Kevin Rau

101289616

CSCI - 3202

Assignment 8 Write-Up

Purpose:

To understand HMM and apply concepts learned in class to, “identify the part-of-speech tag for words in a sequence of words”. As discussed in class, this was ran over a large set of sentences to derive probabilities of tags associated with words given the probabilities measured over the states we were in.

Procedure:

Transition Probabilities:

The data structures that were used consisted of tag objects that were accumulated into a set over the course of the program running after parsing the file. The tags simply had the number of occurrence, type associated, and the probabilities were calculated in the main function given the format of, “P (DT|NN) will be calculated by count (DTi NNj) / count (NNj)”. This was kept track by setting the attributes of a particular tag and the looking at which tag came next, changing its count based on if the tags matched or not.

Emission Probabilities:

This was measured by “(word AND TAG) / count (TAG)” This was done in the main file and in the emission function it was done based on calculating the total count associated with the tags and words.

Viterbi Algorithm:

Once the transition and the emission probabilities were calculated in the main function the last call was made to Viterbi. As commented in the code, there is an initial probability that is established of occupying a state, with V being set up as a dictionary of states. While this algorithm is ran we derive which state is most likely to be transitioned based on the evidence available and get the most probable state and its backtrack. We follow this backtrack all the way to the first observation and print the results.

Data:

The data that was used was the given penntree.tag file. This was pre-processed using was given in the write up when placing the "SSSS" and "EEEE" tokens to break up sentences. The data was formatted in such a way that the words and tags were separated by tabs. Once processing was done, the data was formatted in an array which could be used for the program to work on.

Results:

The results seemed to come out as expected with the given sample results. I also ran some of the tests from the original write up on http://verbs.colorado.edu/~mahu0110/teaching/ling5832/5832-hw3.html and it seemed to be working correctly with the outputs they gave.