Assignment 4, HMM **Deadline : 14th Oct**

In this assignment, you have to implement an HMM-based approach to POS tagging. Specifically, you have to implement the Viterbi algorithm using a bigram tag/state model. As training data, I am providing a POS-tagged section of the BERP corpus. Your systems will be evaluated against an unseen test set drawn from the same corpus.

Training

I am providing you the training data consists of around 15,000 POS-tagged sentences from the BERP corpus. The sentences are arranged as one word-tag pair per line with a blank line between sentences, words and tags are tab-separated. Contractions are split out into separate tokens with separate tags. An example is shown here

I PRP
'D MD
Like VB
french JJ
Food NN

You should assume that the tags that appear in the training data constitute all the tags that exist (no new tags will appear in testing). On the other hand, new words may appear in the test set.

Decoding

For decoding, your system will read in sentences from a file with the same format minus the tags. That is one word per line ending with a period and blank line before the next sentence. As output you should emit an appropriate tag for each word in the same format as the training data.

Assignment

To complete the assignment, you'll need to address the following problems:

Extract the required counts from the training data for the various probability estimates that are needed.
Deal will unknown words in some systematic way.
Do some form of smoothing (for the bigram transition probabilities).
Implement the Viterbi algorithm.

Submit your saved model , running code and report .

. .