Assignment 2: Part-Of-Speech Tagging

In this assignment we had to create a Part of Speech Tagging system, we were given the BERP corpus to form this system. Here, we have first implemented the Baseline system to check the basic prediction. Baseline system works on the basis of word counts for a given tag i.e., for a given word the system checks for the highest tag associated to this word and prints it out.

Baseline Tagger:

Using the basic structure of dictionaries in python, the dictionary wordTagCount[tag][word] (for a given word gives the tag count).

Taking 80-20 data into Training Set & Test Set respectively accuracy obtained was: 91.6% approx.

The problem with Baseline Tagger is it's handling of unknown words and prediction of parts of speech for the next word. Hence, for this problem we use Viterbi Algorithm.

Viterbi Tagger:

Here in Viterbi we try to form paths from different tags and trace back the best possible path of tags to predict the parts of speech 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 parts of speech for a sentence.

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 all the words with occurrence of 1 in the word count and treating them to be unknown words in this corpus.

Taking 80-20 data into Training Set & Test Set respectively, this methodology helped me improve my accuracy to 93% approx.

Thus the system was successfully able to predict the parts of speech of up to 93% of word in a sentence of the same corpus.