CS474

Introduction to Natural Language Processing

Midterm
October 27, 2005

Name:

Netid:

Instructions: You have 1 hour and 15 minutes to complete this exam. The exam is a closed-book exam.

#	description	score		max score
1	word sense disambiguation	____	/	15
2	p-o-s tagging	____	/	15
3	HMMs	____	/	15
4	transformation-based learning / WSD	____	/	20
5	bottom-up chart parsing	____	/	20
6	top-down parsing	____	/	15
	Total score:	_____	/	100

1 Word Sense Disambiguation (15 pts)

Given the WordNet entries below, apply Lesk's dictionary-based word sense disambiguation algorithm to the word **bass** given the context **bass playing maniac**. For full credit, the step-by-step calculations must be shown and each step very briefly described.

bass
bass (the lowest portion of the musical range)
bass, basso (an adult male singer with the lowest voice)
sea bass, bass (the lean flesh of a saltwater fish of the family Serranidae)
bass (the member with the lowest range of a family of musical instruments)

playing
playing (the act of playing a musical instrument)
playing (the action of taking part in a game or sport or other recreation)

maniac
lunatic, madman, maniac (an insane person)
maniac (a person who has an obsession with or excessive enthusiasm for something)

2 Part-of-Speech Tagging (15 pts)

Please read both parts of the problem before answering.

1. (7 pts) Briefly describe one potential problem with the Penn Treebank part-of-speech tagset.

2. (8 pts) Explain why the issue you raise in part (1) is a problem via a concrete example. Be sure to make clear how the example maps on to the problem you specified in part (1), e.g. you might show the relevant part-of-speech tags along with the words of an example sentence.

3 HMM's for Part-of-speech Tagging (15 pts)

In the part-of-speech tagging assignment, you used Hidden Markov Models (HMMs) and the Viterbi algorithm to compute the most probable sequence of tags $T = t_1, t_2, \ldots, t_n$ given the words, $W = w_1 w_2 \ldots w_n$, in a sentence.

1. (8 pts) Assuming a **trigram** part-of-speech tagging model, provide the equation that is maximized by the Viterbi algorithm. (Be sure to define all variables and use clear notation.)

2. (7 pts) Using Maximum Likelihood Estimation (and no smoothing), show (preferably via equations) how each term in the above trigram tagging model (from part 1) would be estimated from a training corpus.

4 Transformation-Based Learning (20 pts)

Think about how to best apply Brill's Transformation-Based Learning (TBL) algorithm (which we studied in the context of part-of-speech tagging) to the Senseval lexical tagging task for word sense disambiguation. (This is just the task you handled in the word sense disambiguation (WSD) assignment.)

1. (5 pts) What would be a reasonable **initial state tagger** for a TBL solution to the task?

2. (5 pts) What **scoring function** should the transformation-based learner use for this task?

3. (10 pts) Design a set of tranformation **templates** for the task.

5 Bottom-up Chart Parsing (20 pts)

Given the grammar and lexicon below, show the **final chart** for the following sentence after applying the bottom-up chart parser from class:

Run the Detroit marathon

Remember that the final chart contains all edges added during the parsing process. You may use either the notation from class (i.e. nodes/links) or the notation from the book to depict the chart.

S → NP VP	Det → the
S → VP	Noun → run | marathon
NP → Det NP	Verb → run
NP → Proper-Noun Noun	Proper-Noun → Detroit
VP → Verb NP	

6 Top-Down Parsing (15 pts)

1. Given the grammar and lexicon below (which is the same as that of question 5), show one possible **top-down derivation** for the sentence:

   Run the Detroit marathon

S → NP VP	Det → the
S → VP	Noun → run — marathon
NP → Det NP	Verb → run
NP → Proper-Noun Noun	Proper-Noun → Detroit
VP → Verb NP	

2. The Earley algorithm introduces top-down predictions into the chart parsing algorithm. What top-down edge(s), if any, would be added to the chart at position 0 for the grammar and sentence above?