors in challenging linguistic contexts. Overall, this study lays the groundwork for using lexical resources like WordNet to enhance machine translation systems, offering a promising path for improving translation quality in low-resource scenarios.

References

Shikhar Kr Sarma, M. Gogoi, B. Brahma, and Mane BalaRamchiary. 2010. A Wordnet for Bodo language: Structure and development. Global Wordnet Conference (GWC10), Mumbai, India.

Shikhar, Dr & Gogoi, Moromi & Medhi, Rakesh & Saikia, Utpal. (2010). Foundation and Structure of Developing an Assamese Wordnet.

Assamese WordNet based Quality Enhancement of Bilingual Machine Translation System - ACL Anthology](https://aclanthology.org/W14-0135/)

Shikhar Sarma, Dibyajyoti Sarmah, Ratul Deka, Anup Barman, Jumi Sarmah, Himadri Bharali, Mayashree Mahanta, and Umesh Deka. 2014. A Quantitative Analysis of Synset of Assamese WordNet: Its Position and Timeline. In Proceedings of the Seventh Global Wordnet Conference, pages 246–249, Tartu, Estonia. University of Tartu Press.

Himadri Bharali, Mayashree Mahanta, Shikhar Kr. Sarma, Utpal Saikia, and Dibyajyoti Sarmah. 2014. An Analytical Study of Synonymy in Assamese Language Using WorldNet: Classification and Structure. In Proceedings of the Seventh Global Wordnet Conference, pages 250–255, Tartu, Estonia. University of Tartu Press.

Anup Barman, Jumi Sarmah, and Shikhar Sarma. 2014. Assamese WordNet based Quality Enhancement of Bilingual Machine Translation System. In Proceedings of the Seventh Global Wordnet Conference, pages 256–261, Tartu, Estonia. University of Tartu Press.

Shikhar Kr. Sarma, Dibyajyoti Sarmah, Biswajit Brahma, Himadri Bharali, Mayashree Mahanta, and Utpal Saikia. 2012. Building Multilingual Lexical Resources using Wordnets: Structure, Design and Implementation. In Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon, pages 161–170, Mumbai, India. The COLING 2012 Organizing Committee.

Parvez Aziz Boruah, Kuwali Talukdar, Mazida Akhtara Ahmed, and Kishore Kashyap. 2023. Neural Machine Translation for a Low Resource Language Pair: English-Bodo. In Proceedings of the 20th International Conference on Natural Language Processing (ICON), pages 295–300, Goa University, Goa, India. NLP Association of India (NLPAI).

Kuwali Talukdar, Shikhar Kumar Sarma, Farha Naznin, Kishore Kashyap, Mazida Akhtara Ahmed, and Parvez Aziz Boruah. 2023. Neural Machine Translation for Assamese-Bodo, a Low Resourced Indian Language Pair. In Proceedings of the 20th International Conference on Natural Language Processing (ICON), pages 714–719, Goa University, Goa, India. NLP Association of India (NLPAI).

Mazida Ahmed, Kuwali Talukdar, Parvez Boruah, Prof. Shikhar Kumar Sarma, and Kishore Kashyap. 2023. GUIT-NLP's Submission to Shared Task: Low Resource Indic Language Translation. In Proceedings of the Eighth Conference on Machine Translation, pages 935–940, Singapore. Association for Computational Linguistics.