Questions for Flipped Classroom Session of COMS 4705 Week 2, Fall 2014. (Michael Collins)

Question 1 In lecture we saw how to build trigram language models using *discounting methods*, and the *Katz back-off* definition. We're now going to build a *four-gram* language model based on these ideas. A four-gram language model gives estimates

$$q(w|t, u, v)$$

where t, u, v, w is any sequence of four words.

Assume we have a corpus, and that $c(t, u, v, w)$ is the number of times the four-gram t, u, v, w is seen in the data. Then take the following definitions:

$$\mathcal{A}(t, u, v) = \{w : c(t, u, v, w) > 0\}$$

and

$$\mathcal{B}(t, u, v) = \{w : c(t, u, v, w) = 0\}$$

Define $c^*(t, u, v, w)$ to be the discounted count for the four-gram (t, u, v, w), as follows:

$$c^*(t, u, v, w) = c(t, u, v, w) - 0.5$$

Assume that for any trigram u, v, w, $q_{BO}(w|u, v)$ is an estimate of the trigram probability, using the backed-off method described in lecture.

Finally, we define the four-gram model as

$$q_{BO}(w|t, u, v) = \begin{cases} \frac{c^*(t,u,v,w)}{c(t,u,v)} & \text{If } w \in \mathcal{A}(t, u, v) \\ \alpha(t, u, v) \times \frac{q_{BO}(w|u,v)}{\sum_{w \in \mathcal{B}(t,u,v)} q_{BO}(w|u,v)} & \text{If } w \in \mathcal{B}(t, u, v) \end{cases}$$

Question: How would you define

$$\alpha(t, u, v)$$

?

Question 2 Recall that the perplexity of a language model on a test corpus is defined as

$$2^{-l}$$

where

$$l = \frac{1}{M} \sum_{i=1}^{m} \log_2 p(x^{(i)})$$

and m is the number of sentences in the corpus, M is the total number of words in the corpus, $\log_2$ is $\log$ base 2, $x^{(i)}$ is the i'th sentence in the corpus, and $p(x^{(i)})$ is the probability of the i'th sentence in the corpus under the language model?

Question 2a: What is the *maximum* value that the perplexity can take?

Question 2b: What is the *minimum* value that the perplexity can take?

Question 2c: Assume that we have a bigram language model, where

$$p(w_1 \dots w_n) = \prod_{i=1}^{n} q(w_i | w_{i-1})$$

and $w_0 =$ *, and $w_n =$ STOP. We estimate the parameters as

$$q(w|v) = \frac{\text{Count}(v, w)}{\text{Count}(v)}$$

Write down a training corpus and a test corpus such that the perplexity of the model trained on the training corpus takes the maximum possible value on the test corpus.

Question 2d: Write down a training corpus and a test corpus such that the perplexity of the model trained on the training corpus takes the minimum possible value on the test corpus. (Assume that we use a bigram language model, as in 2(c).)

Question 3 We define a trigram language model as follows. Take Count(w), Count(v, w) and Count(u, v, w) to be unigram, bigram and trigram counts taken from a training corpus (here w is a single word, v, w is a bigram, and u, v, w is a trigram). Take N to be the total number of words seen in the corpus. Then the unigram, bigram and trigram maximum-likelihood estimates are

$$q_{ML}(w) = \frac{\text{Count}(w)}{N} \qquad q_{ML}(w|v) = \frac{\text{Count}(v, w)}{\text{Count}(v)}$$

$$q_{ML}(w|u, v) = \frac{\text{Count}(u, v, w)}{\text{Count}(u, v)}$$

The final estimate is then defined as

$$\begin{aligned} & q(w|u, v) \\ = \; & \alpha \times q_{ML}(w|u, v) + (1 - \alpha) \times (\beta \times q_{ML}(w|u) + (1 - \beta) \times q_{ML}(w)) \end{aligned}$$

where α and β are smoothing parameters, which satisfy the constraints $0 \leq \alpha \leq 1$ and $0 \leq \beta \leq 1$.

Question 3a: Assume that we define $\alpha = \beta = 0.5$. Show that the model is equivalent to a model of the form

$$q(w|u, v) = \lambda_1 \times q_{ML}(w|u, v) + \lambda_2 \times q_{ML}(w|u) + \lambda_3 \times q_{ML}(w)$$

and calculate the values for $\lambda_1, \lambda_2, \lambda_3$ under these settings for α and β.

Question 3b: Now assume that we define smoothing parameters $\alpha(u, v)$ for every bigram (u, v), and $\beta(u)$ for every unigram u. The new estimate is

$$\begin{aligned} q(w|u, v) \;\; = \;\; & \alpha(u, v) \times q_{ML}(w|u, v) \\ & +(1 - \alpha(u, v)) \times (\beta(u) \times q_{ML}(w|u) + (1 - \beta(u)) \times q_{ML}(w) \end{aligned}$$

Show that providing that $0 \leq \alpha(u, v) \leq 1$ for all (u, v), and $0 \leq \beta(u) \leq 1$ for all u, the estimate satisfies

$$\sum_w q(w|u, v) = 1$$

for all u, v. (For simplicity assume that for all u, v, Count(u, v) > 0, and for all u, Count(u) > 0.

Question 3c: Now say we define

$$\alpha(u, v) = \frac{\text{Count}(u, v)}{\text{Count}(u, v) + C_1} \qquad \beta(u) = \frac{\text{Count}(u)}{\text{Count}(u) + C_2}$$

where $C_1 > 0$ and $C_2 > 0$ are constants.

What is the intuition behind these definitions? What roles do the constants C_1 and C_2 play?

Question 3d: Now say we measure perplexity of the method from question 3c on a test corpus. We assume that for every unigram u seen in the test corpus, $\text{Count}(u) > 0$ where $\text{Count}(u)$ is again the number of times unigram u is seen in the training corpus. Show that the perplexity in this case cannot be infinite.

Question 4 Consider a Katz Bigram model, as defined in lecture. To recap, we define two sets

$$\begin{aligned} \mathcal{A}(w_{i-1}) &= \{w : \text{Count}(w_{i-1}, w) > 0\} \\ \mathcal{B}(w_{i-1}) &= \{w : \text{Count}(w_{i-1}, w) = 0\} \end{aligned}$$

The model is then defined as

$$q_{BO}(w_i \mid w_{i-1}) = \begin{cases} \frac{\text{Count}^*(w_{i-1}, w_i)}{\text{Count}(w_{i-1})} & \text{If } w_i \in \mathcal{A}(w_{i-1}) \\ \\ \alpha(w_{i-1}) \frac{q_{ML}(w_i)}{\sum_{w \in \mathcal{B}(w_{i-1})} q_{ML}(w)} & \text{If } w_i \in \mathcal{B}(w_{i-1}) \end{cases}$$

where

$$\alpha(w_{i-1}) = 1 - \sum_{w \in \mathcal{A}(w_{i-1})} \frac{\text{Count}^*(w_{i-1}, w)}{\text{Count}(w_{i-1})}$$

Which of the following statements is true?

- For all bigrams v, w we have $q_{BO}(w|v) \geq 0$.
- For all unigrams v we have $\sum_w q_{BO}(w|v) = 1$.