CS671A: Introduction to Natural Language Processing

Assignment 3

Ashish Kumar Singh (14142)

Dataset - Configuration

- Dataset was obtained from the official repository https://github.com/UniversalDependencies/UD_English-EWT/tree/master file named **en_ewt-ud-train.conllu**
- It was divided into training and testing dataset in 80/20 ratio
- Some of the lines in dataset had sub-index like 8.1,8.2 ; those were removed as relevant information required was already gathered from index 8
- Proper oracle was defined to convert data into configuration and corresponding transition using stack index and dependency graph

Feature Extraction

- Features taken were – topmost 3 words in stack, front-most 3 words in buffer and Universal POS Tag of the word being removed or shifted
- Words in feature were vectorized using pre-trained glove model
- For Glove Vector representation 'glove.6B.50d.txt' was used
- 50 zeros were used for words not in glove and 50 ones for 'root'
- POS Tag was One-Hot Encoded , 18 dimensionally
- Transition(shift, left,right) was also encode one-hot way , 3 dimension
- Thus, input = 318 dimension & output = 3 dimension

Neural Network

- 3 layer neural network was trained for this assignment
- First layer with 100 neurons, second with 15 and output layer with 3 neurons was trained
- Also a dropout layer with rate 0.3 was used between first and second hidden layer to avoid overfitting the dataset
- optimizer='adam', loss='categorical_crossentropy', batch_size=32, epochs=20
- Keras library with Tensorflow backend was used to train the model

Accuracy

- 95.89% with 319,844 training samples and 79,961 testing samples