CS-671A (Intro to NLP)

Assignment - 3

Shivanshu Singh (14659)

Approach and feature vector:

(i). First part of assignment is to retrieve various configurations and the operation that has to be performed from the train dataset(*en_ewt-ud-train.conllu*) of EWT treebank.

(ii). This is achieved by first parsing the training data-file and storing the required columns(*ID, UPOSTAG, HEAD, DEPREL*) in python lists. This list contains a list for every sentence present in data-file.

(iii). For each element of list (i.e. for each sentence)**,** first a root element is added to the front (of element) and then initial state of stack, buffer and tree is set.

(iv). Using the *HEAD* of each considered word, and values in stack and buffer, choose the next action (left-arc, right-arc or shift).

(v). For each state and known next action, develop its configuration, which is a tuple:

(a[0], b[0], b[1], ldep(a[0]), rdep(a[0]), ldep(b[0]), rdep(b[0]))

Where a[0] is top element of stack, b[0] is leftmost element of buffer and b[1] is second leftmost element of buffer.

(vi). For each of the element of this tuple, store its *word, PoS and dependency relation.* Thus the size of each configuration is 7*3.

(vii). Each of these element are encoded using pre-trained Glove or 1-hot encoding depending upon whether they are english words or tags.

(viii). Concatenate encoding of each element to get a vector.

(ix). For training, encoding of each configuration along with the action is saved for training of classifier.

(x). NN with 2 hidden layer with 5 neurons in 1rst hidden layer (50 neurons) while 2 neuron in 2nd hidden layer (10 neurons). There are total of 4 layers.

(xi). Train this NN with each pair of configuration encoded value and the next action.

Accuray obtained:

On the given data-file (*en_ewt-ud-test.conllu*) of EWT treebank,

Accuray: **90.4125%**