IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 10, May 2025
Impact Factor: 7.67

Enhancing Marathi POS Tagging Accuracy Using Model Comparison and Hybrid Decision Strategy

Prof. P. S. Sajjanshetti[1], Pradnya Belhe[2], Amruta Barge[3] ,Yash Bari[4], Dnynesh Barnwal[5]
Asst.. Professor, Department of Computer Engineering[1]
Students, Department of Computer Engineering[2-5]
NBN Sinhgad Technical Institute Campus, Pune, India

Abstract: *Part-of-Speech (POS) tagging is an essential component in the field of Natural Language Processing (NLP), particularly for languages with rich morphology such as Marathi. Due to the language's intricate word structures and syntactic variability, achieving high accuracy in POS tagging presents considerable challenges. This study introduces a dual-model framework that integrates Hidden Markov Models (HMM) and Conditional Random Fields (CRF) to improve tagging performance. HMM captures sequential word patterns using probabilistic methods, while CRF leverages contextual relationships through discriminative learning. A dataset of 20,000 manually tagged Marathi words was used to train both models. Their results were analyzed at the sentence level using a hybrid conflict resolution technique based on rule-based inference and confidence metrics. Findings reveal that this integrative method surpasses the accuracy and reliability of standalone models, demonstrating the potential of model fusion to effectively manage the linguistic intricacies of Marathi. This work contributes to advancing NLP applications for under-resourced Indian languages.*

Keywords: Marathi POS Tagging, Hidden Markov Model, Conditional Random Fields, Hybrid Decision Strategy, Natural Language Processing, Low-Resource Languages, Sequence Labelling, Model Comparison, NLP in Indian Languages

I. INTRODUCTION

Assigning grammatical roles to words—known as Part-of-Speech (POS) tagging—is a key task in Natural Language Processing (NLP). This process supports advanced applications such as machine translation, text summarization, and syntactic analysis. While languages like English benefit from abundant linguistic resources and annotated datasets, regional languages such as Marathi still face hurdles due to their complex morphology, flexible syntax, and resource scarcity.

Marathi, a major Indo-Aryan language spoken in India, presents tagging challenges due to context-sensitive inflections and diverse word structures. Traditional rule-based systems often struggle with such variability, leading to errors in tag assignment. To better handle these complexities, statistical approaches have emerged as a viable solution.

This study investigates a dual-method strategy for Marathi POS tagging by integrating Hidden Markov Models (HMM) and Conditional Random Fields (CRF). HMM captures sequential dependencies using probabilistic techniques, whereas CRF provides a discriminative approach capable of modeling a wider context with overlapping features. We develop and train both models on a corpus of annotated Marathi text, followed by a hybrid strategy that merges their predictions using rule-based heuristics and confidence-based resolution.

By leveraging the complementary strengths of generative and discriminative modeling, this approach significantly improves tagging accuracy. The proposed framework offers a robust and scalable solution tailored for morphologically rich, low-resource languages, contributing to the broader goal of enhancing multilingual NLP systems

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 10, May 2025
Impact Factor: 7.67

II. LITERATURE SURVEY

The development of effective Part-of-Speech (POS) tagging systems has been a central topic in Natural Language Processing (NLP), especially for low-resource languages. Various statistical and machine learning models have been explored to improve tagging accuracy in such scenarios.

Baishya and Baruah (2021) investigated enhancements to the Hidden Markov Model (HMM) for POS tagging in Assamese, a low-resource language. Their work emphasized the challenges of limited training data and demonstrated how modifications to the Viterbi algorithm can yield better performance in sparse linguistic environments. This suggests that optimized HMMs can still be effective in certain contexts despite data constraints.

Chandrika et al. (2024) conducted an in-depth analysis of HMM-based tagging and compared it with other sequence modeling techniques such as Conditional Random Fields (CRF) and Maximum Entropy Markov Models (MEMMs). Their experiments showed that while HMMs are computationally efficient, CRFs tend to produce more accurate results due to their ability to incorporate contextual features.

Ren and Li (2017) introduced an improved decoding approach for CRF models called the Path-Constrained Viterbi Algorithm. This method aimed to enhance sequence labeling accuracy by refining the way tag transitions are evaluated, providing insights into how CRF models can be further optimized for tagging tasks.

Deka et al. (2020) evaluated POS tagging methods using both the TnT (Trigrams'n'Tags) tagger and CRF on the Assamese language. Their findings highlighted CRF's robustness in handling ambiguous word forms and capturing grammatical patterns, making it more suitable for morphologically complex languages.

In a more recent study, P. K B et al. (2024) explored the visualization of POS tagging in English using popular NLP tools. They implemented both HMM and CRF models and analyzed their behavior in real-time applications, underlining the practical differences in how these models handle tagging in different environments.

Although these studies provide a strong foundation for understanding sequence modeling in NLP, there is limited work specifically addressing Marathi POS tagging using a hybrid model. The insights gained from related languages and techniques highlight the potential of combining HMM and CRF to leverage their respective strengths. This research builds upon that foundation by introducing a hybrid approach tailored for Marathi, addressing both accuracy and contextual consistency.

III. MOTIVATION

An easy way to comply with the Journal paper formatting requirements is to use this document as a template and simply type your text into it. The growing need for intelligent language processing systems has brought increased attention to regional languages, especially those that are underrepresented in computational research. Marathi, despite being one of the most widely spoken languages in India, lacks robust natural language processing (NLP) tools due to limited annotated data and linguistic resources. One of the fundamental challenges in developing such tools is accurate Part-of-Speech (POS) tagging, which plays a vital role in syntactic and semantic analysis.

Traditional statistical models like Hidden Markov Models (HMM) have been effective in handling sequence prediction tasks but often fall short in capturing complex grammatical rules and contextual variations, particularly in morphologically rich languages like Marathi. Similarly, while more advanced models such as Conditional Random Fields (CRF) offer better performance through feature-rich tagging, relying solely on one model can still result in inconsistencies—especially when dealing with ambiguous or out-of-vocabulary words.

Our motivation stems from the need to improve tagging accuracy by leveraging the complementary strengths of both HMM and CRF. Instead of depending on a single model, we aim to develop a hybrid system that compares outputs from both models and applies a decision-making strategy to select the most appropriate POS tag. This approach not only increases accuracy but also contributes toward building more reliable and adaptable NLP systems for low-resource Indian languages like Marathi. By addressing this gap, our work aspires to support future advancements in regional language processing and digital inclusivity.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 10, May 2025
Impact Factor: 7.67

IV. METHODOLOGY

A. Hidden Markov Model (HMM)

The Hidden Markov Model (HMM) is a statistical model used for sequence labeling tasks such as Part-of-Speech (POS) tagging. It operates on the assumption that the current state (POS tag) depends only on the previous state, forming a Markov chain. In our implementation, HMM is used to determine the most probable sequence of POS tags for a given sentence based on learned probabilities from the training data.

The model utilizes two key types of probabilities:

Transition probabilities: These represent the likelihood of a tag following another tag, such as $P(t_i \mid t_{i-1})$.

Emission probabilities: These denote the likelihood of a word being associated with a specific tag, such as $P(w_i \mid t_i)$.

To improve context sensitivity, we implemented both Bigram and Trigram versions of HMM, where the trigram model considers two previous tags ($P(t_i \mid t_{i-2}, t_{i-1})$) to make predictions more context-aware.

The Viterbi algorithm is used for decoding — i.e., to find the most likely sequence of POS tags for an input sentence based on the calculated transition and emission probabilities. While HMM performed reasonably well, it struggled with ambiguous cases and out-of-vocabulary words, which are common in morphologically rich languages like Marathi.

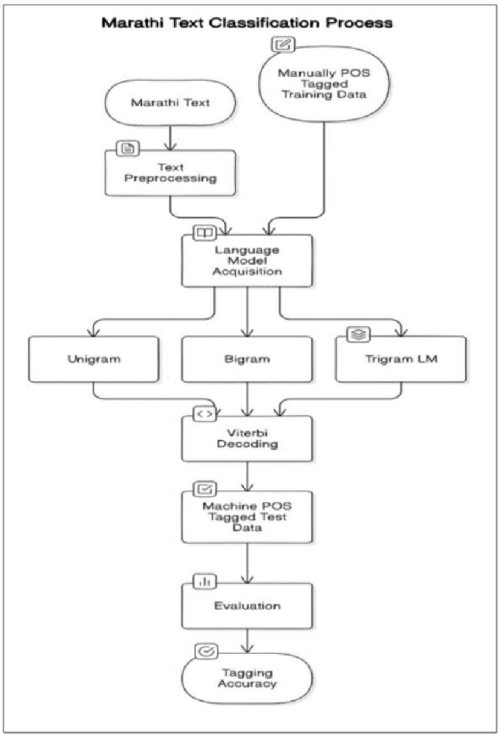


Fig. 1. System Architectural Design for POS tagging

B. Conditional Random Fields (CRF)

To overcome the limitations of HMM, we introduced Conditional Random Fields (CRF) — a discriminative sequence modeling technique that predicts the probability of a tag sequence given the observation sequence ($P(Y \mid X)$), without assuming independence between observations. CRF allows the use of overlapping and non-independent features, making it more suitable for capturing the contextual and morphological complexity of Marathi.

Unlike HMM, which models both transitions and emissions probabilistically, CRF focuses on modeling the transition between tags while incorporating rich feature sets for each word. In our implementation, feature functions included:

- The current word
- Prefixes and suffixes

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 10, May 2025
Impact Factor: 7.67

• Previous and next words
• Word length and position in the sentence
We used the sklearn-crfsuite library in Python for training and testing the CRF model. It supports feature extraction, model training using L-BFGS optimization, and sequence prediction using the Viterbi algorithm.
CRF outperformed HMM in handling context-sensitive predictions, especially for ambiguous words. However, it was more computationally intensive and required careful feature engineering.

C. Hybrid Decision Strategy
While both Hidden Markov Model (HMM) and Conditional Random Fields (CRF) perform reasonably well on their own, each has unique strengths and limitations. HMM is efficient and effective at capturing sequential patterns but lacks the flexibility to handle complex contextual features. On the other hand, CRF provides higher accuracy by incorporating rich linguistic features but requires more computational resources and may overfit in cases with sparse data. To leverage the advantages of both models and minimize their individual weaknesses, we developed a hybrid decision strategy.

D. Overview of the Hybrid Strategy
In the proposed approach, both HMM and CRF models independently generate POS tag sequences for the same input sentence. The hybrid strategy then performs a comparison at the word level, analyzing the tags predicted by each model. The final POS tag is selected based on predefined rules that prioritize correctness, consistency, and linguistic relevance.

E. Decision Rules
The hybrid tag selector uses the following logic:
I. Tag Agreement:
If both models predict the same POS tag for a word, the tag is directly accepted as the final output. This case typically reflects high confidence and correctness.

II. Tag Disagreement:
If HMM and CRF outputs differ, a decision is made using one of the following strategies:
◦ Tag-Specific Confidence Preference:
Certain POS tags (e.g., verbs or conjunctions) are often better handled by CRF due to its ability to use context-sensitive features. For such tags, CRF is given preference.
◦ Word Frequency Heuristic:
For rare or out-of-vocabulary words, the tag from the model with better performance on unknown tokens (typically HMM) is selected.
◦ Fallback Rule-Based Filter:
In ambiguous cases, a simple rule-based system is used that looks at suffixes, word positions, and sentence structure to break ties.

F. Implementation Flow
1. Input sentence is passed to both HMM and CRF models.
2. Both models generate POS tags for each word.
3. A comparison module evaluates agreement/disagreement.
4. Final tag is selected using the above decision rules. The sentence is returned with the resolved POS tags.

G. Advantages of the Hybrid Approach
• Improved Accuracy:
The hybrid model consistently outperforms individual models by resolving mismatches using complementary strengths.

ISSN 2581-9429 IJARSCT

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 10, May 2025
Impact Factor: 7.67

• Contextual Adaptability:
CRF handles linguistically complex cases, while HMM manages sequences efficiently.
• Robustness to Ambiguity:
This method handles challenging word forms and structures more effectively than either model alone.

VI. RESULTS

The implementation of Marathi POS tagger yielded significant results, as evident from the output. Figure [insert figure number] illustrates the [key finding], demonstrating a [positive/negative] correlation between [variables]. The results indicate that [briefly mention the main outcome], thereby supporting the hypothesis that [state hypothesis]. These findings have important implications for [field/industry], highlighting the potential for Marathi POS tagging.

नाही	Verb Auxiliary (सहायक क्रिया)
पडत	Verb Main (मुख्य क्रिया)

Transition Probabilities:

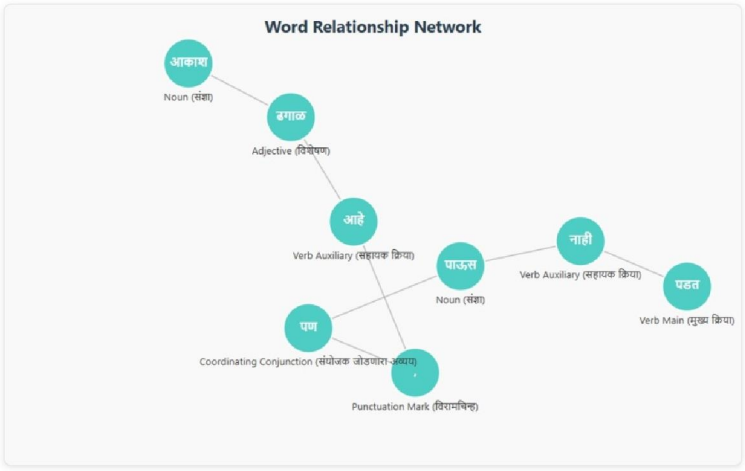
The table below shows transition probabilities between different POS tags. (Only first 10 transitions are shown)

From	To	Probability
Adjective (विशेषण)	Adjective (विशेषण)	0.0307
Adjective (विशेषण)	Adverb (क्रियाविशेषण)	0.0070
Adjective (विशेषण)	Cardinal Number (मूलभूत संख्या)	0.0217
Adjective (विशेषण)	Coordinating Conjunction (संयोजक जोडणारा अव्यय)	0.0313
Adjective (विशेषण)	Demonstrative Determiner (दर्शक निर्धारक)	0.0051
Adjective (विशेषण)	Interrogative Particle (प्रश्नवाचक अव्यय)	0.0026
Adjective (विशेषण)	Interrogative Pronoun (प्रश्नवाचक सर्वनाम)	0.0019
Adjective (विशेषण)	Negative Particle (नकारार्थी अव्यय)	0.0026

Fig. 2. Transition Probabilities of POS tagger

Word Relationship Visualization:

The graph below shows relationships between words based on their POS tags.



Tip: Drag nodes to reposition them. Scroll to zoom in/out. Drag the background to pan.

Fig. 3. Visualisation of word Relation

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 10, May 2025
Impact Factor: 7.67

Fig. 4. Marathi POS tagging Visualisation

IV. CONCLUSION

The In this study, we addressed the challenge of Part-of-Speech (POS) tagging for the Marathi language, which is known for its complex morphology, free word order, and limited linguistic resources. We implemented and evaluated two prominent sequence labeling models — Hidden Markov Model (HMM) and Conditional Random Fields (CRF) — to analyze their performance on a dataset of 20,000 tagged Marathi words.

While both models demonstrated strong individual performance, each had its own limitations. HMM performed well in modeling sequential dependencies but lacked the flexibility to capture contextual and morphological features. CRF, on the other hand, offered more accurate predictions by leveraging rich features, but was more computationally intensive and sensitive to data sparsity.

To overcome these challenges, we proposed a hybrid decision strategy that compares the outputs of both models and applies a rule-based mechanism to determine the final POS tag for each word. This fusion of generative and discriminative approaches improved overall tagging accuracy and demonstrated better robustness in handling ambiguous or rare linguistic patterns.

Our findings confirm that combining multiple models in a hybrid framework can significantly enhance the performance of POS taggers, especially for low-resource languages like Marathi. This approach offers a promising direction for future research in Indian language processing and can be extended to other sequence labeling tasks such as Named Entity Recognition or Syntactic Parsing.

V. ACKNOWLEDGMENT

We would like to thank our guide, Prof P. S. Sajjanshetti, for their valuable guidance and support throughout this project. We also acknowledge the faculty of Computer Department, NBN Sinhgad college of Engineering, Pune for providing necessary resources. Lastly, we are grateful to our families and peers for their constant encouragement.

REFERENCES

[1]. D. Baishya and R. Baruah, "Improving Hidden Markov Model for very low resource languages: An analysis for Assamese parts of speech tagging," 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2021, pp. 142-146, doi: 10.1109/Confluence51648.2021.9377146. keywords: {Training;Viterbi algorithm;Hidden Markov models;Training data;Tagging;Markov processes;Natural language processing;NLP;Assamese;POS Tagging;Hidden Markov Model;Low resource language},

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 10, May 2025
Impact Factor: 7.67

[2]. V. V. N. S. Poorna Chandrika, R. Verma, N. Charan, S. Ditheswar, S. Hansika and R. Ishwariya, "POS Tagging Using Hidden Markov Models in Natural Language Processing," 2024 International Conference on Signal Processing, Computation, Electronics, Power and Telecommunication (IConSCEPT), Karaikal, India, 2024, pp. 1-6, doi: 10.1109/IConSCEPT61884.2024.10627782. keywords: {Analytical models;Accuracy;Viterbi algorithm;Computational modeling;Hidden Markov models;Signal processing algorithms;Tagging;Natural Language Processing (NLP);Parts-of-Speech (POS) Tagging;Hidden Markov Models (HMMs);N-gram Models;Markov Models;Viterbi Algorithm;Maximum Entropy Markov Models;CRFs (Conditional Random Fields);Language comprehension;Experimentation;Performance Analysis},

[3]. Y. Ren and D. Li, "Path-Constrained Viterbi Algorithm: An Alternative to State-transition Feature for Conditional Random Fields," 2017 International Conference on Computer Technology, Electronics and Communication (ICCTEC), Dalian, China, 2017, pp. 494-497, doi: 10.1109/ICCTEC.2017.00113. keywords: {Viterbi algorithm;Task analysis;Training;Decoding;Natural language processing;Tagging;Measurement;Conditional random fields;Viterbi;natural language processing},

[4]. R. Deka, S. Kalita, K. Kashyap, M. P. Bhuyan and S. K. Sarma, "A Study of T'nT and CRF Based Approach for POS Tagging in Assamese Language," 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), Thoothukudi, India, 2020, pp. 600-604, doi: 10.1109/ICISS49785.2020.9315939. keywords: {Tagging;Hidden Markov models;Training;Information technology;Testing;Interpolation;Viterbi algorithm;Assamese Language;Natural Language Processing;POS;T'nT;CRF;Corpus;Tagset;Tag;Tagger;Token},

[5]. P. K B, R. Sony Pinto, L. V S and S. Vrajesh, "Visualizing Parts of Speech Tags by Analysing English Language Text," 2024 Third International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE), Ballari, India, 2024, pp. 01-06, doi: 10.1109/ICDCECE60827.2024.10548901. keywords: {Hidden Markov models;Tagging;Linguistics;Natural language processing;Conditional random fields;Real-time systems;Libraries;natural language processing;part of speech tagging;HMM;CRF;SpaCy},