Homework 3

Sivakarthik Gade

INSTRUCTIONS TO COMPILE AND RUN THE CODE

1. Implemented the assignment in Java. Need Oracle/Sun JDK setup on the target machine to compile and run the code. (Used JDK 1.7.0_45 during development)
2. Extract all java files into some location.
3. Compile instruction through command prompt: 'javac *.java'
4. Run instruction through command prompt:
    a. java HMMPOSTagger -train hw3_train.txt -model hw3_model.txt
    b. java HMMPOSTagger -test hw3_test_00.txt -model hw3_model.txt -o hw3_output.txt
    c. java HMMPOSTagger -ref hw3_test_ref_00.txt -sys hw3_output.txt

REPORT

TRAIN

1. Model file contains the following data:
    a. P(T2/T1) Transition Probabilities. <prob> <t1> <t2>
    b. P(w/T) Observation Probabilities. <prob> <t> <w>
    c. Possible tags for a word. <word> <t1> <t2> …
    d. P(T) Tag unigram probability. This is required while calculating p(w/T) for unseen words while testing. <tag> <prob>

TEST & RESULTS

1. Output file generated is similar to training file. Only exception is – to identify unseen test words, have prefixed them with ~~.
2. For unseen words, have used the suggested 2 approaches in the homework pdf to compute p(t/w) term while computing p(w/T):
    a. p(t/w) = uniform distribution
        i. Overall Accuracy:0.918779402014334
        ii. Known Word Accuracy:0.934968120048947
        iii. Unknown Word Accuracy:0.10662358642972536
    b. P(t/w) = prior distribution p(t)
        i. Overall Accuracy:0.9253622959618603
        ii. Known Word Accuracy:0.9387035486571779
        iii. Unknown Word Accuracy:0.2560581583198708
    c. The submitted code computes p(t/w) using b) option listed above as higher accuracy is reported in that case.
3. For unseen words, have used a small constant as p(w).
4. While training the model using a smaller training set, the overall accuracy of the Viterbi algorithm decreased. This is because as training set size decreased, the number of unseen words in the test increased. Since the methods used to compute p(w/t) for unseen words is not very efficient, the overall accuracy decreased.