Evaluating HMM-Based POS Tagging on Formal and Informal Texts

Finn Hoffmann

December 19, 2024

1 Dataset Description

The datasets used in this project were selected for their linguistic diversity and relevance to the objectives of the experiment:

1.1 EWT and GUM Datasets

The English Web Treebank (EWT) and the Georgetown University Multilayer (GUM) corpus were chosen from the Universal Dependencies (UD) repository. Both datasets are well-suited for training and evaluating Hidden Markov Model (HMM) POS taggers due to their high-quality annotations and linguistic diversity:

- **EWT:** The EWT dataset consists of web-based texts, including blog posts, emails, and online reviews. It provides a rich source of informal and semi-structured language, making it ideal for modeling modern written English.
- **GUM:** The GUM corpus covers a wide range of genres, such as academic texts, interviews, how-to guides, and fiction. Its diverse linguistic styles make it particularly valuable for understanding how the model performs across different domains and registers.

1.2 Movie Reviews Dataset

To evaluate the model's performance on out-of-domain data, a dataset of movie reviews was sourced from the GitHub repository by Bamman et al.[1]. The reviews feature informal, conversational language, including non-standard spellings, colloquialisms, and fragmented sentence structures. These characteristics make the dataset a challenging test case for an HMM-based POS tagger.

2 Experiments and Results

2.1 In-Domain Experiments

For the in-domain experiments, we trained and tested the HMM model on two datasets from UD: English Web Treebank (EWT) and Georgetown University Multilayer (GUM). The results, including accuracy, top error patterns, and error rates by sentence length and word frequency, are summarized in Table 1.

Key Observations:

- With EWT as the training dataset, the model frequently misclassifies words as proper nouns (PROPN).
- With GUM as the training dataset, adjectives (ADJ) are the most common incorrect predictions.
- Longer sentences have lower accuracy, but the proportion of correctly tagged words within these sentences is higher.
- Rare words (1-10 occurrences) exhibit high error rates, with minimal difference between the frequency bands.

[1] `https://github.com/dbamman/anlp21/tree/main/data`

Table 1: In-Domain Results: Accuracy and Error Patterns

Dataset	**EWT (Train/Test)**	**GUM (Train/Test)**
Accuracy	43.99%	34.69%
Top 5 Most Frequent Errors		
Predicted PROPN instead of NOUN	2667	-
Predicted PROPN instead of PUNCT	1932	-
Predicted PROPN instead of VERB	1436	-
Predicted ADJ instead of NOUN	-	3156
Predicted ADJ instead of PUNCT	-	2137
Error Rates by Sentence Length		
1-5 words	32.33%	38.88%
6-10 words	39.71%	44.78%
11-20 words	50.65%	58.28%
21+ words	69.28%	71.21%
Error Rates for Rare Words		
1-2 occurrences	49.58%	64.28%
3-5 occurrences	51.23%	64.67%
6-10 occurrences	49.02%	60.47%

2.2 Out-of-Domain Experiments

For the out-of-domain experiments, the model was trained on EWT and GUM datasets and tested on informal review data. The results are presented in Table 2.

Table 2: Out-of-Domain Results: Accuracy and Error Patterns

Training Dataset	**EWT (Test on Reviews)**	**GUM (Test on Reviews)**
Accuracy	13.66%	24.01%
Top 5 Most Frequent Errors		
Predicted PROPN instead of NOUN	1161	-
Predicted PROPN instead of PUNCT	922	-
Predicted PROPN instead of VERB	651	-
Predicted ADJ instead of NOUN	-	1099
Predicted ADJ instead of PUNCT	-	898
Error Rates by Sentence Length		
1-5 words	22.22%	11.11%
6-10 words	42.06%	38.89%
11-20 words	71.33%	63.53%
21+ words	90.18%	79.22%

Key Observations:

- Out-of-domain accuracy is significantly lower compared to in-domain experiments.
- Training on GUM results in better out-of-domain performance (24.01%) than training on EWT (13.66%). This could be attributed to GUM's greater variability and coverage of informal data structures.

- Errors for longer sentences and rare words are more pronounced in the out-of-domain setting.

3 Discussion

The results of the experiments reveal several interesting patterns and raise important questions regarding the linguistic phenomena and dataset characteristics underlying the observed performance differences. This section explores these findings in greater detail.

3.1 Error Analysis: Linguistic Phenomena

The error analysis highlights that the most frequent misclassification patterns are specific to the training dataset used:

- When training on the EWT dataset, the model frequently misclassifies words as `PROPN` (proper nouns), regardless of their true part of speech (e.g., `NOUN`, `VERB`, or `ADP`). This may stem from the high prevalence of named entities in the EWT corpus, which could bias the model towards over-predicting `PROPN`.
- In contrast, the model trained on the GUM dataset commonly predicts `ADJ` (adjectives) incorrectly for a variety of POS tags, including `NOUN`, `PUNCT`, and `VERB`. This could be due to the broader range of descriptive or modifying constructions in the GUM dataset, potentially reflecting its more diverse genres (e.g., academic texts, interviews, and how-to guides).

These patterns suggest that the model's POS tagging biases are influenced by the linguistic distribution and annotation conventions of the training dataset.

3.2 Sentence Length and Tagging Accuracy

The experiments show that tagging accuracy decreases with increasing sentence length. Specifically:

- While the proportion of fully correct sentences is higher for shorter sentences, the token-level tagging accuracy remains relatively stable or even improves for longer sentences. This implies that errors in longer sentences tend to be concentrated in specific, challenging regions (e.g., complex noun phrases or clauses), rather than being evenly distributed across the sentence.
- The increased complexity and syntactic ambiguity of longer sentences likely contribute to the drop in sentence-level accuracy.

Future work could investigate which specific sentence structures are most prone to errors, potentially by analyzing syntactic trees or dependency relations.

3.3 Rare Words and Error Rates

As expected, rare words (those appearing fewer than 10 times in the training data) show significantly higher error rates. Interestingly:

- The error rate for words appearing 1–2 times is only marginally higher than for words appearing 6–10 times. This suggests that beyond a certain threshold, additional occurrences in the training corpus provide diminishing returns for the model's ability to generalize to these tokens.

- Rare words often include proper nouns, technical terms, or domain-specific vocabulary, which the model struggles to classify correctly without sufficient contextual information.

3.4 In-Domain vs. Out-of-Domain Performance

The stark contrast between in-domain and out-of-domain performance underscores the importance of domain adaptation:

- The model performs significantly worse on the reviews dataset (out-of-domain) than on the EWT and GUM datasets (in-domain), reflecting the informal and conversational nature of the reviews. Features like non-standard spelling, colloquialisms, and sentence fragments in the reviews likely contribute to this performance gap.

- Interestingly, the GUM-trained model outperforms the EWT-trained model on the reviews dataset. This could be attributed to the more varied linguistic styles and genres present in GUM, which may better approximate the diversity of the reviews dataset. Additionally, GUM's inclusion of spoken and informal registers might better prepare the model for tagging conversational data.

3.5 Future Directions

To address these challenges and improve model performance, future research could:

- Incorporate domain adaptation techniques, such as fine-tuning on a small annotated subset of the target domain.

- Explore the use of pre-trained language models with richer contextual embeddings to handle rare words and out-of-domain data more effectively.

- Investigate annotation consistency and linguistic diversity within training datasets to better understand how these factors influence model biases.