Part 1

Name

Simon Raviv

Best Parameters

NER Tagging

- Epochs: 22
- Hidden layer size: 64
- Batch size: 128
- Learning rate: 0.001
- Learning rate scheduling gamma: 1 (no scheduling in this case)
- Learning rate scheduling step size: 10 (no scheduling in this case)
- Dropout probability: 0.5
- Weight decay: 0.0001

POS Tagging

- Epochs: 45
- Hidden layer size: 80
- Batch size: 128
- Learning rate: 0.009
- Learning rate scheduling gamma: 0.8
- Learning rate scheduling step size: 10
- Dropout probability: 0.2
- Weight decay: 0.00001

Considerations

We add an unknown word token to the train set vocabulary i.e., “<UNKNOWN>”. When building the embedding matrix, this token will be assigned to an embedding vector, as all other words, as well as trained during the training process. When building the dev/test datasets, we replace all the words didn’t appear in the train set with the unknow token.

We add additional token to the train set vocabulary i.e., “<PADDING>”. When building the embedding matrix, this token will be assigned to an embedding vector, as all other words, as well as trained during the training process.

To improve the vocabulary, I have done preprocessing on the data that did 2 things:

- Change numbers to a special token “<NUMBER>”
- Change date to a special token “<DATE>”

This indeed improved the accuracy by several percent. It make sense, because the with this, the model thinks of numbers and dates as numbers and dates, and not as multiple different tokens.

Accuracy:

NER Tagging:

The dev accuracy (without considering common tag O) is ~82 [%].

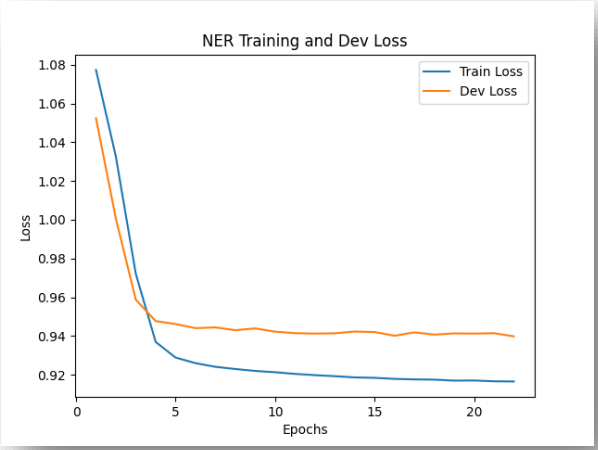
POS Tagging:

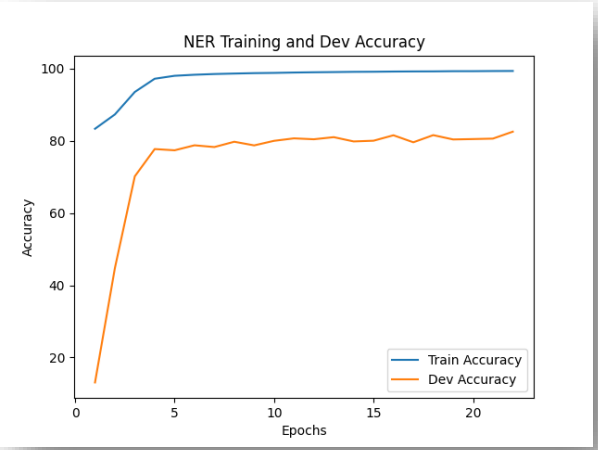
The dev accuracy is ~94 [%].

Training Graphs

NER Tagging

Note: The calculation for the dev done without considering the common label ‘O’ as requested. The calculation for the train done with it.





POS Tagging