International Journal of Religion
2024
Volume: 5| Number 4 | pp. 163 – 170
ISSN: 2633-352X (Print) | ISSN: 2633-3538 (Online)
ijor.co.uk
DOI: https://doi.org/10.61707/nn1dfz44

Deep Learning based UPoS Tagger for Assamese Religious Text

Kuwali Talukdar[1], Shikhar Kumar Sarma[2], Farha Naznin[3] and Ratul Deka[4]

Abstract

Religious texts are known to be with specific patterns of writing, and also involve specific vocabularies. These are also known to follow specific style of writing. Thereby these texts are enriched with typical semantic and syntactic characteristics, demanding special attention for Natural Language Processing (NLP) tasks. This research paper focuses on the application of Deep Learning (DL) techniques for Parts of Speech (PoS) tagging focusing on Assamese language religious texts. We have created a specialized dataset comprising approximately 11,000 sentences extracted from various sources including web crawling and filtering religious texts from existing corpora. The dataset was manually validated by linguists to ensure accuracy, errors, and corrections required. A performance matrix was constructed to analyze the performance of the initial tagging using a pre-existing DL-based model trained for Assamese Universal Parts of Speech (UPoS) tagger. Following this, we utilized a subset of the dataset for manual evaluation, and the validated dataset is then considered as a gold standard training dataset for training other DL models using GRU, RNN and Bidirectional LSTM (BiLSTM) architectures. Training accuracies were recorded and presented, demonstrating the effectiveness of the proposed approach. Accuracies, Precision, and Recall were recorded for all the three Models. F1 scores also have been calculated. Comparison of training and testing accuracies are depicted with performance graphs.

Keywords: *Universal Parts of Speech (UPoS), GRU, RNN, BiLSTM*

INTRODUCTION

Parts of Speech (PoS) tagging is a fundamental task in natural language processing (NLP), which involves assigning a grammatical category to each word in a sentence. This task is crucial for various NLP applications such as machine translation, information retrieval, and sentiment analysis etc. In the context of the Assamese language, PoS tagging plays a vital role in understanding and analyzing text. Assamese is an Indo-Aryan language spoken primarily in the Indian state of Assam. It holds significant cultural and historical importance in the region and has a rich literary tradition. Despite its importance, the development of NLP tools and resources for Assamese has been relatively limited compared to other languages.

The motivation behind this research originates from the need to develop robust NLP tools specifically tailored for the Assamese language, particularly focusing on religious texts. Religious texts often contain unique linguistic features and vocabulary, making them challenging for conventional PoS tagging models. By developing specialized techniques for PoS tagging in Assamese religious texts, we aim to enhance the overall NLP capabilities for this language. Religious texts are increasingly available in digital platforms, and started demanding automatic analysis and processing, including machine translation, information trivial, question answering, sentiment analysis, summarization etc. This has created a new dimension of looking at these data with their special characteristics for computational processing.

The main objectives of this research are as follows:

To create a specialized dataset comprising religious Assamese texts for PoS tagging.

To validate the dataset manually to ensure accuracy and identify errors.

To evaluate the performance of existing DL models for PoS tagging in Assamese for this dataset.

[1] Department of Information Technology, Gauhati University, Guwahati, Assam, India, kuwalitalukdar@gmail.com

[2] Department of Information Technology, Gauhati University, Guwahati, Assam, India, sks001@gmail.com

[3] Department of Information Technology, Gauhati University, Guwahati, Assam, India, farha.gu@gmail.com
[4] Department of Information Technology, Gauhati University, Guwahati, Assam, India, rdeka8258@gmail.com

To train DL models using GRU, RNN and BiLSTM architectures on the validated dataset.

To analyze and discuss the results, identifying areas for improvement and future research directions.

The successful application of DL techniques for PoS tagging in Assamese religious texts has significant implications for NLP research in the language. It opens up possibilities for developing more advanced NLP tools and applications related to specific domains, such as religious literature, thereby enhancing the accessibility and usability of NLP technologies for Assamese speakers.

LITERATURE REVIEW

This section provides an overview of existing literature on NLP works in various languages, with a specific focus on religious texts. Previous studies have highlighted the challenges associated with NLP tasks including PoS tagging in religious texts due to their unique linguistic characteristics and vocabulary.

While there have been efforts to develop NLP tools for the Assamese language, limited research has been conducted specifically on PoS tagging, especially in the context of religious texts. This section provides a summary of previous NLP works including PoS tagging in Assamese and identifies gaps in the existing literature.

Celeno et. al. in their work [1] experimented with five PoS taggers, the Mate tagger, the Hunpos tagger, RFTagger, the OpenNLP tagger, and NLTK Unigram tagger. They had considered an Ancient Greek Dependency Treebank as a dataset and cross validation of 10-folds has been performed. An accuracy score of 88% has been obtained by Mate tagger which outperforms all the other taggers. Emad et. al. [2] has experimented with both segmentation and PoS tagging considering a small religious Arabic corpus. They had obtained the segmentation accuracy and PoS tagging accuracy separately which had outperformed the Arabic Treebank which was a larger corpus. Out of multiple experiments the maximum segmentation accuracy rate is 95.70% and PoS tagging accuracy rate is 91.26%. Alashqar et. al. in their work [3] has compared the performance of different PoS tagging techniques such as N-Gram, Brill, HMM, and TnT taggers using Quran corpus for Arabic. Performance analysis has been done via multiple experiments on diacritized and undiacritized classical Arabic. A work by S. Dipper et. al. [4] performed multiple sets of experiments in different levels that includes token normalization on dialect-specific subcorpora. PoS tagging has been done in the manuscripts written in middle high German and accuracy rate greater than 91% has been obtained. Hadni et. al. in their work [5] used hybrid approach PoS tagging for Arabic language where they used two corpora, one is the holy Quran and the other is Kalimat. Rule based methods results in misclassified and unanalysed word along with ambiguity issues. Using Holy Quran they had obtained accuracy rate of 97.6%, 96.8% and 94.4% respectively for hybrid tagger, HMM tagger and rule-based tagger whereas, for Kalimat corpus accuracy rate is 94.60%, 97.40% and 98%. Gaikwad et. al. in their paper [6] has done survey on various computational techniques such as PoS tagging, *Sandhi* splitting, *Alankaar* finder, *Samaas* finder for Indo-Aryan as well as Dravidian language. They had observed that for *Samaas* finder no work has been done yet and for *Alankaar* finder one technique has been used. Comparatively more work could be seen in PoS tagging. Željko et. al. in their paper [7] has develop simple methods which can learn PoS taggers for 100 of languages and also evaluation of cross lingual model on 25 languages has been done. Their model has outperformed the Bible translation induced state-of-art PoS tagger. Christian et. al. in their paper [8] has presented a PoS tagger which has been developed using character level neural network based on historical text. In another work by Kuwali et. al. [9], authors have covered analysis of various PoS tagging techniques and approaches, also the evolutionary journey of PoS taggers. The performance analysis of those taggers specific to Indo Aryan languages along with their accuracy rates have been depicted. K.K. Baruah et. al. in their paper [10] has done Statistical machine translation considering Assamese and English parallel corpus. Translation tool such as Moses, IRSTLM, GIZA were also used. To check the performance of the translation, BLEU matrix is used. Jumi et. al. in their papers [11] [12] included literature survey on Word sense disambiguation and has observed that in language like Hindi, Nepali, Manipuri etc. WSD system has been developed whereas in Assamese no such automated work has been reported.

Although NLP works for Assamese language is relatively new, works in various contemporary aspects could be traced. Tools and resources required for developing neural computational model for Assamese language, various stages of pre processing for NMT models including Assamese language, next-word-suggestions for the word which someone is intended to write in composing text etc. are few works enriching Assamese NLP [13, 14]. Jumi et. al. in their other works [15] [16] have also developed stemmer for Assamese language which gives accuracy of 85%. Brahma et. al. in their paper [17] have described about the corpus building of 1.5 million words for Bodo language which is a low resource language spoken in the same geography to the Assamese speaking region. Nomi et al in their work [18] have discussed about the text compaction of Assamese language. They had used Recall-Oriented Understudy for Gisting Evaluation (ROUGE) framework for summary condensing and recorded F-measure of 0.633. Sarma et. al. in their paper [19] have developed implementation of multilingual lexical resources on Assamese and Bodo languages with the help of Hindi Wordnet. A multilingual dictionaries for language such as Hindi, Bodo and Assamese has also been developed. Mazida et. al. in their paper [20] had worked on three language pairs English-Mizo, English-Khasi, and English-Assamese and various NMT techniques alongwith subword tokenisation and model configuration were explored with different hyperparameters. In another work by Kuwali et. al. [21] on Assamese NLP , authors have discussed the significance and impact of data quality and quantity while training an NMT model. Performance enhancement of an NMT model not only requires quantity data but it also requires quality data. In their paper [22] Basumatary et. al. had developed a deep learning based Bodo tagger and had employed various models such as RNN, GRU, LSTM and BiLSTM. On analyzing the performance of the models it has been observed that BiLSTM yields more accuracy i.e. 93%. Kuwali et. al. in their paper [23] had developed a deep learning based tagger for Assamese. Both LSTM and BiLSTM models have been developed and UPoS tagset has been used which is the first attempt in the field of NLP for Assamese language. In another work [24] on low resource language Bodo, Sarma et. al. have developed the Bodo wordnet where the characteristics of the Bodo language in terms of morphology and syntax have also been explained. Kashyap et. al. in their paper [25] have developed two base NMT models for two languages i.e. Assamese and English and had analyzed the translation performance in terms of BLEU. From English to Assamese reported BLEU score is 9.01 and from Assamese to English it is 14.71.

While there have been efforts to develop NLP tools for the Assamese language, limited research has been conducted specifically on PoS tagging, especially in the context of religious texts. Here we tried to provide a summary of previous NLP works on Assamese, and also few on Bodo, which is an associate official language in the state of Assam, including PoS tagging in Assamese.

DATASET CREATION

Methodologies

To create a dataset suitable for PoS tagging of Assamese religious texts, we employed several methodologies. One of the primary methods involved has been- web crawling, where we systematically searched online sources for religious texts in Assamese. Additionally, we filtered existing corpora to extract relevant religious content. The religious Assamese texts used in our dataset were sourced from various sources, including religious websites, digital libraries, and online forums dedicated to Assamese literature. These texts encompassed a wide range of religious themes, including scriptures, prayers, hymns, and religious discourses. The basic information on the collected dataset is given in Table I. In the dataset, texts on various religions including Hinduism, Islam, Christian, Buddhism are present.

Table 1. Basic Dataset Information

Total Number of sequences (sentences):	11038
Total Number of Tokens (words):	155559
Average Sequence (sentence) length:	14.1

Table 2. Religion wise number of sequences and tokens

Sequence Length	#Sequences	#Tokens
Hinduism	2997	45177
Islam	2596	40592
Christian	3851	46333
Buddhism	769	11921
Mixed	825	11536
Total:	11038	155559

Table 3. Sequence length wise frequencies

Sequence Length	#Sequences
<=5	781
6-10	3500
11-20	4942
21-30	1399
31-40	309
>40	107
Total	11038

Pre-processing Steps

Before proceeding with PoS tagging, we performed pre-processing steps to clean and format the dataset. This involved removing HTML tags, punctuation marks, and non-textual elements. We also tokenized the text into sentences and words to facilitate the tagging process.

Dataset Validation

To ensure the accuracy of the dataset, we conducted manual validation with the assistance of linguists proficient in Assamese. The validation process involved systematically reviewing a subset of the dataset to identify errors and inconsistencies in the initial tagging. During validation, linguists examined each sentence in the dataset and manually corrected any mislabeled tags. They also recorded errors and inconsistencies for further analysis. The validation methodology focused on achieving high accuracy and reliability in the dataset.

MODEL TRAINING

In this chapter, we discuss the training of Deep Learning (DL) models for Parts of Speech (PoS) tagging in Assamese religious texts. We utilized the popular architectures: Long Short-Term Memory (LSTM). LSTM cells utilize gates to regulate the flow of information, allowing them to retain relevant information over long sequences. A previously trained LSTM model has been used for tagging PoS in the Assamese Religious text dataset.

Table 4. Performances of existing Model

Model	**Accuracy (A)**	**Precision (P)**	**Recall (R)**	**F1 Score**
LSTM	97.19%	97.57%	96.92%	97.24%
BiLSTM	97.36%	97.61%	97.47%	97.54%

Training Process

We divided our manually validated dataset into training, validation, and test sets. The training set was used to train the DL models, while the validation set helped in tuning hyperparameters and preventing overfitting. The test set was kept separate for evaluating the final performance of the trained models.

BiLSTM networks process input sequences in both forward and backward directions, enabling them to capture dependencies from both past and future contexts.

Parameter Settings

During training, we experimented with various hyperparameters such as learning rate, batch size, and number of hidden units in the LSTM layers. Three models have been trained with-GRU, RNN and BiLSTM.

The training accuracies of the GRU, RNN and BiLSTM models were recorded at regular intervals during the training process. These accuracies provide insights into the convergence behaviour of the models and their ability to learn from the training data.

We analyzed the convergence behaviour of the trained models by plotting training and validation loss curves. This analysis helped us determine whether the models were learning effectively and whether further training iterations were necessary.

RESULTS AND DISCUSSION

After training the GRU, RNN and BiLSTM models on the validated dataset, we evaluated their performance using various metrics such as accuracy, precision, recall, and F1-score. These metrics provide a comprehensive assessment of the models' ability to correctly assign PoS tags to words in Assamese religious texts.

We compared the performance of the trained GRU, RNN and BiLSTM models with the initial tagging results obtained using the pre-existing DL-based model. This comparison allowed us to assess the effectiveness and improvements achieved by training the models on the specialized religious text dataset.

Table 5. Performances of GRU, RNN and BiLSTM Model

Model (11038 Sequences)	**Accuracy (A)**	**Precision (P)**	**Recall (R)**	**F1 Score**
GRU	98.05	98.24	97.90	98.07
RNN	97.94	98.31	97.67	97.99
BiLSTM	97.98	98.11	97.86	97.98

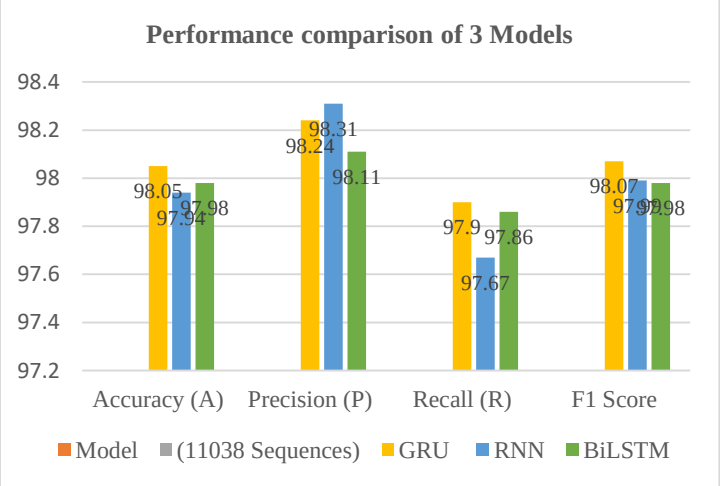

Figure 1: Performance comparison of Accuracy, Precision and F1 Scores

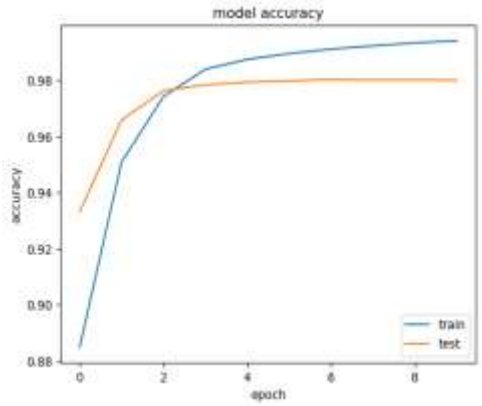

Figure 1: Model Accuracy vs Epoch for GRU Training

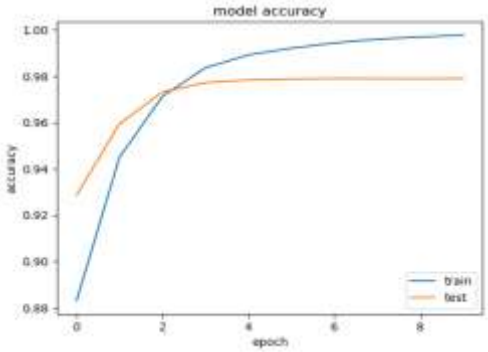

Figure 2: Model Accuracy vs Epoch for RNN Training.

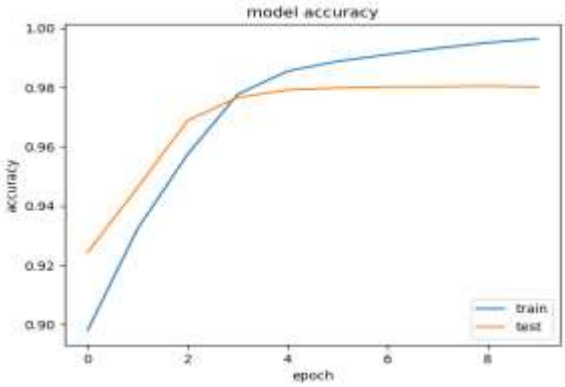

Figure 3: Model Accuracy vs Epoch for BiLSTM Training

The findings of our study indicate that the all the models-GRU, RNN and BiLSTM have resulted comparable performances with initial tagging results, demonstrating the effectiveness of training on a specialized dataset. This is evident from all the performance matrix values including Accuracy, Precision, Recall, and F1 scores. A slight improvement of the BiLSTM model trained with this specially prepared tagset in comparison to that the original model justifies that DL model trained for Assamese PoS tagging works fine even when the corpus is of specialized domain. We observed significant improvements in accuracy and error reduction, particularly in handling complex linguistic structures and rare vocabulary terms present in religious texts.

CONCLUSION

In this study, we aimed to develop and evaluate Deep Learning (DL) techniques for Parts of Speech (PoS) tagging in Assamese religious texts. We created a specialized dataset comprising approximately 11,000 sentences

sourced from various religious texts, and manually validated the dataset to ensure accuracy. We then trained DL models using GRU, RNN and Bidirectional LSTM (BiLSTM) architectures on the validated dataset. While our study has made a start in PoS tagging for Assamese religious texts, there are still several avenues for future research. These include exploring the effectiveness of other DL architectures, such as Transformer-based models, investigating techniques for handling out-of-vocabulary words, and expanding the dataset to include a broader range of religious texts and genres.

Our findings indicate that the BiLSTM models trained on the specialized dataset outperformed the initial tagging results obtained using a pre-existing DL-based model. The models demonstrated improved accuracy and reduced errors in assigning PoS tags to words in Assamese religious texts. This research contributes to the advancement of Natural Language Processing (NLP) capabilities for the Assamese language, particularly in the domain of religious literature. By developing specialized techniques for PoS tagging in religious texts, we enhance the accessibility and usability of NLP technologies for Assamese speakers.

REFERENCES

Celano, Giuseppe G. A., Crane, Gregory and Majidi, Saeed. "Part of Speech Tagging for Ancient Greek" Open Linguistics, vol. 2, no. 1, 2016. https://doi.org/10.1515/opli-2016-0020

Emad Mohamed. 2012. Morphological Segmentation and Part of Speech Tagging for Religious Arabic. In Fourth Workshop on Computational Approaches to Arabic-Script-based Languages, pages 65–71, San Diego, California, USA. Association for Machine Translation in the Americas.

A. M. Alashqar, "A comparative study on Arabic POS tagging using Quran corpus," 2012 8th International Conference on Informatics and Systems (INFOS), Giza, Egypt, 2012, pp. NLP-29-NLP-33.

Stefanie Dipper, POS-Tagging of Historical Language Data: First Experiments, https://www.linguistics.ruhr-uni-bochum.de/~dipper/pub/konvens10_preprint.pdf

Hadni, Meryeme & Ouatik El Alaoui, Said & LACHKAR, Abdelmonaime & Meknassi, Mohammed. (2013). Hybrid Part-Of-Speech Tagger for Non-Vocalized Arabic Text. International Journal on Natural Language Computing. 2. 1-15. 10.5121/ijnlc.2013.2601.

Gaikwad, Hema & Saini, Jatinderkumar. (2021). On State-of-the-art of POS Tagger, 'Sandhi' Splitter, 'Alankaar' Finder and 'Samaas' Finder for Indo-Aryan and Dravidian Languages. International Journal of Advanced Computer Science and Applications. 12. 10.14569/IJACSA.2021.0120455.

Željko Agić, Dirk Hovy, and Anders Søgaard. 2015. If all you have is a bit of the Bible: Learning POS taggers for truly low-resource languages. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 268–272, Beijing, China. Association for Computational Linguistics.

Christian Hardmeier. 2016. A Neural Model for Part-of-Speech Tagging in Historical Texts. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 922–931, Osaka, Japan. The COLING 2016 Organizing Committee.

Kuwali Talukdar and Shikhar Kumar Sarma, "Parts of Speech Taggers for Indo Aryan Languages: A critical Review of Approaches and Performances," IEEE Explore, 2023 4th International Conference on Computing and Communication Systems (I3CS), Shillong, India, 2023, pp. 1-6, doi: 10.1109/I3CS58314.2023.10127336.

Baruah, Kalyanee & Das, Pranjal & Hannan, Abdul & Sarma, Shikhar. (2014). Assamese-English Bilingual Machine Translation. International Journal on Natural Language Computing. 3. 10.5121/ijnlc.2014.3307.

Jumi Sarmah, Shikhar Kumar Sarma, "Survey on Word Sense Disambiguation: An Initiative towards an Indo-Aryan Language", International Journal of Engineering and Manufacturing(IJEM), Vol.6, No.3, pp.37-52, 2016.DOI: 10.5815/ijem.2016.03.04

Jumi Sarmah and Shikhar Kumar Sarma, "Word Sense Disambiguation for Assamese," 2016 IEEE 6th International Conference on Advanced Computing (IACC), Bhimavaram, India, 2016, pp. 146-151, doi: 10.1109/IACC.2016.36.

Mazida Akhtara Ahmed, Kishore Kashyap and Shikhar Kumar Sarma, "Pre-processing and Resource Modelling for English-Assamese NMT System," 2023 4th International Conference on Computing and Communication Systems (I3CS), Shillong, India, 2023, pp. 1-6, doi: 10.1109/I3CS58314.2023.10127567.

Manash Pratim Bhuyan and Shikhar Kumar Sarma, "Automatic Formation, Termination & Correction of Assamese word using Predictive & Syntactic NLP," IEEE Explore, 2018 3rd International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2018, pp. 544-548, doi: 10.1109/CESYS.2018.8724023.

Jumi Sarmah, Shikhar Kumar Sarma, and Anup Kumar Barman. 2019. Development of Assamese Rule based Stemmer using WordNet. In Proceedings of the 10th Global Wordnet Conference, pages 135–139, Wroclaw, Poland. Global Wordnet Association.

A. K. Barman, J. Sarmah and Shikhar Kumar Sarma, "WordNet Based Information Retrieval System for Assamese," 2013 UKSim 15th International Conference on Computer Modelling and Simulation, Cambridge, UK, 2013, pp. 480-484, doi: 10.1109/UKSim.2013.90.

Brahma, Biswajit, Anup Kr. Barman, Shikhar Kr Sarma and Bhatima Boro. "Corpus Building of Literary Lesser Rich Language-Bodo: Insights and Challenges." ALR@COLING (2012), ACL Anthology

Nomi Baruah, Shikhar Kr. Sarma, Surajit Borkotokey, Evaluation of Content Compaction in Assamese Language, Procedia Computer Science, Volume 171, 2020, Pages 2275-2284, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2020.04.246.

Sarma, Shikhar & Sarmah, Dibyajyoti & Brahma, Biswajit & Bharali, Himadri & Mahanta, Mayashree & Saikia, Utpal. (2012). Building Multilingual Lexical Resources using Wordnets: Structure, Design and Implementation. 161-170, Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon. 2012. S. 161-170, ACL Anthology

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.

Kuwali Talukdar, Shikhar Kumar Sarma, Farha. Naznin and Kishore Kashyap, "Influence of Data Quality and Quantity on Assamese-Bodo Neural Machine Translation," IEEE Explore, 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India, 2023, pp. 1-5, doi: 10.1109/ICCCNT56998.2023.10306702.

B. Basumatary, M. Rahman, Shikhar Kumar Sarma, P. A. Boruah and K. Talukdar, "Deep Learning Based Bodo Parts of Speech Tagger," IEEE Explore, 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India, 2023, pp. 1-5, doi: 10.1109/ICCCNT56998.2023.10308365.

Kuwali Talukdar and Shikhar Kumar Sarma, "UPoS Tagger for Low Resource Assamese Language: LSTM and BiLSTM based Modelling," 2023 IEEE International Conference on Machine Learning and Applied Network Technologies (ICMLANT), San Salvador, El Salvador, 2023, pp. 1-6, doi: 10.1109/ICMLANT59547.2023.10372865.

Shikhar Kumar Sarma, Biswajit Brahma, Moromi Gogoi, Manebala Ramchiary, A wordnet for Bodo language: Structure and development, 2010 Proceedings of Global Wordnet Conference (GWC10), IIT Mumbai, India.

Kashyap, K., Sarma, S. K. & Ahmed, M.A. Improving translation between English, Assamese bilingual pair with monolingual data, length penalty and model averaging. Int. j. inf. tecnol. (2024). https://doi.org/10.1007/