NLP Assignment 3

Om Khare
112003066
Div 1, Batch B4

Feature Selection and Analysis for POS Tagging

Introduction

Part-of-Speech (POS) tagging is a fundamental task in Natural Language Processing (NLP). The accuracy of this task largely depends on the features selected for the model. In this study, we discuss the selection, analysis, and the effect of switching off subsets of features for POS tagging using a Conditional Random Fields (CRF) model.

Feature Selection

The selected features for our CRF model include:

- Bias: A default feature to ensure non-zero predictions.
- word.lower(): Lowercase representation of the current word.
- word[-3:]: Last three characters of the word, capturing common word suffixes.
- word.istitle(): Boolean indicating if the word is title-cased.
- word.isdigit(): Boolean indicating if the word is a digit.
- -1:word.lower() and +1:word.lower(): Lowercase representation of the previous and next word, respectively.
- -1:word.istitle() and +1:word.istitle(): Boolean indicating if the previous or next word is title-cased.

Analysis of Selected Features

The chart titled "Performance degradation by removing each feature" provides an insight into the importance of each feature:

- The Baseline (with all features) F1 score is 0.9676, which indicates a high performance of the model with all the features.
- On switching off individual features, the F1 score showed minimal degradation for most features, indicating that while these features contribute to the overall performance, the model is relatively robust against their removal.
- The features word.istitle() and +1:word.istitle() show a very similar degradation pattern, highlighting the importance of title casing in predicting POS tags, both for the current word and the subsequent word.

Effect of Switching Off Subsets of Features

We further analyzed the performance by removing multiple features:

- Removing '+1:word.lower()' and '+1:word.istitle()': The prediction for the sentence "The Army Corps..." showed slight differences in tags, indicating that context (next word features) plays a role in determining tags of certain words.
- Only using 'bias', 'word.lower()', 'word[-3:]', and 'word.istitle()': Here, the reduced feature set led to more noticeable differences in the predicted tags across the provided sentences, showcasing the significance of context and other removed features.

Conclusion

Feature selection plays a crucial role in the performance of the POS tagging model. While the CRF model with all features performed commendably, the degradation study revealed the importance of certain features, especially those related to title casing and context. It's evident that a balanced combination of lexical, morphological, and contextual features is vital for achieving optimal performance in POS tagging tasks.