A Comparative Analysis of Sanskrit to Malayalam Machine Translation of Shlokas and Sentence input using NMT Models

Sreedeepa H S
Department of Computer Science
Cochin University of Science and Technology

Abstract

This paper presents a comparative analysis of performance of Sanskrit- Malayalam translation model developed using encoder-decoder models with Long Short-Term Memory (LSTM) and attention mechanisms on two types of inputs- Sanskrit shlokas and sentence text. The model leverages the power of neural networks to capture the complex linguistic relationships between the two languages, offering a potential solution to the challenges posed by Sanskrit's intricate grammatical structure and Malayalam's rich morphological system. The complexity of Sanskrit grammar and the relative scarcity of computational resources for Malayalam cause unique challenges. By constructing a robust parallel corpus and employing state-of-the-art neural network architectures, the paper demonstrates significant improvements in translation quality compared to traditional rule-based and statistical methods. The promised work mainly focused on the translation of Sanskrit Shlokas and texts. A parallel corpus for Shlokas is created from ancient text books such as Bhagavad Gita and Ramayana and an updated Sanskrit-Malayalam corpus for sentence text are used for training and testing. Here the evaluation of the performance of the model is done on a curated dataset of Sanskrit-Malayalam parallel sentence texts corpus and created domain based Shloka parallel corpus. The LSTM with attention model is out performed for sentence text input rather than for the direct Shloka input. Finally get into the conclusion that if convert shlokas into sentence form it gives more accurate results. The evaluation is done considering metrics such as BLEU score and human evaluation. The findings highlight the challenges of using shlokas as input to the model and, providing valuable insights for future research in Sanskrit-Malayalam machine translation.

1 Introduction

Sanskrit, an ancient Indian language with a rich literary heritage, has gained renewed interest in recent years. However, its complex grammatical structure and the dearth of annotated resources have hindered the development of effective machine translation systems. This paper aims to address these challenges by exploring the application of encoder-decoder models, a class of neural network architectures that have achieved significant success in machine translation tasks.

1.1 Challenges in Processing Sanskrit Shlokas

Sanskrit shlokas, with their intricate grammatical structures and rich cultural context, pose several challenges for computational processing. Here are some key difficulties:

Morphological Complexity:

- **Inflectional and Derivational Morphology:** Sanskrit has a highly complex morphological system with extensive inflectional and derivational processes. This makes it difficult to accurately analyze and understand the meaning of words and their relationships within a shloka.
- **Sandhi Rules:** Sanskrit employs intricate sandhi rules that govern the combination of words at word boundaries. These rules can obscure the original form of words, making it challenging to identify and process them correctly.

Ambiguity and Polysemy:

- **Multiple Meanings:** Many Sanskrit words have multiple meanings, making it difficult to determine the intended interpretation within a shloka.
- **Contextual Dependence:** The meaning of a word or phrase often depends on the surrounding context, making it challenging to accurately interpret shlokas without considering the broader semantic context.

Semantic Complexity:

- **Figurative Language:** Sanskrit shlokas frequently employ figurative language, such as metaphors, similes, and allusions, which can make it difficult to understand the underlying meaning.
- **Cultural References:** Shlokas often contain cultural references that may be unfamiliar to modern readers, requiring specialized knowledge to interpret accurately.

Data Scarcity:

- **Limited Annotations:** There is a lack of annotated Sanskrit shlokas, making it difficult to train and evaluate machine learning models for tasks like machine translation, sentiment analysis, and question answering.
- **Dialectal Variations:** Sanskrit has numerous dialects and regional variations, which can introduce additional challenges in processing shlokas from different regions.

Encoding and Standardization:

- **Character Encoding:** Ensuring consistent encoding of Sanskrit characters, especially those with diacritics and special characters, is crucial for accurate processing.
- **Standardization:** Establishing standardized formats and conventions for representing Sanskrit text can help improve interoperability and facilitate data sharing.

Addressing these challenges requires a combination of linguistic expertise, advanced computational techniques, and large-scale annotated datasets. By overcoming these obstacles, we can unlock the rich cultural and linguistic heritage encoded in Sanskrit shlokas. The proposed an encoder-decoder model trained and tested for both Sanskrit shlokas and sentence texts.

The proposed LSTM with attention model employs an LSTM network as the encoder to capture the sequential nature of Sanskrit sentences. The attention mechanism is used to selectively focus on relevant parts of the encoded representation during decoding, improving the model's ability to handle long-range dependencies.

2 Related Works

The study builds upon previous work in MT, focusing on rule-based, statistical, and neural approaches. Rule-based systems, while accurate for syntactically rigid languages, fall short for highly inflectional languages like Sanskrit and Malayalam. Statistical methods, hough more flexible, require extensive parallel corpora, which are often unavailable. Recent advances in neural machine translation (NMT) have shown promise, particularly for low-resource languages.

2.1 NMT for Low-Resource Languages

Recent studies have focused on extending NMT to low-resource languages through techniques such as transfer learning, multilingual models, and unsupervised learning. Johnson et al. [6] demonstrated the effectiveness of multilingual NMT models that share parameters across multiple language pairs, thereby improving performance for low-resource languages. Similarly, Lample et al. [8] explored unsupervised NMT, which requires only monolingual corpora and has shown promise for languages with limited parallel data.

There are several advancements in machine translation, particularly in handling low-resource languages and addressing the challenges of specific language pairs:

Adapting Transformers for Low-Resource Languages: Recent works have adapted transformer models for low-resource settings. For instance, Fan et al. [9] introduced a multilingual approach using mBART (Multilingual BART), which pre-trains a sequence-to-sequence model on a large corpus of text in multiple languages before fine-tuning it on specific language pairs. This approach has shown substantial improvements for low-resource languages. Large language models can be used for developing machine translation of low resources languages using transfer learning techniques .

Data Augmentation and Back-Translation: Data augmentation techniques, such as back-translation (Sennrich et al.,[10]), where synthetic parallel data is generated by translating monolingual data, have been effectively employed. Gao et al. [11] demonstrated the efficacy of these methods in

improving translation quality for underrepresented languages.

Few-Shot and Zero-Shot Learning: Advances in few-shot and zero-shot learning have enabled MT systems to handle language pairs with minimal or no parallel data. For example, the work by Lin et al. [12] on few-shot learning for MT showed that with just a few examples, models could learn to translate new language pairs.

Efficient Pre-training Techniques: Researchers have explored efficient pre-training techniques to enhance the performance of MT models for low-resource languages. Lewis et al. [13] introduced the BERT-like pre-training for seq2seq models, significantly boosting performance by leveraging large-scale monolingual corpora.

Specific to Sanskrit-Malayalam Translation

Specific to Sanskrit-Malayalam translation, there have been limited but noteworthy efforts:

Hybrid Approaches: The work by Anoop et al. [14] on Sanskrit-English translation using a hybrid approach combining RBMT and SMT methods laid the groundwork for more advanced models.

Deep Learning Techniques: Recent applications of deep learning for Indian languages, as explored by Kunchukuttan et al. (2020), have provided valuable insights into the challenges and potential solutions for the proposed Sanskrit-Malayalam NMT. They utilized models like IndicTrans, a multilingual transformer-based model fine-tuned for Indian languages.

2.2 Neural Machine Translation (NMT) for Sanskrit:

Sanskrit-English Translation: Several studies have focused on translating Sanskrit to English using NMT models. These works have explored different encoder-decoder architectures, attention mechanisms, and data augmentation techniques.

Sanskrit-Hindi Translation: There have been fewer studies on Sanskrit-Hindi translation due to the limited availability of parallel data. However, existing research has demonstrated the feasibility of using NMT for this task.

Encoder-Decoder Architectures: Previous research has compared different encoder-decoder architectures, such as LSTM, GRU, and Transformer, for Sanskrit-English translation. These studies have highlighted the advantages and disadvantages of each architecture in terms of performance and computational efficiency.

Attention Mechanisms: Various attention mechanisms, including global attention, local attention, and hierarchical attention, have been explored in NMT for Sanskrit. Comparative studies have shown that the choice of attention mechanism can significantly impact translation quality.

Data Augmentation: Data augmentation techniques, such as backtranslation and noise injection, have been used to address the scarcity of parallel data for Sanskrit. Comparative studies have evaluated the effectiveness of these techniques in improving NMT performance.

Challenges and Limitations:

- **Data Scarcity:** The limited availability of high-quality parallel data for Sanskrit remains a significant challenge. This can hinder the development of accurate and robust NMT models.
- **Morphological Complexity:** The complex morphological structure of Sanskrit poses challenges for NMT models. Handling inflectional and derivational morphology requires specialized techniques.
- **Lack of Standardized Evaluation Metrics:** There is a lack of standardized evaluation metrics specifically designed for Sanskrit-related tasks. This makes it difficult to compare the performance of different models across studies.

Overall, while progress has been made in NMT for Sanskrit, there is still room for improvement. Addressing the challenges of data scarcity, morphological complexity, and evaluation metrics is crucial for developing more accurate and effective Sanskrit translation systems.

3 Methodology

3.1 Dataset

Two parallel corpora were created for the training and testing of the created model. Curated a parallel corpus of Sanskrit-Malayalam sentence texts, consisting of a diverse range of genres such as literature, philosophy, and religious texts. Also, shloka parallel corpus is created using Bhagavadgita and Ashtanga hrudaya.

That was a challenging phase as there is no digitized version of Malayalam is available. So the dataset was created and verified manually to ensure accuracy and consistency. The size of the sentence text corpus is around 57K and Shloka corpora contains almost 12K shlokas and its parallel meaning in Malayalam.

3.2 Model Architecture

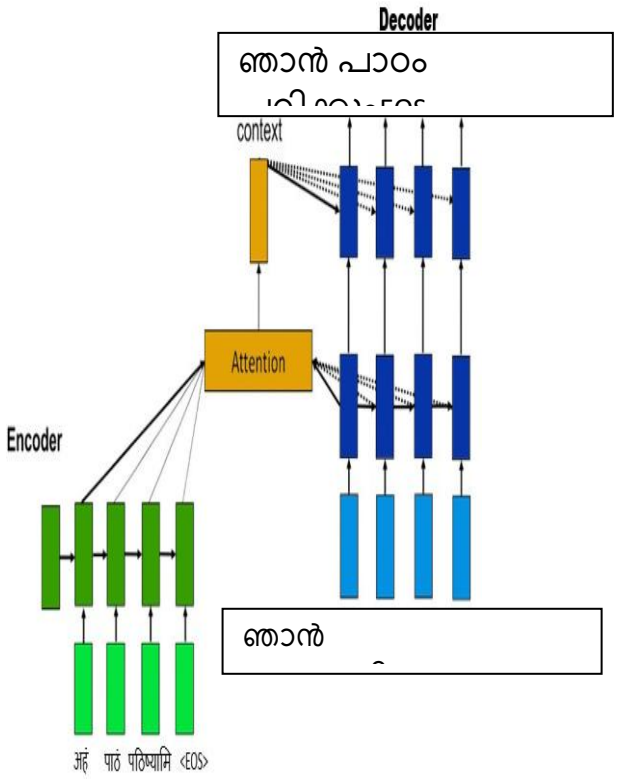

Figure1: Architecture of NMT with Attention

LSTM with Attention: Encoder: A bidirectional LSTM network processes the Sanskrit input sequence, both sentence text and shlokas. Decoder: A unidirectional LSTM network generates the Malayalam translation, guided by the attention mechanism. Attention: The attention mechanism calculates weights for each element in the encoder's hidden states, allowing the decoder to focus on relevant parts of the input. The architecture of the proposed system is shown I figure1.

3.3 Training and Evaluation

The model was trained using the Adam optimizer and a cross-entropy loss function. Initially the model overfitted for Bhagavad Gita shlokas so employed early stopping to prevent overfitting. The models were evaluated using standard metrics such as BLEU score and human evaluation. Hyperparameters are essential in shaping the performance of neural machine translation (NMT) models. Key parameters for the sequence-to-sequence (seq2seq) architecture with attention include the learning rate, batch size, number of layers, hidden units, and dropout rates. In the proposed model, four layers were utilized, which enabled the network to learn more intricate patterns but also increased the computational cost. A hidden layer size of 256 units was used to enhance the model's ability to capture meaningful data representations. The system employed a batch size of 32, a learning rate of 0.01, and a dropout rate of 0.1. The learning rate, critical for the speed of convergence, must be carefully tuned to avoid either overshooting the optimal solution or slow convergence. The dropout technique was used to mitigate overfitting. Finally, the optimal combination of these hyperparameters, yielding the highest BLEU score, was selected for the final model.

3.4 Results and Discussion

Model Performance Comparison

Model performance comparison is done interms of BLEU scores and types of inputs given. The comparison is given in the Table1and Figure3.

Table1: Comparative Analysis of model based on BLEU Score and type of input.

	Input	BLEU-Unigram	BLEU-bi gram	BLEU-tri gram	Average BLEU
LSTM with Attention	Shloka	56.47	43.73	40.32	46.84
LSTM With Attention	Sentence Text	65.35	57.45	50.23	57.67

Sample output: Sample out put obtained is given in the Figure 2.

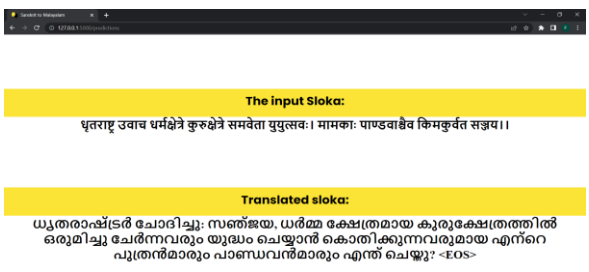

Figure2: Sample Output

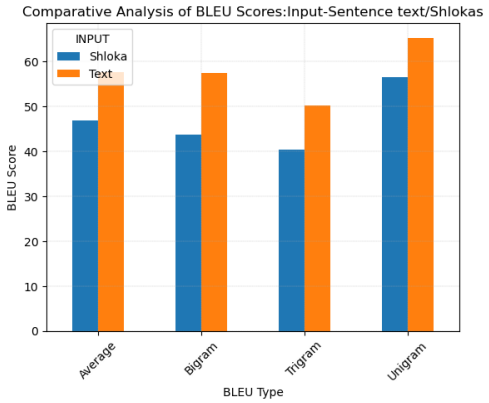

Figure 3: Comparative Analysis of model based on BLEU Score and type of input

The above table shows the comparative analysis of BLEU scores given by the model when the type of input changed. When the input given is sentence text the model consistently outperformed with the input as shlokas in terms of both BLEU and ROUGE scores, indicating its inferior ability to capture long-range dependencies and generates less accurate translations. This can be overcome by using transformer-based model which has superior ability to capture longrange dependencies and self-attention. The human evaluations revealed that the Transformer sometimes produced more generic translations, suggesting that it might benefit from incorporating more domain-specific knowledge. Future research could explore the integration of domain-specific knowledge, such as using pre-trained language models or incorporating external information sources, to improve the quality of the translations. Additionally, experimenting with different attention mechanisms or using larger datasets could further enhance the performance of both models.

Conclusion

This paper has presented a comparative analysis of performance of encoder-decoder model, LSTM with attention for Sanskrit shlokas and sentence text to Malayalam translation. The findings demonstrate the effect of inputs to the model in capturing the complex linguistic relationships between the two languages. While the LSTM with attention model offers a balance between performance and computational efficiency with text sentences, it lacks a little these in case of the shloka input.

As the transformer architecture exhibits superior performance, especially for longer sequences it may give better results for shlokas. Future research can explore that and the integration of domain-specific knowledge, transfer learning techniques, and larger datasets to further improve the quality of Sanskrit to Malayalam translation.

References

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (NeurIPS).

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Proceedings of the 3rd International Conference on Learning Representations (ICLR).

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. In Proceedings of the 3rd International Conference on Learning Representations (ICLR).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, S., Vo, G., & Sutskever, I. (2019). Language models are few-shot learners. OpenAI Blog.

Bharati, A., Chaitanya, V., & Sangal, R. (1996). Natural Language Processing: A Paninian Perspective. Prentice-Hall of India.

Kulkarni, A., Goyal, A., & Shukl, D. (2010). Sanskrit Computational Linguistics. Springer.

Koehn, P., Och, F. J., & Marcu, D. (2003). Statistical Phrase-Based Translation. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (pp. 48-54).

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. arXiv preprint arXiv:1409.0473.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., ... & Dean, J. (2017). Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. Transactions of the Association for Computational Linguistics, 5, 339-351.

Fan, A., Bhosale, S., Schwenk, H., Ma, M., El-Kishky, A., Goyal, S., ... & Edunov, S. (2021). Beyond English-Centric Multilingual Machine Translation. Journal of Machine Learning Research, 22, 1-48.

Gao, Q., Lample, G., & Hashimoto, T. B. (2021). Residual Vector Quantile Networks for Consistent Generative Query Networks. Advances in Neural Information Processing Systems, 34.

Kumar, V., Bhattacharyya, P., & Sasikumar, M. (2012). Challenges in Developing SMT Systems for Indian Languages. In Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages (pp. 51-66).

Kunchukuttan, A., Mehta, P., & Bhattacharyya, P. (2020). The IndicNLP Library. arXiv preprint arXiv:2009.09218.

Lample, G., Conneau, A., Denoyer, L., & Ranzato, M. (2018). Unsupervised Machine Translation Using Monolingual Corpora Only. arXiv preprint arXiv:1711.00043.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2020). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. arXiv preprint arXiv:1910.13461.

Lin, Y., Wang, S., & Ruder, S. (2021). Few-Shot Learning for Neural Machine Translation. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (pp. 2433-2444).

Sennrich, R., Haddow, B., & Birch, A. (2016). Improving Neural Machine Translation Models with Monolingual Data. arXiv preprint arXiv:1511.06709.