CS 6320.002: Natural Language Processing
Fall 2020
Midterm
19 Oct. 2020

Please clearly number each question, and circle/bold/italicize or otherwise emphasize your answers.

1

Suppose we have a trained first-order hidden Markov model for part-of-speech tagging with the parameters below (assume the value is 0 if the parameter is not shown in the tables).

Tag transition probabilities:

	DT	JJ	NN	VB
⟨S⟩	0.5	0.1	0.3	0.1
DT	0.0	0.5	0.5	0.0
JJ	0.0	0.4	0.6	0.0
NN	0.1	0.0	0.2	0.7
VB	0.4	0.2	0.2	0.2

Word emission probabilities:

	the	old	man	ships
DT	1.0	0.0	0.0	0.0
JJ	0.0	0.8	0.2	0.0
NN	0.0	0.2	0.4	0.4
VB	0.0	0.0	0.5	0.5

What is the predicted tag sequence for the sentence "the old man the ships," decoded using beam search Viterbi with beam size 3, and what is the predicted probability of that tag sequence? Show your work. You can leave the answer as a product of decimals.

2

Which of the following underlined phrases is a constituent? Choose all that apply.

A The cat walked across the porch with a confident air.
B They arrived at the concert more quickly than they expected.
C I am very fond of my nephew.

3

Suppose you want to build a system to perform graph-based dependency parsing, and you have access to a working word sense disambiguation (WSD) system. How could you use the information provided by the WSD system to try to improve the performance of your model? Which parameter(s) of the model would you modify, and how? (2-3 sentences.)

4

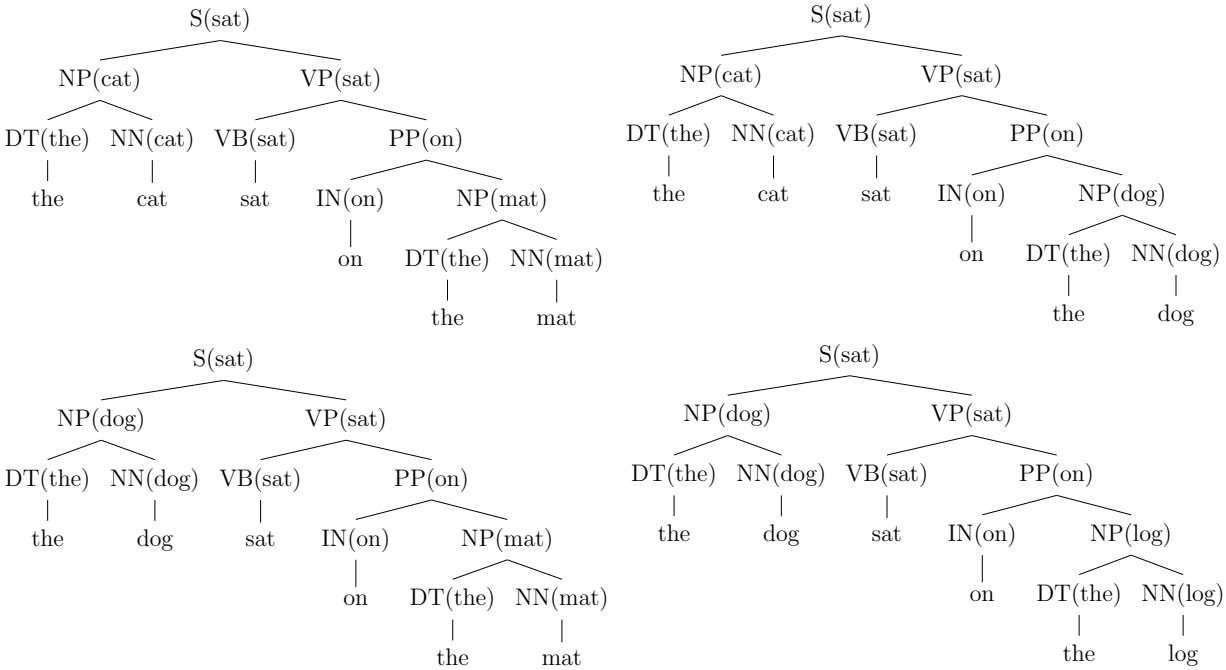
Suppose we train a language model on the following training sequences:

- the man saw the dog with the telescope
- the dog saw the man in the park
- the man with the telescope saw the park

We use linear interpolation between a bigram model and a unigram model, with $\lambda_1 = \lambda_2 = 0.5$. The bigram model is smoothed using Kneser-Ney smoothing with $\beta = 1$, and the unigram model is not smoothed. We do not use any start or end tokens. What is $p(\text{dog}|\text{the})$ under this model? Show your work. You can leave the answer as a sum of fractions.

5

Suppose we want to estimate a lexicalized PCFG using these four training trees:



What is the probability of the rule S(sat) $\rightarrow$ NP(dog) VP(sat), estimated using Charniak's method with uniform λs for smoothing? You can assume that the position of the head is fixed for each rule, so there is only one rule, not two different rules depending on head position. Show your work.

6

Suppose we have a trained first-order hidden Markov model for part-of-speech tagging with the parameters below (assume the value is 0 if the parameter is not shown in the tables).

Tag transition probabilities:

	DT	NN	VB	⟨/S⟩
⟨S⟩	0.5	0.3	0.2	0.0
DT	0.0	0.9	0.1	0.0
NN	0.1	0.2	0.3	0.4
VB	0.4	0.2	0.2	0.2

Word emission probabilities:

	the	can	see	⟨/s⟩
DT	1.0	0.0	0.0	0.0
NN	0.0	0.9	0.1	0.0
VB	0.0	0.5	0.5	0.0
⟨/S⟩	0.0	0.0	0.0	1.0

What is the predicted tag sequence for the sentence "can the can see ⟨/s⟩", decoded using beam search Viterbi with beam size 2, and what is the predicted probability of that tag sequence? Show your work. You can leave the answer as a product of decimals.

7

Suppose we have the following probabilistic context-free grammar:

NP	$\rightarrow$	DT NBAR	1.0
NBAR	$\rightarrow$	NN	0.4
NBAR	$\rightarrow$	JJ NBAR	0.3
NBAR	$\rightarrow$	NBAR NBAR	0.3
DT	$\rightarrow$	the	1.0
JJ	$\rightarrow$	blue	1.0
NN	$\rightarrow$	fountain	0.3
NN	$\rightarrow$	ink	0.3
NN	$\rightarrow$	pen	0.4

How many possible parses are there for the phrase "the blue fountain pen ink" under this grammar? Which has the highest probability, and what is that probability? Show your work.

8

Suppose you want to build a system to perform phrase-based machine translation, and you have access to a working word sense disambiguation (WSD) system. How could you use the information provided by the WSD system to try to improve the performance of your model? Which parameter(s) of the model would you modify, and how? (2-3 sentences.)

9

Suppose we have a training corpus of two sentences:

1.	grand	cheval	2.	grand
	big	horse		big

Now suppose we have a word-level statistical machine translation model with the parameters $q(j|i, l, m)$ and $t(f_i|e_j)$ initialized uniformly. Perform one iteration of expectation maximization to update the parameters using the training corpus. What is $t(\text{cheval}|\text{horse})$ at the end of this iteration? Show your work.

10

Suppose we have the following probabilistic context-free grammar:

S	$\rightarrow$	NP VP	1.0
VP	$\rightarrow$	VB VP	0.3
VP	$\rightarrow$	VB NP	0.7
NP	$\rightarrow$	DT NN	1.0
DT	$\rightarrow$	the	1.0
NN	$\rightarrow$	can	0.9
NN	$\rightarrow$	see	0.1
VB	$\rightarrow$	can	0.5
VB	$\rightarrow$	see	0.5

How many possible parses are there for the sentence "the can can see the can" under this grammar? Which has the highest probability, and what is that probability? Show your work.

11

Suppose we have the following context-free grammar:

S	$\rightarrow$	NP VP
VP	$\rightarrow$	VB NP
NP	$\rightarrow$	NP and NP
NP	$\rightarrow$	JJ NP
NP	$\rightarrow$	NN
JJ	$\rightarrow$	black
NN	$\rightarrow$	I
NN	$\rightarrow$	cats
NN	$\rightarrow$	dogs
VB	$\rightarrow$	like

What new rule(s) would you have to add to the grammar to be able to parse the sentences "I think I like black cats" and "I heard cats like dogs"? If you need new symbols, use X, Y, and Z.

12

Suppose we train a Naive Bayes classifier on the following training sequences (the letter after the comma is the class label):

- the cat sat on the mat, A
- the cat sat in the hat, A
- the dog sat on the log, B
- the dog sat on the cat, B
- the fish sat in the dish, C
- the fish in the hat sat, C

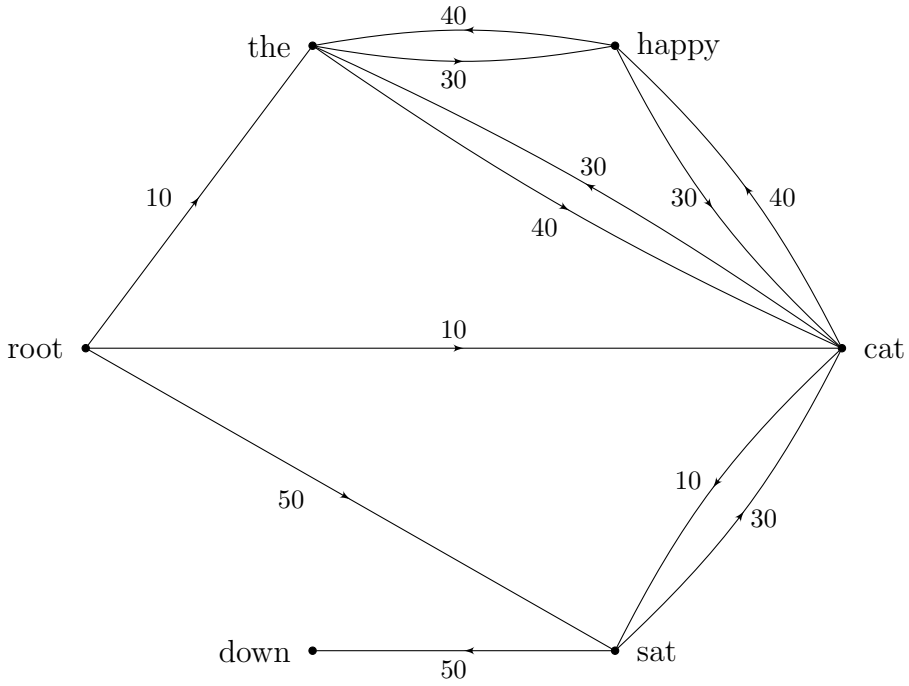
Our classifier uses skipgram count features, no start or end tokens, and Laplace smoothing with $\delta = 1$. Given the test sentence "the cat in the hat," what is the predicted probability of this sentence belonging to class A under this model? Show your work. You can leave the answer as a product of fractions.

13

"Projectivize" the dependency parse tree of the sentence "What is that dog chewing on over there?" using Nivre and Nilsson's method. What arc(s) would you need to lift, and what new arc(s) would you replace them with? You can give your answer using arrows (A $\rightarrow$ B) or ordered pairs (A, B).

14

Suppose we have the following graph model of candidate dependencies for the sentence "the happy cat sat down":



You can assume that any edges not shown have score 0. What is the highest-scoring dependency parse tree based on this graph (you can give your answer using arrows (A $\rightarrow$ B) or ordered pairs (A, B)), and what is its total score? Show your work.

15

Suppose we have a language model that uses linear interpolation to handle rare/out-of-vocabulary words. Is this strategy for handling rare/out-of-vocabulary words sufficient, or do we also need to use unknown word token? Briefly explain your answer (1-2 sentences).

16

Suppose we train a hidden Markov model to do part-of-speech tagging using the following training sequences:

- the/DT cat/NN is/VB in/PP the/DT box/NN
- the/DT cat/NN sat/VB on/PP the/DT mat/NN
- the/DT black/JJ cat/NN sat/VB

We use a start tag $\langle \mathrm{S} \rangle$, but no end tag, bigram tag transitions, and we smooth the transition probabilities using interpolation with $\lambda_1 = \lambda_2 = \frac{1}{2}$. What is the predicted tag sequence for the sentence "the black box cat" under this model, and what is its probability? Show your work. You can leave your answer as a product of fractions.

17

Which of the following underlined phrases is a constituent? Choose all that apply.

A These black cats detest those green peas.

B These black cats detest those green peas.

C Put it over on the table.

D Put it over on the table.

18

Suppose we have a training corpus of two sentences:

1.	souris	blanche	2.	souris	
	white	mouse		mouse	

Additionally, suppose we already have a trained IBM Model 1 statistical machine translation system, and now we want to train an IBM Model 2. The parameters $q(j|i,l,m)$ are initialized uniformly, and the parameters $t(f_i|e_j)$ are initialized using the Model 1 as follows:

- $t(\text{souris}|\text{white}) = \frac{1}{2}$
- $t(\text{souris}|\text{mouse}) = \frac{3}{4}$
- $t(\text{blanche}|\text{white}) = \frac{1}{2}$
- $t(\text{blanche}|\text{mouse}) = \frac{1}{4}$

Perform one iteration of expectation maximization to update the parameters using the training corpus. What is $t(\text{souris}|\text{white})$ at the end of this iteration? Show your work.

19

Suppose we have a dataset of restaurant reviews, and we want to perform binary sentiment classification. However, the restaurant corpus does not come with gold standard labels. We have another dataset consisting of movie reivews that does come with gold standard labels. How could you use the movie corpus, along with bootstrapping, to develop a system to perform sentiment classification on the restaurant corpus? Describe how you would design and train the system (3-4 senctences).

20

Suppose we have two unigram probability distributions:

$$p(\text{cat}) = \tfrac{1}{2} \quad p(\text{dog}) = \tfrac{1}{4}$$
$$p(\text{mat}) = \tfrac{1}{8} \quad p(\text{log}) = \tfrac{1}{8}$$

and

$$q(\text{cat}) = \tfrac{1}{8} \quad q(\text{dog}) = \tfrac{1}{4}$$
$$q(\text{mat}) = \tfrac{1}{2} \quad q(\text{log}) = \tfrac{1}{8}$$

What is the cross-entropy $H(p, q)$? Show your work.