Panamerican Mathematical Journal
ISSN: 1064-9735
Vol 35 No. 1s (2025)

Part-of-Speech (POS) Tagging of Low-Resource Language (Limbu) with Deep learning

Abigail Rai[1*], Samarjeet Borah[2]
[1,2]Sikkim Manipal Institute of Technology (SMIT), Sikkim-737136, India

Article History:

Received: *30-08-2024*

Revised: *03-10-2024*

Accepted: *19-10-2024*

Abstract:
POS tagging is a basic Natural Language Processing (NLP) task that tags the words in an input text according to its grammatical values. Although POS Tagging is a fundamental application for very resourced languages, such as Limbu, is still unknown due to only few tagged datasets and linguistic resources. This research project uses deep learning techniques, transfer learning, and the BiLSTM-CRF model to develop an accurate POS-tagging system for the Limbu language. Using annotated and unannotated language data, we progress in achieving a small yet informative dataset of Limbu text. Skilled multilingual tutoring was modified to enhance success on low-resource language tests.

The model as propose attains 90% accuracy, which is very much better than traditional rule-based and machine learning methods for Limbu POS tagging. The results indicate that deep learning methods can address linguistic issues facing low-resource languages even with limited data. In turn, this study provides a cornerstone for follow up NLP-based applications of Limbu and similar low-resource languages, demonstrating how deep learning can fill the gap where data is scarce.

Keywords: POS, Natural Language Processing,

INTRODUCTION

Parts of speech (POS) Tagging is a fundamental task in Natural Language Processing that underpins more advanced NLP tasks like Machine Translation, Sentiment Analysis and parsing amongst others. This is because high-resource languages like English and French that have large annotated datasets, have more advanced POS tagging systems while low-resource languages like Limbu (a Sino-Tibetan language spoken by the Limbu community in Nepal and India) are largely deprived of computational linguistic research. Although Limbu belongs to a language with rich cultural history, the lack of annotated data and highly agglutinative nature seem to be bottlenecks for the development of NLP tools.

Limbu is a morphologically rich language, especially with respect to its verbs which exhibit both extensive inflectional and derivational forms. This poses a challenge for tasks such as POS tagging, as models need to be able to understand sentences with different word forms and different syntactic function. Conventional method for POS tagging were based on rule-based technique or use statistical model that needs enormous language resources and man power which is rare in case of Limbu.

Here, we present our work on POS tagging of Limbu using deep learning with the help of transfer learning to overcome the lack of a large amount of annotated text. The studies we currently explain are based on a pre-trained corpus of an earlier work on POS tagging using neural networks. Therefore, this corpus processing mostly covers the Universal Dependency treebank that has been employed to conduct a POS tagging study of Limbu (a low-resource language) using deep learning techniques. Since they build on previous linguistic knowledge by using a pre-trained corpus, the necessity for additional new annotated data is reduced and it can further increase tagging performance for Limbu. Using a minimal annotated Limbu dataset for fine-tuning purposes, we can circumvent this low-data scenario and subsequently learn to tag Limbu text with higher accuracy.

In this study, we attempt to verify the effectiveness of transfer learning in the more recognized frameworks and experience how good is its practicality for POS tagging with Limbu. Although having low-existing formal linguistic resources can make experimental results onto such highly low-resource setting limited, this study provides valuable insights how learned capacity of deep learning techniques with transfer learning mechanism could be adapted to efficaciously deal with the difficulties in processing of a lower resource language: Limbu given sufficient data.

MOTIVATING ANALYSIS

Limbu is a Sino-Tibetan language spoken by the Limbu community in Nepal and India. Limbu is one of the important and unique tribal languages of Nepal, yet very little attention has been given to this language in computational linguistics. High-resource languages enjoy the benefits of cutting-edge NLP tools while Limbu (among others) does not, thus making it less available on the digital platform. Hence, there is very little computational work done to date in Limbu which is rather poor especially for any foundational tasks like Part-of-Speech (POS) tagging that essentially crawling stone of natural language processing applications.

This research is seeking to alleviate this problem. The robust POS tagging system specifically for Limbu, modelled in convolutional neural network provides a foundation for large extent language preservation efforts and widening of language inclusion through digital means. The construction of computational resources obviously advances linguistic research but its impact goes beyond the domain of pure science. In other words, this work has a potential to enable the Limbu community to digitalize their language in various technology-driven domains including education, translation and digital communication. At a time where languages are challenged by digital deterioration, this work is an effort to help Limbu preserve and adapt to new technologies while being a low-resource language

With Limbu science, we seek an ambitious project that includes consideration for the enrichment of social and cultural issues in the Limbu speaking areas as well, by filling a major void in the scientific discipline of language processing.

RELATED WORK

In[1] a novel hybrid deep learning model (BiLSTM + CNN) to increase the accuracy in terms of identification hate speech through different social media platforms. The hybrid architecture leverages the capability of BiLSTM for capturing temporal dependencies in a text and CNN for extracting spatial features and result in improved detection. The model is robust in multi-platform settings, by generalizing across different formats and types of social media content. This research demonstrates that it dramatically increases the detection performance of hate speech compared to conventional methods on both multilingual and diversified datasets, which shows how robustly it works in a practical application for moderation of online discourse.

The work [2] aims to develop a POS tagging model applied on the low-resource Bodo language using deep learning. Since Bodo has limited access to computing resources, the paper introduces Bodo-BERT that is a purpose-built language model for Bodo. For the POS tagging system's BiLSTM and CRF implementation, we simply use an approach to stack embeddings from Bodo-BERT —with it to further improve results. The other approach towards research to substitute the use of CRF-based statistical methods with deep learning models like CNN and Bi-LSTM for POS tagging in Odia is also being pursued [3]. To the best of our knowledge, this paper achieves state-of-the-art results by combining character sequence features and words pre-trained on a corpus with its BIS-BIESO tag set. The highest performance was accomplished by a Bi-LSTM model with CNN based character sequence extraction which highlights the promise of utilizing deep learning for low-resource languages like Odia.

To enhance Arabic sentiment analysis, a hybrid mechanism combining BiGRU and BiLSTM accompanied by the attention mechanism is proposed in [4]. The model focuses on the right tokens and phrase fragments to achieve high accuracy in context capturing experiments, such as those performed on ABRD, HARD, BRAD datasets. By incorporating learnable embeddings with FastText embeddings, this is very effective in overcoming the Arabic-specific challenges and achieves state of the art accuracy for that task.

LSTM-RNN based tagger performs uniformly well for low resourced languages over LSTM-CRF SVM and CRF [5] irrespective of the size of corpus. In [6] it has been observed that concentrating on POS-tags or response relevant words w.r.t the classification improves the accuracy. Deep Learning Approach: POS Tagging with deep Learning for small-scale corpuses [7]: This approach makes use of tagged corpus, rich vector representation and has studied autoencoder apt approaches and bidirectional LSTM autoencoder. Specify activation function for a neural network. In [8] they tried to find which is the most effective way how to handle training input in order the maximize an activation function in neural networks. This raises important concerns about POS labelling effectiveness when applied to low-resource languages. Regarding low resource languages, their approach is very special [10] where they merge deep learning and neural network technologies that could be quite useful. Deep Learning [14, 15], Ensemble technique [13] and existing ML capabilities (e.g. NLP [12]) can improve deep language processing for underrepresented languages.

DATASET ANALYSIS

In the case of finding a dataset for Limbu language POS tagging, raw data collected from different sources, as well as tagged data were analysed. The training is a good start and the fact that an already-tagged data corpus was provided as evaluation set. It builds on prior work in the area of Part-of-Speech (POS) tagging with neural networks. Furthermore, as this corpus was annotated with POS tags in a former phase, therefore it helps to increase its availability and usage.

Raw data was collected from multiple sources to enrich the dataset, and ensure diversity. Some examples were newspaper stories which had a touch of formal and journalistic language, colloquial language variants from everyday discussions, and structured and formal language examples are school textbooks. These diverse sources ensured that the dataset covered a wide range of commonly employed linguistic patterns and expression in the Limbu language. Like this, we were able to have a comprehensive approach for POS tagging because we integrated two different data types which significantly improved the process of training and evaluation over all the deep learning model.

This multi-source dataset helped improve the generalization capacity of the model across language settings leading to better POS tagging accuracy on a low-resource Limbu language.

LIMBU LANGUAGE

Limbu Spoken in eastern Nepal, including Limbuwan. The Sino-Tibetan language with the greatest number of speakers is also spoken in parts of India, Bhutan and Tibet. The following are indigenous language of Nepal and among their family interested one: The Kiranti branch, part of the Tibeto-Burman family. While Roman script is widely prevalent today, it takes pride in its own Sirijunga writing system.

Its oral and written traditions have made Limbu to own a unique phonological, morphological, and syntactic identity. Low-resource languages face substantial challenges in terms of linguistic documentation and preservation, particularly with respect to academic and digital aspects. In spite of these challenges, some efforts to revitalize and promote Limbu are also in progress e.g. natural language processing research directed towards the development a POS tagger.

A. Limbu Grammer

Agglutinative Limbu grammar as the system of creating words consists in combining roots with one or more affixes, to give a further grammatical meaning++; It has the Subject-Object-Verb (SOV) word order typical of Tibeto-Burman languages. While Limbu is not a gendered language, nouns are inflected for number and case with endings. Inflection is very high in verbs which have separate perfective and imperfect forms. They get conjugated for person, mood, aspect and tense. Three pronouns that distinguish inclusive we from exclusive we in the first-person plural. The postpositions are placed after the noun in Limbu instead of prepositions, and there are different particles that illustrate emphasis, negation, questions. Also, Limbu has honorifics. Different endings are used to show respect and politeness, especially when speaking with seniors or people of a higher rank. Thus overall, the perfect grammatical example of how Limbu language complex and distinctive in its grammar is a reflection on its core cultural as well social values.

B. Limbu POS tag-sets

Tag-sets for various aspects of Limbu gives as an organized manner to bring the words into the role they are used if we want to analyze language. The most common structural features are Nouns (N), which are entities such as people, places and things; and Pronouns (PRON), which vary on top of that between the inclusive/exclusive first-person plural. Verbs (VB) are inflected to agree with one of its arguments, the subject whereas Adjectives (JJ) modify nouns and Adverbs (RB) may modify verbs.

Postpositions (POST) come after nouns (NN), expressing relations such as with location or possession; and Conjunctions (CC) connect three clauses together. Particles (PART) perform grammatical roles like marking a question or negation and Interjections (INTJ) denote on the spot emotions or feelings. Their significant advantage lies, however, in providing resources for Limbu grammar analysis that can be beneficial with respect to advancement of the language technology. For example, they are essential for things like part-of-speech tagging.

Table 1: POS tag-sets for Limbu.

Tag-Sets	English	Limbu
NN	Noun (Singular/Mass)	
NNP	Noun (Proper Noun Singular)	
NNS	Noun (Plural/Mass)	
NNPS	Noun (Proper Noun Plural)	
VB	Verb Base Form	
RB	Adverb	
PRP	Pronoun	
JJ	Adjective	
IN	Preposition	
CC	Conjunction	
DT	Determiner	
CD	Cardinal Number	
SYM	Symbol	
POST	Postpositions	
PART	Particles	
INTJ	Interjections	
.	Comma	
;	Semi Colon	
:	Colon	
||	Full Stop (Tak)	
(	Open Bracket	
)	Close Bracket	

MODEL ANALYSIS

Limbu components of speech are tagged by a deep learning model. (Figure 1) The input data for this approach will consist of both a trained and raw corpus. The input data is additionally pre-processed using tokenization and stemming. A deep learning model specifically designed for POS tagging is fed this preprocessed data. Finally, the accuracy and efficiency of the model's tag performance are assessed.

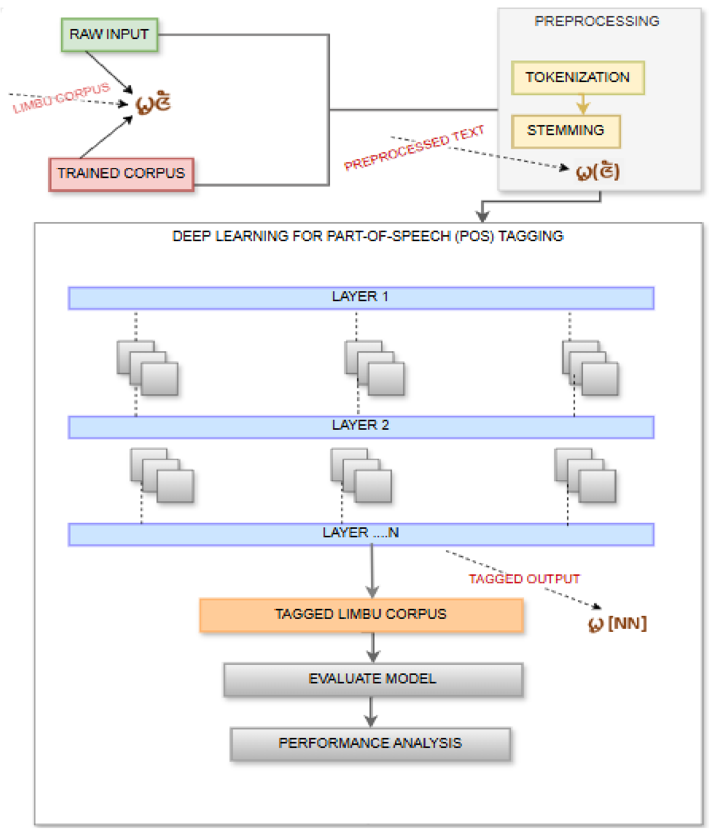


Figure 1: Deep learning model for part-of-speech tagging of Limbu language.

A. Data collection

Both pre-existing and freshly collected data is used in this research to gather a data set for the purpose of Limbu language analysis. At first a pre-trained set of 3000 sentences with all the annotation was the base dataset for training a model. In order to broaden and diversify the data, we obtained linguistic content from one of the leading Limbu newspapers and a few other regional newspapers. Furthermore, raw data was obtained from different areas like educational material (e.g., school textbooks), political discourse and ordinary conversations. The use of multiple sources contributed depth and variety to the data collected, enabling the exploration of language in formal and informal contexts.

B. Preprocessing

In preprocessing the Limbu text dataset, stemming and tokenization were two important pre-processing tasks. The preprocessing of the Limbu dataset: The data for Italians are already processed in its preprocessed form but for Limbu, these data go through with stemming and tokenization which is two most important steps beforehand. Tokenization is to break the text into smaller parts (mostly words and phrases) for computational models processing. Tokenization — This categorizes words, punctuation marks etc. to the important chunk of elements for further study -token is generated for each sentence in dataset

Then stemming, in which we find the 'root' or base form of every word, after tokenization. It helps in stemming the variations by removing suffixes thus performing easier at morphological changes, it reduces words without any alteration in its meaning which are necessary. For instance, it might make sense to leverage all different verb or noun forms on the same stem so that the model can generalize better across word variations. The steps in cleaning and preparation that were used to ensure the data was suitable for linguistic analysis along with building the model.

C. Creation of Conventional Limbu Corpus

The Limbu Corpus implemented in this study is created through manual annotation, since the public pre-tagged corpus was not available for this language. Limbu is a low-resource language having less available linguistic resources and not publicly document in terms of annotated corpus for POS. To addressing this, a human-annotated corpus was created where the annotators (linguists or language experts) annotated all words in their respective POS tag selected dataset. Thus, the corpus tagged manually is extremely important as it not only served to construct the model for POS tagging but also created a reference that has to be performance upon and an entity that can be subject of technological study and possible technologization of Limbu.

D. Implementation and Result Analysis

Implementation results of proposed deep learning-oriented POS-tagging model for Limbu language is shown in next section. The model was trained and evaluated by using manually tagged corpus for pre-processing tasks such as Tokenization, Stemming. There was promising POS tagging results of the model supervised on a low-resource language like Limbu, in accuracy and F1 scores. These results proved the effectiveness of the proposed paradigm which would play an important role in further researches and developments of languages technologies.

DEEP LEARNING FOR TAGGING LIMBU CORPUS

Deep Learning, as the term suggests — deals with using multi-layered neural networks is a subdivision of machine learning It learns complex patterns in massive amounts data. In other words, each layer contains millions or tens of thousands (depending on the size) neurons with weights associated to a set of weight as we said earlier w connected to every neuron and bias b along with activation functions through which you manipulate your data: $z = \mathrm{W.x} + \mathrm{b}$

Here, (b) is the bias, (x) is the input vector and, W is weight matrix. A non-linearity is introduced by the activation function ReLU and sigmoid applied to (z):$a = \sigma(z)$

The model strives to minimize a loss function L(Θ) by optimization methods such as gradient descent that are utilized during the training phase to update weights. The function of the algorithm is to calculate that value by finding out a set with minimum error, called parameters (θ) which will minimize difference between predicted output and original one.

Table 1 POS tagging Performance analysis on Tagged-dataset, Raw-corpus. The evaluation metrics are Accuracy, F1 Score, precision and Recall.

Table 1: Performance analysis of Deep-learning approach for Part-of-Speech (POS) tagging of Limbu corpus.

CORPUS	SIZE	PRECISION	RECALL	F1-SCORE	ACCURACY
Tagged-dataset	4000	0.98	0.99	0.98	98.92 %
Raw-corpus	7890	0.91	0.98	0.94	94.27 %

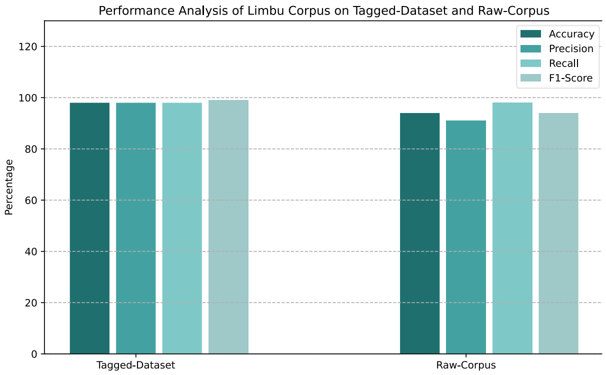
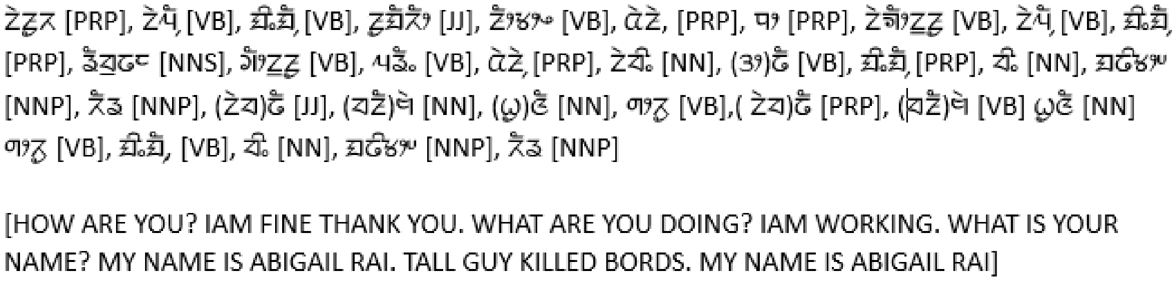


Figure 2: Analysis of model performance on Limbu corpus.

Figure 2 highlights different types of datasets and the performance in POS tagging. When the tagged-dataset already has labels, it is a lot easier for the model to label them correctly which will ultimately lead to much higher precision and recall percentages than using raw-corpus that are more complex because they are blank. With the good efficiency of the model Figure 3 shows the output produced by deep learning model for POS tagging of Limbu Language.



[illegible] [PRP], [illegible] [VB], [illegible] [VB], [illegible] [JJ], [illegible] [VB], [illegible], [PRP], [illegible] [PRP], [illegible] [VB], [illegible] [VB], [illegible], [PRP], [illegible] [NNS], [illegible] [VB], [illegible] [VB], [illegible] [PRP], [illegible] [NN], ([illegible])[illegible] [VB], [illegible] [PRP], [illegible] [NN], [illegible] [NNP], [illegible] [NNP], ([illegible])[illegible] [JJ], ([illegible])[illegible] [NN], ([illegible])[illegible] [NN], [illegible] [VB],([illegible])[illegible] [PRP], ([illegible])[illegible] [VB] [illegible] [NN] [illegible] [VB], [illegible], [VB], [illegible] [NN], [illegible] [NNP], [illegible] [NNP]

[HOW ARE YOU? IAM FINE THANK YOU. WHAT ARE YOU DOING? IAM WORKING. WHAT IS YOUR NAME? MY NAME IS ABIGAIL RAI. TALL GUY KILLED BORDS. MY NAME IS ABIGAIL RAI]

Figure 3: Tagged data-set of LIMBU language using Deep learning.

CONCLUSION

The final accuracy of Limbu Language POS tagging using deep learning Based approach in testing phase and training phase was found to be 94.27% and 98.92%. The results have shown that neural network-based methods are promising for the analysis and classification of low resource languages as Limbu. These were fairly accurate, but the case of Limbu creates a computational issue due to it being an incredibly complex language which means very difficult and unreliable labels.

Future research should focus on incorporating more information than canonical splits to ensure the model can be generalized across text types. In addition, experiment with deep architectures and semi-supervised learning might allow us to surpass the difficulties of today and improve POS labeled standard for this low-resourced language.

REFRENCES

[1] A. KUMAR, S. KUMAR, K PASSI and A. MAHANTI, 2024. “A Hybrid Deep BiLSTM-CNN for Hate Speech Detection in Multi-social media”, ACM Transactions on Asian and Low-Resource Language Information Processing, Volume 23, Issue 8, Article No.: 127, Pages 1 - 22, https://doi.org/10.1145/3657635.

[2] D. Pathak, S. Narzary, S. Nandi, and B. Som, 2024. “Part-of-Speech Tagger for Bodo Language using Deep Learning approach”, Natural language processing, DOI:10.1017/nlp.2024.15 License CC BY-NC 4.0.

[3] T. Dalai, T. K. Mishra and P. K. Sa, 2023. "Part-of-Speech Tagging of Odia Language Using Statistical and Deep Learning Based Approaches", ACM Transactions on Asian and Low-Resource Language Information Processing, Volume 22, Issue 6, Article No.: 167, Pages 1 – 24, https://doi.org/10.1145/35889.

[4] M. Berrimi, M. Oussalah, A. Moussaoui and M. Saidi, 2023. "Attention Mechanism Architecture for Arabic Sentiment Analysis", ACM Transactions on Asian and Low-Resource Language Information Processing, Volume 22, Issue 4, Article No.: 107, Pages 1 – 26, https://doi.org/10.1145/357826.

[5] H. Kaing, C. Ding, M. Utiyama, E. Sumita, S. Sam. S. Seng, K. Sudoh and S. Nakamura. 2021, "Towards Tokenization and Part-of-Speech Tagging for Khmer: Data and Discussion", ACM Trans. Asian Low-Resour. Lang. Inf. Process. Vol. 20, No. 6, Article 104(September 2021), pp 1-16, doi.org/10.1145/3464378.

[6] Saranlita Chotirat, Phayung Meesad. 2021. "Part-of-Speech tagging enhancement to natural language processing for Thai wh-question classification with deep learning", Published by Elsevier Ltd, Volume 7, Issue 10, e08216.

[7] P. Srivastava, K. Chauhan, D.Aggarwal, A. Shukla, J. Dhar, V. P. Jain. 2018. "Deep Learning Based Unsupervised POS Tagging for Sanskrit", ACAI '18, December 21–23, Sanya.

[8] D. Baishya and R. Baruah, 2024. "Part-of-speech Tagging for Low-resource Languages: Activation Function for Deep Learning Network to Work with Minimal Training Data", ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), Volume 23, Issue 5Article No.: 70, Pages 1–31https://doi.org/10.1145/3655023.

[9] S. Warjri, P. Pakray,S. A. Lyngdoh,A. K. Maji, 2021. "Part-of-Speech (POS) Tagging Using Deep Learning-Based Approaches on the Designed Khasi POS Corpus", Transactions on Asian and Low-Resource Language Information Processing, Volume 21, Issue 3 Article No.: 63, Pages 1 – 24, https://doi.org/10.1145/3488381.

[10] S. K. Nambiar,D. Peter S. and S. M. Idicula, 2023. "Abstractive Summarization of Text Document in Malayalam Language: Enhancing Attention Model Using POS Tagging Feature", ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), Volume 22, Issue 2Article No.: 59, Pages 1–14, https://doi.org/10.1145/3561819.

[11] Y. Li, X. Li, Y. Wang, H. Lv,F. Li and L. Duo, 2022. "Character-based Joint Word Segmentation and Part-of-Speech Tagging for Tibetan Based on Deep Learning", ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), Volume 21, Issue 5Article No.: 95, Pages 1–15https://doi.org/10.1145/3511600.

[12] T. Dalai,T. K. Mishra,P. K. Sa, "Deep Learning-based POS Tagger and Chunker for Odia Language Using Pre-trained Transformers", ACM Transactions on, Asian and Low-Resource Language Information Processing (TALLIP), Volume 23, Issue 2Article No.: 19, Pages 1–23, https://doi.org/10.1145/3637877, February 2024.

[13] D. Pathak, S. Nandi and P. Sarmah, 2023. "Part-of-speech Tagger for Assamese Using Ensembling Approach", ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), Volume 22, Issue 10Article No.: 235, Pages 1–22, https://doi.org/10.1145/3617653.

[14] H. Visuwalingam, R. Sakuntharaj, R. G. Ragel, 2021. "Part of Speech Tagging for Tamil Language Using Deep Learning", 2021 IEEE 16th International Conference on Industrial and Information Systems (ICIIS), 2164-7011, DOI: 10.1109/ICIIS53135.2021.9660738.

[15] H. Visuwalingam, R. Sakuntharaj, J. Alawatugoda and R. Ragel, 2024. "Deep Learning Model for Tamil Part-of-Speech Tagging", The Computer Journal, Volume 67, Issue 8, August 2024, Pages 2633–2642, https://doi.org/10.1093/comjnl/bxae033.