CS4740/CS5740
Introduction to Natural Language Processing

Midterm
October 8, 2015

Exam number:

Name:

Netid:

Instructions (**PLEASE PLEASE PLEASE READ****):**

- IF YOU HAVE ANY QUESTIONS ABOUT A QUESTION ON THE TEST, PLEASE JUST INDICATE THIS IN YOUR ANSWER. STATE ANY ASSUMPTIONS THAT YOU NEED TO MAKE IN ORDER TO ANSWER THE QUESTION AND THEN JUST ANSWER THE QUESTION AS BEST AS YOU CAN.
- GIVEN THE TIGHT SEATING FOR THE TEST (AND THE FACT THAT YOUR SCORE ON THE TEST DOES NOT MATTER), NO QUESTIONS WILL BE ENTERTAINED DURING THE EXAM.
- FOR THE SAME REASON, WE ASK THAT YOU DO NOT LEAVE THE CLASSROOM DURING THE FINAL 10 MINUTES OF THE EXAM...FROM 2:30PM ONWARD.

Exam number: _____

There are 5 questions on the test.

#	description	score		max score
1	some basics	____	/	20
2	lexical semantics	____	/	20
3	word sense disambiguation	____	/	15
4	language modeling	____	/	25
	Total score:	_____	/	80

1 Some Basics (20 points)

1. (5 pts) Stating whatever assumptions are necessary, indicate the number of **word tokens** and **word types** in the following text:

   It always rains on tents. Rainstorms will travel miles and miles to rain on tents.

2. (5 pts) (True or False. Explain your answer.) Trigrams are better than bigrams for language modeling tasks.

3. (10 pts) **Describe** and **Name** (e.g., lexical, syntactic, semantic, discourse, pragmatic) **three** different **ambiguities** that an NLP system would need to handle in order to understand the following sentence:

```
James called the man from Detroit.
```

2 Lexical Semantics (20 points)

1. (10 pts) Show the lexeme (or lexemes) for the word **walk**.

2. (2 pts) Give an example of **homonymy**?

3. (8 pts) Why might homonyms cause a problem for speech recognition systems?

3 Word Sense Disambiguation (15 points)

1. (5 pts) For the Senseval **all words** task, a word sense disambiguation (WSD) system must determine the correct sense of (most of) the content words in a document of substantial length. Describe one advantage of a dictionary-based approach (i.e. the Lesk algorithm) to WSD over a machine learning approach for this task.

2. (5 pts) Give one example of an **extrinsic** evaluation of a word sense disambiguation system. Explain why it is an extrinsic evaluation.

3. (5 pts) The word **ball** has (at least) two word senses — the basket**ball**, base**ball** sense and the party or dance sense (i.e., the Homecoming Ball). Provide 6 potentially useful **co-occurrence features** for a machine learning approach to word sense disambiguation.

4 Language Modeling (25 points)

```
 You do not like them . So you say .
Try them ! Try them ! And you may .
Try them and you may , I say .
```

Assume that the above text is provided as the (entire) training corpus for a **bigram** language model. For preprocessing, assume that **all words are converted to lower case**.

(a) (10 pts) Using Maximum Likelihood Estimation and the above training data, show **the fractions** that correspond to the bigram probability estimates for all bigrams starting with "you". (That is, just show the numerator and denominator of the fraction; i.e. you do **not** need to show the decimal equivalent.) Be sure to include the *unseen bigrams*.

(b) (5 pts) Do the very same thing as the previous question (just above), but use Laplacian (add-one) smoothing.

(c) (10 pts) Describe one method for handling **unknown words** when training (n-gram based) language models.