VIETNAM NATIONAL UNIVERSITY, HO CHI MINH CITY
HO CHI MINH CITY UNIVERSITY OF TECHNOLOGY
Faculty of Computer Science and Engineering

Lab 3 - NLP 242

Math Exercise

Supervisor: Mr. Bui Khanh Vinh
Student: Huynh Ngoc Van 2252898

Ho Chi Minh City, February 13, 2025

Contents

1 Problem 1

- **Step 1: Add 2 padding both ends to generate trigrams for each sentence**

  The `I am Sam` corpus becomes:

  `<s> <s> I am Sam </s> </s>`

  `<s> <s> Sam I am </s> </s>`

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

- **Step 2: Count** (w_{i-2}, w_{i-1}, w_i)

Trigram	Count
(<s>, <s>, I)	2
(<s>, I, am)	1
(I, am, Sam)	1
(am, Sam, </s>)	1
(Sam, </s>, </s>)	2
(<s>, <s>, Sam)	1
(<s>, Sam, I)	1
(Sam, I, am)	1
(I, am, </s>)	1
(am, </s>, </s>)	1
(<s>, I, do)	1
(I, do, not)	1
(do, not, like)	1
(not, like, green)	1
(like, green, eggs)	1
(green, eggs, and)	1
(eggs, and, Sam)	1
(and, Sam, </s>)	1

Table 1: *Count the frequency of trigram in corpus*

- **Step 3: Count** (w_{i-2}, w_{i-1})

Trigram context	Count
(<s>, <s>)	3
(<s>, I)	2
(I, am)	2
(am, Sam)	1
(Sam, </s>)	2
(<s>, Sam)	1
(Sam, I)	1
(am, </s>	1
(I, do)	1
(do, not)	1
(not, like)	1
(like, green)	1
(green, eggs)	1
(eggs, and)	1
(and, Sam)	1

Table 2: *Count the frequency of the context of trigram*

- **Step 4: Equation for trigram probability estimation**

$$P(w_i|w_{i-2}, w_{i-1}) = \frac{count(w_{i-2}, w_{i-1}, w_i)}{count(w_{i-2}, w_{i-1})}$$

- **Step 5: Calculate all non-zero probabilities**

$P(\texttt{I|<s>,<s>}) = \frac{2}{3} \approx 0.67$

$P(\texttt{am|<s>,I}) = \frac{1}{2} = 0.5$

$P(\texttt{Sam|I,am}) = \frac{1}{2} = 0.5$

$P(\texttt{</s>|am,Sam}) = \frac{1}{1} = 1$

$P(\texttt{</s>|Sam,</s>}) = \frac{2}{2} = 1$

$P(\texttt{Sam|<s>,<s>}) = \frac{1}{3} \approx 0.33$

$P(\texttt{I|<s>,Sam}) = \frac{1}{1} = 1$

$P(\texttt{am|Sam,I}) = \frac{1}{1} = 1$

$P(\texttt{</s>|I, am}) = \frac{1}{2} = 0.5$

$P(\texttt{</s>|am,</s>}) = \frac{1}{1} = 1$

$P(\texttt{do|<s>,I}) = \frac{1}{2} = 0.5$

$P(\texttt{not|I,do}) = \frac{1}{1} = 1$

$P(\texttt{like|do,not}) = \frac{1}{1} = 1$

$P(\texttt{grren|not,like}) = \frac{1}{1} = 1$

$P(\texttt{eggs|like,green}) = \frac{1}{1} = 1$

$P(\texttt{and|green,eggs}) = \frac{1}{1} = 1$

$P(\texttt{Sam|eggs,and}) = \frac{1}{1} = 1$

$P(\texttt{</s>|and,Sam}) = \frac{1}{1} = 1$

2 Problem 2

$$\texttt{<s> i want chinese food </s>}$$

- **Unsmoothed probabilities:**

  $P_{\text{unsmoothed}} = P(\texttt{I|<s>}) \times P(\texttt{want|I}) \times P(\texttt{chinese|want}) \times P(\texttt{food|chinese}) \times P(\texttt{</s>|food}) = 0.19 \times 0.33 \times 0.0065 \times 0.52 \times 0.4 = 8.47704 \times 10^{-5}$

- **Smoothed probabilities:**

  $P_{\text{smoothed}} = P(\texttt{I|<s>}) \times P(\texttt{want|I}) \times P(\texttt{chinese|want}) \times P(\texttt{food|chinese}) \times P(\texttt{</s>|food}) = 0.19 \times 0.21 \times 0.0029 \times 0.52 \times 0.4 = 2.406768 \times 10^{-5}$

3 Problem 3

Clearly, $P_{\text{unsmoothed}} > P_{\text{smooth}}$. The reason is the smoothed probability add 1 to the numerator increases the count of rare or non-existent bigrams, while V (the vocabulary size) increases the denominator significantly, resulting in reducing the probability.

4 Problem 4

- **Add-one Smoothing Probability for bigram**

$$P(w_i|w_{i-1}) = \frac{count(w_{i-1}, w_i) + 1}{count(w_{i-1}) + V}$$

- For the given corpus, we have:

$$count(\texttt{am,Sam}) = 2,\ count(\texttt{am}) = 3,\ V = 11$$

$$P(\texttt{Sam|am}) = \frac{count(\texttt{am,Sam}) + 1}{count(\texttt{am}) + V} = \frac{2+1}{3+11} = \frac{3}{14} \approx 0.2142857$$

5 Problem 5

- Using **linear interpolation smoothing** between a maximum-likelihood bigram model and a maximum-likelihood unigram model, we have:

$$P(w_i|w_{i-1}) = \lambda_1 P_{bi}(w_i|w_{i-1}) + \lambda_2 P_{uni}(w_i)$$

  where:

  - $P_{bi}(w_i|w_{i-1})$ is maximum likelihood estimate of bigram
  - $P_{uni}(w_i)$ is maximum likelihood estimate of unigram
  - $\lambda_1 = \lambda_2 = 0.5$

- **Calculate** $P(\texttt{Sam|am})$

$$P_{bi}(\texttt{Sam|am}) = \frac{count(\texttt{am, Sam})}{count(\texttt{am})} = \frac{2}{3}$$

$$P_{uni}(\texttt{Sam}) = \frac{count(\texttt{Sam})}{N} = \frac{4}{25}$$

$$P(\texttt{Sam|am}) = 0.5 \times \frac{2}{3} + 0.5 \times \frac{4}{25} = \frac{31}{75} \approx 0.4133$$

6 Problem 6

- The unigram probability using maximum likelihood estimate is calculated as:

$$P(w) = \frac{count(w)}{N}$$

where: N is the token count in the training set

- **Calculate the Probability**

From the traning set, we have:

$$N = 100$$
$$count(0) = 91$$
$$count(1) = count(2) = ... = count(9) = 1$$

Therefore, the unigram probability will be:

$$P(0) = \frac{count(0)}{N} = \frac{91}{100} = 0.91$$

$$P(3) = \frac{count(3)}{N} = \frac{1}{100} = 0.01$$

Given the test set: `0 0 0 0 0 3 0 0 0 0`, the probability of given test set based on unigram model is:

$$P(\texttt{0 0 0 0 0 3 0 0 0 0}) = P(0)^9 \times P(3) = 0.91^9 \times 0.01 \approx 0.004279$$

- **Calculate the Perplexity** We have the formula for perplexity is:

$$PP(W) = P(W)^{\frac{-1}{N}}$$

where N is the number of token in test set. In this case, $N = 10$

So, we have the perplexity is:

$$PP(W) = 0.004279^{\frac{-1}{10}} = 1.7253$$