COMS 6998-7, Spring 2010
Statistical Methods for NLP

Homework 2

Due: March 11 (Thursday), 2010 (11:59pm)

Total Points: 100

Parts of Speech Tagging with Hidden Markov Models

You are given two sets of corpora. First corpus is Wall Street Journal text that has all the words tagged with Parts of Speech (POS). You should use this corpus for Q1. For Q2 you are given a large amount of text from Gigaword corpus. You can use this text for unsupervised training of your HMM. You cannot use these corpora for any other purpose besides these homework questions.

Q1. Viterbi Algorithm [40]

Data directory: /home/smaskey/CS6998/hw2/labeled_data

(1) Use the labeled training data provided to estimate the transition and emission matrix for a fully connected HMM. [10]
(2) Implement Viterbi algorithm. [20]
(3) Test your sequence classification accuracy using the test file in the test directory. Compute the accuracy of your classifier. [10]

Q2. Baum-Welch (Forward-Backward) Algorithm [60]

Data directory: /home/smaskey/CS6998/hw2/unlabeled_data

(1) Implement Forward Algorithm [10]
(2) Implement Baum-Welch algorithm [30]
(3) Train your HMM model and provide the number of iterations vs. log likelihood table. How many iterations of forward-backward are run before convergence? [10]
(4) Using Viterbi algorithm you implemented in Q1 to test your POS tagging accuracy on the test set. Compute the accuracy of the classifier. [10]

Extra Credit [5]

Find the best topology for your HMM and explain how you found it.