Final Exam for Natural Language Processing
May 2017

Name: ____

Instructions

There are 8 questions, each will be worth 12.5 points. The maximum score on the test will be 100. You will have approximately 1:50 minutes to complete this test (50% more time than you had for the midterm). 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. However, please do NOT: communicate with others (texts, email, etc.) or run actual programs to calculate answers.

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.

Question 1. Annotate the text in the chart on the next page with BIO tags for the following Named Entity classes. Only mark a token as **O** if it is not part of any of the listed types of named entities. If a token is part of a named entity use one the following tags: B-GPE, I-GPE, B-FAC, I-FAC, B-Date, I-Date, and B-Event, I-Event. If a token is part of two named entities, mark it with two tags (e.g., the first token of *New York Coliseum* would be marked B-GPE and B-FAC and the second token would be marked I-GPE and I-FAC. The third token would simply be marked I-FAC. Note that B and I tokens should only be marked for parts of names (capitalized, unique references, etc.), not for pronouns or common noun phrases. **Important:** Only mark names as named entities. Do not mark pronouns (*it*, *they*, *he*, *she*) and do not mark common noun phrases, e.g., *the party* is an event, but not a named entity for an event. The Named Entity Classes are:

- **Event** – a name or NP containing a name that refers to an activity or a set of activities that occur at a particular time. This includes holidays, contests, sporting events, birthday celebrations, etc.
- **GPE** – a location that has a government. This includes countries, cities, states, provinces, villages, etc.
- **Facility** – a man-made structure that can function as a location, such as a building, a street, a bridge, or a part thereof (a room, a closet).
- **Date** – an expression representing a time that contains one or more of the following: year, month, day of the month.

If you find any item difficult to mark. Include a short explanation of why it is difficult to mark. There are at least two types of difficulties that you may point out: 1) classification – it is difficult to figure out if this instance is a name (proper noun) or a common noun phrase; or 2) it may be unclear which category a phrase falls into.

The following world knowledge might be helpful to annotate this passage. The *Tour de France* is an internationally known bicycle race held in France. *Mont Saint-Michel* is a town in France. *Champs-Élysées* is an avenue in Paris.

Token	**BIO Tag1**	**BIO Tag2**	**Token**	**BIO Tag1**	**BIO Tag2**
The	B-Event		took	O	
2016	I-Event	B-Date	place	O	
Tour	I-Event		from	O	
de	I-Event		2	B-Date	
France	I-Event	B-GPE	July	I-Date	
was	O		to	O	
the	O		24	B-Date	
103rd	O		July	I-Date	
edition	O		2016	I-Date	
of	O	,		O	
the	O		starting	O	
race	O		in	O	
,	O		Mont	B-GPE	
one	O		Saint-Michel	I-GPE	
of	O		in	O	
cycling	O		Normandy	B-GPE	
’s	O		finishing	O	
Grand	B-Event		the	O	
Tours	I-Event		Champs-Élysées	B-FAC	
.	O		in	O	
The	O		Paris	B-GPE	
21-stage	O		.	O	
race	O				

Some Variation allowed for: 1) including *the* or not including *the* in *The 2016 Tour de France* and *the Élysées*; and 2) Whether or not *Grand Tours* is considered a Named Entity at all.

Tag	Description	Tag	Description
CC	Coordinating conjunction	PU	punctuation
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
PRP$	Possessive pronoun		

Table 1: **Penn Treebank POS tags**

Question 2. Draw a phrase structure tree for *The 2016 Tour de France was the 103rd edition of the race, one of cycling's Grand Tours.*. You can assume that the period is a child of the root of the sentence. Commas are part of the smallest constituent that includes the items to the left and right of the comma. Also remember that two NPs can be linked together with a comma to form a new NP (NP → NP , NP). When the meaning of the expression is something like "NP1 which is an instance of NP2", this construction is called apposition, e.g., in "Mary Jones, Vice President of Sales", the NP *Mary Jones* combines with a comma and the NP *Vice President of Sales*. This means something like "Mary Jones, who is the Vice President of Sales". Use the standard Penn parts of speech as in the table above for terminal nodes. Use PU for any instance of punctuation.

Either labeled bracketing or tree

```
(S  (NP  (DT The)
         (CD 2016)
         (NNP Tour)
         (NNP de)
         (NNP France))
    (VP  (VBD was)
         (NP  (NP  (DT the)
                   (JJ 103rd)
                   (NN edition)
                   (PP   (IN of)
                         (NP  (DT the)
                              (NN race))))
              (PU ,)
              (NP  (CD one)
                   (PP  (IN of)
                        (NP  (POSSP  (NP  (NN cycling))
                                     (POS ’s)
                             )
                             (NNP  Grand)
                             (NNPS  Tours))))))
    (PU .))
```

Variation: 1 Are the words in “Tour de France” FW or NNP?; 2) For PP modification of nouns, do we use a flat structure (as above) or do we add an Noun-Group level, i.e., NP → NP PP?; 3) Should Grand Tours be a constituent, i.e., NP → PossP NP?

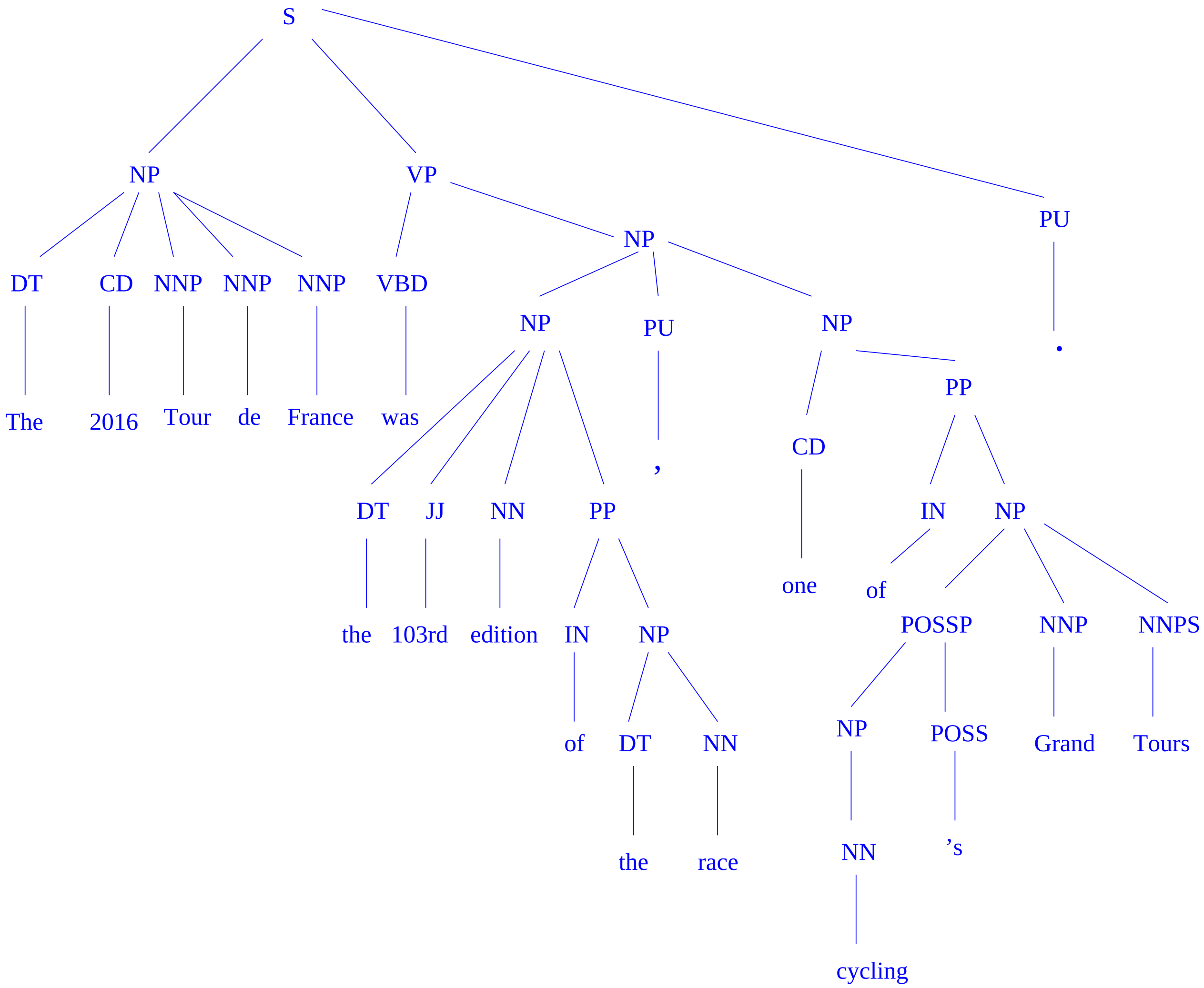

Figure 1: Tree for Tour de France Sentence

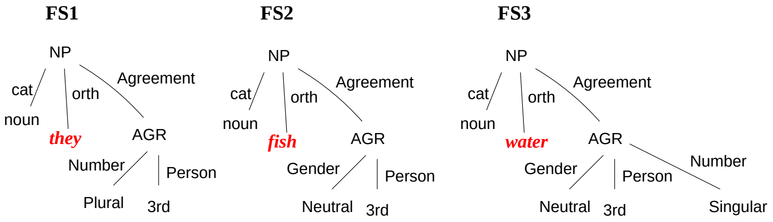

Figure 2: Feature Structure Representing Agrement Properties for Coreference

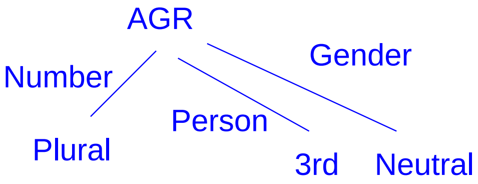

Figure 3: Answer to 3a

Question 3a. Draw the the feature structure resulting from unifying the two Agreement values of the feature structures **FS1** and **FS2** in figure 2. The root should be AGR (you should not draw any other parts of these graphs). The feature structure **FS1** represents the pronoun *they* and the feature structure **FS2** represents the noun phrase *fish*. This unfication is one possible implementation of an agreement test to see if the pronoun *they* can be coreferential with *fish*, which for example, could be used in conjunction with a Hobbs search. Answer is in Figure 3.

Question 3b. Briefly explain why this method would predict that **water** (as in **FS3**) cannot be an antecedent of *they* (as in **FS1**). **Water has a number of singular and that conflicts with the plural number of *they*.**

(4) Calculate B-cubed precision, recall and F-score given the coreference entities in the system output and answer key. For this example, both the answer key and the system has found exactly two entities. **Note: Only consider the NPs that are part of the answer key and system output. Other NPs in the passage are irrelevant**. A with subscripts indicate one answer key entity and B with subscripts indicate another answer key entity. The actual context is provided below with the entities marked in square brackets.

Answer:

```
Precision = ((6/7 * 6) + 1/7 + (4/4 * 4))/11 = 84.44% (or 65/77)
Recall = ((6/7 * 6) + (4/6 * 4)+ 1/6)/13 = 61.36% (or 335/536)
F-measure = 2/(1/.8444 + 1/.6136) = 71.1%
```

Answer Key:

1. A_1 =**Belgium**, A_2 =**Belgium**, A_3 =**their**, A_4 =**Belgium**, A_5 =**the home team**, A_6 =**They**, A_7 =**The Belgians**
2. B_1 =**Great Britain**, B_2 =**Great Britai**n, B_3 =**Great Britain**, B_4 =**their**, B_5 =**Great Britain**, B_6 =**who**

System Output:

1. A_1 =**Belgium**, A_2 =**Belgium**, A_3 =**their**, B_4 =**their**, A_4 =**Belgium**, A_6 =**They**, A_7 =**The Belgians**
2. B_1 =**Great Britain**, B_2 =**Great Britain**, B_3 =**Great Britain**, B_5 =**Great Britain**

The above entities were taken from the following passage from the wikipedia page: `https://en.wikipedia.org/wiki/2015_Davis_Cup`. The NPs that are part of the entities are in bold.

> **Belgium**[A_1] and **Great Britain**[B_1] won through to the final by winning semi-final ties against Argentina and Australia respectively. This meant that **Belgium**[A_2] would participate in **their**[A_3] first Davis Cup final since 1904 (a 50 defeat against **Great Britain**[B_2]), and **Great Britain**[B_3] in **their**[B_4] first since 1978. It also marked a remarkable recovery in fortunes for **Great Britain**[B_5], **who**[B_6] had been in danger of relegation to the lowest division of the Davis Cup in 2010.
>
> **Belgium**[A_4] were drawn as **the home team**[A_5] under the rotation policy used by the organizers. **They**[A_6] chose to play the tie on a clay surface in the Flanders Expo, an indoor arena in Ghent. **The Belgians**[A_7] opted for a clay surface . . .

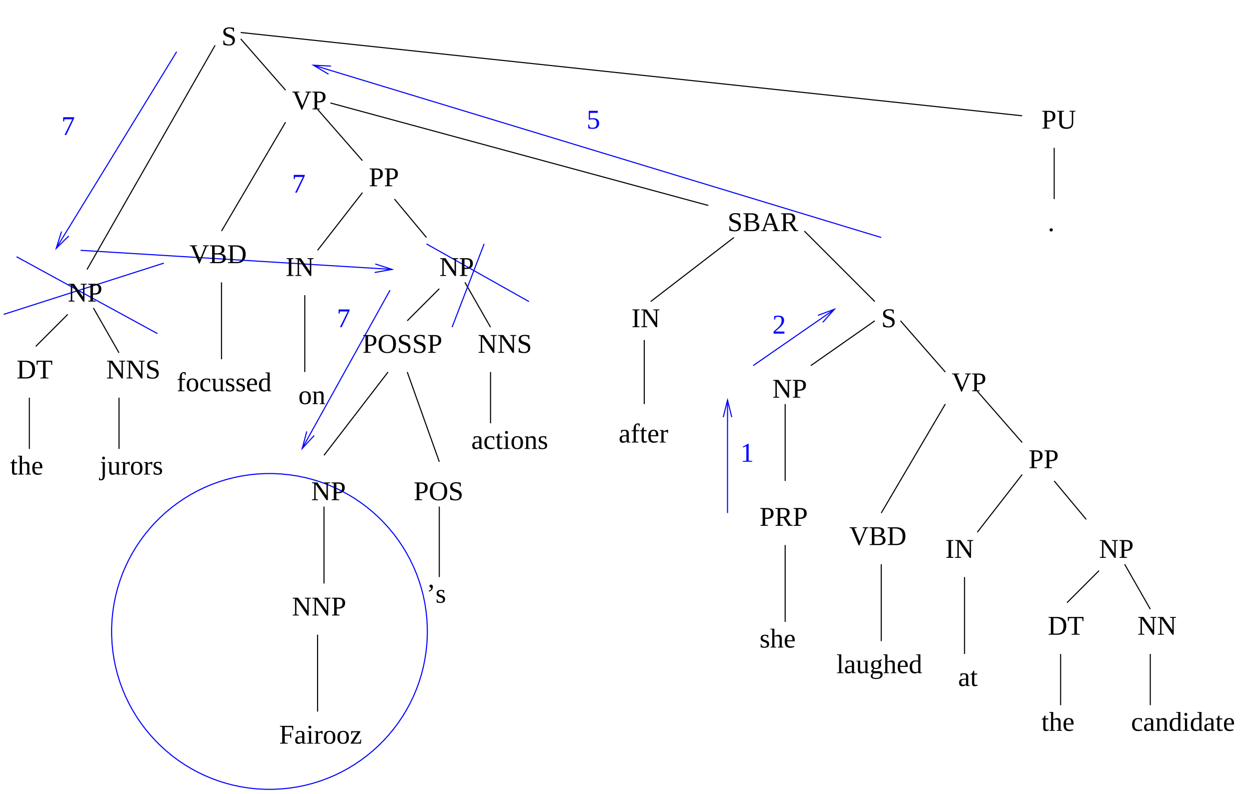

Figure 4: **The jurors were focused on Fairooz's actions after she laughed at the candidate.**

Question 5. Figure 4 is a parse tree for the sentence, **The jurors were focused on Fairooz's actions after she laughed at the candidate.** Annotate the tree with arrows indicating the search for the antecedent of the pronoun **she** using the Hobbs search algorithm, diagrammed as Figure 5. 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 5). If multiple steps apply in a row (e.g, step 4a and 5) and the search goes to a node during the last step, you should only list the last step in the sequence. Also, remember that step 6, circled in a dashed red line, should not be used.

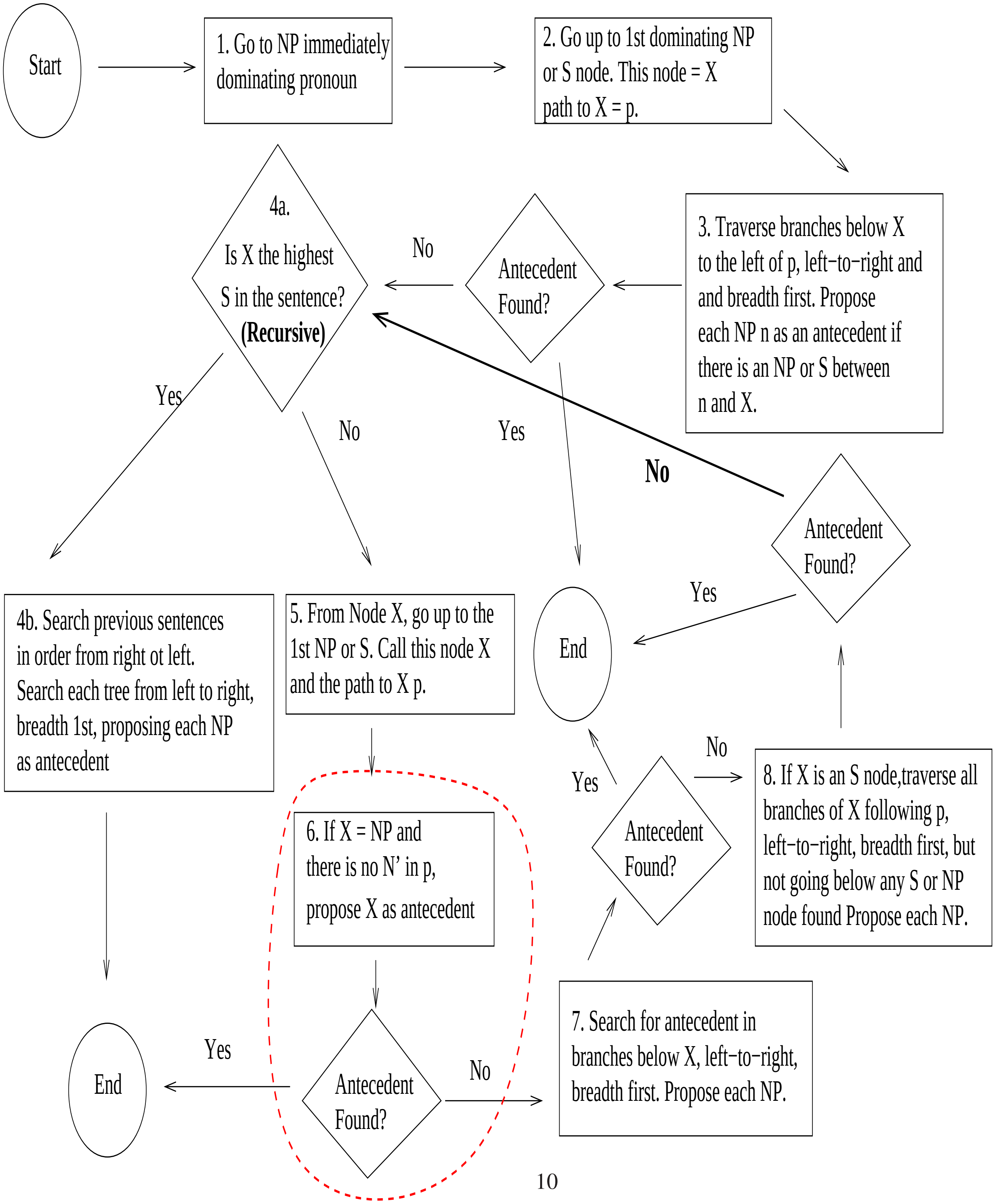
Start
1. Go to NP immediately dominating pronoun
2. Go up to 1st dominating NP or S node. This node = X path to X = p.
3. Traverse branches below X to the left of p, left−to−right and and breadth first. Propose each NP n as an antecedent if there is an NP or S between n and X.
Antecedent Found?
No
4a.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

Figure 5: Hobbs Search Algorithm for Finding Antecedents of Pronouns

(6) Given the likelihood probabilities and the Markov chain in Figure 6, fill in the chart for an HMM assigning POS tags to the string **hope floats**. Feel free to round off to one or two decimal places (as appropriate).

Likelihood Probabilities

- **hope**: NN = 7.6×10^{-5}, VB = 4.1×10^{-4}, VBP = 3.3×10^{-3}, NNP = 1.4×10^{-5}
- **floats**: NNS = 1.4×10^{-5}, VBZ = 1.2×10^{-4}

	0	**1: hope**	**2: floats**	**3**
Q_0: START				
Q_1: NNP		$.39 \times 1.4 \times 10^{-5} = 5.46 \times 10^{-6}$	0	
Q_2: NN		$.19 \times 7.6 \times 10^{-5} = 1.44 \times 10^{-5}$	0	
Q_3: NNS		0	(from NNP) ~~0~~ (from NN) $1.44 \times 10^{-5} \times .35 \times 1.4 \times 10^{-4} = 7.08 \times 10{-}11$	
Q_4: VB		0	0	
Q_5: VBZ		0	(from NNP) ~~$5.46 \times 10^{-6} \times .51 \times 1.2 \times 10^{-4} = 3.34 \times 10^{-10}$~~ (from NN) $1.44 \times 10^{-5} \times .31 \times 1.2 \times 10^{-4} = 5.36 \times 10^{-10}$	
Q_6: VBP		0	0	
Q_F: END		0	0	(from NNS) ~~$7.08 \times 10{-}11 \times .44 = 3.11 \times 10{-}11$~~ (from VBZ)$5.36 \times 10^{-10} \times .52 = 2.79 \times 10^{-10}$

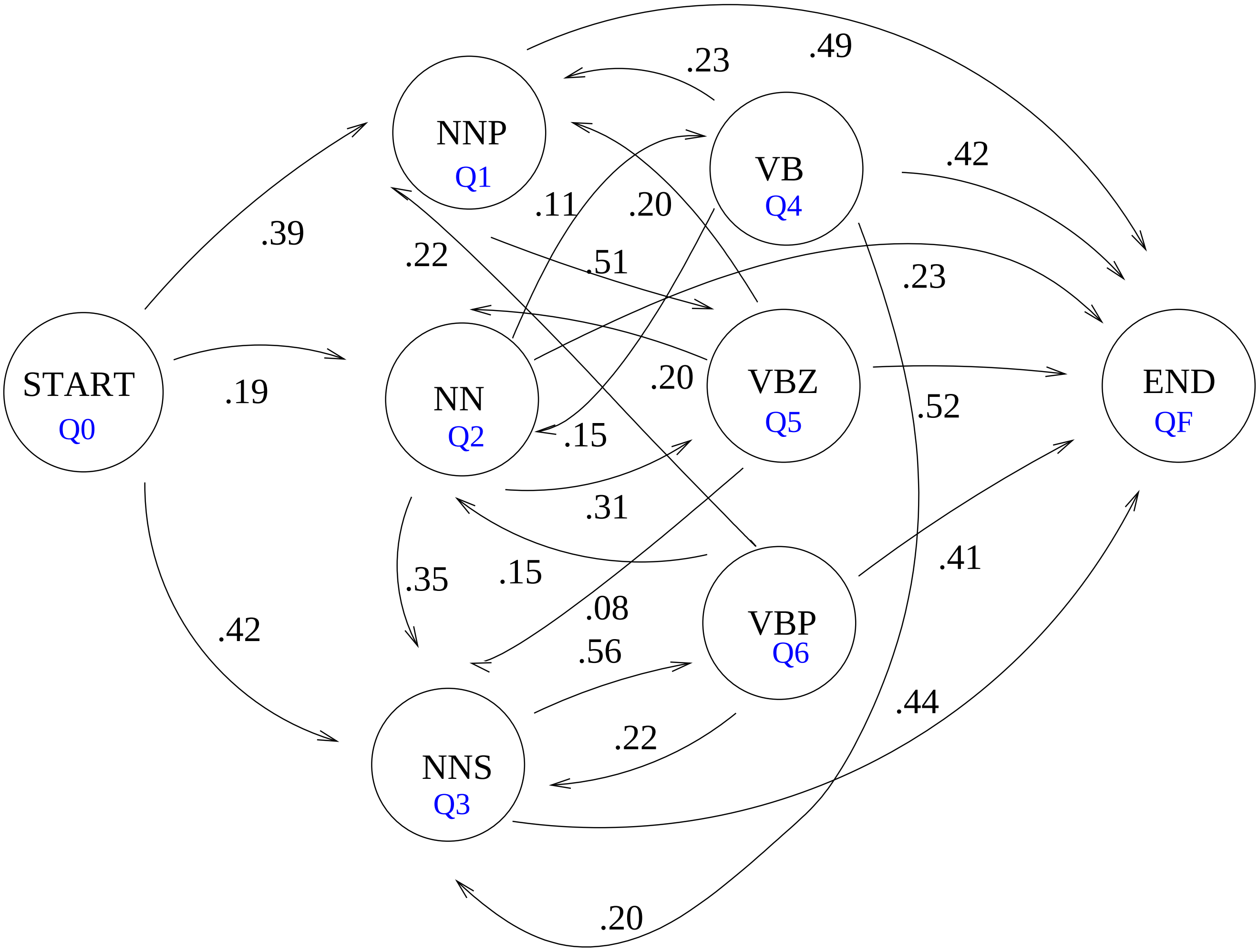

Figure 6: POS Transitions

Question 7. This question has parts a and b. Part b continues onto the next page. This question assumes the expectation maximization algorithm for training an a one to one word alignment model (IBM model 1) like the one discussed in class. In the first iteration of the procedure. we assume that each source language (Spanish) word is equally likely to translate with any target language word (English). Thus if there are only 3 words in each language in our bitext, we intially assume that a given source language word has a 1/3 chance of translating as each of the target language words. This yields the following probabilities of the translation alignments.

- **la pelota → the ball**
  - **la/the pelota/ball** = $1/3 \times 1/3 = 1/9 = .11$
  - **la/ball pelota/the** = $1/3 \times 1/3 = 1/9 = .11$
- **la naranja → the orange**
  - **la/the naranja/orange** = $1/3 \times 1/3 = 1/9 = .11$
  - **la/orange the/naranja** = $1/3 \times 1/3 = 1/9 = .11$
- **la pelota naranja → the orange ball**
  - **la/the pelota/orange naranja/ball** = $1/3 \times 1/3 \times 1/3 = 1/27 = .037$
  - **la/the pelota/ball naranja/orange** = $1/3 \times 1/3 \times 1/3 = 1/27 = .037$
  - **la/orange peolota/ball naranja/the** = $1/3 \times 1/3 \times 1/3 = 1/27 = .037$
  - **la/orange peolota/the naranja/ball** = $1/3 \times 1/3 \times 1/3 = 1/27 = .037$
  - **la/ball pelota/orange naranja/the** = $1/3 \times 1/3 \times 1/3 = 1/27 = .037$
  - **la/ball pelota/the naranja/orange** = $1/3 \times 1/3 \times 1/3 = 1/27 = .037$

Question 7a. Compute the probabilities for the second stage for each of these alignments. Fill this table with the new probabilities for each source target word pair:

		Source Words		
		la	**pelota**	**naranja**
	the	$(.5 + .5 + .33)/3 = .44$	$(.5 + 0 + .33)/3 = .28$	$(0 + .5 + .33)/3 = .28$
Target Words	**orange**	$(.5 + .33)/2 = .42$	$.33/2 = .17$	$(.5 + .33)/2 = .42$
	ball	$(.5 + .33)/2 = .42$	$(.5 + .33)/2 = .42$	$.33/2 = .17$

Question 7b. Use the probabilities from the table you created to recalculate the probability of each possible alignment. Fill in the probability for each alignment after the equals (=) signs below.

- **la pelota → the ball**
  - **la/the pelota/ball =** $.44 \times .42 = .123$
  - **la/ball pelota/the =** $.42 \times .28 = .0784$
- **la naranja → the orange**
  - **la/the naranja/orange =** $.44 \times .42 = .123$
  - **la/orange naranja/the =** $.42 \times .28 = .0784$
- **la pelota naranja → the orange ball**
  - **la/the pelota/orange naranja/ball =** $.44 \times .17 \times .17 = .013$
  - **la/the pelota/ball naranja/orange =** $.44 \times .42 \times .42 = .078$
  - **la/orange pelota/ball naranja/the =** $.42 \times .42 \times .28 = .022$
  - **la/orange pelota/the naranja/ball =** $.42 \times .28 \times .17 = .052$
  - **la/ball pelota/orange naranja/the =** $.42 \times .17 \times .28 = .020$
  - **la/ball pelota/the naranja/orange =** $.42 \times .28 \times .42 = .049$

Note that in the third and subsequent stages, the relative probabilities of alignments will weight further alignments scores, causing the probability of more probable alignment pairings to increase.

Question 8. Fill in the CKY chart below for sentence **the advanced biology textbook** assuming the phrase structure rules below.

Phrase Structure Rules for CKY

1. **NP → DT NBAR**
2. **NBAR → JJ NBAR**
3. **NBAR → NBAR NN**
4. **NBAR → NN**
5. **DT → the**
6. **JJ → advanced**
7. **NN → biology**
8. **NN → textbook**

CKY Chart to Fill in

	The	**advanced**	**biology**	**textbook**
	1	2	3	4
0	DT		NP	NP
1	XXXXXXXXXXX XXXXXXXXXXX XXXXXXXXXXX XXXXXXXXXXX XXXXXXXXXXX	JJ	NBAR	NBAR
2	XXXXXXXXXXX XXXXXXXXXXX XXXXXXXXXXX XXXXXXXXXXX XXXXXXXXXXX	XXXXXXXXXXX XXXXXXXXXXX XXXXXXXXXXX XXXXXXXXXXX XXXXXXXXXXX	NN, NBAR	NBAR
3	XXXXXXXXXXX XXXXXXXXXXX XXXXXXXXXXX XXXXXXXXXXX XXXXXXXXXXX	XXXXXXXXXXX XXXXXXXXXXX XXXXXXXXXXX XXXXXXXXXXX XXXXXXXXXXX	XXXXXXXXXXX XXXXXXXXXXX XXXXXXXXXXX XXXXXXXXXXX XXXXXXXXXXX	NN, NBAR