HW 4: Named Entity Recognition

In this assignment we had to implement a learning based approach to named entity recognition, here the named entities are given the IOB (Inside–Outside–Beginning) tags.

This task of labelling can be achieved through HMM-based solution, which is Viterbi Algorithm. I have used Viterbi for the implementation of this assignment as it proved to me better in performance than logistic regression.

Viterbi Implementation:

Here in Viterbi we try to form paths from different tags and trace back the best possible path of tags to predict the IOB tags for a given sentence. For this system, we would need two major values:

1. Word->Tag counts: For a given word the probabilities of its tags
2. Tag1->Tag2 counts: For a given tag the probabilities of its next tags

Using these probabilities we form a matrix of all possible transitions from one tag to another given its probability of word at the ending tag. After forming the system we backtrack the matrix using the previous best tag for a given word. In this manner we are able to form a system of IOB tags for a sentence.

Handling limitations:

The transition from one tag to another may not happen always, thus we may get zeros for these values, if these zeros are filling the maximum of the table then the chances of finding the best possible path may not be achieved. To overcome this problem we have used Laplace smoothing on the Transition table (Tag->Tag transition table).

The next limitation of the system is handling the unknown words, to overcome this problem I have considered an unknown word to get the least value that a word given tag in the observation matrix can get, thus this word will have the least probabilities of the tags assigned to it.

Testing:

Taking 80%-20% data into Training Set & Test Set respectively, the system was tested using this data. By using the program of evalNER.py we could measure our Precision, Recall and F1-measure. This gave us an idea about what could go wrong and where should we correct the program.