CS74040

Name:_____

Midterm exam

November 4, 2021

Question 1: We have a vocabulary V of size 300. How many parameters will you have to estimate for:

1. an MLE unigram language model
2. a bigram language model with add-one smoothing
3. a maximum likelihood trigram language model

Question 2: Consider the following training corpus T of sentences:

1. <s> I am Sam </s>
2. <s> Sam I am </s>
3. <s> I do not like green eggs and ham </s>

Compute the following maximum likelihood parameters:

- $p(I| < s >) =$
- $p(< /s > |Sam) =$
- $p(I|do) =$
- $(p(ham|eggs, and) =$

Question 3: We have the following training corpus:

1. the green book STOP
2. my blue book STOP
3. his green house STOP
4. book STOP

Assume we have a trigram language model with linear interpolation based on this corpus, with $\lambda_i = \frac{1}{3}$ for all i. Compute the value of the parameter $p(book|the, green)$ under this model. Assume STOP as part of your unigram model.

Question 4: Given a vocabulary V of size 1000 and a tagset T of size 30, suppose you wish to train a *bigram HMM tagger*. How many transition and emission parameters will your model have? Explain.

Question 5:
Let V={Karlsson,lives,happily} Let T={N,V}
We have a *trigram HMM tagger* that has the following non-zero parameters (all other parameters are zero):
Transition parameters:

- p(N|START)=0.5
- p(V|START)=0.5
- p(N|START,N)=1.0
- p(N|START,V)=0.9
- p(V|START,V)=0.1
- p(STOP|N,N)=1.0
- p(STOP|V,V)=1.0
- p(STOP|V,N)=1.0
- p(s|u,v)=0 for all other p parameters

Emission parameters:

- e(Karlsson|N)=0.8
- e(happily|N)=0.2
- e(lives|V)=0.7
- e|happily|V)=0.3

Under this model, how many pairs of sequences $x_1 \dots x_n, y_1 \dots y_n$ have non-zero probability $p(x_1, \dots x_n, y_1 \dots y_{n+1}) \geq 0$? Show all of these possible sequences.

Question 6: Find one tagging error in each of the following sentences that are tagged with the Penn Treebank tagset:

1. I/PRP need/VBP a/DT flight/NN from/IN Atlanta/NN
2. Does/VBZ this/DT flight/NN serve/VB dinner/NNS
3. I/PRP have/VB a/DT friend/NN living/VBG in/IN Denver/NNP
4. Can/VBP you/PRP list/VB the/DT nonstop/JJ afternoon/NN flights/NNS

Question 7 :
We have a vocabulary V={Hello} and a constant $N \geq 1$. For any $x_1 \dots x_n$ such that $x_i \in V$ for $i = 1 \dots (n-1)$ and $x_n = STOP$, we define

$$p(x_1, \dots, x_n) = \begin{cases} \frac{1}{N}, & \text{if } n \leq N. \\ 0, & \text{otherwise.} \end{cases}$$

Is this a valid language model? Explain.

Question 8 :
Now we have a vocabulary V={Hello, Goodbye} and a constant $N \geq 1$. For any $x_1 \ldots x_n$ such that $x_i \in V$ for $i = 1 \ldots (n-1)$ and $x_n = STOP$, we define

$$p(x_1, \ldots, x_n) = \begin{cases} \frac{1}{2}, & \text{if n=2.} \\ 0, & \text{otherwise.} \end{cases}$$

Is this a valid language model? Explain.

Question 9: Consider the task of classifying the word *bass* using the Naïve Bayes algorithm. The features used are bag-of-word features. Assume the following likelihoods for each word being part of a "fish" and "music" class, and equal prior probabilities for each class.
What class will Naïve Bayes assign to the sentence "I eat fresh bass after music lesson"? Show your work.

	fish	**music**
I	0.09	0.16
eat	0.29	0.06
fresh	0.10	0.05
after	0.07	0.06
music	0.04	0.15
lesson	0.08	0.11