Practice Final Exam for Natural Language Processing

Name: ____

Instructions

There are 10 questions, each will be worth 10 points. The maximum score on the test will be 100. You will have approximately 1:50 minutes to complete this test. If you feel that your test is complete, you may hand in your test and leave early. It is essential that you **PUT YOUR NAME ON ALL TEST MATERIALS**. It can be difficult identify the author of an unsigned test and it would be better to avoid this problem.
The test materials will include this printout and one blank test booklet. I suggest that you fill in all answers directly on this printout and use the blank test booklet as scrap paper. However, if you run out of space, you have the option of using the test booklet. However, please include a clear note on the test so I know where to look for your answer.
This test is an open book/open notes test: Please feel free to bring your text book, your notes, copies of class lectures and other reading material to the test. A calculator is also permitted.

Answer all questions on the test. If you show your work and you make a simple arithmetic mistake, but it is clear you knew how to do it, you will get partial credit.
Note that this is a sample test. It is intended to approximate the content of the actual test. Of course questions on the actual test may end up being more or less difficult for you.

Questions

(1) Write a single regular expression for identifying social security numbers in text. The social security numbers consists of 9 digits and must be followed and preceded by spaces, beginning's of lines or ends of lines. For example, one should not find a valid social security number as a substring of a larger number. In addition there are certain restrictions:

- The first three digits cannot be all zeros and cannot be all sixes
- The fourth and fifth number cannot both be zeros
- The seventh, eighth and ninth number cannot all be zeros
- The nine digits can all be next to each other or there can be a delimiter between the third and fourth digit and the fifth and sixth digit. This delimiter can be a space or a hyphen.

The following are well formed social security numbers: 123456789, 123-45-6789, 123 456 789.
The following are ill-formed social security numbers: 000-23-4567, 123-00-4567, 123-45-0000, 666-12-7788.
There is no valid social security number on the following line:

12345678910 is a big number, 345-678-910 is a lotto number and 3333333334 is a 10 digit number.

Tag	Description	Tag	Description
CC	Coordinating conjunction	PRP$	Possessive pronoun
CD	Cardinal number	RB	Adverb
DT	Determiner	RBR	Adverb, comparative
EX	Existential there	RBS	Adverb, superlative
FW	Foreign word	RP	Particle
IN	Preposition or subordinating conjunction	SYM	Symbol
JJ	Adjective	TO	to
JJR	Adjective, comparative	UH	Interjection
JJS	Adjective, superlative	VB	Verb, base form
LS	List item marker	VBD	Verb, past tense
MD	Modal	VBG	Verb, gerund or present participle
NN	Noun, singular or mass	VBN	Verb, past participle
NNS	Noun, plural	VBP	Verb, non-3rd person singular present
NNP	Proper noun, singular	VBZ	Verb, 3rd person singular present
NNPS	Proper noun, plural	WDT	Wh-determiner
PDT	Predeterminer	WP	Wh-pronoun
POS	Possessive ending	WP$	Possessive wh-pronoun
PRP	Personal pronoun	WRB	Wh-adverb

Table 1: **Penn Treebank POS tags**

(2) Assign Penn parts of speech tags (as per Table 1) to all the words in the following two sentences using the notation word/POS:

a. *John and Mary bought a refrigerator with three doors .*

b. *It was purchased from a very small store near their house .*

(3) Mark the noun groups in the following sentence using BIO (beginning, intermediate, other) tags.

```
Mary
has
a
room
with
a
view
and
a
bottle
of
beer
```

(4) Draw a Phrase Structure Tree representing one parse of the following sentence. Make a list of the phrase structure rules that you are assuming.

John and Mary bought a refrigerator with three doors .

(5) Calculate precision, recall and f-score (aka, f-measure) for the following system output and answer key. In this task, politicians are automatically extracted from a collection of documents and classified with a label (*left*, *right* or *center*) or are left unclassified. A correct instance is one in which the system output and answer key agree on a label. If no label is assigned in either the answer key or the system output, it will not be listed below and will not be considered in the score.

Example	**System Output**	**Answer Key**
Barack Obama	center	left
Bill Clinton	center	left
Bugs Bunny	left	
George W. Bush	right	right
Hillary Clinton	center	left
Howard Stern		center
Jason Brown	center	
John Major	right	center
Jonathan Swift	left	
Karl Marx	left	
Mitt Romney	right	center
Noam Chomsky	left	left
Pippi Longstocking	center	
Ralph Nader	left	left
Richard Cheney	right	right
Rush Limbaugh		right
Sarah Palin	right	right
Shaquille O'Neal	center	

(6) Given the training data below, execute the following 3 steps, the first two create the HMM and the third uses this HMM for decoding: (a) calculate the likelihood probabilities for each word given each POS; (b) draw a finite state machine where states are POS and edges are labeled with transition probabilities; (c) draw a chart where the columns are positions in the sentence and the rows are names of states (start, end, POS tags) and fill in the probability scores assigned by the Viterbi algorithm assigning POS tags to the string *flying planes*.

Training Data:

- *buffalo/NNS flying/VBG is/VBZ dangerous/JJ*
- *flying/JJ planes/NNS are/VBZ numerous/JJ*
- *I/PRP saw/VBZ Mary/NNP flying/VBG planes/NNS*
- *He/PRP planes/VBZ shelves/NNS*

(7) Fill in the CKY chart below for sentence *The rain rains down* assuming the following rules:

1. **S → NP VP**
2. **NP → N**
3. **NP → DT N**
4. **VP → V ADVP**
5. **VP → V**
6. **ADVP → ADV**
7. **DT → the**
8. **N → rain**
9. **N → rains**
10. **V → rain**
11. **V → rains**
12. **ADV → down**

	The	*rain*	*rains*	*down*
	1	2	3	4
0				
1				
2				
3				

(8) Some defining characteristics of organization and facility as per the ACE guidelines are as follows:

- *An Organization entity must have some formally established association. Typical examples are businesses, government units, sports teams, and formally organized music groups. Industrial sectors and industries are also treated as Organization entities.* (ACE Entity Guidelines v6.6, page 7)
- *A facility is a functional, primarily man-made structure. These include buildings and similar facilities designed for human habitation, such as houses, factories, stadiums, office buildings, gymnasiums, prisons, museums, and space stations; objects of similar size designed for storage, such as barns, parking garages and airplane hangars; elements of transportation infrastructure, including streets, highways, airports, ports, train stations, bridges, and tunnels. Roughly speaking, facilities are artifacts falling under the domains of architecture and civil engineering.* (ACE Entity Guidelines v6.6, page 22)

In the following text from the May 3, 2012 New York Times (*A House Tour: Yes, That House*) mark the organizations by underlining them and writing an **ORG** immediately above them; mark the facilities by underlining them and writing **FAC** immediately above. If a particular piece of text is difficult to mark only **ORG** or only **FAC**, mark it **ORG/FAC**.

After the 9/11 attacks, the system changed radically. Now, anyone who wants to tour the White House must apply through the office of his or her representative in Congress, which forwards the names to the White House for clearance...

Once they get the green light, visitors show up at the appointed time on 15th Street between E and F Streets and join the line to enter through the southeast gate.

Anyone who has flown on an airline in recent years will recognize the familiar territory of identity checks and electronic scans, although here you do get to keep your shoes on. At the head of the line, rangers from the National Park Service check photo IDs against a list of names.

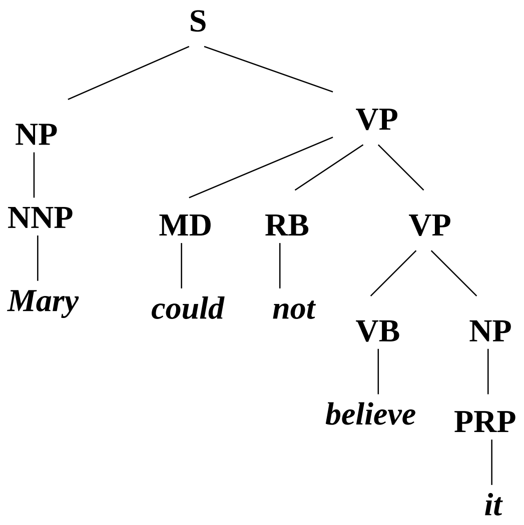

Figure 1: The 1st sentence: *Mary could not believe it*

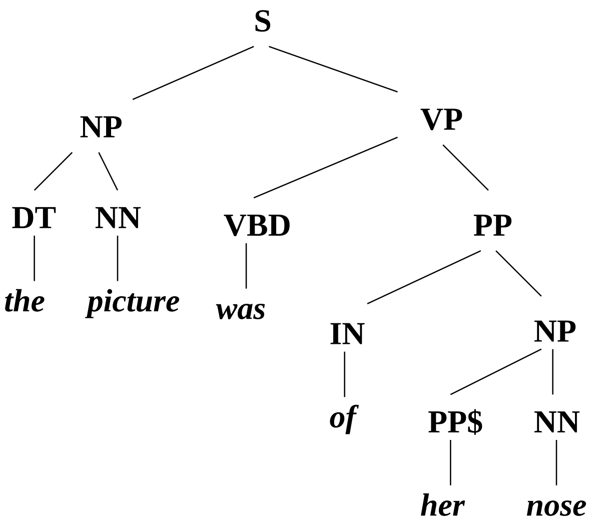

Figure 2: The 2nd sentences: *The picture was of her nose*

(9) Figures 1 and 2 provide the parse trees for two consecutive sentences. Annotate the trees with arrows indicating the search for the antecedent of the possessive pronoun *her* using the Hobbs search algorithm, diagrammed as figure 3. Put an X through each NP that was considered as an antecedent, but was rejected (due to agreement, semantics, etc.). Circle the antecedent found by the Hobbs Search. Next to each arrow, indicate the number of the step associated with this move (as per figure 3).

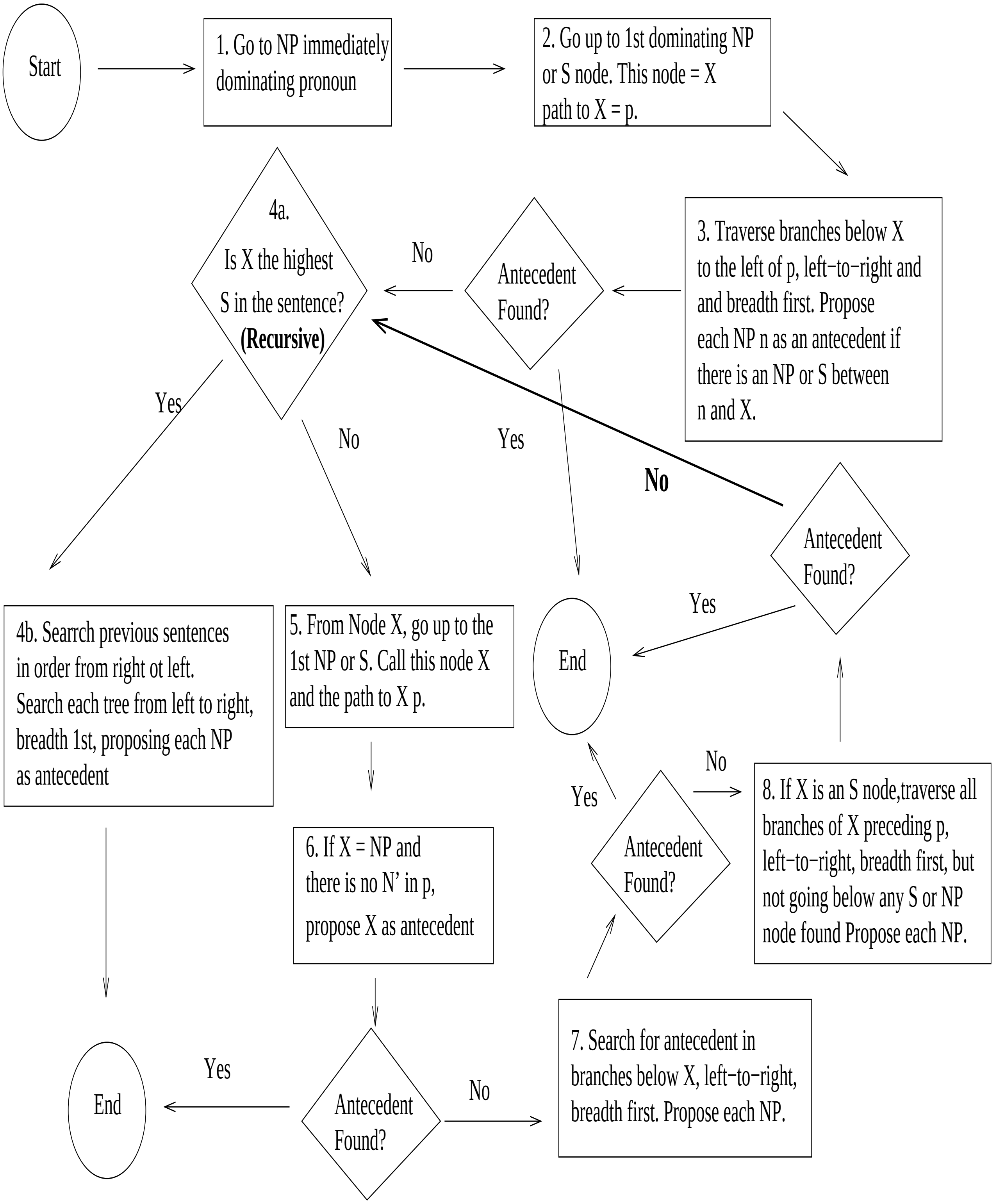
Start
1. Go to NP immediately dominating pronoun
2. Go up to 1st dominating NP or S node. This node = X path to X = p.
3. Traverse branches below X to the left of p, left−to−right and and breadth first. Propose each NP n as an antecedent if there is an NP or S between n and X.
Antecedent Found?
No
4a. Is X the highest S in the sentence? (Recursive)
Yes
No
Yes
No
Antecedent Found?
Yes
4b. Searrch previous sentences in order from right ot left. Search each tree from left to right, breadth 1st, proposing each NP as antecedent
5. From Node X, go up to the 1st NP or S. Call this node X and the path to X p.
End
No
Yes
Antecedent Found?
8. If X is an S node,traverse all branches of X preceding p, left−to−right, breadth first, but not going below any S or NP node found Propose each NP.
6. If X = NP and there is no N' in p, propose X as antecedent
End
Yes
Antecedent Found?
No
7. Search for antecedent in branches below X, left−to−right, breadth first. Propose each NP.

Figure 3: Hobbs Search Algorithm for Finding Antecedents of Pronouns

(10) Given the following unigram and translation probabilities, what is the overall probability scores that a decoder would assign to *dog bad* being a translation of the Spanish *perro malo*?

What would need to be added to the model to get the correct word order?

Unigram Probabilities for English:

- *dog*: 1.7×10^{-5}
- *cat*: 8×10^{-6}
- *bird*: 5×10^{-6}
- *bad*: 1.6×10^{-4}
- *good*: 4.9×10^{-4}
- *stupid*: 1.1×10^{-5}

Translation probabilities

	Spanish					
English	*perro*	*gato*	*pjaro*	*malo*	*bueno*	*estupido*
dog	.6	.03	.01	.12	.15	.09
cat	.04	.55	.04	.15	.11	.11
bird	.01	.02	.8	.03	.02	.12
bad	.01	.03	.02	.87	.05	.02
good	.03	.01	0	.11	.83	.02
stupid	.05	.05	.05	.05	.75	.05