# Part-of-Speech Tagging using Hidden Markov Models: A Statistical Sequence Modeling Approach

Sarthak Thakare

May 20,2025

**Abstract**

This document presents an implementation and evaluation of a Hidden Markov Model (HMM) based Part-of-Speech (POS) tagger trained on the Brown Corpus with Universal POS tags. The model estimates transition and emission probabilities and applies a Viterbi-like algorithm to predict tags. Evaluation includes accuracy, F-score metrics, and a detailed confusion matrix.

## 1 Introduction

Part-of-Speech (POS) tagging is a fundamental task in Natural Language Processing (NLP), aiming to assign grammatical categories to words. Hidden Markov Models (HMMs) are probabilistic models well-suited for sequential data and have been widely applied in POS tagging tasks.

## 2 Dataset and Preprocessing

- **Corpus:** Brown Corpus from the NLTK library
- **Tagset:** Universal POS Tagset
- **Preprocessing:**
  - Sentences are padded with special tokens: `#####` (start) and `$$$$$` (end)
  - All words are converted to lowercase

## 3 Methodology

The system employs the following steps:

- Estimate transition probabilities $P(t_i|t_{i-1})$ from training data

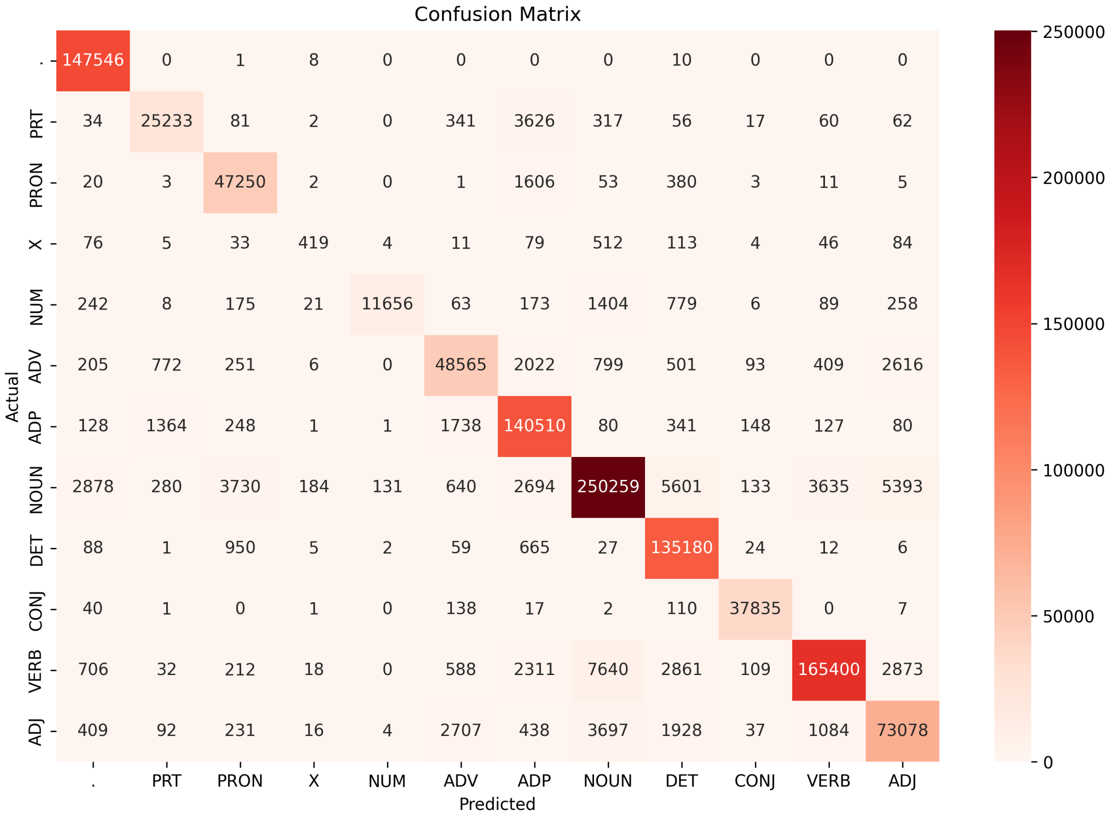
- Estimate emission probabilities $P(w|t)$ using add-one (Laplace) smoothing
- Apply a custom Viterbi-like dynamic programming algorithm to decode the most probable tag sequence

## 4 Evaluation

The evaluation is done using 5-fold cross-validation. Accuracy and F-score metrics are computed, along with a confusion matrix to analyze class-level performance.

### 4.1 Confusion Matrix

Figure 1 displays the confusion matrix between predicted and actual POS tags.



|  | . | PRT | PRON | X | NUM | ADV | ADP | NOUN | DET | CONJ | VERB | ADJ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| . | 147546 | 0 | 1 | 8 | 0 | 0 | 0 | 0 | 10 | 0 | 0 | 0 |
| PRT | 34 | 25233 | 81 | 2 | 0 | 341 | 3626 | 317 | 56 | 17 | 60 | 62 |
| PRON | 20 | 3 | 47250 | 2 | 0 | 1 | 1606 | 53 | 380 | 3 | 11 | 5 |
| X | 76 | 5 | 33 | 419 | 4 | 11 | 79 | 512 | 113 | 4 | 46 | 84 |
| NUM | 242 | 8 | 175 | 21 | 11656 | 63 | 173 | 1404 | 779 | 6 | 89 | 258 |
| ADV | 205 | 772 | 251 | 6 | 0 | 48565 | 2022 | 799 | 501 | 93 | 409 | 2616 |
| ADP | 128 | 1364 | 248 | 1 | 1 | 1738 | 140510 | 80 | 341 | 148 | 127 | 80 |
| NOUN | 2878 | 280 | 3730 | 184 | 131 | 640 | 2694 | 250259 | 5601 | 133 | 3635 | 5393 |
| DET | 88 | 1 | 950 | 5 | 2 | 59 | 665 | 27 | 135180 | 24 | 12 | 6 |
| CONJ | 40 | 1 | 0 | 1 | 0 | 138 | 17 | 2 | 110 | 37835 | 0 | 7 |
| VERB | 706 | 32 | 212 | 18 | 0 | 588 | 2311 | 7640 | 2861 | 109 | 165400 | 2873 |
| ADJ | 409 | 92 | 231 | 16 | 4 | 2707 | 438 | 3697 | 1928 | 37 | 1084 | 73078 |

Figure 1: Confusion matrix of predicted vs actual POS tags

### 4.2 Accuracy Metrics

- **Overall Accuracy:** 93.62%
- **Baseline Accuracy (Most frequent tag):** 32.65%
- **Accuracy Improvement:** +60.97 percentage points

### 4.3 F-beta Scores

- Weighted $F_{0.5}$ Score: 0.9354
- Weighted $F_{1.0}$ Score: 0.9362
- Weighted $F_{2.0}$ Score: 0.9372

### 4.4 Qualitative Observations

- Frequent confusions:
  - ADJ ↔ NOUN
  - VERB ↔ NOUN
  - ADV ↔ ADJ
- Diagonal dominance in confusion matrix indicates high tag prediction fidelity

## 5 Conclusion

The HMM-based POS tagger shows strong predictive capability with over 93% accuracy and substantial improvement over a simple baseline. Common tag confusions are consistent with linguistic ambiguities. This system offers a strong statistical baseline for POS tagging tasks.

## Future Work

- Integrating morphological features
- Using a trigram HMM or CRF for context-rich modeling
- Expanding to multilingual corpora