COMS W4705x: Natural Language Processing
FINAL EXAM
December 18th, 2008

DIRECTIONS

This exam is closed book and closed notes. It consists of four 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 multiple choice. The second part is short answer. The third part is problem solving. The fourth part is essay.

Important: Answer Part I by circling answers on the test sheets and turn in the test itself. Answer Part II, Part III and Part IV in separate blue books. In other words, you should be handing in the test and at least three blue books. Put your NAME on the EXAM ITSELF and each TEST BOOKLET.

Part I – Short Answer. 18 points total. 20 minutes.
Answer all questions below on the test itself:

1. Dynamic programming approaches to parsing address the inefficiencies of backtracking state-space parsers by
a. enabling parsers to make the right decisions when ambiguous choices are encountered during parsing
b. recording the decisions made so far in a chart
c. dynamically allocating memory as needed for the parse tree
d. none of the above

2. The pyramid method
a. is a method for summarization evaluation
b. determines semantic content units through overlap between sentences
c. weights units that occur in more than one human summary more heavily
d. all of the above

3. Hearst’s topic segmentation algorithm uses three features to determine a score of lexical cohesion between sentences. Which of the following is *not* one of the three features?

a. lexical similarity between sentences computed by the cosine metric
b. anaphora between sentences
c. introduction of new terms
d. lexical chains

4. Rhetorical structure theory represents
a. topic segmentation
b. good speaker rhetoric
c. hierarchical discourse structure using relations between sentences
d. nucleus and satellite of each word in a sentence

5. True or False: Developing an efficient information extraction system using pattern matching would be difficult.

6. Consider the example: Kathy *and Donia bought coffee and donuts. They shared them.* Which of the following constraints help in determining the antecedent of *they* and *them*? Circle all that apply.
a. number agreement
b. case agreement
c. gender agreement
d. selectional restrictions

7. True or False: Translational divergence is a term used in MT to refer to cases where one word may be translated by many words in the target language.

8. True or False: A finite state automaton can be used to characterize a formal language that consists of an infinite number of strings.

9. In the sentence *Sally sent the book to Florida.* Which thematic role does *the book* play in the sentence?
a. theme
b. patient
c. stimulus
d. instrument

Part II. Short Answer. 25 minutes. 20 points total. Answer 4 out of 5 questions in one blue book.

1. Show an **event-oriented, first order logic** representation for the following example sentence: (don't worry about how it might be produced; just show what you think the final representation should be.)

Kathy decided to grade HW4.

2. State two constraints on lexical choice in language generation and in one sentence per constraint, describe how they would be used.

3. There are a number of problems with WordNet that it make it difficult to use for Words Eye. Choose two of the following three problems and explain why they are problems for Words Eye, giving an example for each:

Lack of multiple inheritance
Lacks relations other than is-a
Cluttered with obscure words and senses

4. What is the difference between homonymy and polysemy? Give an example of each that illustrates your point.

5. The plural of potato is potatoes. Describe the general technique for using finite-state machines for morphological analysis that is suggested by this and other related facts.

Part III. Problem Solving. 44 points. 80 minutes.
There are three problems in this section worth 20 points each. Do all 3 problems. Use a separate blue book for this part.

1. [12 points total] Syntax. Consider the following grammar:

S ->NP VP	**NP->Kathy**	**Verb->flew**
VP -> Verb NP	**NP->London**	**P->in**
VP -> Verb PP	**NP->Paris**	**P->to**
VP ->VP PP	**NP->February**	**CONJ -> and**
NP ->NP PP		
NP -> NP and NP		
PP-> P NP		

a. [6 points] Show three parse trees that would be derived for the sentence *Kathy flew to London and Paris in February*.

b. [3 points] Given a treebank how would you determine probabilities for the grammar rules given in Question 1 (for use with a basic PCFG parser)?

c. [3 points] To improve performance on sentences like the example in Question 1, advanced probabilistic parsers make use of probability estimates other than those based on grammar rules alone. Describe one such probability estimate that might have helped with determining the best parse of part a.

2. [20 points total] Semantics. The preposition “of” is semantically ambiguous and, in fact, has many different possible meanings. This was graphically demonstrated in Words Eye in class. Consider the following four noun phrases:

house of brick
Bowl of cherries
Deck of cards
Height of the horse

a. [4 points] For each of the noun phrases, describe the meaning that is intended.

b. [6 points] If you were to write a program to do lexical disambiguation of “of” using semantic features derived through WordNet lookup (note: WordNet has no entry for “of” so the program must be based on look up of other nouns), how would you do it? For which of the noun phrases above do you think you could get a correct answer?

c. [6 points] If you were to write a program using a supervised statistical approach to lexical disambiguation how would you do it?

d. [4 points] In your opinion, which approach would work better for this problem and why?

3. [12 points]. Information Extraction and bootstrapping. Suppose you are building an information extraction system to identify the city and state in which a person was born. You want to use bootstrapping to do this.

a. [6 points] You know where Barack Obama was born (Honolulu, Hawaii) but you don’t know where any other famous person was born. Describe how you could use this information, along with a combination of Google and Wikipedia, to find patterns that could be used in general to determine place of birth. Be specific.

b. [6 points] Why is it better to use both Google and Wikipedia rather than one corpus alone? What advantage does each corpus have?

Part IV. 18 points. 30 minutes.
Essay. Put your answers in one blue test book. NOTE: ANSWER TWO OUT OF 1a, 1b, and 1c.

1. In class we discussed the question of whether the state of the art in natural language processing is mature enough to allow summarization from intermediate representations and still allow robust processing of domain independent material. Over the course of the class, we considered the following types of natural language processing that would result in intermediate representations: POS tagging, parsing, semantic interpretation, and discourse processing (e.g., anaphora resolution, discourse segmentation, rhetorical structure theory). On your homework, you addressed the question of whether parsing was robust enough that it could be used for summarization in general, so we will not consider that type of processing here.

   Choose **two** of the following three questions addressing the remaining types of processing:

   a. POS tagging: Do you think POS tagging is robust enough to be used as part of a domain independent summarization system? Justify your response. How could it be used to help summarization (assuming that it could eventually be made robust enough if it is not now)? Use 1-2 pages of your blue book.

   b. Semantic processing: Do you think word sense disambiguation is robust enough to be used as part of a domain independent summarization system? Justify your response. How could it be used to help summarization (assuming that it could eventually be made robust enough if it is not now)? Use 1-2 pages of your blue book.

   c. Discourse processing: Do you think topic segmentation is robust enough to be used as part of a domain independent summarization system? Justify your response. How could it be used to help summarization (assuming that it could eventually be made robust enough if it is not now)? Use 1-2 pages of your blue book.