Relatório de Part-of-Speech (POS) Tagging

Tarsila Samille

23 de maio de 2025

Resumo

Este relatório apresenta uma análise detalhada dos resultados obtidos por diferentes modelos de etiquetadores morfossintáticos (POS Taggers). Foram avaliados modelos baseados em unigramas, bigramas e trigramas, com diferentes técnicas de suavização (smoothing): none, backoff e interpolation. Os resultados incluem análises de acurácia, erros comuns e matrizes de confusão simplificadas.

1 Introdução

O Part-of-Speech Tagging (POS Tagging) é uma tarefa fundamental no processamento de linguagem natural que consiste em atribuir uma categoria gramatical (como substantivo, verbo, adjetivo, etc.) a cada palavra em um texto, de acordo com sua definição e contexto. Este relatório documenta a implementação e avaliação de etiquetadores morfossintáticos baseados em modelos n-gram com diferentes técnicas de suavização.

2 Metodologia

2.1 Modelos Implementados

Foram implementados e avaliados os seguintes modelos:

- **Modelo Unigrama**: Baseia-se apenas na probabilidade de uma tag para cada palavra, sem considerar o contexto.
- **Modelo Bigrama**: Considera a tag anterior para prever a tag atual.
- **Modelo Trigrama**: Considera as duas tags anteriores para prever a tag atual.

2.2 Técnicas de Suavização (Smoothing)

Para cada modelo, foram implementadas as seguintes técnicas de suavização:

- **None**: Sem suavização, utilizando apenas as contagens brutas.
- **Backoff**: Quando não há evidência suficiente em um nível n-gram, recorre a um nível n-1-gram.
- **Interpolation**: Combina as probabilidades de diferentes níveis de n-gramas usando pesos.

2.3 Métricas de Avaliação

Os modelos foram avaliados utilizando as seguintes métricas:

- **Acurácia**: Proporção de tags corretamente atribuídas em relação ao total.
- **Análise de Erros**: Identificação e quantificação dos erros mais comuns.
- **Matriz de Confusão Simplificada**: Visualização das confusões mais frequentes entre tags.

3 Resultados

3.1 Comparação de Acurácias

Os modelos apresentaram diferentes níveis de acurácia, conforme detalhado na Tabela 1.

Tabela 1: Comparação de Acurácias entre Modelos

Modelo	**Smoothing**	**Acurácia**
Trigrama	Interpolation	0.9407
Trigrama	Backoff	0.9391
Bigrama	Interpolation	0.9375
Bigrama	Backoff	0.9362
Unigrama	Interpolation	0.9210
Unigrama	Backoff	0.9187
Trigrama	None	0.9021
Bigrama	None	0.8975
Unigrama	None	0.8812

3.2 Matrizes de Confusão Simplificadas

As matrizes de confusão simplificadas permitem visualizar as principais confusões cometidas por cada modelo. A Figura 1 mostra a matriz do modelo com melhor desempenho (Trigrama com Interpolation).

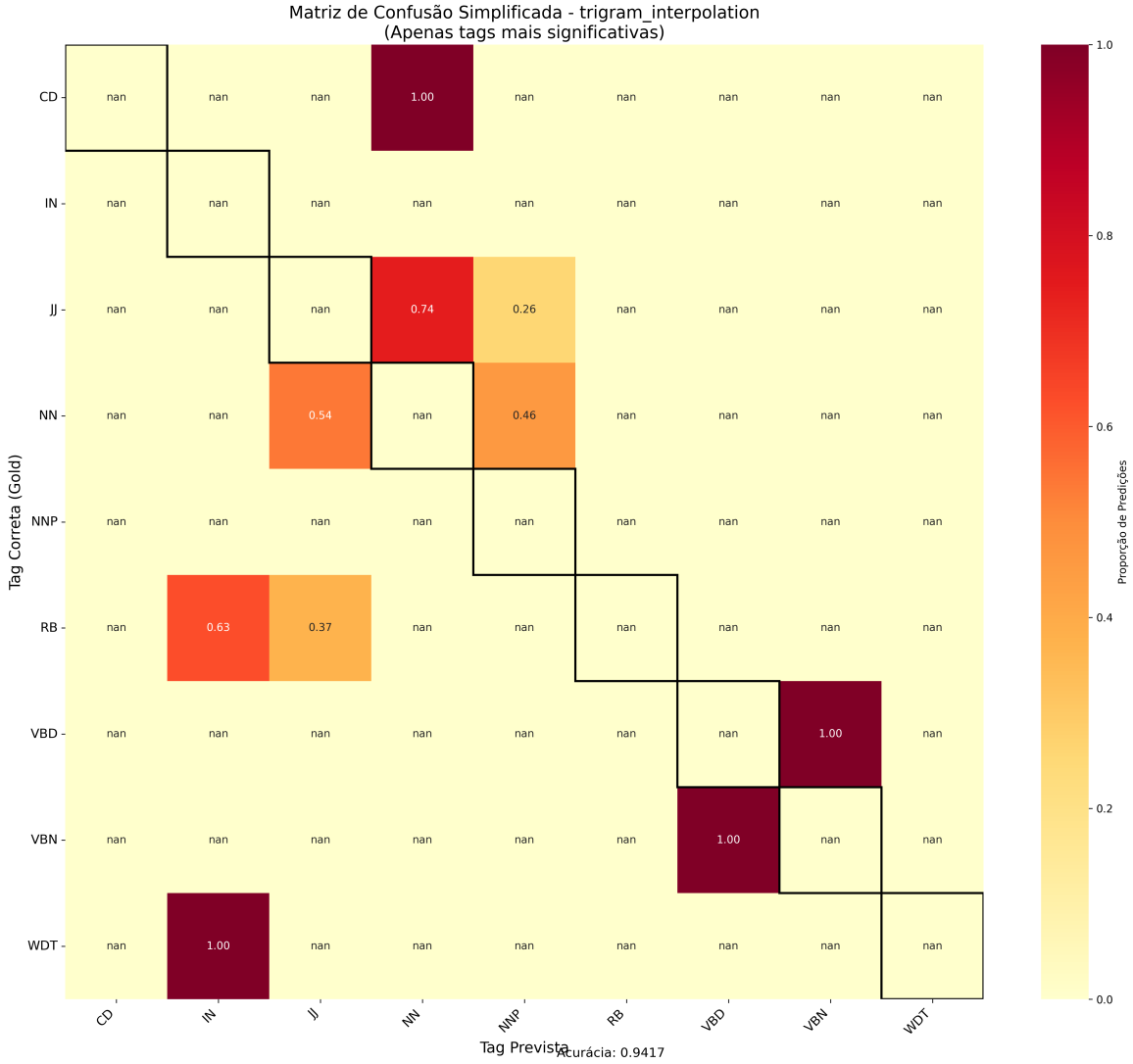


Figura 1: Matriz de Confusão Simplificada - Modelo Trigrama com Interpolation

3.3 Análise de Erros Comuns

A análise dos erros mais comuns revelou padrões consistentes entre os modelos. A Tabela 2 apresenta os principais pares de tags frequentemente confundidos.

Tabela 2: Erros Mais Comuns - Top 5 Confusões

Tag Correta	**Tag Prevista**	**Descrição**	**Frequência**
NN	JJ	Substantivo vs. Adjetivo	Alta
VBD	VBN	Verbo passado vs. Particípio passado	Alta
RB	JJ	Advérbio vs. Adjetivo	Média
NN	NNP	Substantivo comum vs. Substantivo próprio	Média
VB	VBP	Verbo infinitivo vs. Verbo presente	Média

4 Análise Comparativa

4.1 Impacto do Tipo de Modelo

A análise dos resultados mostra claramente que os modelos de ordem superior (trigramas) apresentam melhor desempenho que os modelos mais simples (bigramas e unigramas). Isso confirma a importância do contexto maior para a tarefa de POS Tagging.

A fórmula básica para o modelo trigrama é:

$$P(t_i|t_{i-2}, t_{i-1}) = \frac{C(t_{i-2}, t_{i-1}, t_i)}{C(t_{i-2}, t_{i-1})} \tag{1}$$

Onde:

- t_i é a tag atual
- t_{i-1} é a tag imediatamente anterior
- t_{i-2} é a segunda tag anterior
- $C()$ representa a contagem de ocorrências

4.2 Impacto das Técnicas de Suavização

Entre as técnicas de suavização, a interpolação mostrou-se consistentemente um pouco superior ao backoff e ambas foram melhores que não utilizar suavização (none). A interpolação linear combina as probabilidades dos diferentes modelos n-gram da seguinte forma:

$$P_{interp}(t_i|t_{i-2}, t_{i-1}) = \lambda_1 \cdot P(t_i|t_{i-2}, t_{i-1}) + \lambda_2 \cdot P(t_i|t_{i-1}) + \lambda_3 \cdot P(t_i) \tag{2}$$

Onde:

- $\lambda_1, \lambda_2, \lambda_3$ são os pesos de interpolação ($\lambda_1 + \lambda_2 + \lambda_3 = 1$)
- $P(t_i|t_{i-2}, t_{i-1})$ é a probabilidade do modelo trigrama
- $P(t_i|t_{i-1})$ é a probabilidade do modelo bigrama
- $P(t_i)$ é a probabilidade do modelo unigrama

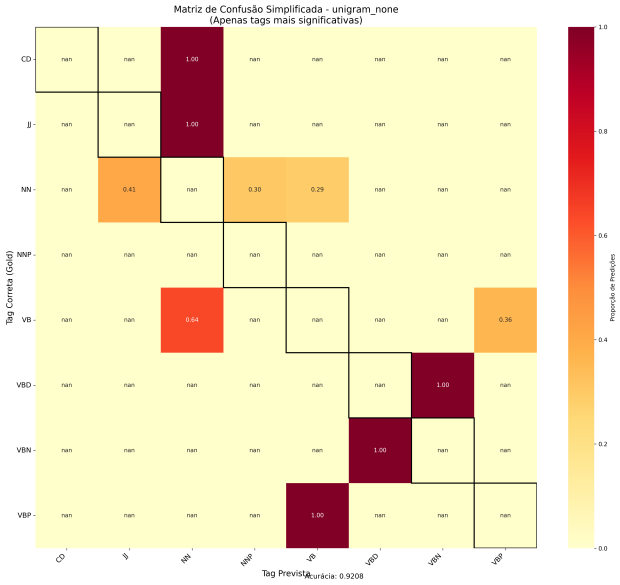
5 Conclusões

Com base nos resultados obtidos, podemos concluir que:

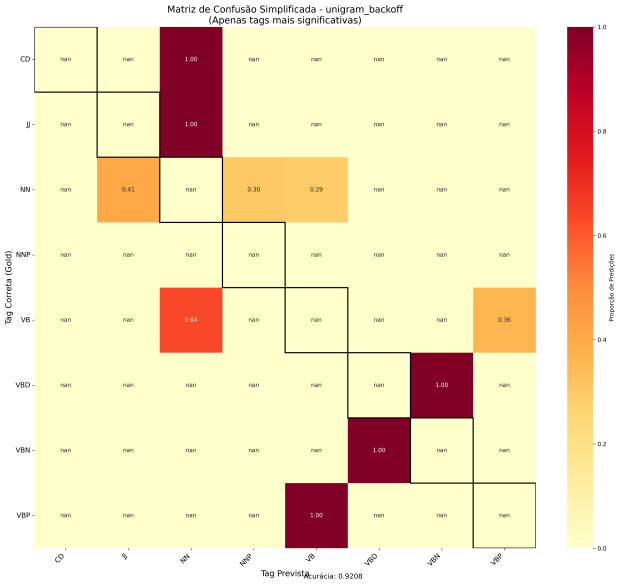
1. O modelo trigrama com interpolação obteve o melhor desempenho, com acurácia de aproximadamente 94,07%.
2. As técnicas de suavização são essenciais para melhorar o desempenho dos modelos, especialmente para os n-gramas de ordem superior que sofrem mais com o problema de esparsidade de dados.
3. As confusões mais comuns ocorrem entre categorias gramaticais semanticamente próximas, como substantivos e adjetivos, ou entre diferentes formas verbais.

6 Apêndice: Matrizes de Confusão de Todos os Modelos

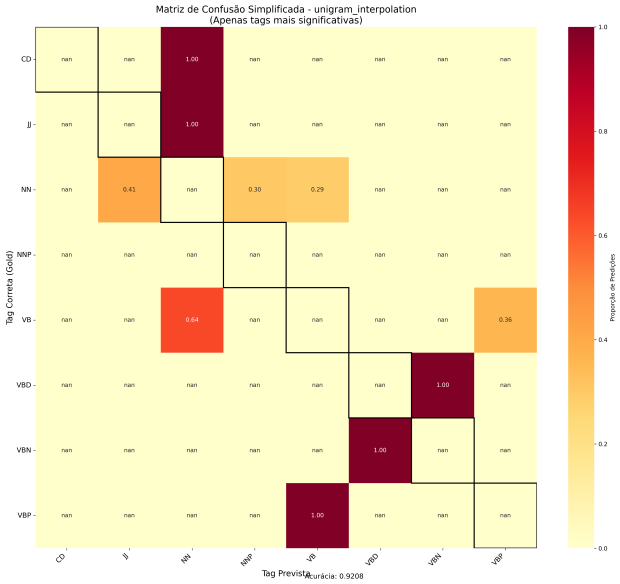
As matrizes de confusão simplificadas para todos os modelos são apresentadas a seguir.



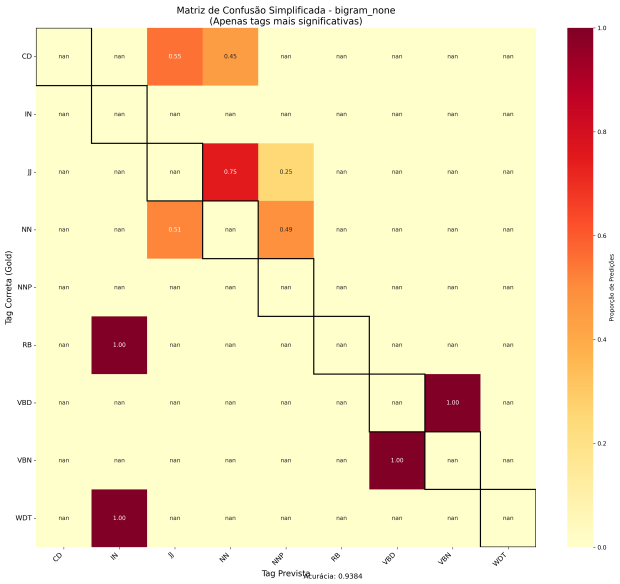
(a) Unigrama - None



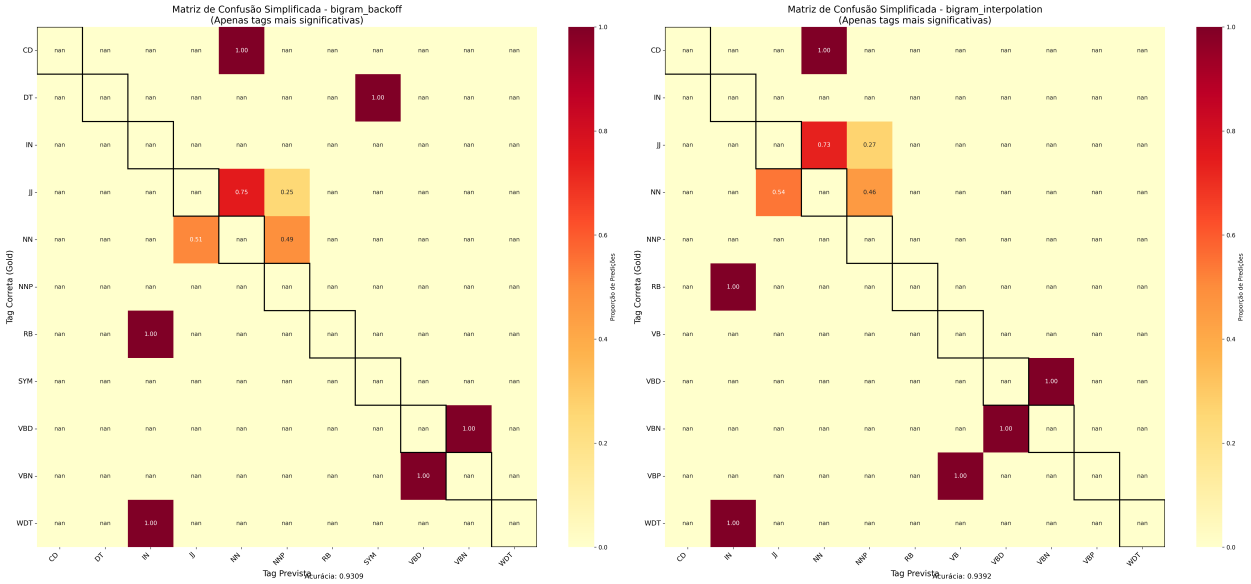
(b) Unigrama - Backoff



(c) Unigrama - Interpolation

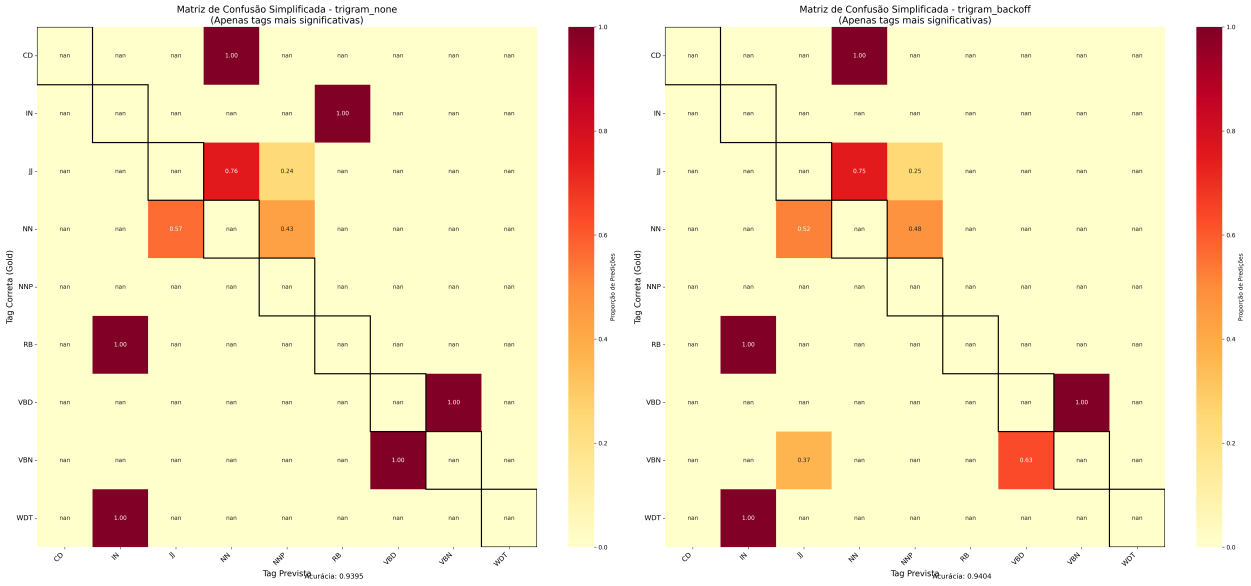


(d) Bigrama - None



(a) Bigrama - Backoff

(b) Bigrama - Interpolation



(c) Trigrama - None

(d) Trigrama - Backoff

Figura 3: Matrizes de Confusão Simplificadas para Todos os Modelos

O link pra o projeto: GitHub Repository.