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.

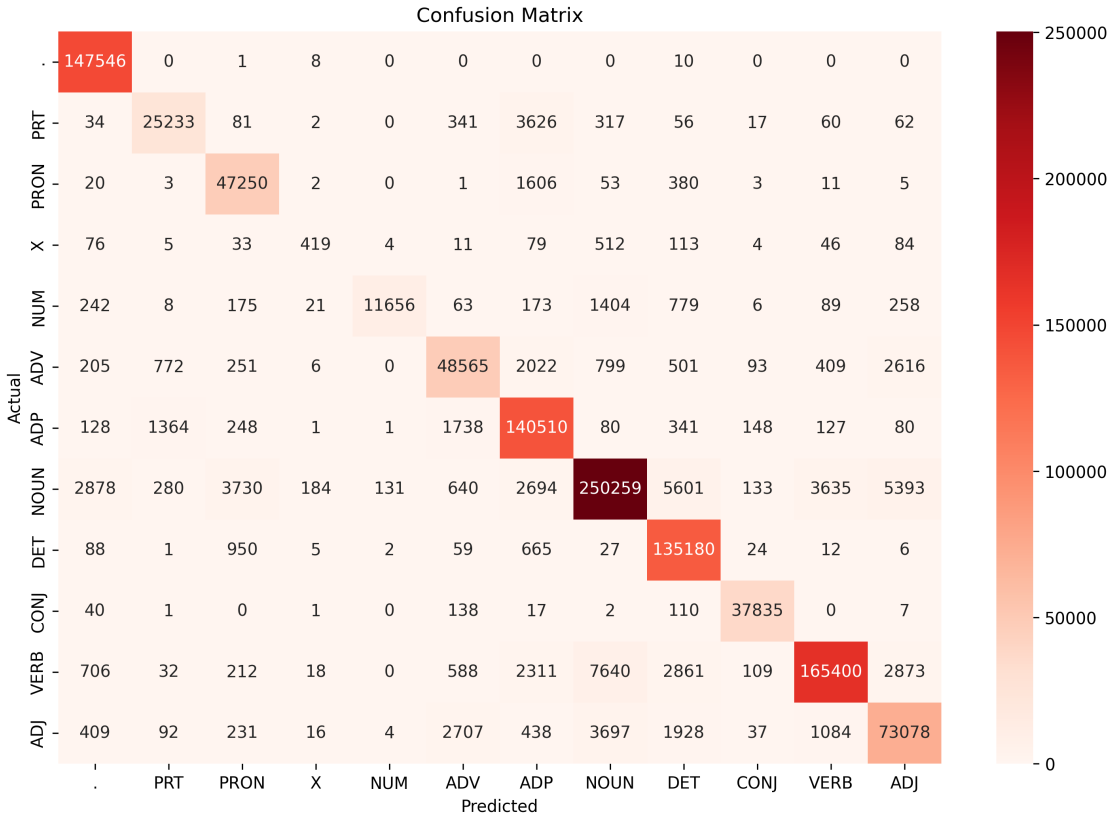

Figure 1: Confusion matrix of predicted vs actual POS tags

4.2 Accuracy Metrics

- **Overall Accuracy:** 93.32%
- **Baseline Accuracy (Most frequent tag):** 23.79%
- **Accuracy Improvement:** +69.53 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 $\leftrightarrow$ NOUN
  - VERB $\leftrightarrow$ NOUN
  - ADV $\leftrightarrow$ 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