INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Parts-of-Speech Tagging for Unknown Words in Assamese using Viterbi Algorithm

Rituraj Phukan[1]*, Nomi Baruah[1], Shikhar Kr Sarma[2], Darpanjit Konwar[1]

1 Dibrugarh University, Dibrugarh, Assam, India
2 Gauhati University, Guwahati, Assam, India

OPEN ACCESS

Received: 23-03-2023
Accepted: 26-06-2023
Published: 02-11-2023

Editor: Guest Editor: Dr. Madhuryya Saikia & Dr. Niranjan Bora

Citation: Phukan R, Baruah N, Sarma SK, Konwar D (2023) Parts-of-Speech Tagging for Unknown Words in Assamese using Viterbi Algorithm. Indian Journal of Science and Technology 16(SP2): 53-59. https://doi.org/10.17485/IJST/v16iSP2.8203

* **Corresponding author**.

rituraj phukan01@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: This study aims to explore the application of the Viterbi algorithm for Part-of-Speech (POS) tagging in the Assamese language, focusing on tagging out-of-vocabulary words. The objective of this paper is to assess the algorithm's performance using various training and testing data ratios of 50:50, 70:30, and 90:10. **Methods:** The study utilizes the dynamic programming capabilities of the Viterbi algorithm to determine the most likely sequence of hidden states based on observable events. A corpus comprising approximately 50,000 words is employed to train the algorithm, with different ratios of this data utilized for training and testing purposes. **Findings:** The Viterbi algorithm achieves an accuracy of 86.34%, surpassing the state-of-the-art POS taggers for the Assamese language. The experimental evaluation demonstrates that the proposed approach outperforms previously existing research work by achieving 6.14% higher accuracy in tagging out-of-vocabulary words, highlighting its effectiveness in addressing the challenges associated with less-resourced languages like Assamese. **Novelty:** The results of this study contribute to the understanding and development of POS tagging techniques in less-resourced languages like Assamese. The proposed approach not only achieves superior performance in terms of accuracy but also showcases its potential for improving POS tagging in similar linguistic contexts, surpassing the achievements of previous research efforts.

Keywords: Assamese; NLP; Outofvocabulary; POS Tagging; Viterbi Algorithm

1 Introduction

Parts-of-Speech (POS) Tagger is one of the most important components in the development of applications in different fields of Natural Language Processing (NLP). POS tagging, also known as grammatical tagging is the process of labeling a word in a sentence as relating to a part of speech, based on both its definition and its context[(1)]. POS tagging is employed as a preprocessing activity by other Natural language processing (NLP) applications like machine translation and relation extraction to enhance their performance, it is a crucial study area to achieve high-quality research

in other NLP areas [2]. An example of POS tagging is given below,

For example, For example, তাই<PPR>বহুত<JINT>পলমকৈ<AMN>ভাবি<VM>পালে<VA>।<PUN> / Tai bohut polom koi bhabi pale/She understood very late.

Here, "তাই/tai" is a pronoun, "পলমকৈ/polomkoi" is an adverb, "ভাবি/bhabi" is a main verb, "পালে/pale" is an auxiliary verb, and "।/." is a punctuation mark. Pronouns, adverbs, verbs, and punctuation marks are the POS tag of a sentence.

In the context of Assamese language processing, an important aspect of POS tagging is handling out-of-vocabulary (OOV) words. OOV words are those that are not present in the predefined vocabulary or training data used by a POS tagger. The presence of OOV words poses a significant challenge for accurate tagging. To address this challenge, this study focuses on the development of a POS tagger for handling OOV words using the Viterbi algorithm. The Viterbi algorithm uses contextual information to assign the most likely POS tags to OOV words based on their surrounding words [3]. It improves tagging accuracy by considering the context of a particular word.

Assamese is the most eastern Indo-Aryan language spoken in India, with 30 million people speaking it primarily in Assam [4]. Assamese is a subject-object-verb (SOV) structured language with a strong morphological base. POS tagging in Assamese is a challenging task because of the ambiguity present in Assamese words. For example, the word **"ৰবি/robi"** neans "Sun", It could also be the name of someone. In these cases, POS of the word**"ৰবি/robi"** is noun. But another meaning of **"ৰবি/robi"** s to tell someone to stop, where the POS tag of the same word changes to a Verb. A POS tagger should identify this kind of ambiguity and be able to tag each word accurately depending on its contextual meaning.

In the domain of POS tagging for the Assamese language, there is a noticeable scarcity of research work and high-performing taggers specifically designed to handle unknown or OOV words. OOV words pose a significant challenge in the accurate tagging of words, as they are not included in the training corpus and lack predefined tags. The limited existing research on OOV word tagging in Assamese signifies the need for further exploration and advancements in this particular area. While various POS taggers have been developed for Assamese, their focus has primarily been on well-represented words rather than addressing the challenges posed by OOV words. This research gap highlights the significance of developing specialized taggers that can effectively handle OOV words in Assamese. Constructing a comprehensive corpus that includes every word in a language is an arduous task, given the continuous creation of new words or variations of existing ones. Currently, there is only one existing POS tagger for Assamese that addresses OOV words, but it uses a very small corpus. Recent research in POS tagging does not specifically focus on tagging OOV words [5]. In comparison, POS taggers developed for other languages achieve higher accuracy due to extensive research efforts and the utilization of larger datasets [6]. In the course of developing this paper, POS tagging in other low-resource languages, such as Dogri [7] and Nyishi [8], has also been studied.

The primary contribution lies in the development of a tailored POS tagger that effectively handles OOV words, improving the accuracy of POS tagging in the Assamese language. Additionally, this research contributes to the linguistic resources available for Assamese language processing by creating a corpus comprising approximately 50,000 words. This corpus serves as a valuable asset for further research and development in Assamese, enabling researchers to explore various aspects of language processing and analysis.

In this paper, we present POS tagging for unknown words in Assamese. Section 2 of this paper describes the methodology of the presented research work. In Section 3, the result obtained from this work and an analysis of the performance is presented, and a comparison between the proposed work and already existing research work in POS tagging of the unknown word for Assamese is shown. Section 4 is the conclusion of this paper, and the future scope of this research work is also discussed.

2 Methodology

In this section, we present different steps that have been followed in our proposed work. Figure 1 shows a pictorial diagram of the presented POS tagger using the Viterbi algorithm.

2.1 Tagset

The tagset we used was developed by CIIL Mysore Assamese tagset. There is a total of 13 primary tags in the tagset. The primary tags are then expanded into further sub-categories which combine up to total of 38 tags. The List of tags that are present in CIIL Mysore tagset are, Common Noun (NC), Proper Noun (NP), Verbal (NV), Spatio-temporal Noun (NST), Pronominal Pronoun (PPR), Reflexive Pronoun (PRF), Reciprocal Pronoun (PRC), Relative Pronoun (PRL), Wh-pronoun (PWH), Absoluteive Demonstrative (DAB), Relative Demonstrative (DRL), Wh-demonstrative (DWH), Adjective Nominal Modifier (JJ), Quantifier Nominal Modifier (JQ), Intensifier Nominal Modifier (JINT), Main Verb (VM), Auxiliary Verb (VA), Manner Adverb (AMN), Location Adverb (ALC), Post- position (PP), Co-ordinating Particle (CCD), Subordinating Particle (CSB), Interjection Particle (CIN), (Dis)Agreement Particle (CAGR), Delimitive Particle (CDLIM), Dedative Particle

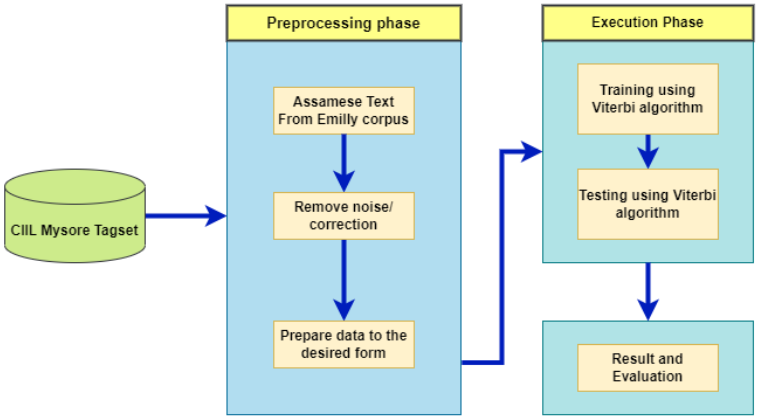

Fig 1. Pictorial diagram of the proposed work

(CDED), Dubitative Particle (CDUB), Similative Particle (CSIM), Other Particle (CX), Real Numeral (NUMR), Serial Numeral (NUMS), Calendric Numeral (NUMC), Ordinal Numeral (NUMO), Reduplication (RDP), Residual (RD), Unknown (UKN), Punctuation (PU).

2.2 Data

The corpus that we are using is developed in EMILLE project, which is a machine-generated corpus[9]. We manually annotated a portion of the data from that corpus because it contains several incorrectly tagged words and some noisy data. The Assamese tagset created by CIIL Mysore is the foundation of the corpus. A linguistic expert from the Department of Assamese, Dibrugarh University, India, reviewed and corrected the dataset after it was manually annotated. The inaccurate or noisy words are marked with a red box inFigure 2, which displays some of the incorrectly tagged words and noise found in the machine-generated corpus.

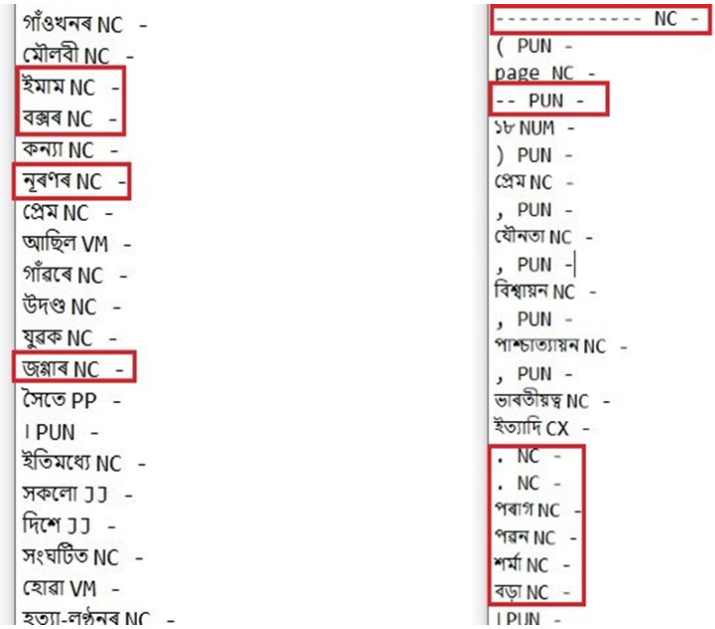

Fig 2. Incorrect and noisy data in the corpus

After correction, the data were transformed using some Python code into the necessary form so that they could be utilized as input for the POS tagger. The input data used to train the POS tagging model is depicted in Figure 3. The dataset that we used for our research included about 50,000 tagged words.

Fig 3. Corrected data used for training and testing

2.3 Viterbi Algorithm

The Viterbi algorithm is a statistical approach that aims to find the optimum role sequence from a given set of data[10]. The Viterbi algorithm uses dynamic programming that avoids the polynomial expansion of breadth-first search (BFS). This algorithm computes the most likely tag sequence by computing the probability of each search state. It mainly works based on the number of assumptions it makes[6].

Each state at time t is examined by the Viterbi algorithm, along with all of the transitions leading up to that state. It chooses the transition with the highest metric, or the one that is most likely to happen, out of all the possible transitions. One of the transitions is selected at random as the most likely transition if two or more transitions are discovered to have the same optimum matrices. This greatest metric is then assigned to the survivor path metric for the state. Following that, the Viterbi algorithm discards all other transitions into that state and adds this state to the survivor path of the state at (t – 1), where the transition originated. Hereafter, this becomes the survivor path of the state under consideration at time t. The same operation is performed on all the states at time t, after which the VA moves onto the states at (t + 1) and performs the same operations on those states. When we reach time (t = T) (the truncation length), the VA determines the survivor paths again, but this time it must also decide which of these survivor paths is the most likely. This is accomplished by determining the survivor with the highest metric; if more than one survivor has the highest metric, the most likely path taken is chosen at random. The Viterbi algorithm then returns the survivor path as well as the survivor metric[11].

2.4 Testing and validating data

The developed POS tagging model is fed with the training data. For testing purposes, first of all, we calculated the number of the unknown words present in the testing data. Then we test the model with the testing data and calculate the accuracy of tagged unknown words. We divided the corpus into different ratios for testing and training. For testing and training, we divided the corpus into several ratios. We evaluate the model using precision, recall, f1-score, and accuracy metrics for each case.

3 Results and Discussion

The outcome of the suggested work is discussed in this section. Our corpus was divided into three different ratios: 50:50, 70:30, and 90:10. We used 50% of the complete corpus as training data and 50% as testing data for the first experiment. We used 70% and 90% of the whole corpus as training data for the second and third experiences, and 30% and 10% as testing data, respectively. For the evaluation of the result, we used evaluation metrics such as precision, recall, F1-score, and accuracy. Formulae for

calculating precision, recall, F1-score, and accuracy are:

$$Precision = \frac{TP}{TP+FP} \tag{1}$$

$$Recall = \frac{TP}{TP+FN} \tag{2}$$

$$F1-score = \frac{Precision * Recall}{Precision + Recall} \tag{3}$$

$$Accuracy = \frac{TN+TP}{TN+FP+TP+FN} \tag{4}$$

Where,

TP=True Positive

FP=False Positive

TN=True Negative

FN=False Negative

The result achieved from our research work is shown in Table 1. It is observed from the result that, the highest accuracy i.e., 86.34% was obtained when the training and testing data ratio was 90:10.

Table 1. Result

Sl. No.	Ratio of training and testing data	Precision	Recall	F1- score	Accuracy
1	50:50	83.70%	83.70%	83.70%	83.70%
2	70:30	85.25%	85.25%	85.25%	85.25%
3	90:10	86.34%	86.34%	86.34%	86.34%

The performance of the developed POS tagger from our research work is tested by dividing the corpus for training and testing data into three different ratios as mentioned above. In the first case, when the testing and training data ratio was 50:50, the tagger shows less accuracy than in the other two cases. Whereas, dividing the corpus into 90:10 ratio shows the highest accuracy of the POS tagger. The occurrence of each tag in the training data when the ratio is 90:10 is shown in Table 2.

Table 2. Occurrence of each tag in the training data

Tag	Occurrence
VM	0.0777873010725289
NC	0.6128240172345897
PRL	0.0025481248117862356
NP	0.020222845097176215
CSB	0.0006486135884546781
JINT	0.0011350737797956867
NUMR	0.006648289281660451
NV	0.0020384998494289886
NST	0.00419282355393917
NUMC	0.00011582385508119252
PP	0.00850147096295953
AMN	0.02230767448863768
PWH	0.0015983692001204568
CCD	0.04334128657138224
RDP	0.0006717783594709166
NUMO	0.00039380110727605457

Continued on next page

Table 2 continued

RDF	0.00023164771016238503
CX	0.01053997081238852
DWH	0.00027797725219486205
JQ	0.004679283745280178
AV	0.00018531816812990804
VA	0.01348189673145081
MV	4.632954203247701e-05
PPR	0.04642220111654196
JJ	0.03875466191016702

The proposed POS tagger based on the Viterbi algorithm for unknown words achieves higher accuracy than the previously existing POS tagger for unknown words in Assamese. Although there is only one state-of-the-art study available for the Assamese language that provides accuracy measurements for OOV words using their tagger, this area of study remains relatively unexplored. Therefore, there is a significant research gap regarding the specific challenges associated with accurately tagging OOV words in Assamese. A comparison is shown in Table 3.

Table 3. Comparison between existing work and proposed work

Reference	Accuracy
[12]	80.20%
Proposed Approach	86.34%

The quality of the data used for training and testing the proposed POS tagger plays a crucial role in determining its performance. Also, the size of the dataset used in the proposed POS tagger, i.e., 50,000 words, is higher than the previously existing work [12] which used a dataset consisting of 25,000 words. The importance of dataset size and quality cannot be overstated, as they directly impact the effectiveness of the POS tagger. The dataset used in this research is meticulously annotated and can serve as a valuable resource for future Assamese NLP research. The availability of such a dataset contributes to the overall development of resources and tools for Assamese language processing, fostering further advancements in the field.

However, it is important to acknowledge the limitations of our POS tagger. Firstly, since there is only one previous work available for comparison, it becomes challenging to provide a comprehensive assessment of our tagger's performance against a range of different approaches. The lack of multiple baselines and comparative studies restricts our ability to draw more robust conclusions about the effectiveness of our tagger in comparison to a wider range of methodologies. Moreover, the accuracy of the proposed approach for tagging OOV words could be improved using deep learning approaches.

4 Conclusion

This study emphasizes for the construction of a Viterbi algorithm-based POS tagger specifically tailored for the Assamese language to handle OOV words. The experimental evaluation is conducted using a meticulously annotated corpus of approximately 50,000 words along with their tags, and various training and testing data ratios are explored. Different combinations of training and testing data were examined, such as 50:50, 70:30, and 90:10 splits. The results clearly demonstrated that the proposed POS tagger outperformed existing Assamese taggers in tagging unknown words. Notably, it achieved an accuracy of 86.34%, with the best results obtained using a 90:10 training and testing data ratio.The results demonstrate the superior performance of the proposed POS tagger compared to existing Assamese POS taggers in terms of tagging unknown words. This research contribution not only benefits POS tagging in Assamese but also holds potential for application in other NLP research fields within the language.

The result emphasizes the importance of using accurate and high-quality data for training and testing, which significantly contributes to the POS tagger's enhanced performance. However, it also acknowledges the scope for further enhancement. Future directions include expanding the size of the corpus and exploring alternative approaches to develop a POS tagger that can provide even more precise and accurate tagging results for unknown words in Assamese. These future endeavors aim to address the existing limitations and advance the state-of-the-art in Assamese POS tagging for improved language processing and analysis.

5 Declaration

Presented in Fourth Industrial Revolution and Higher Education (FIRHE 2023) during 23rd-25th Feb 2023, organized by DUIET, Dibrugarh University, India. The Organizers claim the peer review responsibility.

References

1) Chiche A, Yitagesu B. Part of speech tagging: a systematic review of deep learning and machine learning approaches. *Journal of Big Data*. 2022;9(10):1–25. Available from: https://doi.org/10.1186/s40537-022-00561-y.
2) Tehseen A, Ehsan T, Liaqat HB, Ali A, Al-Fuqaha A. Neural POS tagging of shahmukhi by using contextualized word representations. *Journal of King Saud University - Computer and Information Sciences*. 2023;35(1):335–356. Available from: https://doi.org/10.1016/j.jksuci.2022.12.004.
3) Gore T, Khatavkar V. Development of Part-of-Speech tagger for a low-resource endangered language. In: 2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), 16-17 December 2022, Greater Noida, India. IEEE. 2022;p. 1531–1535. Available from: https://doi.org/10.1109/ICAC3N56670.2022.10074031.
4) Das R, Singh TD. Image–Text Multimodal Sentiment Analysis Framework of Assamese News Articles Using Late Fusion. *ACM Transactions on Asian and Low-Resource Language Information Processing*. 2023;22(6):1–30. Available from: https://doi.org/10.1145/3584861.
5) Pathak D, Nandi S, Sarmah P. AsPOS: Assamese Part of Speech Tagger using Deep Learning Approach. In: 2022 IEEE/ACS 19th International Conference on Computer Systems and Applications (AICCSA), 05-08 December 2022, Abu Dhabi, United Arab Emirates. IEEE. 2022;p. 1–8. Available from: https://doi.org/10.1109/AICCSA56895.2022.10017934.
6) Dutta D, Halder S, Gayen T. Intelligent Part of Speech tagger for Hindi. *Procedia Computer Science*. 2023;218:604–611. Available from: https://doi.org/10.1016/j.procs.2023.01.042.
7) Jamwal SS. Development of POS tag set for the Dogri language using SMT. *International Journal of Electronics Engineering*. 2021;13(1):12–15. Available from: http://www.csjournals.com/IJEE/PDF13-1/3.%20Shub.pdf.
8) Siram J, Sambyo K, Sarkar A. Parts of Speech Tagging of the Nyishi Language Using Hmm. *Advanced Engineering Science*. 2022;54(02):3873–3880. Available from: https://advancedengineeringscience.com/article/pdf/3873.pdf.
9) Ali MNY, Rahman ML, Chaki J, Dey N, Santosh KC. Machine translation using deep learning for universal networking language based on their structure. *International Journal of Machine Learning and Cybernetics*. 2021;12:2365–2376. Available from: https://doi.org/10.1007/s13042-021-01317-5.
10) Nugraha DW, Richasdy D, Ihsan AF. Tagging Efficiency Analysis of Part of Speech Taggers on Indonesian News. *Jurnal Media Informatika Budidarma*. 2023;7(1):214–222. Available from: http://dx.doi.org/10.30865/mib.v7i1.5384.
11) Bharti SK, Gupta RK, Patel S, Shah M. Context-Based Bigram Model for POS Tagging in Hindi: A Heuristic Approach. *Annals of Data Science*. 2022;p. 1–32. Available from: https://doi.org/10.1007/s40745-022-00434-4.
12) Baishya D, Baruah R. Highly Efficient Parts of Speech Tagging in Low Resource Languages with Improved Hidden Markov Model and Deep Learning. *International Journal of Advanced Computer Science and Applications*. 2021;12(10):82–94. Available from: https://doi.org/10.14569/IJACSA.2021.0121011.