CS630 Representing and Accessing Digital Information
Fall 2004
Assignment 3

Due at the beginning of class, on Monday, November 22

Overall Goal

Implement an automatic system that performs limited part-of-speech tagging. To simplify the problem, this assignment focuses only on one individual part-of speech tag. In particular, the goal is to build a system that recgnizes nouns.

Given is a sample of training and test documents from the Penn Treebank Wall Street Journal corpus. The training set consists of sections 2-21, including an overall number of 947286 words. The test set consists of section 22, including 39995 words. The datasets contain one word per line. The tag is appended to the word after a "/". The tag for nouns is "NN".

Instead of using HMMs as discussed in class, phrase the problem as a binary classification task. Nouns are positive examples, all other words are negative examples. Describe each word by a set of features. However, you MUST NOT use the part-of-speech tags to generate the features (that would be cheating, since they are obviously unknown in a realistic setting).

You can use any software tools, as long as you acknowledge their use.

Task 1

Design a set of features that will be useful for classifying nouns. Be creative. List and explain the features you inventedd.

Task 2

Construct a classification rule that is maximally accurate. You can construct the rule by hand, or use any machine learning method. You can use the training set to tune your rule. Evaluate your rule on the test set using error rate as a performance measure. Report the error rate on the training set and on the test set. Describe the approach you took and include all source code in your writeup.

If your method is not efficient enough to handle the full training set, select a suitable subset.

Task 3

Repeat the experiment for one other part-of-speech tag of your choice.

Task 4

Discuss in which respects this classification approach is inferior or superior to an HMM approach.