

Report

Name: Advaith Malladi

Method (Telugu):

For this POS tagging purpose, I considered only continuous sequences of length 3. Given past 2 words, I would try to assign the POS tag to the 4th word. I chose this number because most chunks in a sentence are not longer than 4 words. Apart from the reasons mentioned, the sentences in Telugu corpus are quite smaller, around 3.

Telugu Corpus:

Scores:

On Training:

- Average accuracy: 0.967387036369295
- Average F1 Score: 0.9539225504937929
- Average Recall: 0.956848689202816
- Average Precision: 0.9565866797493937

On Testing:

- Average accuracy: 0.8589041224897724
- Average F1 Score: 0.8076397042835399
- Average Recall: 0.8322401609045442
- Average Precision: 0.8163676886279625

On Validation:

- Average accuracy: 0.8552238950088843
- Average F1 Score: 0.8189647003079836
- Average Recall: 0.8189647003079836
- Average Precision: 0.8136353944562897

Hyperparameters Used:

- epochs: 15
- batch size = 10
- sequence length = 3
- hidden dimension = 200
- layers = 1
- embedding dimension = 200
- hidden2tag layer: input: 3*256, output: target size

Let TP = True Positives

FP = False Positives

FN = False Negatives

Precision = $TP / (TP + FP)$

Recall = $TP / (TP + FN)$

Analysis:

- From the analysis, it is evident that the model did not underfit as the validation scores are quite close to the training scores.
- The testing and validation metrics are both relatively close in value, indicating that the model is generalizing well to new data.
- The F1 score, which is a harmonic mean of precision and recall, is a good measure of the overall performance of the model. The F1 score is high on all three datasets, which indicates that the model has a good balance between precision and recall.
- The task of POS taggings isn't too complex as we were able to achieve these scores using the above mentioned hyperparameters.
- The models seems to have overfit slightly as it does not perform on validation and testing as well as it does on training.
- The model seems to slightly underperform compared to the English model as the dataset in Telugu is quite small compared compared to the English dataset.
- The model is also quite simpler compared to the English model as the dataset is much more simpler, making the model more complex resulted in overfitting.
- **Given all these factors, I feel that these scores are fine for POS tagging in Telugu as Telugu is an agglutinative language and involves a lot of morpheemes and state changes.**

THE END