Assignment 3 - Named Entity Recognition

Group Members:

1. Ragul Venkataraman Ravisankar - 109772226 - rave2101@colorado.edu
2. Aravindakumar Vijayasri Mohan Kumar - 109788237 - arvi7401@colorado.edu

Approach:

We took an RNN based approach where we used Bi-directional LSTM and LSTM to perform NER.

Model Architecture:

Layer 1: Embedding Layer - 512 (Embeddings learned from train)
Layer 2: Bi-LSTM (200 units, dropout and recurrent dropout 0.2)
Layer 3: Dropout Layer (0.2)
Layer 4: LSTM (100 units, dropout and recurrent dropout 0.1)
Layer 5: Dropout Layer (0.1)
Layer 6: Dense Softmax Layer (3 => (B, I, O))

Optimizer used: Adam
Loss function: Categorical Cross Entropy

Data Preparation and Training:

1. Train Validation Split => 85:15
2. Randomly selected 30 words from train data and replaced with UNK token. This helps in handling unknown data that we encounter in the Validation and Test data set.
3. Convert the train sentences to encodings with post padding of maxlength 300.
4. Convert the train labels to one hot encoded format.
5. Fit the model with training data.

Testing:

1. Read the test data from the file
2. Replace the words not in train data with UNK token
3. Encode and pad the test sentences such that max length is 300.
4. Predict the labels for the test data
5. Write the results into a text file in the given format (same as train)