CSC 248/448 Assignment #4
Due: November 25th, 2003

1. Consider the example in the Lecture 17 notes for identifying sentences that involve making appointments. Say we want to build a vector representation that consists only of 5 values.
a) Develop good method of identifying the most useful words to use in a vector representation to perform this task. Describe your method and your rationale for why you think it will be effective. Apply your method to identify the five words you will use in your representation.
b) Describe whatever normalization and other criteria you will apply to produce your final vector representation of each document. Justify your decisions.
c) Develop a method to evaluate how well your model does using on the 8 training instances provide. Do you classify each sentence correctly? Is your test a fair test of your model (this may involve a discussion of pros and cons rather than just a yes of no).

2. Programming Part
This assignment involves building and testing a word-to-phoneme conversion which could be used as input to a speech synthesizer. You have available approximately 3000 words of training data, especially selected so that only 10 phonemes are used throughout.

I suggest you first implement a program that given a word of length n and a phoneme sequence of length m, generates a list of all possible alignment functions mapping the n symbols to a phoneme or the empty symbol, which we'll write as <e>. For instance, given MIDST and M IH S T, we would generate <e> M IH S T, M <e> IH S T, M IH <e> S T, M IH S <e> T, and M IH S T <e>, The output of this program should be a data structure that is suitable for use in the rest of the assignment.

a) Given some initial probability distribution, implement an EM algorithm that computes a new probability distribution given the training data. Demonstrate your algorithm running on the first 20 instances in the training data. Show the probability distribution for P(* | A) and P(* | H) for three iterations of the EM algorithm starting from the uniform distribution.

b) Implement two versions of the letter to phoneme conversion. One that simply tries to maximize $\prod P(S_i \mid T_{a(i)})$ and the other that uses this formula plus the prior distribution $P(T_{1,s})$ as well, i.e., the full formula derived in the notes. Design and implement one or more evaluation criteria for determining how well each model does, and then, using only the data you have been provided, determine whether having the prior in the model increases your accuracy.

c) Submit a program that reads in a single word and produces the best translation of this word into phoneme. This program will be tested on previously unseen data.

d) (CSC 448 Students, or 248 for extra credit)

Design an alternate version of this program that uses a more complex alignment function, say one in which either single letters or pairs of letters may be associated with single phonemes. Derive an appropriate formulation of this model and describe it carefully, paying particular attention to how to capture the alignments. Evaluate this new model against the model developed in part (c) and submit it in the same way for evaluation.