CS 585 Natural Language Processing
Fall 2004
Homework 4

Out: Tuesday, November 30, 2004
Due: Tuesday, December 7, 2004

1. [10 points] What problem with generative probabilistic models is addressed by conditional maximum entropy models? Why are practitioners of NLP especially interested in addressing this problem?

2. [15 points] You are told that there are three classes of email messages, $C \in \{Spam, Ham, Salami\}$. Half the messages belong to *Ham*. Ten percent of the messages contain the word "Money" and 100% of those belong to *Spam*. What is the maximum entropy distribution distribution for $P(C, M)$, (where M indicates contains "Money", or does not contain "Money")?

3. [15 points] NLP practitioners need some measures for evaluating the performance of their parsers. Say that the correct parse for a sentence is
((The astronomers) (saw (the (stars)) (with telescopes)))
but an automatic parser returned instead
((The astronomers) (saw (the (stars (with telescopes)))))
Give the precision, recall, and crossing-bracket percentages. Name one problem with these standard measures.

4. (a) [8 points] Give two examples of (English or other natural language) linguistic phenomena not captured by CFGs. (b) [8 points] For one of these phenomena, explain how NLP researchers have augmented the basic CFG model to capture it.

5. [20 points] Calculate the string edit distance between "processing" and "parsing" using Levenshtein distance. Create the dynamic programming table, fill it in completely with costs, state the optimal cost, and show the edit operations that achive that cost.