CSCI381/CSCI780

Name:_____

Final exam

Please write clearly. If the response is not clearly written, it will not get full credit. Please put away your cell phones, calculators, textbooks, and notebooks. Only a pen (or pencil) is allowed.

December 18, 2017

Question 1: POS tagging: 12 points

• **6 points** Tag the sentence below with parts-of-speech.

Karlsson lives on the roof of a perfectly ordinary house

Assume the following tags from the Penn Treebank: {**NNP, NN, NNS, VB, VBD, VBG, VBZ, RB, IN, DT, JJ**}

• **6 points** Identify **one** word in the above sentence that is ambiguous with respect to part-of-speech. Provide an example sentence that shows its ambiguity and explain why it is ambiguous.

Question 2: Syntactic parsing: 16 points

• **8 points** Give an example sentence with **PP attachment ambiguity**. Draw syntactic trees that correspond to each interpretation.

• **8 points** Give an example sentence that has **coordinating conjunction ambiguity**. Draw syntactic trees that correspond to each interpretation.

Question 3:Probabilistic context-free grammars: 6 points

• **6 points** Given the two sentences below assume the most likely parse trees are marked with square brackets. Name one problem that a PCFG would have when finding the most likely parse trees (as indicated) corresponding to each of the sentences:

1. Workers dumped [NP sacks] [PP into a bin]
2. Fishermen caught [NP tons [PP of herring]]

Question 4: Lexicalized PCFGs: 6 points
Suppose we have a PCFG with 10 non-terminals: $|N| = 10$ and a vocabulary of size 20: $|\Sigma| = 20$.
• **6 points** How many lexicalized rules can we have s.t.:

- $X(h) \rightarrow_1 Y_1(h) Y_2(w)$
- $X(h) \rightarrow_2 Y_1(w) Y_2(h)$?

Question 5: Machine Translation and alignments: 14 points Suppose we have the following alignment from f to e for the sentence pair below: $a_1 = 1, a_2 = 2, a_3 = 0, a_4 = 3, a_5 = 4, a_6 = 5$.
• **6 points** What is the value of p(a|e,m) for this alignment under IBM Model 1? **Show all your work and explain.**
e="*The dog ate his homework*"
f="*Le chien a mangé son devoir*"

- **8 points** Which of the following alignments are valid under IBM Model 1? **Explain.**

1. $a_1 = 1, a_2 = 1, a_3 = 0, a_4 = 3, a_5 = 4, a_6 = 5$
2. $a_1 = 1, a_3 = 0, a_4 = 3, a_5 = 4, a_6 = 5$
3. $a_1 = 1, a_2 = 1, a_3 = 0, a_3 = 1, a_4 = 3, a_5 = 4, a_6 = 5$
4. $a_1 = 1, a_2 = 1, a_3 = 1, a_4 = 1, a_5 = 1, a_6 = 1$

Question 6: Machine Translation and IBM Model 1: 16 points
Consider the following sentence pair:
e: This is a cat
f: C'est un chat
Say we have an alignment $a1 = 2, a2 = 3, a3 = 4$
The parameters are:

$$t(c'est|this) = 0.5$$
$$t(un|this) = 0.5$$
$$t(chat|this) = 0$$
$$t(c'est|is) = 0.1$$
$$t(un|is) = 0.2$$
$$t(un|is) = 0$$
$$t(chat|is) = 0.7$$
$$t(c'est|a) = 0$$
$$t(un|a) = 0.5$$

$$t(chat|a) = 0.5$$

$$t(c'est|cat) = 0$$

$$t(un|cat) = 0$$

$$t(chat|cat) = 0.9$$

• **8 points** What is the value of p(f|a,e,m) for this example under IBM model 1? **Please show your work and provide the final calculation.**

• **8 points** What is the value of p(f,a|e,m) for this example under IBM model 1? **Please show your work and provide the final calculation.**

Question 7: Log-linear models: 12 points

• **5 points** What is the key advantage of a log-linear POS tagger over an HMM-based tagger?

• **7 points** Consider the label set Y consisting of POS tags, s.t. $|Y| = 20$, and the set X consisting of histories of the form $< t_1, t_2, \ldots, t_{i-1}, w_1, \ldots, w_n, i >$. Let the number of words in the vocabulary be 100. Say that our features are of the form (where N maps a word/tag pair to an integer:

$$f_{N(u,t)} = \begin{cases} 1, & \text{if current word } w_i = u \text{ and } y = t. \\ 0, & \text{otherwise.} \end{cases}$$

How many possible features are there in this model?

Question 8: Computing class probabilities in log-linear models: 12 points Consider the label set Y={cat,dog,hat,cot} with three features:

$$f_1(x, y) = \begin{cases} 1, & \text{if } x = \text{the and } y \text{ ends with at} \\ 0, & \text{otherwise} \end{cases}$$

$$f_2(x, y) = \begin{cases} 1, & \text{if } x = \text{the and } y \text{ starts with c} \\ 0, & \text{otherwise} \end{cases}$$

$$f_3(x, y) = \begin{cases} 1, & \text{if } x = \text{the and } y \text{ has second letter o} \\ 0, & \text{otherwise} \end{cases}$$

The weight vector v=<3,1,1>.

- **12 points** Compute the value of p(hat|the;v). **Show your work.**

Question 9: Brown clustering: 6 points

- **3 points** Explain what is Brown clustering.

- **3 points** Give an example of a task where it might be useful.