CS4740/CS5740
Introduction to Natural Language Processing

Midterm
March 13, 2014

Name:

Netid:

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

Exam number: _____

There are 5 questions on the test.

#	description	score		max score
1	some basics	____	/	25
2	lexical semantics	____	/	15
3	word sense disambiguation	____	/	15
4	language modeling	____	/	20
	Total score:	_____	/	75

1 Some Basics (25 points)

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

   Mary had a little lamb, little lamb, little lamb. Mary had a little lamb. Its fleece was white as snow.

2. (5 pts) (True or False. Explain your answer.) 4-grams are better than trigrams for language modeling tasks.

3. (5 pts) (True or False. Explain your answer.) Zipf's law ~~that~~ accounts for the fact that each corpus has a fairly small "working vocabulary".

4. (10 pts) Describe three (3) different **ambiguities** that an NLP system would need to handle in order to understand the following sentence:

   ```
   John saw the plane flying to New York.
   ```

2 Lexical Semantics (15 points)

Consider the highlighted word in each following two sentences:

(a) Next **fall**, Marseille and I will be on sabbatical.
(b) How did Marseille manage to **fall** off of the stage?

1. (5 pts) Based only on these examples, show what the lexeme or lexemes for "fall" might look like in a lexicon.

2. (5 pts) Give an example of polysemy.

3. (5 pts) (True or False. Explain your answer.) WordNet distinguishes polysemy from homonymy.

3 Word Sense Disambiguation (15 points)

Consider again the same two sentences:

(a) Next **fall**, Marseille and I will be on sabbatical.
(b) How did Marseille manage to **fall** off of the stage?

1. (5 pts) Assuming a supervised learning approach to word sense disambiguation for the word "fall", show what the **collocation** features might be for sentence (b). (State any assumptions that you are relying on.)

2. (5 pts) Assuming a supervised learning approach to word sense disambiguation for the word "fall", show five (5) potentially useful co-occurrence features. (State any assumptions that you are relying on.)

3. (5 pts) Is Lesk's word sense disambiguation algorithm a supervised machine learning algorithm, an unsupervised learning algorithm, a bootstrapping approach, or none of the above? Explain your answer.

4 Language Modeling (20 points)

```
Mary had a little lamb , little lamb ,
little lamb .  Mary had a little lamb .
```

Assume that the above text is provided as the (entire) training corpus for a **bigram** language model. (I know; this is not enough training.) Further, assume that no additional preprocessing is applied other than the tokenization provided.

1. (5 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 "Mary" and with "lamb". (That is, just show the numerator and denominator of the fraction; i.e. you do **not** need to show the decimal equivalent.) `Include unseen bigrams.`

2. (5 pts) Do the very same thing as the previous question (just above), but use Laplacian (add-one) smoothing.

Note: Questions 3 and 4 below do **not** refer to the "Mary had a little lamb" example.

3. (5 pts) Assuming a **trigram** language model, show the equation for computing the probability of a sequence of n words, $P(w_1\ w_2\ w_3\ \ldots\ w_n)$.

4. (5 pts) Assuming a **bigram** language model, show the equation for computing the probability of the next word, w_n, in a sequence of $n-1$ words, $P(w_n \mid w_1\ \ldots\ w_{n-1})$.