N-gram Language Model Exercises: Solution

Pham Quang Nhat Minh
Natural Language Processing

February 20, 2023

Problem 1. (20 points) Write out the equation for trigram probability estimation (modifying Eq. 3.11). Now write out all the non-zero trigram probabilities for the I am Sam corpus on page 4.

Trigram probability estimation:

$$P(w_n|w_{n-2}w_{n-1}) = \frac{C(w_{n-2}w_{n-1}w_n)}{C(w_{n-2}w_{n-1})} \tag{1}$$

We need to add special sentence start in the beginning of sentences in the corpus:

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

Applying the Equation 1, we calculate non-zero trigram probabilities as follows.

$$P(\text{<s> <s> I}) = \frac{C(\text{<s> <s> I})}{C(\text{<s> <s>})} = \frac{2}{3} = 0.67$$

We have non-zero trigram probabilities calculated from the corpus as follows.

```
P(<s> <s> I) = 2/3 = 0.67
P(<s> I am) = 0.5
P(I am Sam) = 0.5
P(am Sam </s>) = 1.0
P(<s> <s> Sam) = 1/3 = 0.33
P(<s> Sam I) = 1.0
P(Sam I am) = 1.0
P(I am </s>) = 1/2 = 0.5
P(<s> I do) = 0.5
P(I do not) = P(do not like) = P(not like green)
= P(like green eggs) = P(green eggs and)
= P(eggs and ham) = P(and ham </s>) = 1.0
```

Problem 2. (20 points) Calculate the probability of the sentence `i want chinese food`. Give two probabilities, one using Fig. 3.2 and the 'useful probabilities' just below it on page 6, and another using the add-1 smoothed table in Fig. 3.6. Assume the additional add-1 smoothed probabilities `P(i|<s>)=0.19` and `P(</s>|food)= 0.40`.

The unsmoothed probability of the sentence `i want chinese food` is calculated as follows.

```
P(<s> i want chinese food </s>)
    = P(i|<s>)P(want|i)P(chinese|want)P(food|chinese)P(</s>|food)
    = .25 x .33 x 0.0065 x 0.52 x 0.68
    = 0.00019
```

The add-1 smoothed probability of the sentence `i want chinese food` is calculated as follows.

```
P(<s> i want chinese food </s>)
    = P(i|<s>)P(want|i)P(chinese|want)P(food|chinese)P(</s>|food)
    = 0.19 x 0.21 x 0.0029 x 0.052 x 0.40
    = 0.0000024
```

Problem 3. (20 points) We are given the following corpus, modified from the one in the chapter:

```
<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.

We construct a vocabulary:
$V = \{$`<s>, I, am, Sam, do, not, like, green, eggs, and, </s>`$\}$
So there are $|V| = 11$ words in our vocabulary
Apply Equation 3.23 in page 14 of the chapter 3, we have the smoothed probability for `P(Sam|am)`?

$$P(\texttt{Sam}|\texttt{am}) = \frac{C(\texttt{am Sam}) + 1}{C(\texttt{am}) + |V|} = \frac{2+1}{3+11} = \frac{3}{14} = 0.214$$

Problem 4. (20 points) We are given the following corpus, modified from the one in the chapter:

```
<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 $\lambda_1 = \frac{1}{2}$ and2 and $\lambda_2 = \frac{1}{2}$, what is `P(Sam|am)`? Include `<s>` and `</s>` in your counts just like any other token.

Applying linear interpolation formula, we have:

$$\hat{P}(\texttt{Sam}|\texttt{am}) = \lambda_1 P_{MLE}(\texttt{Sam}|\texttt{am}) + \lambda_2 P_{MLE}(\texttt{Sam})$$

From the corpus we will calculate

```
P(Sam|am) = 2/3 = 0.67
P(Sam) = 4/25 = 0.16
```

Thus:

$$\hat{P}(\texttt{Sam}|\texttt{am}) = 0.5 \times 0.67 + 0.5 \times 0.16 = 0.415$$

Problem 5. (20 points)

You are given a training set of 100 numbers consists of 91 zeros and 1 each of the other digits 1-9. Now we see the following test set: 0 0 0 0 0 3 0 0 0 0. What is the unigram perplexity?

Given the training corpus we will have following unigram probabilities

```
P(0) = 91/100
P(1) = P(2) = P(3) = ... = P(9) = 1/100
```

We calculate the perplexity in two different ways.

Solution 1: Applying the Entropy formula (refer to the slide 45 of the Ngram language modeling). We use Python code so that you can understand better.

```
H = 1/10 * (9 * -math.log2(P(0)) + -math.log2(P(3)))
  = 0.786841013595898
P = 2**H = 1.7252925496828493
```

Solution 2: Applying the Equation 3.14 of the chapter 3.

$$PP(W) = P(w_1 w_2 ... w_N)^{-\frac{1}{N}}$$

The probability of the test set using the unigram language model is:

$$P(0\ 0\ 0\ 0\ 0\ 3\ 0\ 0\ 0\ 0) = P(0)^9 \times P(3) = (\frac{91}{100})^9 \times \frac{1}{100}$$

You can confirm that we can get the same perplexity as the result we obtained with the solution 1.