CS572 Assignment 2 Report

Austin Brown

March 5 2024

1 Introduction

In this paper, we seek to improve the CRF that we implemented. Currently it is being trained for 10 epochs, learning rate of .2, and weight decay of .000001. The features that are being used are the ones that were given to us from the start.

2 Extensions and Modifications

2.1 Extension 1: More Features, More Epochs

The first change that I decided to make was increasing the number of features. I introduced a set of new features. A list of them can be found at the end. These features collectively contribute to a more comprehensive understanding of the linguistic properties of tokens as opposed to the basic lexical features that were included in the base mode; thus enabling the model to make more accurate predictions. I felt that these new features would allow the model to be more expressive, however with there now being more features to learn, I decided to increase the number of epochs to 20.

The prefixes and suffixes that I used came from the following source: `http://www.uefap.com/vocab/build/building.htm#:~:text=Nouns,%2C%20%2Dage%2C%20%2Dery.`

2.2 Extension 2: Weight Decay

In response to the increased complexity introduced by the expanded feature set, I observed signs of overfitting during model training, as depicted in the graph in section 5. To mitigate this issue, I opted to modify the regularization mechanism by increasing the weight decay parameter to 1e-5. This adjustment helped strike a better balance between model complexity and generalization performance, ultimately enhancing the model's ability to generalize well to unseen data. After seeing promising results with 1e-5, I attempted to increase it again 1e-4, however this change was not as effective as the prior one.

3 Conclusion

In this paper, we aimed to enhance our Conditional Random Field (CRF) model's performance by enriching its feature set and adjusting regularization parameters. Initially trained for 10 epochs with a learning rate of 0.2 and weight decay of 1e-6, our CRF utilized a basic set of features yielding a validation F1 score of .800. We extended the feature set to capture more nuanced linguistic properties, introducing prefixes, suffixes, and additional token characteristics, boosting our F1 to 0.867 . To address signs of overfitting observed during training, we increased the weight decay parameter to 1e-5, effectively balancing model complexity and leading to a validation score of .870. While further increases in weight decay did not yield significant improvements, our experimental results demonstrated notable enhancements in model performance.

4 Experimental Results

Table 1: Summary of Experimental Results

Model	**Validation F1**
Baseline CRF	0.800
More Feature, 20 Epochs	0.867
More Feature, 20 Epochs, Decay 1e-5	0.870
More Feature, 20 Epochs, Decay 1e-4	0.864

5 Figures

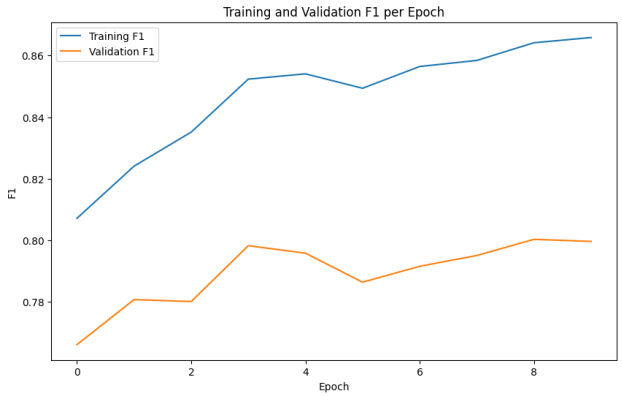

Figure 1: Base CRF

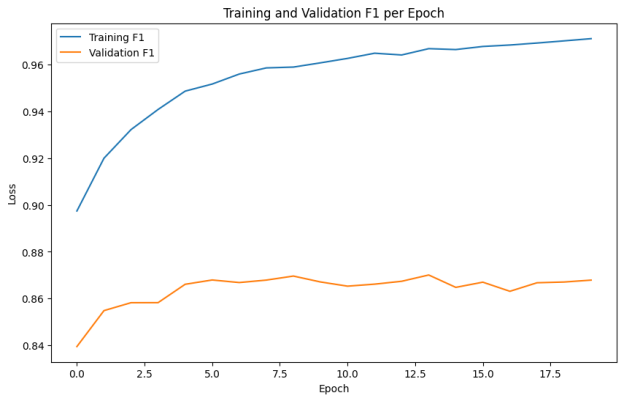

Figure 2: Best CRF: More Spam, 20 Epochs, Decay 1e-5

5.1 Added Features

- **is_title_case**: Indicates if the token is in title case
- **is_short_word and is_long_word**: Classifies tokens based on length
- **prefix and suffix**: Extracts the first two and last two characters of the token

- **len**: Provides the length of the token
- **longprefix and longsuffix**: Extracts longer substrings from the token
- **has_punctuation**: Checks if the token contains punctuation characters
- **has_nounprefix, has_nounsuffix, has_verbprefix, has_verbsuffix**: Indicates if the token begins or ends with noun or verb-related affixes
- **prevhas_nounprefix, nexthas_nounsuffix, nexthas_verbprefix, prevhas_verbsuffix**: Extends the analysis to neighboring tokens