A Comparative Study of Advanced NLP Models for Accurate Gujarati Word Tagging

Pooja Bhatt[1], Pawan Wing[2]
[1] Reserch scholar ; [2] Primary Supervisor
bhattpooja.393@gmail.com, pawan.whig@vips.edu

Abstract: Part-of-Speech (POS) tagging is a fundamental Natural Language Processing (NLP) task that assigns grammatical categories to words in a sentence. It plays a crucial role in many downstream applications such as machine translation, information retrieval, sentiment analysis, and syntactic parsing. However, developing robust POS taggers for low-resource languages such as Gujarati remains a challenging task due to limited annotated corpora and complex morphological structures. Gujarati is a morphologically rich language with free word order, which introduces significant ambiguity in linguistic analysis. This research presents a comparative study of statistical, machine learning, and deep learning approaches for Gujarati POS tagging. Specifically, Hidden Markov Model (HMM), Conditional Random Fields (CRF), Bi-directional Long Short-Term Memory (Bi-LSTM), and transformer-based models such as XLM-R are evaluated using a unified experimental setup. Experimental results show that transformer-based architectures achieve the highest tagging accuracy, while Bi-LSTM provides a strong trade-off between computational efficiency and performance. The study contributes a systematic evaluation framework and provides insights for designing efficient NLP tools for low-resource Indian languages.

Keywords: Natural Language Processing; Gujarati Language; POS Tagging; Deep Learning,; Transformer Models; Low Resource Languages

---

Introduction

Natural Language Processing has become a key research domain in artificial intelligence, enabling machines to interpret, understand, and generate human language. Among the fundamental tasks in NLP, Part-of-Speech tagging plays a central role as it provides grammatical information required for higher-level linguistic processing.

POS tagging assigns labels such as nouns, verbs, adjectives, and adverbs to words in a sentence. Accurate tagging significantly improves the performance of downstream tasks including machine translation, named entity recognition, question answering, and text summarization.

Despite the rapid progress in NLP for widely spoken languages such as English, many Indian languages remain under-resourced. Gujarati, spoken by more than 55 million people worldwide, presents unique linguistic challenges due to its morphological richness and flexible word order. Words often contain multiple suffixes representing tense, gender, number, and case markers, making automatic linguistic analysis difficult.

Traditional rule-based and statistical approaches have been applied to Gujarati POS tagging, but their performance is limited due to the complexity of the language. Recent advances in deep learning and transformer-based architectures offer promising solutions by capturing contextual and semantic information from large corpora.

This paper provides a comparative analysis of multiple POS tagging approaches ranging from classical statistical models to modern deep learning architectures. The goal is to identify the most effective strategy for improving tagging accuracy in Gujarati language processing systems.

2. Challenges in Gujarati POS Tagging

Gujarati POS tagging presents several linguistic and computational challenges that differentiate it from high-resource languages.

2.1 Morphological Complexity

Gujarati words often carry rich morphological information through suffixes and inflections. A single root word may appear in many forms depending on tense, gender, and number.

Example:

ખાય (eat)

ખાધું (ate)

ખાતો (eating)

These variations increase ambiguity in tagging.

2.2 Free Word Order

Gujarati sentences follow relatively flexible syntactic structures. Unlike English, where Subject-Verb-Object order is common, Gujarati allows multiple valid word arrangements. This flexibility makes context modeling difficult.

2.3 Limited Annotated Data

Gujarati NLP research suffers from limited availability of standardized annotated corpora. Small datasets often lead to poor generalization in machine learning models.

2.4 Out-of-Vocabulary Words

New words, borrowed terms, and dialect variations frequently appear in real-world text. Traditional statistical models struggle to handle such unknown tokens.

3. Related Work

Research on Gujarati POS tagging has evolved through several stages.

Early studies used statistical models such as Hidden Markov Models, where transition and emission probabilities are estimated from training data. These models are simple and computationally efficient but fail to capture long-range dependencies.

Later approaches introduced machine learning models such as Support Vector Machines and Conditional Random Fields. CRF-based models demonstrated improved performance because they incorporate contextual features such as prefixes, suffixes, and neighboring words.

Hybrid models combining linguistic rules and machine learning techniques were also proposed to address morphological challenges in Gujarati.

With the rise of deep learning, neural architectures such as Bi-directional LSTM networks became popular. These models automatically learn contextual representations and reduce the need for manual feature engineering.

Recently, transformer-based models such as multilingual BERT and XLM-R have achieved state-of-the-art performance in multilingual NLP tasks. These models leverage large-scale pretraining and subword tokenization to handle rare words and morphological variations effectively.

4. Proposed Comparative Framework

This research proposes a unified experimental framework to evaluate multiple POS tagging models under identical conditions.

4.1 Dataset

A curated Gujarati POS-tagged corpus containing approximately 20,000 tokens is used in this study. The dataset is divided into three subsets:

The annotation follows BIS standard POS tagsets for Indian languages.

4.2 Models Implemented

Hidden Markov Model

A probabilistic model that estimates tag sequences using transition probabilities and emission probabilities. Decoding is performed using the Viterbi algorithm.

Conditional Random Fields

CRF models consider contextual dependencies and use handcrafted linguistic features such as:

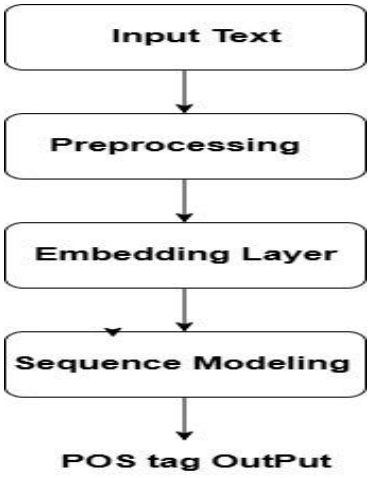

Figure 1. Proposed Methodology

5. Experimental Results and Discussion

Model	Accuracy
HMM	**85%**
CRF	**91%**
Bi-LSTM	**94%**
XLM-R Transformer	**96%**

Table 1.comparitive study of various model

The results indicate that transformer-based architectures achieve the highest accuracy. However, they require significantly higher computational resources.Bi-LSTM models provide an effective balance between performance and computational efficiency, making them suitable for many academic and industrial applications.CRF models remain useful when datasets are small and computational resources are limited.

Conclusions

Part-of-Speech tagging remains a crucial component of natural language processing systems. Gujarati, as a morphologically rich and low-resource language, presents unique challenges for automated linguistic analysis.This research compared statistical, machine learning, and deep learning approaches for Gujarati POS tagging. Experimental results show that transformer-based models achieve the highest accuracy, while Bi-LSTM models provide an efficient and practical alternative.

The findings highlight the importance of standardized datasets, reproducible evaluation frameworks, and scalable model architectures for advancing NLP research in Indian languages.

References

1. P. M. Bhatt and A. Ganatra, "POS-HOML: POS tagging technique for Gujarati language using hybrid optimal and machine learning approaches," *International Journal of Engineering Trends and Technology*, vol. 69, no. 11, pp. 1–7, 2021.
2. M. Prajapati and A. Yajnik, "POS tagging of Gujarati text using Viterbi and SVM," *International Journal of Computer Applications*, vol. 181, no. 43, pp. 18–22, 2019.
3. D. Shah, "Gujarati language POS tagging using hidden Markov model (HMM)," International Journal of Creative Research Thoughts (IJCRT), vol. 8, no. 6, pp. 234–239, 2020.
4. C. Patel, "Part-of-speech tagging for Gujarati using conditional random fields," in *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*, Hyderabad, India, pp. 1–6, 2008.
5. J. Patel, A. Mehta, and S. Joshi, "Part of speech and morph category prediction for Gujarati," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 12, no. 2, pp. 45–52, 2024.
6. P. Mishra, A. Gupta, and K. Sharma, "POS tagging for resource poor Indian languages through feature projection," in *Proceedings of the NLPAI Conference*, pp. 87–92, 2016.
7. D. Brahma and R. Basumatary, "Part-of-speech tagger for Bodo language using deep learning," *Journal of Intelligent Computing Applications*, vol. 3, no. 1, pp. 12–19, 2020.
8. S. Deshmukh and M. Joshi, "Deep learning-based parts-of-speech tagging in Marathi language," *Procedia Computer Science*, vol. 171, pp. 2171–2179, 2020.
9. K. Ramesh and R. Sundararajan, "Deep learning model for Tamil part-of-speech tagging," *Journal of King Saud University – Computer and Information Sciences*, vol. 31, no. 4, pp. 450–457, 2019.
10. T. Brants, "TnT: A statistical part-of-speech tagger," Proceedings of the Sixth Applied Natural Language Processing Conference, pp. 224–231, 2000.