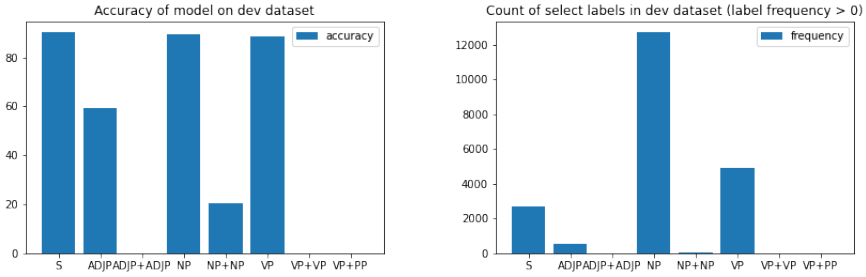
CS288 HW3 Writeup

Will Lavanakul

1 Model performance on labels

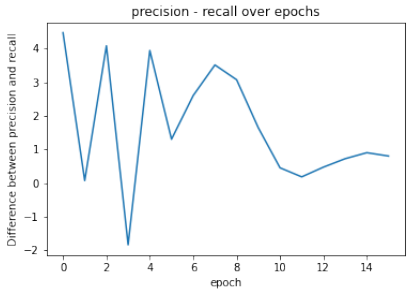
I evaluate the parsing model performance on different labels on the validation dataset. I chose the labels with the idea to see its performance on common and uncommon labels. For example, the label 'NP' occurs a lot more frequently than the label 'NP+NP'. Below are plots that display the accuracy of the model for chosen tags and the frequency of the tag. Each chosen tag has a frequency > 0 in the dataset.



As expected, the higher the frequency of the tag in the dataset, the more accurate the model is as it has trained on more examples with those labels. I think it is also interesting to see that the model performs better on the S tag over the NP tag despite S having a much lower frequency. We can also see the lack of performance on flattened labels like 'VP+PP'. I think the inherent training structure of combining the tags sequentially leads to poor accuracy on these labels. Instead, it could be an improvement to somehow induce an 'invariance' on combined tags so the model performs better in these situations. Specifically, instead of labeling the node 'VP+PP', we can allow both 'VP+PP' and 'PP+VP' to be correct predcitions.

2 Precision - recall curve

Another metric I explored is the difference between precision and recall on the parsing model. Below is the plot.



We can see that precision is on average higher than recall. What this suggests is that the parsing model performs better at classifying the label of a span than classifying if a span has a label. Specifically, when the model predicts a label for a span, it predicts well. When compared to the truth labels, it performs slightly worse as it is missing on some specific labels. This is also in line with the metrics in section (1) where we see it predicts poorly on small frequency labels.