COMS W4705x: Natural Language Processing
MIDTERM EXAM
October 21st, 2008

DIRECTIONS

This exam is closed book and closed notes. It consists of three parts. Each part is labeled with the amount of time you should expect to spend on it. If you are spending too much time, skip it and go on to the next section, coming back if you have time.

The first part is short answer. The second part is problem solving. The third part is essay.

Important: Answer Part I on the test sheets and turn in the test itself. Answer Part II using a separate blue book for each problem. Answer Part III in a separate testbook. In other words, you should be handing in the test and at least four blue books.

Part I – Short Answer. 25 points total. 15 minutes.

Provide 2 or 3 sentences for *five out of the following seven* questions. Each question is worth 5 points. Write your answers on the test.

NOTE: Choose FIVE.

1. State the difference between inflectional and derivational morphology, noting their relation with word class.

2. What is the difference between a FSA and a FST?

3. How would you build a dictionary for the most frequent POS tagging algorithm?

4. Give the FOPC representation for the following two sentences, providing two representations if it is ambiguous, one representation if it is not and describing any problems that FOPC might have.

John asked every girl a question.
Someone in 4705 likes language.
Most students are happy at the end of the semester.

5. Give examples of three different types of structural ambiguities and say why they are ambiguous.

6. Assuming the grammar below, show the parse tree for the sentence *The big yellow dog sat under the house.*

S-> NP VP
VP -> VP PP
VP -> verb NP
VP -> verb
NP -> DET NOM
NOM -> ADJ NOM
NOM -> NOUN
PP -> PREP NP
DET -> the
ADJ -> big
ADJ -> yellow
NOUN -> dog
VERB -> sat
PREP -> under
NOUN -> house

7. Show how you would have to modify the grammar above to handle the sentence *The dog in the white hat ran under the house.*

Part II. Problem Solving. 45 points. 40 minutes.
There are three problems in this section worth 15 points each. Do all 3 problems. Use a separate blue book for each problem.

1. **Finite State Transducers**. Eliza was a program developed to answer questions as a psychologist would. For this question, you are asked to write an FST that will respond to questions.
   a. Write an FST that will respond as in the following dialog. Make sure that your FST will handle other similar dialogs but with different input statements and give an example of a similar dialog it can handle.

   Input: I am very unhappy.
   Eliza: Why are you very unhappy?
   Input: My girlfriend thinks I'm mean
   Eliza: And are you mean?

   b. Now create an FST to handle the input "*I am very unhappy*": as well as the two paraphrases: 1. ``*I am not very happy*'' 2. ``*I am just not happy*''

   c. Create a new FST to handle the input with ``I am sorry to hear that you… ''. Show an example of input and output.

2. **Context Free Grammar:** You are given the grammar below. Show how it would be used to derive a parse tree for the sentence below. Show the order in which rules would be applied if using a top down search strategy and the order in which the rules would be applied if using a bottom-up grammar. Identify when disambiguation would occur, all partial structures that would be built and the parse tree that would be returned.

   The complex houses first-year students.

   S -> NP VP
   NP -> DET NOM
   NOM -> ADJ NOM
   NOM -> NOUN
   VP -> VERB NP
   VP -> VERB
   DET -> the
   ADJ -> complex
   ADJ -> first-year
   NOUN -> complex
   NOUN -> first-year
   NOUN -> houses
   NOUN -> students
   VERB -> houses

3. **Hidden Markov Models:** You are given the sentence below and the tables of probabilities show in Table 3a (this page) and Table 3b (next page).

I promise to back the bill.

a. Describe how the probabilities would be obtained using the Penn Treebank.

b. A hidden markov model includes states, observations, transition probabilities, observation likelihoods. Describe what each one of these would correspond to when using an HMM for POS tagging.

c. Given the sentence ``I *promise to back the bill.*'' show how you would compute the probability of ``*back*'' as a verb versus the probability of ``*back*'' as a noun using the probabilities in Tables 3a and 3b using the Viterbi algorithm. You are given the values for the third column of the Viterbi table which correspond to observation 3 or ``to''. They are VB: 0, TO: .00000018, NN: 0, PPSS: 0. Thus, you will show two computations both of which will use these values. You do not need to do the arithmetic; just show the formula that would be computed.

	I	***promise***	***to***	***back***
VB	0	.0093	0	.00008
TO	0	0	.99	0
NN	0	.0085	0	.00068
PPSS	.37	0	0	0

Table 3a: Observation Likelihoods

	VB	**TO**	**NN**	**PPSS**
<s>	.019	.0043	.041	.067
VB	.0038	.035	.047	.0070
TO	.83	0	.00047	0
NN	.0040	.016	.087	.0045
PPSS	.23	.00079	.0012	.00014

Table 3b: Tag transition probabilities. The rows are labeled with the conditioning event. Thus, P(VB|<s>) = .019.

Part III. 30 points. 20 minutes.
Essay. Answer *2 out of the following 4 questions*, worth 15 points each. Use no more than 2 paragraphs for each question. Put your answers in one blue test book.

NOTE: CHOOSE 2

1. This question concerns the rule-based POS tagger, ENGTWOL, and Brill's transformational POS tagger, TBL. Give a brief description of how each tagger works and then state one similarity and one difference between the two.

2. In each of the following sentences, identify the semantic roles selecting from *agent, patient, theme, experiencer, stimulus, goal, recipient, benefactive, source, instrument, location, temporal*. Justify your choice.

The company wrote me a letter.
Jack opened the lock with a paper clip.
The river froze during the night
Kathy ran to class every day at Columbia.
I felt the warmth from the fire.

3. Consider the sentences *President George Bush has re-invigorated the economy by providing a bail-out program for failing Wall Street firms*. and *President George Bush has caused a disastrous economic situation by failing to provide regulations on Wall Street firms.* You'd like to compute the likelihood of these sentences given a corpus of NY Times, Wall Street Journal and the New York Post gathered over the last year. You develop a bi-gram language model. Describe how you would: 1. Build the language model, 2. Compute the likelihood of these sentences and 3. Evaluate your language model.

4. Describe how probabilities would be computed from a corpus for the rule VP -> verb NP PP in a PCFG grammar. How is the probability of a parse tree computed in a PCFG? Contrast this with the computation of probabilities from a corpus for rules capturing lexical dependencies that might be used to disambiguate sentences such as *The woman dropped the stamped letter into the mailbox*. and *The woman opened the letter from a friend.* How would disambiguation occur?