POS Tagging Assignment

This is an assignment which tests you on the various POS tagging concepts we learnt in class. You can use your favourite programming language to do this assignment.

I. Corpus

Divide the Brown corpus into 2 distinct files. The first 2000 sentences should be stored in a file called brown-test.txt. Store the remaining sentences in another file called brown-train.txt

II. Tagger Implementation

1. Implement a simple unigram tagger i.e. for each word in the training data (brown-train.txt), extract the most frequent tag and store it it to file. Then read in this file to tag each word in (brown-test.txt).

2. Implement a simple bigram tagger i.e. previous-tag-current-tag frequencies corresponding to each word in the training data (brown-train.txt), and store it it to file. Then read in this file to tag each word in (brown-test.txt).

3. Explain your strategy for dealing with unknown words. (5 points)

III. Tagger Evaluation

1. For each of the taggers above, print out the overall POS tagging accuracy. (2 points per tagger)

2. Print out a confusion matrix for each tagger. (5 points per tagger)

3. Print out accuracy for unknown words i.e. words which are present in the test data (brown-test.txt), but absent from the training data (brown-train.txt). (5 points per tagger)

Extra credit (5 points in total)

Experiment with an alternate strategy for tagging unknown words using a bigram tagger and report performance.