Incremental Learning for GRU and RNN-based Assamese UPoS Tagger

Kuwali Talukdar, Shikhar Kumar Sarma
Department of Information Technology, Guwahati University, Guwahati, India

***Abstract*—This research paper introduces a novel approach to enhance the performance of Universal Part-of-Speech (UPoS) tagging for the low-resource language Assamese, employing Recurrent Neural Networks (RNNs) and Gated Recurrent Units (GRUs). The novelty added in this study is the experimentation with Incremental Learning, a dynamic paradigm allowing the models to continually refine their understanding as they encounter new set of linguistic data. The proposed model utilizes the strengths of GRUs and traditional RNNs to capture long range sequential dependencies and contextual information within Assamese sentences. Incorporation of Incremental Learning ensures the model's adaptability to evolving linguistic patterns, particularly crucial for under-resourced languages like Assamese. Experimental results showcase the superiority of the proposed approach, achieving state-of-the-art accuracy in Assamese UPoS tagging. The research not only contributes to the field of natural language processing but also addresses the specific challenges posed by under-resourced languages. The significance of Incremental Learning is highlighted, showcasing its role in dynamically updating the model's knowledge base with new UPoS-tagged data. This feature proves essential in real-world scenarios where language evolves, ensuring sustained optimal performance in Assamese UPoS tagging. The paper presents the details of the innovative framework for UPoS tagging in Assamese, combining the significance of Incremental Learning with Deep Learning techniques, pushing the boundaries of natural language processing models for low resource languages exploring the importance of dynamic learning paradigms.**

Keywords*—*Assamese UPoS; PoS tagger; RNN; GRU; incremental learning

I. Introduction

The landscape of natural language processing (NLP) continues to evolve rapidly, presenting both challenges and opportunities for understanding and analyzing diverse languages, particularly in the context of digital world and Artificial Intelligence. For under-resourced languages, such as Assamese, the need for robust language processing models becomes particularly challenging. This paper introduces an exploration into advancing the effectiveness of Universal-part-of-speech (UPoS) tagging in Assamese, a language with unique syntactic characteristics and limited linguistic resources, through experimenting Incremental Learning in a Deep Learning paradigm.

Assamese, spoken by a significant population of around 15 million native speakers in the North Eastern part of India, faces the inherent challenges of being a low-resource language in the NLP domain. The scarcity of annotated data and linguistic resources pose hurdles for developing accurate and adaptable language models. Part-of-speech tagging, a fundamental task in NLP, forms the cornerstone of syntactic analysis, providing crucial insights into the structure and meaning of sentences.

Motivated by the imperative to bridge the gap in language technology for underrepresented languages, this research concentrates into the intricacies of Assamese, aiming to enhance the accuracy and adaptability of UPoS tagging. The significance of this endeavor lies not only in its contribution to advancing NLP capabilities, but also in addressing the broader need for tailored solutions for languages with limited linguistic resources.

PoS tagging experiments in Assamese is relatively new, and only very few previous works could be traced. Experiments using ML/DL techniques have been done by few researchers, but with limited resources. With the advent of NLP applications, including Machine Translation, Sentiment Analysis, Summarization etc., demand for linguistically embellished corpus has started increasing. And Corpus with PoS tagged embellishment is a need for most of the preprocessing and transfer learning pipelines. The bottleneck is the low amount of required resources, including tagged corpus. Another concern is the non-universality of tagset. We standardized the PoS tagset for Assamese language mapping the BIS tagset to the Universal PoS tagset, opening a new dimension for Assamese NLP with universal adaptation for inter-linguistic NLP works. The resources, generated so, shall be of benefits to the Assamese NLP research community for advancing different tasks. Low-resource constraints have been tried to overcome with the integration of incremental learning concept. Both UPoS tagging, as well as experimenting with incremental learning paradigm, are novel to the Assamese NLP, and significantly contribute to the overall resource and experiment scenario.

The primary objective of this research is to introduce Incremental Learning to Assamese UPoS tagging by leveraging the strengths of Recurrent Neural Networks (RNNs) and Gated Recurrent Units (GRUs). Incorporation of Incremental Learning, a dynamic paradigm is used for enabling the models to continuously refine their understanding as they encounter new sets of UPoS annotated linguistic data. The paper includes a comprehensive review including Assamese NLP works, and few on PoS, UPoS experiments. Next section discusses the methodology of the entire work including dataset and model architecture. How the incremental learning has been integrated, also included as part of this section. Here, the experimental flow alongwith the details of

the dataset have been elaborated. Next chapter includes the complete results with summarization of the analysis reflecting the efficacy of the incremental learning for low resource situation in doing DL based automatic PoS tagging.

II. Scope and Structure

This study encompasses a comprehensive exploration of UPoS tagging in Assamese, emphasizing the resource constraints of the language. The proposed models' adaptability through Incremental Learning is positioned as a pivotal aspect, catering to the evolving nature of the language. The scope extends beyond the immediate task, aiming to contribute insights and methodologies that can be easily replicated through similar experimentations to other low-resource languages facing similar challenges.

The remainder of this paper unfolds as follows: next provides a comprehensive review of related works in the field, highlighting existing approaches and challenges faced in context of under resourced languages. Then the methodologies are described, detailing the architecture of the proposed model, and explaining the incorporation of Incremental Learning with GRU and RNN. Next chapter presents the experimental setup and results, offering a comparative analysis of the experimented models against existing methods. Finally, we conclude the paper discussing implications, and suggesting avenues for future research.

III. Literature Review

This literature review surveys pivotal research to enrich our investigation into refining Assamese Universal-part-of-speech tagging through Incremental Learning with GRU and RNN models. Foundational works by Elman (1990) [1] on Recurrent Neural Networks (RNNs) and Cho et al. (2014) [2] on Gated Recurrent Units (GRUs) elucidate the architecture of sequential data processing, with recent optimizations by Chung et al. (2014) [3] further shaping the application of these models in natural language processing. Schmidhuber's exploration of Incremental Learning in NLP (1991) [4] and Fernando et al.'s work on gradient descent in super neural networks (2017) [5] lay the groundwork for our dynamic learning paradigm. There are several fundamental works initiating and standardizing the Assamese NLP tasks, that provides insights into Assamese NLP resources, as well as design and development of tools and technologies. Fundamental resources for Assamese NLP tasks have been created at various levels. These are impressive starting although not sufficient. A structured Assamese Corpus (GUIT Corpus) was built by Sarma et al. (2012) [6], covering a wide variety of domains, and collecting standard written texts in a much longer timeline. Different approaches have been already implemented for almost all major Indian languages for PoS tagging tasks (Kuwali and Shikhar, 2023) [7]. PoS tagging experimentations also have been carried out for Assamese language both using traditional approaches (Barman et al., 2013) [8] as well as using contemporary Machine Learning techniques. Assamese is relatively new for language processing research, still various preliminary works such as Wordnet development (Sarma et al., 2010) [9], statistical Machine Translation (Baruah et al., 2014) [10], Word Sense Disambiguation (Sarmah et al., 2016) [11, 12], Neural Machine Translation (Ahmed et al., 2023) [13] etc. could be traced in recent years. Application oriented works like Wordnet enhanced MT (Barman et al., 2014) [14], Word corrections (Bhuyan and Sarma, 2018) [15], development of rule based stemmer (Sarmah et al., 2019) [16], Assamese Information Retrieval system using Wordnet (Barman et al., 2013) [17] etc. also could be seen.

PoS tagging in Indian languages are predominantly using the Bureau of Indian Standard tagset (BIS, 2021) [18], although Universal Parts of Speech tagset (Marie et al., 2021) [19] as defined in the Universal Dependency also gained pace very recently (Das et al., 2023) [20]. This is promising in the sense that a Universally accepted tagset for across the language landscape shall facilitate transparency and adaptability in multilingual NLP research and application development.

IV. Methodology

Here we detail the technical methodology behind our approach for enhancing Assamese part-of-speech tagging using UPoStagset through Incremental Learning with GRU and RNN models. This chapter establishes the technical framework supporting our approach, providing a detailed insight into the dataset, model architecture, training procedures, and evaluation metrics that form the core of our investigation.

A. Dataset Preparation

We begin by curating a comprehensive dataset for Assamese, comprising diverse linguistic structures and contexts. This dataset includes three chunks of 10000 gold standard sentences each, extracted from the GUIT corpus. While extracting the sentences from the GUIT Corpus, special attention has been given to include diversified domains. The statistical information on the dataset is given in the Table I. The raw corpus of approximately 35000 sentences has been subjected to data cleaning and filtering to arrive at a cleaned dataset of fair quality. The dataset is considered a gold standard, as the corpus has been created with expert linguists, and the sources are standard written texts. Thus, the corpus represents the Assamese linguistic behaviours reflecting the syntactic and lexical coverages, and assumed to be embedded with varied linguistic phenomena inherent in Assamese.

TABLE I. Statistical Information on the Dataset

Dataset	Seq. Length	Frequency	% Frequency	Total	Dataset Label
Data Chunk-1	5-10	1926	19.26	10000	D1
	11-20	3560	35.60		
	21-30	2896	28.96		
	31-35	1618	16.18		
Data Chunk-2	5-10	1723	17.23	10000	D2
	11-20	3669	36.69		
	21-30	2635	26.35		
	31-35	1973	19.73		
Data Chunk-3	5-10	1867	18.67	10000	D3
	11-20	2970	29.7		
	21-30	3345	33.45		
	31-35	1818	18.18		

The following cleaning and filtering processes have been adopted:

1) Removing all blank lines

2) Excluding all sequences of less than 5 token and more than 35 tokens. These values are considered from expert linguistic inputs to consider only realistic Assamese sentences. As PoS tagging is a sequence labelling task, a fair length of Assamese text sequences shall only contribute to the machine learning in a better way.

3) Removing all unwanted and foreign sequences. This includes sequences written in scripts other than Assamese, as well as html segments and only-symbol segments.

The cleaning and filtering have been done with a customized Python script run over the raw corpus. The raw corpus, originally in the form of xml tagged text file, has been converted into excel file. An intermediary text file was again created with the sentences, and the Python script has been run over this text file to perform the cleaning and filtering operations. Fig. 1 and Fig. 2 shows the frequency of number of sentences and percentage distribution of sentences based on the Sequence-Lengths. The stages of dataset preparation are depicted inFig. 3.

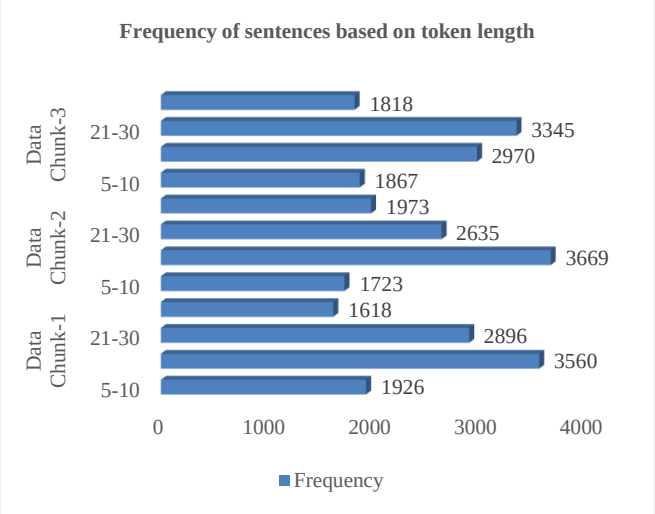

Fig. 1. Token-Length wise frequency distribution of sentences.

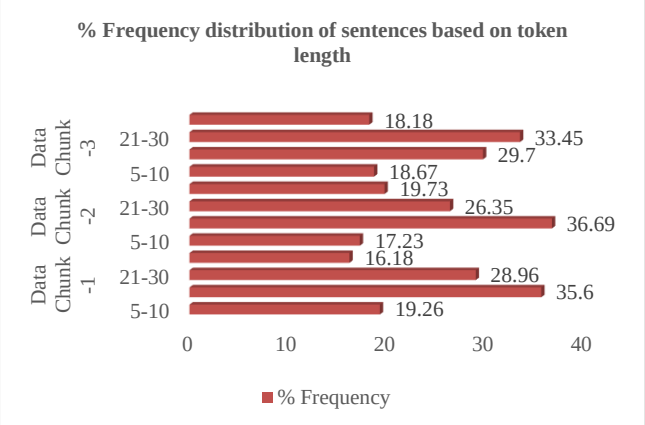

Fig. 2. Token-Length wise percentage frequency distribution of sentences.

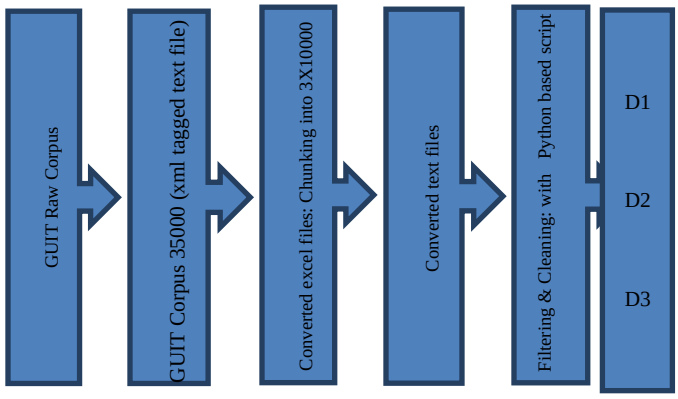

Fig. 3. Dataset preparation stage diagram

B. Model Architecture

Our experimented model leverages the strengths of Gated Recurrent Units (GRUs) and Recurrent Neural Networks (RNNs) for capturing sequential dependencies within Assamese sentences. The GRU's ability to selectively update information and its computational efficiency, and the RNN's contextual understanding, synergistically contribute to the effective architectures for part-of-speech tagging performance enhancement with Incremental Learning. The Incremental Learning is based on the baseline GRU and RNN models that were achieved against a similar set of standalone modelling with a single-shot dataset of 29501 Assamese UPoS tagged sequences containing 200022 tokens. These Assamese UPoS trained GRU and RNN models are subjected to D1, D2, and D3 in an incremental manner of training and tagging. Performances of the base GRU and RNN models are shown in Table II.

TABLE II. PERFORMANCES OF THE BASE GRU AND RNN MODELS

Models	Accuracy	Precision	Recall	F1
RNN (UPoS)	93.78%	94.75	93.28	**94.01**
GRU (UPoS)	94.38%	95.44	93.70	**94.56**

C. Incremental Learning Integration

A novel aspect of our methodology involves the integration of Incremental Learning to adapt the model dynamically to evolving linguistic patterns. For training the models with incremental learning approach, the base model already trained with UPoS tagged Assamese data is considered as the pretrained model, and the first chunk of dataset D1 is tagged with this model. This 10000 tagged sentences dataset is then added to the previous dataset, and fresh training is subjected through the GRU and RNN. The three chunks are incrementally added to enhance the size of the dataset, and performances have been recorded. In this approach, as new data are encountered, models are trained afresh with larger chunks of dataset, contributing positively to the inherent data-hungry nature of deep learning. The incremental dataset sizes against the training iterations are shown in Table III.

The model undergoes training with careful consideration of hyperparameter tuning. Training performance is monitored through relevant metrics such as accuracy, precision, recall, and F1 score. The experiments also include comparisons with

baseline models and existing methodologies for a comprehensive assessment.

TABLE III. INCREMENTAL DATASET SIZES AGAINST ITERATIONS

Training Iteration	Dataset	#Sequence	Training Model
0	D0 (base)	29501	Base GRU, RNN Model
1	base+ D1-first 10K chunk	39501	Model 1
2	base+ D2-second 10K chunk	49501	Model 2
3	base+ D3-third 10K chunk	59501	Model 3

Our evaluation metrics encompass standard measures such as accuracy, precision, recall, and F1 score to analyse the model's effectiveness in part-of-speech tagging. The impact of Incremental Learning on the model's adaptability over the trajectory are recorded and presented.

V. EXPERIMENTAL SETUP AND RESULTS

In this chapter, our experimental design is systematically outlined, encompassing phases of training, validation, and testing. Batch sizes, number of layers are considered as key system parameters for optimal model performance.

We have configured the deep learning pipeline in a laboratory environment with local server. The server configuration is outlined below:

- 64 bit Intel Xeon CPU,
- 16 GB main memory,
- NVIDIA Quadro P1000 GPU
- 640 CUDA Cores and
- 4096 MB of GPU memory.
- System's graphics clock speed: min-136 MHz, max-5010 MHz
- Graphics RAM 4 GB.
- System storage: 256 GB SSB and 1 TB HDD.

The entire training setup was done with Tensorflow. The pipeline for training includes *keras*and *sklearn* packages. The powerful python library *Keras* has been used to import deep learning models-RNN and GRU. Splitting of training and testing dataset has been done by *Sklearn* package. Two other important tools of python-*pandas* and *numpy* also have been used in the framework.

Both default and customized set of hyperparameters are part of the experiments. Both RNN and GRU were configured with 64 Cell Model layers. The model parameters are depicted in the Table IV. The architecture pipeline framework is given in Fig. 4.

TABLE IV. MODEL HYPERPARAMETERS

Models	Embedding Layer	Model Layer	Dense Layer
RNN	Input dim = vocab size, Output dim = 300, Input length = 100	RNN Layer = 64 Cell	36 no. of classes
GRU	Input dim = vocab size, Output dim = 300, Input length = 100	GRU Layer = 64 Cell	36 no. of classes

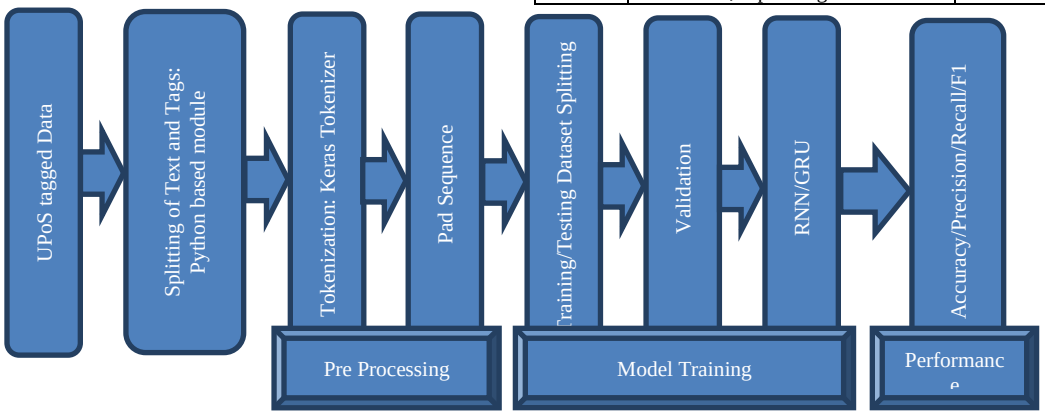

Fig. 4. Architecture pipeline framework.

Conducting a comparative analysis against state-of-the-art baseline models, we establish a benchmark for evaluating the advancements achieved through our Incremental Learning approach. Key evaluation metrics, including accuracy, precision, recall, and F1 score, are analyzed. The focus is on the model's performance in part-of-speech tagging and the impact of Incremental Learning on continuous improvement over time.

Experimental results are systematically presented here through Tables V to IX, and Fig. 5 to 9, emphasizing the unique contributions of our models, and the implications of Incremental Learning in the context of Assamese universal-part-of-speech tagging. Accuracy, precision, and recall against different experiments with incremental Dataset chunks have been presented both quantitatively and graphically.

TABLE V. PERFORMANCE OF MODEL 1

Models	Accuracy	Precision	Recall	F1
RNN (UPoS):Model1	94.22	95.57	93.82	**94.69**
GRU (UPoS): Model1	95.42	96.21	95.00	**95.60**
Base RNN (UPoS)	93.78	94.75	93.28	**94.01**
Base GRU (UPoS)	94.38	95.44	93.70	**94.56**
Increase in Model1 RNN (UPoS)	0.44	0.82	0.54	**0.68**
Increase in Model1 GRU (UPoS)	1.02	0.77	1.30	**1.04**

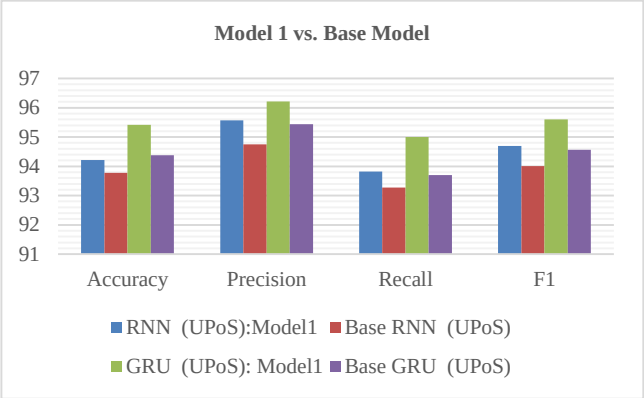

Fig. 5. Model 1 vs. Base model.

TABLE VI. PERFORMANCE OF MODEL 2

Models	Accuracy	Precision	Recall	F1
RNN (UPoS):Model2	94.75	96.45	94.78	**95.61**
GRU (UPoS):Model2	96.53	97.27	96.08	**96.67**
RNN (UPoS):Model1	94.22	95.57	93.82	**94.69**
GRU (UPoS):Model1	95.42	96.21	95.00	**95.60**
Increase in Model2 RNN (UPoS)	0.53	0.88	0.96	**0.92**
Increase in Model2 GRU (UPoS)	1.11	1.06	1.08	**1.07**

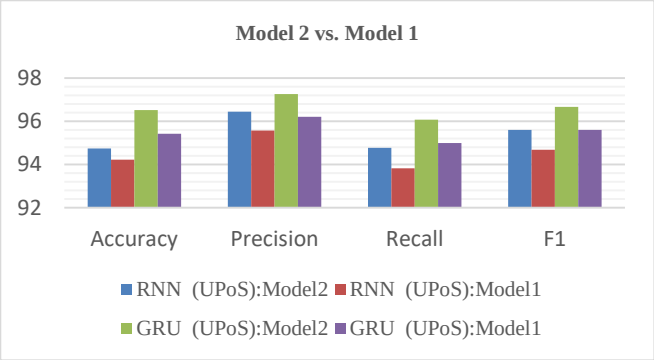

Fig. 6. Model 2 vs. Model 1.

TABLE VII. PERFORMANCE OF MODEL 3

Models	Accuracy	Precision	Recall	F1
RNN (UPoS):Model3	95.63	97.23	95.86	**96.54**
GRU (UPoS):Model3	97.56	97.98	97.26	**97.62**
RNN (UPoS):Model2	94.75	96.45	94.78	**95.61**
GRU (UPoS):Model2	96.53	97.27	96.08	**96.67**
Increase in Model2 RNN (UPoS)	0.88	0.78	1.08	**0.93**
Increase in Model2 GRU (UPoS)	1.03	0.71	1.18	**0.95**

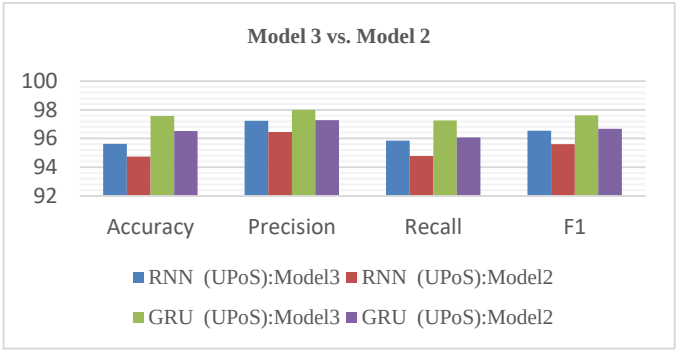

Fig. 7. Model 3 vs. Model 2.

TABLE VIII. PERFORMANCE ANALYSIS OF ALL MODELS

Models	Accuracy	Precision	Recall	F1
Base RNN (UPoS)	93.78	94.75	93.28	**94.01**
Model 1 RNN (UPoS)	94.22	95.57	93.82	**94.69**
Model 2 RNN (UPoS)	94.75	96.45	94.78	**95.61**
Model 3 RNN (UPoS)	95.63	97.23	95.86	**96.54**
Base GRU (UPoS)	94.38	95.44	93.70	**94.56**
Model 1 GRU (UPoS)	95.42	96.21	95.00	**95.60**
Model 2 GRU (UPoS)	96.53	97.27	96.08	**96.67**
Model 3 GRU (UPoS)	97.56	97.98	97.26	**97.62**

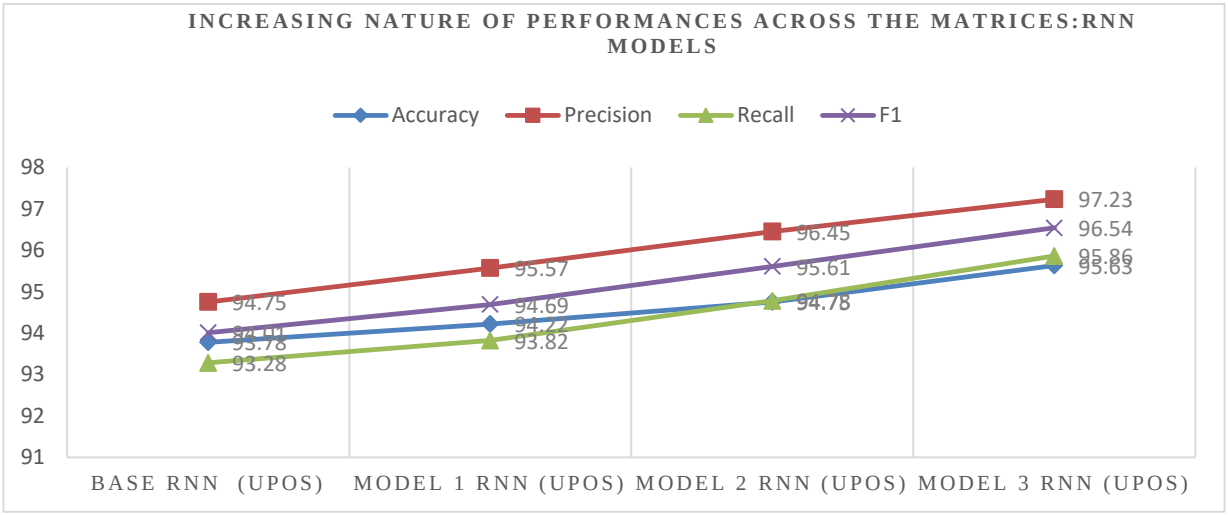

Fig. 8. Increasing nature of performances of RNN models.

TABLE IX. PERFORMANCE INCREMENT ACROSS MODELS

Models	Accuracy	Precision	Recall	F1
Base to Model 1 RNN (UPoS)	0.44	0.82	0.54	0.68
Model 1 to Model 2 RNN (UPoS)	0.53	0.88	0.96	0.92
Model 2 to Model 3 RNN (UPoS)	0.88	0.78	1.08	0.93
Base to Model 1 GRU (UPoS)	1.04	0.77	1.30	1.04
Model 1 to Model 2 GRU (UPoS)	1.11	1.06	1.08	1.07
Model 2 to Model 3 GRU (UPoS)	1.03	0.71	1.18	0.95

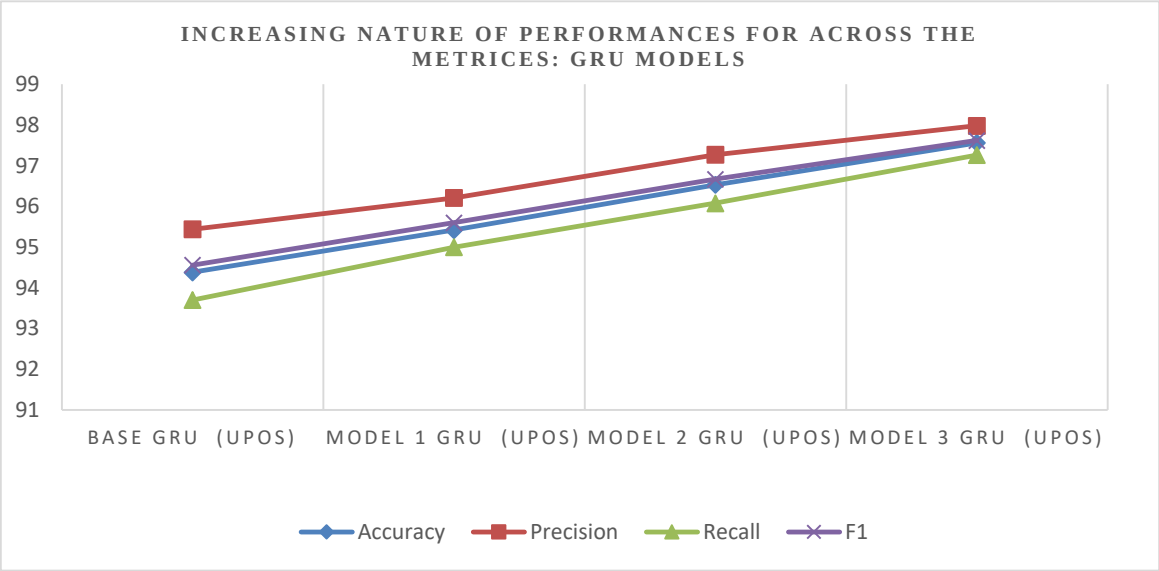

Fig. 9. Increasing nature of performances: GRU models.

F1 scores are calculated for all models both for GRU and RNN based experiments. The results have reflected positive performance enhancement through incremental training. The impact of incremental learning for the DL based UPoS tagger modeling has been investigated to be contributory, and is the major contribution of this study. Low resource languages with contraints of having sizable dataset may be greatly benifited from the experience of the current study and the significance of incremental learning.

It is evident from the result analysis that Incremental Learning has positively contributed to the performances of both GRU and RNN. The baseline GRU and RNN models for Assamese UPoS tagging are added with increased performances of upto 3.18 points for Accuracy and 3.06 points for F1 score. Performances of the models through base model to the model 3 are graphically represented in Fig. 8 and 9. Steady percentage increase of performances across the matrices suggest that increased dataset size, and subsequent incremental training of Deep Learning model successfully contribute to the enhanced performance of the models. This could be best utilized for low resource language PoS tagging tasks in most effective manner.

VI. CONCLUSION

The impact of Incremental Learning on adaptability and continuous improvement over a series of incremental dataset are highlighted in this paper. Here, we reflect on the broader contributions of our research to the field of natural language processing, particularly in the context of under-resourced languages, by adopting Incremental Learning in sequence labelling task, and the experimental finding that Incremental Learning could be best utilized as an efficient approach for resource poor situation. Our models' ability to dynamically adapt to evolving linguistic patterns and the advancements achieved through Incremental Learning are underscored as pivotal contributions. Practical implications of the current research, as we visualize, particularly in the context of resource constraint situation, the experimented approach could be best exploited for, opening avenues for the development of language processing tools for other underrepresented languages. The findings could be well replicated with experiments across the low resource languages, as this has proved to be effective in enhancing performances. Low resource languages are always with constraints of having sizable dataset, and at the same time ML/DL training requires good amount data reflecting maximum possible patterns like syntax, well-covered vocabulary etc. in order to tap all possible linguistic phenomenon. Incremental learning with larger data chunks may be experimented in future for more effective performances. Transfer learning with pre trained models shall be of immense potentiality, to experiment for probable enhanced performance in similar situations.

REFERENCES

[1] Elman, J. L. (1990). "Finding Structure in Time." Cognitive Science, 14(2), 179-211.

[2] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., &Bengio, Y. (2014). "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation." arXiv preprint arXiv:1406.1078.

[3] Chung, J., Gulcehre, C., Cho, K., &Bengio, Y. (2014). "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling." arXiv preprint arXiv:1412.3555.

[4] Schmidhuber, J. (1991). "A Possibility for Implementing Curiosity and Boredom in Model-Building Neural Controllers." In Proc. of the International Conference on Simulation of Adaptive Behavior.

[5] Fernando, C., Banarse, D., Blundell, C., Zwols, Y., Ha, D., Rusu, A. A., ... & Wierstra, D. (2017). "PathNet: Evolution Channels Gradient Descent in Super Neural Networks." arXiv preprint arXiv:1701.08734.

[6] Shikhar Kr. Sarma, HimadriBharali, AmbeswarGogoi, RatulDeka, and Anup Kr. Barman. 2012. A Structured Approach for Building Assamese Corpus: Insights, Applications and Challenges. In Proceedings of the 10th Workshop on Asian Language Resources, pages 21–28, Mumbai, India. The COLING 2012 Organizing Committee.

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

[8] A.K. Barman, J. Sarmah and S. K. Sarma, "POS Tagging of Assamese Language and Performance Analysis of CRF++ and fnTBL Approaches," 2013 UKSim 15th International Conference on Computer Modelling and Simulation, Cambridge, UK, 2013, pp. 476-479, doi: 10.1109/UKSim.2013.91.

[9] SarmaShikhar, GogoiMoromi, Medhi Rakesh and SaikiaUtpal. 2010. Foundation and Structure of Developing an Assamese Wordnet. Proceedings of 5th International Conference of the Global WordNet Association (GWC-2010)

[10] 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.

[11] JumiSarmah, 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.

[12] J. Sarmah and S. K. 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.

[13] M. A. Ahmed, K. Kashyap and S. K. 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.

[14] Anup Barman, JumiSarmah, and ShikharSarma. 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.

[15] M. P. Bhuyan and S. K. Sarma, "Automatic Formation, Termination & Correction of Assamese word using Predictive & Syntactic NLP," 2018 3rd International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2018, pp. 544-548, doi: 10.1109/CESYS.2018.8724023.

[16] JumiSarmah, 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.

[17] A. K. Barman, J. Sarmah and S. K. 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.

[18] Bureau of Indian Standards.(2021) “Linguistic Resources-POS Tag Set for Indian Languages-Guidelines for Designing Tagsets and Specification.” www.bis.gov.in, www.standardsbis.in

[19] Marie-Catherine de Marneffe, Christopher D. Manning, JoakimNivre, and Daniel Zeman. 2021. Universal Dependencies. Computational Linguistics, 47(2):255–308.

[20] Das, A., Choudhury, B., Sarma, S.K. (2023). POS Tagging for the Primitive Languages of the World and Introducing a New Set of Universal POS Tagging for Sanskrit. In: Fong, S., Dey, N., Joshi, A. (eds) ICT Analysis and Applications. Lecture Notes in Networks and Systems, vol 517. Springer, Singapore. https://doi.org/10.1007/978-981-19-5224-1_3