Feedback — Language Modeling and Spelling Correction

Help

You submitted this quiz on **Tue 20 Mar 2012 10:50 AM PDT**. You got a score of **5.00** out of **5.00**.

Question 1

We are given the following corpus, similar to the one in lecture but with "ham" replaced by "Sam" and "I am Sam" included twice:

- <s> I am Sam </s>
- <s> Sam I am </s>
- <s> I am Sam </s>
- <s> I do not like green eggs and Sam </s>

Using a bigram language model with add-one smoothing, what is P(Sam | am)? Include <s> and </s> in your counts just like any other token.

Your Answer		Score	Explanation
○ $\frac{2}{3}$			
○ $\frac{2}{14}$			
◉ $\frac{3}{14}$	✔	1.00	This is the correct answer.
○ $\frac{3}{28}$			
Total		1.00 / 1.00	

Question Explanation

Using a bigram language model with add-one smoothing, $P(Sam|am) = \frac{c(am,Sam)+1}{c(am)+V} = \frac{2+1}{3+11} = \frac{3}{14}$.

Question 2

Suppose we want to smooth the likelihood term of a noisy channel model of spelling. We are given two words, x and w, where x is the same as w, except the letter w_{i-1} in w has been miss typed as $w_{i-1}x_i$ in x. Specifically, we want to apply add-one smoothing to $P(x|w)$, the probability of typing $w_{i-1}x_i$ instead of w_{i-1}, where x_i and w_{i-1} are single letters. For insertions, $P(x|w) = \frac{ins[w_{i-1},x_i]}{c(w_{i-1})}$, where $ins[w_{i-1},x_i]$ is the number of times that x_i is inserted after w_{i-1} in the corpus, and $c(w_{i-1})$ is the number of times letter w_{i-1} appears in our corpus. Again, please note that here x_i and w_{i-1} are individual letters, not words.

What is the formula for $P(x|w)$ if we use add-one smoothing to the insertion edit model? Assume the only characters we use are lowercase a-z, that there are V word types in our corpus, and n total characters, not counting spaces.

Your Answer		Score	Explanation
$\frac{ins[w_{i-1},x_i]+1}{c(w_{i-1})+n}$			
$\frac{ins[w_{i-1},x_i]+1}{c(w_{i-1})+V}$			
$\frac{ins[w_{i-1},x_i]}{c(w_{i-1})}$			
$\frac{ins[w_{i-1},x_i]+1}{c(w_{i-1})+26}$	✔	1.00	This is the correct answer.
Total		1.00 / 1.00	

Question Explanation

The distribution $P(x|w)$ has 26 entries, one for each possible value of x_i. Thus, we add 26 total fictional counts to our data, which means we must add 26 to the denominator.

Question 3

We are given the following corpus, similar to the one in lecture but with "ham" replaced by "Sam" and "I am Sam" included twice:

- <s> I am Sam </s>
- <s> Sam I am </s>
- <s> I am Sam </s>
- <s> I do not like green eggs and Sam </s>

Using interpolated Kneser-Ney smoothing, what is $P_{KN}(Sam|am)$ if we use a discount factor of $d = 1$?

Here are some quantities of interest to make this less tedious:

- $c(am, Sam) = 2$
- $c(am) = 3$
- $c(Sam) = 4$
- $|\{w : c(am, w) > 0\}| = 2$
- $|\{(w_{j-1}, w_j) : c(w_{j-1}, w_j) > 0\}| = 14$
- $|\{w_{i-1} : c(w_{i-1}, \text{Sam}) > 0\}| = 3$

As a reminder, here is the formula for P_{KN}:
$P_{KN}(w_i|w_{i-1}) = \frac{\max(c(w_{i-1},w_i)-d,0)}{c(w_{i-1})} + \lambda(w_{i-1})P_{CONTINUATION}(w_i)$ where
$\lambda(w_{i-1}) = \frac{d}{c(w_{i-1})}|\{w : c(w_{i-1}, w) > 0\}|$ and $P_{CONTINUATION}(w_i) = \frac{|\{w_{i-1}:c(w_{i-1},w_i)>0\}|}{|\{(w_{j-1},w_j):c(w_{j-1},w_j)>0\}|}$

Your Answer		Score	Explanation
$\frac{2-1}{3} + \frac{2}{3} \cdot \frac{3}{14}$	✔	1.00	This is the correct answer.
$\frac{1-1}{3} + \frac{2}{3} \cdot \frac{3}{14}$			

$\frac{2-1}{3} + \frac{2}{3} \cdot \frac{4}{21}$		
$\frac{2-1}{3} + \frac{2}{3} \cdot \frac{3}{21}$		
Total	1.00 / 1.00	

Question Explanation

$$P_{KN}(Sam|am) = \frac{\max(c(am,Sam)-1,0)}{c(am)} + \frac{1}{c(am)}\left|\{w : c(am,w) > 0\}\right| \frac{|\{w_{i-1}:c(w_{i-1},Sam)>0\}|}{|\{(w_{j-1},w_j):c(w_{j-1},w_j)>0\}|} = \frac{2-1}{3} + \frac{2}{3} \cdot \frac{3}{14}$$

Question 4

We are given the following corpus, similar to the one in lecture but with "ham" replaced by "Sam" and "I am Sam" included twice:

- <s> I am Sam </s>
- <s> Sam I am </s>
- <s> I am Sam </s>
- <s> I do not like green eggs and Sam </s>

If we use linear interpolation smoothing between a maximum-likelihood bigram model and a maximum-likelihood unigram model with $\lambda_1 = \frac{1}{2}$ and $\lambda_2 = \frac{1}{2}$, what is $P(\text{Sam}|\text{am})$? Include <s> and </s> in your counts just like any other token.

Your Answer		Score	Explanation
$\frac{4}{25}$			
$\frac{1}{2} \cdot \frac{4}{17} + \frac{1}{2} \cdot \frac{2}{3}$			
$\frac{1}{2} \cdot \frac{4}{25} + \frac{1}{2} \cdot \frac{2}{3}$	✔	1.00	This is the correct answer.
$\frac{1}{2} \cdot \frac{4}{25} + \frac{1}{2} \cdot \frac{2}{2}$			
Total		1.00 / 1.00	

Question Explanation

$$\frac{1}{2}P(\text{Sam}) + \frac{1}{2}P(\text{Sam}|\text{am}) = \frac{1}{2}\frac{C(\text{Sam})}{\sum_{w \in V} C(w)} + \frac{1}{2}\frac{C(\text{am, Sam})}{C(\text{am})} = \frac{1}{2} \cdot \frac{4}{25} + \frac{1}{2} \cdot \frac{2}{3}.$$

Question 5

Suppose we train a bigram language model with add-one smoothing on a given corpus. The corpus contains V word types. What is $P(w_2|w_1)$, where w_2 is a word which follows w_1? We use the notation $c(w_1, w_2)$ to denote the number of times that bigram (w_1, w_2) occurs in the corpus, and $c(w_i)$ is the number of times word w_i occurs.

Your Answer	Score	Explanation

$\frac{c(w_1,w_2)+V}{c(w_1)+V^2}$			
$\frac{c(w_1,w_2)+1}{c(w_1)+V}$	✔	1.00	This is the correct answer.
$\frac{c(w_1,w_2)}{c(w_1)}$			
$\frac{c(w_1,w_2)+1}{c(w_1)+V^2}$			
Total		1.00 / 1.00	

Question Explanation

Although there are V^2 possible bigrams in the corpus, we are only interested in bigrams which start with w_1. Thus, we add one to each of the V values for $P(w_2|w_1)$, meaning we add V to the denominator.