A Novel Approach for Speech Tagging for Hindi using NLP

*** Lalan Kumar**
**** Asst. Prof. Ayush Kumar**

ABSTRACT

This paper presents a POS tagging approach for Hindi, a morphologically rich language, to demonstrate that strong morphology can offset limited training data. The methodology uses a modestly-sized, locally annotated corpus (15,562 words), detailed morphological analysis, a high-coverage lexicon and a CN2 decision-tree-based learning algorithm. The system's performance was evaluated using 4-fold cross-validation on news data from BBC Hindi. The POS tagger currently achieves an accuracy of 93.45%, with potential for further improvement.

Keywords:- *Part of Speech, Hindi, Emission probabilities, Transition probability.*

*Lalan Kumar, Research Scholar, Department of Computer Science & Engineering, Radharaman Engineering College, Bhopal , India, kumar.lalan2021@gmail.com
**Ayush Kumar, Asst. Professor, Department of Computer Science & Engineering, Radharaman Engineering College, Bhopal, India

I. INTRODUCTION

In Part of Speech (POS) tagging, words in a sentence are assigned grammatical categories based on their meaning and context. This foundational process in NLP supports various applications, including Machine Translation, Text Classification, and Sentiment Analysis. POS tagging requires tokenized text, meaning that both words and punctuation marks are classified based on context. However, applying POS tagging methods across languages, particularly morphologically rich ones like Hindi, presents unique challenges.

Hindi, spoken widely across Northern India, shares some grammatical similarities with English but also has distinct structures due to its morphological features and flexible word order. This means that methods developed for English POS tagging often fall short for Hindi, as they don't account for Hindi's rich morphology or sentence structure. For example, stochastic taggers, which rely on probabilistic models, generally don't perform well because they don't incorporate morphological information specific to Hindi. Additionally, Hindi corpora often lack the diversity needed for robust tagging, making accurate POS tagging challenging.

The primary issue lies in the lack of large, annotated datasets for Indian languages, limiting the accuracy of POS tagging algorithms. Moreover, designing a common POS tagset across India's linguistically diverse languages is challenging, as each language has unique grammatical rules. Thus, a tailored approach is essential for each language, incorporating language-specific tags

within broader frameworks. Hindi's free word order and unique morphological features, such as suffix changes depending on context, require an approach that adapts to unknown words or infrequent forms, which are common in Hindi corpora.

To address these limitations, researchers are developing improved Hindi POS tagging algorithms that combine methods like morphological analysis, a high-coverage lexicon and a CN2 decision-tree-based learning algorithm. This approach enhances tagging accuracy by handling unknown words more effectively than traditional methods. With better tagging, NLP applications in Hindi, including Sentiment Analysis and Machine Translation, can achieve higher accuracy, ultimately enabling more robust processing of Hindi text across various domains.

II. LITERATURE REVIEW

He and Choi (2023) explore advancements in sequence-to-sequence (Seq2Seq) models for sequence tagging and structure parsing [1]. They discuss enhancements in model architectures and training techniques that have improved performance across various NLP tasks, including named entity recognition, syntactic parsing, and semantic role labeling. The paper emphasizes the integration of attention mechanisms and transformer architectures to handle long-range dependencies and improve the accuracy of sequence tagging tasks.

Nunsanga (2023) conducts a detailed analysis of part-of-speech tagging specifically for the Mizo language [2]. The study investigates challenges unique to less-resourced languages and proposes language-specific adaptations to existing tagging models. Nunsanga's work contributes insights into improving accuracy and applicability of NLP tools for underrepresented languages.

Rudrappa et al. (2023) present HiTEK, a preprocessing framework tailored for speech and text in natural language processing [3]. The paper discusses techniques for noise reduction, feature extraction, and data normalization, enhancing the robustness and efficiency of NLP systems operating in diverse linguistic environments.

Jahan and Oussalah (2023) provide a systematic review of automatic hate speech detection methods employing NLP techniques [4]. They analyze current approaches, including supervised and unsupervised learning models, and highlight challenges such as data bias and cultural context sensitivity. The review outlines directions for future research aimed at improving the effectiveness and fairness of hate speech detection systems.

Srivastava et al. (2023) introduce an AI-powered voice bot for Sanskrit based on natural language processing techniques [5]. The paper details the development of voice interaction capabilities, semantic understanding, and contextual processing tailored for the Sanskrit language, demonstrating applications of NLP in preserving and promoting linguistic heritage.

Prajna et al. (2024) present methods for visualizing parts-of-speech tags through the analysis of English language texts [6]. Their study explores techniques to extract and visualize syntactic structures, aiding linguistic analysis and facilitating insights into language usage patterns across different textual genres.

Smidt et al. (2024) address the intricacies of fine-grained part-of-speech tagging, focusing on the challenges posed by ancient languages encoded in cuneiform scripts [7]. Their paper underscores the complexities involved in linguistic annotation and the adaptation of contemporary natural language processing (NLP) methodologies to effectively preserve and interpret ancient textual artifacts.

The study emphasizes the unique linguistic features and syntactic structures inherent in cuneiform texts, presenting hurdles such as sparse linguistic resources and ambiguous lexical forms. Smidt et al. explore how modern NLP techniques, including deep learning models and computational linguistics approaches, can be tailored to handle these challenges and extract meaningful linguistic insights from ancient scripts.

Their research highlights the importance of developing specialized tools and methodologies for fine-grained part-of-speech tagging in historical linguistics, contributing to the broader effort of preserving and understanding ancient languages through advanced computational techniques.

Woldemariam (2024) investigates NLP methods to enhance user rating systems in crowdsourcing forums and improve speech recognition for less-resourced languages [8]. The study proposes adaptive learning approaches and data augmentation strategies to mitigate challenges related to data scarcity and linguistic diversity.

Pradhan and Yajnik (2024) conduct a comparative study on various models for part-of-speech (POS) tagging tailored for Nepali texts [9]. Their research evaluates the performance of Bidirectional LSTM, Conditional Random Fields (CRF), and Hidden Markov Models (HMM) in handling the morphological complexities of the Nepali language.

The study addresses the challenges posed by Nepali's rich morphology, which includes diverse inflectional patterns and syntactic structures. Pradhan and Yajnik highlight how each model excels in capturing different aspects of Nepali linguistic features: Bidirectional LSTM for its ability to model contextual dependencies effectively, CRF for its sequential labeling capabilities, and HMM for its probabilistic framework suited for tagging sequences.

Through empirical evaluations, they provide nuanced insights into the strengths and limitations of each model, offering recommendations to enhance tagging accuracy and efficiency in Nepali NLP applications. Their findings underscore the importance of selecting appropriate modeling techniques that align with the linguistic characteristics of morphologically complex languages like Nepali, aiming to advance the development of robust and adaptable POS tagging systems tailored for diverse linguistic contexts.

Jatta (2024) delves into the forefront of automatic speech recognition (ASR) tailored for maritime settings, employing artificial intelligence (AI) techniques to bolster transcription precision and operational utility amidst formidable acoustic challenges [10]. His study underscores the critical need for robust ASR solutions capable of navigating the complex acoustic environments inherent to maritime operations.

By harnessing AI methodologies, Jatta enhances transcription accuracy, ensuring clearer communication and operational efficiency in maritime contexts where ambient noise and variable acoustic conditions pose significant hurdles. His research contributes insights into optimizing ASR systems, emphasizing adaptive algorithms and noise suppression techniques that enhance speech signal clarity amidst challenging maritime acoustic landscapes.

Jatta's work highlights the transformative potential of AI-driven ASR advancements in maritime domains, offering practical applications in navigation, communication, and operational safety. His findings advocate for ongoing innovation in ASR technologies tailored for specific environmental contexts, aiming to elevate performance standards and expand the scope of automated speech recognition across diverse operational settings.

Li et al. (2024) undertake a bibliometric analysis aimed at comprehensively mapping the landscape of part-of-speech (POS) tagging research [11]. Their study offers a systematic exploration of methodologies, applications, and emerging trends within the field, providing a detailed overview of recent advancements and future trajectories.

Through their bibliometric approach, Li et al. identify and analyze key themes and research methodologies employed in POS tagging studies. They highlight the evolution from traditional rule-based approaches to modern machine learning and deep learning techniques, emphasizing the shift towards neural network architectures like Bidirectional LSTMs and transformer models.

III. PROPOSED WORK

This paper introduces a POS tagger for Hindi, the national language of India with over 500 million speakers and the fourth most spoken language globally. It outlines a tagging methodology suited to resource-limited languages lacking large annotated corpora. This approach utilizes a locally annotated corpus (15,562 words), comprehensive morphological analysis, a high-coverage lexicon, and a CN2 decision-tree learning algorithm [12]. The method includes in-depth linguistic analysis, efficient suffix handling, accurate verb group identification, and disambiguation rule learning, marking a novel approach for Hindi POS tagging.

This approach can be applied to other inflectional languages by integrating language-specific resources, such as suffix replacement rules (SRRs), lexicons, group identification guidelines, and morpheme analysis rules, while maintaining the same core processes illustrated in Figure 1.

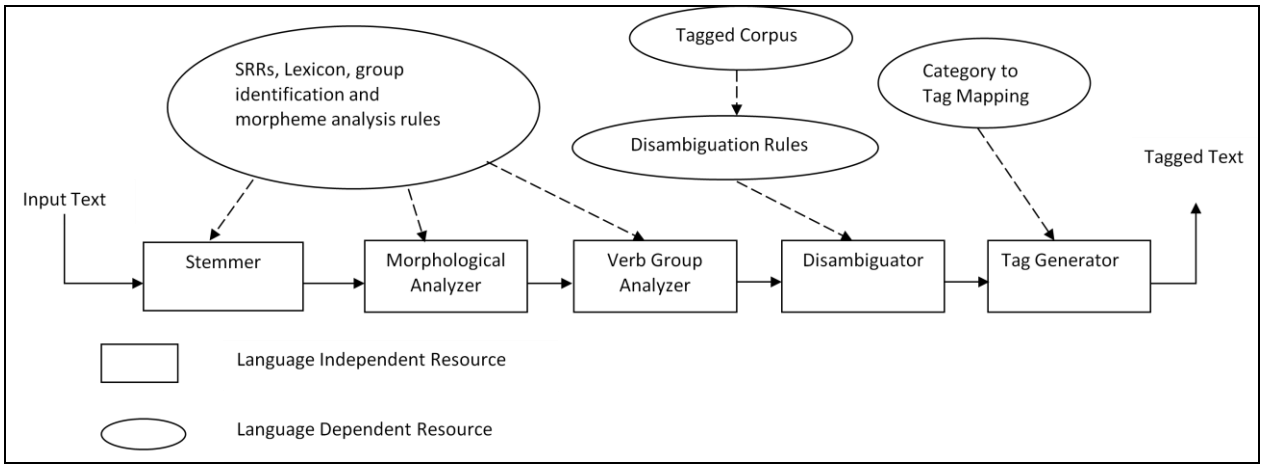

Figure 1: Architecture of the Tagger

IV. DESIGN OF HINDI POS TAGGER

4.1 Morphology Driven Tagger

The morphology-driven tagger uses word affix information to assign POS tags without relying on contextual clues, except when identifying the main verb and auxiliaries within a verb group (VG).

Other POS categories are tagged through a lexicon lookup of root forms. The process is fully rule-based, with no learning or disambiguation, relying solely on hand-crafted morphological rules. The architecture of this tagger is illustrated in Figure 1, and the methodology operates on two levels.

1. **Word Level**: A stemmer works alongside a lexicon and Suffix Replacement Rules (SRRs) to generate potential root-suffix pairs with POS labels for each word. If the word is not in the lexicon and lacks an inflectional suffix, derivational rules apply.

2. **Group Level**: A Morphological Analyzer (MA) adds morphological information based on suffixes. For nouns, suffixes indicate "Number," while "Case" is inferred from nearby words. For verbs, GNP values are determined at the word level, and TAM values during verb group identification. Each suffix component is analyzed separately using a morpheme analysis table.

For verbs, GNP values are found at the word level, while TAM values are identified during the VG Identification phase, described later. The analysis of the suffix is done in a discrete manner, i.e., each component of the suffix is analyzed separately. A morpheme analysis table comprising individual morphemes with their paradigm information and analyses is used for this purpose. MA's output for the word खाऊंगी {khaaoongii} (will eat) looks like –

Stem: खा (eat)

Suffix: ऊंगी Category: Verb

Morpheme 1: ऊं Analysis: 1 Per, Sg

Morpheme 2: ग Analysis: Future

Morpheme 3: ई Analysis: Feminine

4.1.1 Verb Group Identification

The structure of a Hindi VG is relatively rigid and can be captured well using simple syntactic rules. In Hindi, certain auxiliaries like रह {rah}, पा {paa}, सक, {sak} or पड़ {pad} can also occur as main verbs in some contexts. VG identification deals with identifying the main verb and the auxiliaries of a VG while discounting for particles, conjunctions and negation markers. The VG identification goes left to right by marking the first constituent as the main verb or copula verb and making every other verb construct an auxiliary till a non-VG constituent is encountered. Main

verb and copula verb can take the head position of a VG and can occur with or without auxiliary verbs. Auxiliary verbs, on the other hand, always come along with a main verb or a copula verb. This results in a very high accuracy of 99.5% for verb auxiliaries. Ambiguity between a main verb and a copula verb remains unresolved at this level and asks for disambiguation rules.

4.2 Need for Disambiguation

The simple lexicon lookup approach (LLB) achieves 61.19% accuracy, while the morphology-driven tagger improves upon this but still leaves considerable ambiguity. These findings emphasize the importance of detailed morphological analysis, a point also observed for Japanese by Uchimoto et al. (2001). When a word has multiple possible POS tags, the tagger's accuracy drops to 73.62%, as shown in Table 1.

Table 1: Average Accuracy(%) Comparison of Various Approaches

LLB	LLBD	MD	BL	LB
61.19	86.77	73.62	82.63	93.45

To resolve ambiguity, baseline (BL) tagging can assign the most frequent tag, reaching an accuracy of 82.63%. However, BL tagging lacks room for improvement, while the morphology-driven (MD) tagger shows potential for accuracy enhancement through disambiguation. Nearly 30% of words remain ambiguous or unknown in the MD tagger, underscoring the need for a disambiguation step.

Commonly, POS disambiguation involves machine learning techniques, particularly decision trees (e.g., ID3, AQR, ASSISTANT) and neural networks. Among these, CN2 is known for robust performance on noisy data (Clark & Niblett, 1989). Given the absence of such techniques for Hindi, CN2 was selected for this work due to its effective handling of noisy data.

V. EXPERIMENTAL SETUP

The experiment focused on two main tasks: optimizing the CN2 algorithm's parameters and assessing the effectiveness of the disambiguation rules it produced for POS tagging.

5.1 CN2 Parameters: The CN2 algorithm includes parameters like rule type (ordered or unordered), star size, significance threshold, and training instance window size. The best results

were achieved with ordered rules, a star size of 1, a significance threshold of 10, and a window size of 4 (considering two neighboring words on each side for training).

5.2 Evaluation: Testing was conducted on a corpus of 15,562 words, split into 75% for training and 25% for testing.

$$Accuracy = \frac{no.of\ singal\ correct\ tages}{total\ no.of\ tokens} \quad (1)$$

The results are obtained by performing a 4- fold cross validation over the corpora. Figure 2 gives the learning curve of the disambiguation module for varying corpora sizes starting from 1000 to the complete training corpora size. The accuracy for known and unknown words is also measured separately.

VI. RESULTS AND DISCUSSION

These The learning-based (LB) tagger achieved an average accuracy of 93.45% after 4-fold cross-validation, setting a notable benchmark for Hindi. Table 1 shows that the disambiguation module raised accuracy by 25% for the lexicon-based approach (LLBD) and improved the morphology-driven (MD) tagger's accuracy by about 20%, confirming the benefits of combining morphology with disambiguation.

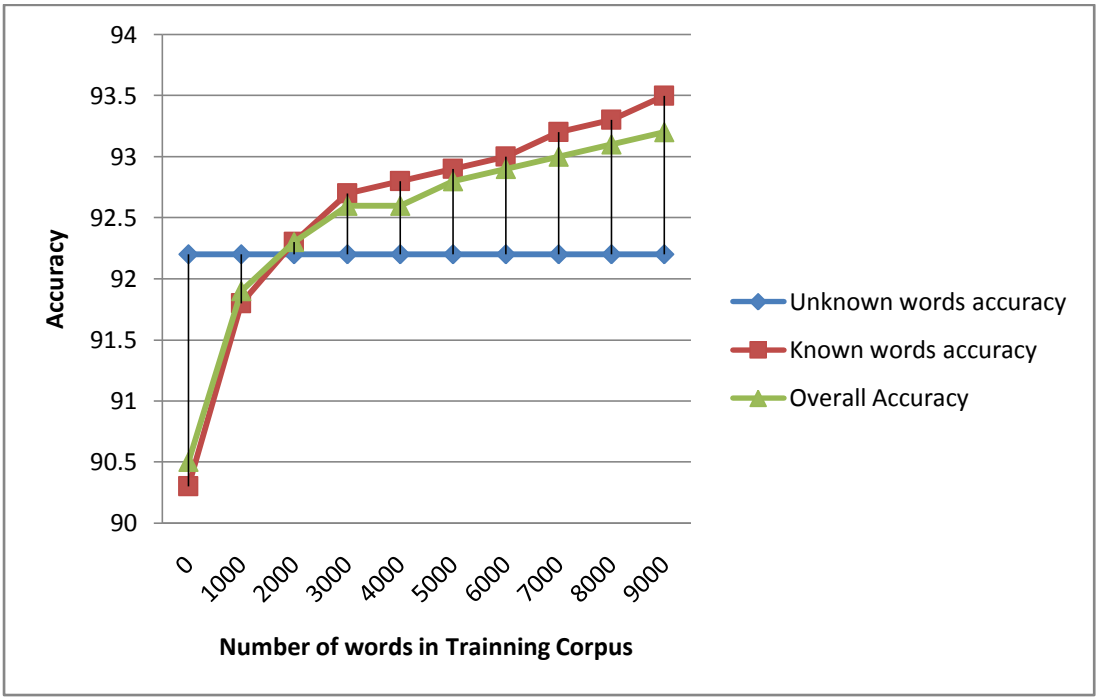

Figure 2: POS Learning Curve

Interestingly, the LB tagger outperformed both MD and BL taggers across most POS categories, effectively disambiguating and tagging unknown words with a high accuracy of 92.08%. However, certain POS categories—like pronouns, intensifiers, demonstratives, and verb copulas—remain challenging due to their rarity, high ambiguity, and the need for semantic information, suggesting a need for further analysis.

One interesting observation is the performance of the tagger on individual POS categories. Figure 3 shows clearly that the per POS accuracies of the LB tagger highly exceeds those of the MD and BL tagger for most categories. This means that the disambiguation module correctly disambiguates and correctly identifies the unknown words too. The accuracy on unknown words, as earlier shown in Figure 2, is very high at 92.08%. The percentage of unknown words in the test corpora is 0.013. It seems independent of the size of training corpus because the corpora is unbalanced having most of the unknowns as proper nouns. The rules are formed on this bias, and hence the application of these rules assigns PPN tag to an unknown which is mostly the case.

VII. CONCLUSION

In In this paper, we presented a Part-of-Speech (POS) tagger for Hindi that addresses the challenge of limited annotated corpora by leveraging the language's rich morphology and relatively fixed word order within a Verb Group (VG). The primary focus of our work was to identify and eliminate factors that negatively impact the accuracy of verb tagging. A comprehensive analysis of accuracy distribution across POS tags highlighted areas that required detailed disambiguation rules. A key strength of this approach lies in the automatic learning of disambiguation rules, which would traditionally need to be manually coded, requiring extensive analysis of linguistic phenomena. Achieving an accuracy rate of nearly 94% with a corpus of only 15,562 words supports the idea that "morphological richness can offset resource scarcity." This work could pave the way for developing POS taggers for other morphologically rich languages with limited annotated corpora.

Several promising avenues for future research emerge from this work. One such direction is exploring statistical methods like Conditional Random Fields, where feature functions would be derived from morphological data. The natural next step after the POS tagger is the development of a chunker for Hindi.

REFERENCES

1. He, Han, and Jinho D. Choi. "Unleashing the true potential of sequence-to-sequence models for sequence tagging and structure parsing." Transactions of the Association for Computational Linguistics 11 (2023): 582-599.
2. Nunsanga, Morrel VL. "Analysis of Part of Speech Tagging for Mizo Language." PhD diss., Mizoram University, 2023.
3. Rudrappa, N. T., M. V. Reddy, and M. Hanumanthappa. "HiTEK Pre-processing for Speech and Text: NLP." Indian Journal of Science and Technology 16, no. 19 (2023): 1413-1421.
4. Jahan, Md Saroar, and Mourad Oussalah. "A systematic review of hate speech automatic detection using natural language processing." Neurocomputing 546 (2023): 126232.
5. Srivastava, Vedika, Arti Khaparde, Akshit Kothari, and Vaidehi Deshmukh. "NLP-Based AI-Powered Sanskrit Voice Bot." Artificial Intelligence Applications and Reconfigurable Architectures (2023): 95-124.
6. Prajna, K. B., Reenal Sony Pinto, V. S. Lakshmishree, and S. Vrajesh. "Visualizing Parts of Speech Tags by Analysing English Language Text." In 2024 Third International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE), pp. 01-06. IEEE, 2024.
7. Smidt, Gustav Ryberg, Els Lefever, and Katrien de Graef. "At the Crossroad of Cuneiform and NLP: Challenges for Fine-grained Part-of-speech Tagging." In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pp. 1745-1755. 2024.
8. Woldemariam, Yonas Demeke. "NLP methods for improving user rating systems in crowdsourcing forums and speech recognition of less resourced languages." PhD diss., Umeå University, 2024.
9. Pradhan, Ashish, and Archit Yajnik. "Parts-of-speech tagging of Nepali texts with Bidirectional LSTM, Conditional Random Fields and HMM." Multimedia Tools and Applications 83, no. 4 (2024): 9893-9909.
10. Jatta, Lamin. "Maritime Automatic Speech Recognition: Improving the Quality of Transcriptions using Artificial Intelligence." (2024).
11. Li, Xinye, Bingliang Zhang, Litong Wu, Xiaoyi Du, and Feng Hu. "The Scope of Part of Speech Tagging: A Bibliometric Study." Lecture Notes on Language and Literature 7, no. 4 (2024): 51-58.
12. Swe, Su Myo, and Khin Myo Sett. "Approaching rules induction CN2 algorithm in categorizing of biodiversity." Int J Trend Sci Res Dev 3, no. 4 (2019): 1581-1584.