Natural Language Processing - CC01 (Math Exercise - Lab 3)

Le Thi Phuong Thao - 2252757

February 2025

1 Problem 1

1.1 The equation for trigram probability estimation

$$P(w_n|w_1, w_2, \ldots, w_{n-1}) \approx \prod_{i=1}^{n} P(w_i|w_{i-2}, w_{i-1})$$

$$P(w_n \mid w_{n-2}w_{n-1}) = \frac{count(w_{n-2}, w_{n-1}, w_n)}{count(w_{n-2}, w_{n-1})}$$

1.2 Write out all the non-zero trigram probabilities for the I am Sam corpus from

1.2.1 Corpus

`<s> I am Sam </s>`
`<s> Sam I am </s>`
`<s> I do not like green eggs and Sam </s>`
We need to add two start symbols (`<s> <s>`) at the beginning and two end symbols (`</s> </s>`) at the end of each sentence to make sure we can compute trigram probabilities correctly.

Padded Corpus:
`<s> <s> I am Sam </s> </s>`
`<s> <s> Sam I am </s> </s>`
`<s> <s> I am Sam </s> </s>`
`<s> <s> I do not like green eggs and Sam </s> </s>`

The non-zero trigram probabilities:

$$P(I \mid \langle s\rangle, \langle s\rangle) = \frac{\text{count}(\langle s\rangle, \langle s\rangle, I)}{\text{count}(\langle s\rangle, \langle s\rangle)} = \frac{2}{3} \approx 0.6667$$

$$P(am \mid \langle s\rangle, I) = \frac{\text{count}(\langle s\rangle, I, am)}{\text{count}(\langle s\rangle, I)} = \frac{1}{2} = 0.5$$

$$P(Sam \mid I, am) = \frac{\text{count}(I, am, Sam)}{\text{count}(I, am)} = \frac{1}{2} = 0.5$$

Trigram	**Count**
$\langle s\rangle, \langle s\rangle, I$	2
$\langle s\rangle, I, am$	1
(I, am, Sam)	1
(am, Sam, $\langle /s\rangle$)	1
(Sam, $\langle /s\rangle$, $\langle /s\rangle$)	2
$\langle s\rangle, \langle s\rangle, Sam$	1
$\langle s\rangle, Sam, I$	1
(Sam, I, am)	1
$I, am, \langle /s\rangle$	1
$am, \langle /s\rangle, \langle /s\rangle$	1
$\langle s\rangle, I, do$	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, $\langle /s\rangle$)	1

Table 1: Trigram Counts (with padding)

$$P(\langle /s\rangle \mid am, Sam) = \frac{\text{count}(am, Sam, \langle /s\rangle)}{\text{count}(am, Sam)} = \frac{1}{1} = 1$$

$$P(\langle /s\rangle \mid Sam, \langle /s\rangle) = \frac{\text{count}(Sam, \langle /s\rangle, \langle /s\rangle)}{\text{count}(Sam, \langle /s\rangle)} = \frac{2}{2} = 1$$

$$P(Sam \mid \langle s\rangle, \langle s\rangle) = \frac{\text{count}(\langle s\rangle, \langle s\rangle, Sam)}{\text{count}(\langle s\rangle, \langle s\rangle)} = \frac{1}{3} = 0.3333$$

$$P(I \mid \langle s\rangle, Sam) = \frac{\text{count}(\langle s\rangle, Sam, I)}{\text{count}(\langle s\rangle, Sam)} = \frac{1}{1} = 1$$

$$P(am \mid Sam, I) = \frac{\text{count}(Sam, I, am)}{\text{count}(Sam, I)} = \frac{1}{1} = 1$$

$$P(\langle /s\rangle \mid I, am) = \frac{\text{count}(I, am, < /s >)}{\text{count}(I, am)} = \frac{1}{2} = 0.5$$

$$P(langles\rangle \mid I, am) = \frac{\text{count}(I, am, < /s >)}{\text{count}(I, am)} = \frac{1}{2} = 0.5$$

$$P(\langle /s\rangle \mid am, \langle /s\rangle) = \frac{\text{count}(am, \langle /s\rangle), \langle /s\rangle}{\text{count}(am, \langle /s\rangle)} = \frac{1}{1} = 1$$

$$P(do \mid \langle s\rangle, I) = \frac{\text{count}(\langle s\rangle, I, do)}{\text{count}(\langle s\rangle, I)} = \frac{1}{2} = 0.5$$

$$P(not \mid I, do) = \frac{\text{count}(I, do, not)}{\text{count}(I, do)} = \frac{1}{1} = 1$$

Bigram	**Count**
$\langle s\rangle, \langle s\rangle$	3
$\langle s\rangle, I$	2
(I, am)	2
(am, Sam)	1
(Sam, $\langle /s\rangle$)	2
($\langle /s\rangle$, $\langle /s\rangle$)	3
$\langle s\rangle, Sam$	1
(Sam, I)	1
$am, \langle /s\rangle$	1
(I, do)	1
(do, not)	1
(not, like)	1
(like, green)	1
(green, eggs)	1
(eggs, and)	1
(and, Sam)	1

Table 2: Bigram Counts (with padding)

$$P(\mathit{like} \mid \mathit{do}, \mathit{not}) = \frac{\text{count}(\mathit{do}, \mathit{not}, \mathit{like})}{\text{count}(\mathit{do}, \mathit{not})} = \frac{1}{1} = 1$$

$$P(\mathit{green} \mid \mathit{not}, \mathit{like}) = \frac{\text{count}(\mathit{not}, \mathit{like}, \mathit{green})}{\text{count}(\mathit{not}, \mathit{like})} = \frac{1}{1} = 1$$

$$P(\mathit{eggs} \mid \mathit{like}, \mathit{green}) = \frac{\text{count}(\mathit{like}, \mathit{green}, \mathit{eggs})}{\text{count}(\mathit{like}, \mathit{green})} = \frac{1}{1} = 1$$

$$P(\mathit{and} \mid \mathit{green}, \mathit{eggs}) = \frac{\text{count}(\mathit{green}, \mathit{eggs}, \mathit{and})}{\text{count}(\mathit{green}, \mathit{eggs})} = \frac{1}{1} = 1$$

$$P(\mathit{Sam} \mid \mathit{eggs}, \mathit{and}) = \frac{\text{count}(\mathit{eggs}, \mathit{and}, \mathit{Sam})}{\text{count}(\mathit{eggs}, \mathit{and})} = \frac{1}{1} = 1$$

$$P(\langle /s\rangle \mid \mathit{and}, \mathit{Sam}) = \frac{\text{count}(\mathit{and}, \mathit{Sam}, \langle /s\rangle)}{\text{count}(\mathit{and}, \mathit{Sam})} = \frac{1}{1} = 1$$

2 Problem 2

Assume the additional Laplace smoothed probabilities:

$$P(i \mid \langle s\rangle) = 0.19, \quad P(\langle /s\rangle \mid \text{food}) = 0.40$$

Calculate the probability of the sentence:

$$\langle s\rangle \; i \; want \; chinese \; food \; \langle /s\rangle$$

($\langle s\rangle$ and $\langle /s\rangle$ are not smoothed)

The probability using bi-gram probabilities table:

$$\begin{aligned}P &= P(i \mid \langle s\rangle) \times P(want \mid i) \times P(chinese \mid want) \times P(food \mid chinese) \times P(\langle /s\rangle \mid food)\\ &= 0.19 \times 0.33 \times 0.0065 \times 0.52 \times 0.4\\ &= 8.47704 \times 10^{-5}\end{aligned}$$

The probability using Laplace smoothed probabilities:

$$\begin{aligned}P &= P(i \mid \langle s\rangle) \times P(want \mid i) \times P(chinese \mid want) \times P(food \mid chinese) \times P(\langle /s\rangle \mid food)\\ &= 0.19 \times 0.21 \times 0.0029 \times 0.52 \times 0.4\\ &= 2.406768 \times 10^{-5}\end{aligned}$$

3 Problem 3

Which of the two probabilities you computed in the previous problem is higher, unsmoothed or smoothed? Explain why.

The unsmoothed probability (0.000085) is higher than the smoothed probability (0.000024). This is because the unsmoothed probability directly reflects the actual observed frequencies of bigrams in the dataset. On the other hand, Laplace smoothing redistributes probability mass to ensure that no probability is zero, which helps handle unseen bigrams. However, this process also lowers the probabilities of frequently occurring bigrams, spreading the probability more evenly across all word pairs.

4 Problem 4

We are given the following corpus:

$\langle s\rangle$ I am Sam $\langle /s\rangle$
$\langle s\rangle$ Sam I am $\langle /s\rangle$
$\langle s\rangle$ I am Sam $\langle /s\rangle$
$\langle s\rangle$ I do not like green eggs and Sam $\langle /s\rangle$

Using a bigram language model with add-one smoothing, what is $P(\text{Sam} \mid \text{am})$? Include $\langle s\rangle$ and $\langle /s\rangle$ in your counts just like any other token.

Vocabulary size ($V = 11$):

Unique tokens: { $\langle s\rangle$, I, am, Sam, $\langle /s\rangle$, do, not, like, green, eggs, and }
Using a bigram language model with add-one smoothing:

$$P(w_n \mid w_{n-1}) = \frac{\text{count}(w_{n-1}, w_n) + 1}{\text{count}(w_{n-1}) + V}$$

$$P(\text{Sam} \mid \text{am}) = \frac{\text{count}(\text{am}, \text{Sam}) + 1}{\text{count}(\text{am}) + V} = \frac{2 + 1}{3 + 11} = \frac{3}{14} \approx 0.214$$

5 Problem 5

We are given the following corpus, modified from the one in the chapter:

$$\langle s\rangle \text{ I am Sam } \langle /s\rangle$$
$$\langle s\rangle \text{ Sam I am } \langle /s\rangle$$
$$\langle s\rangle \text{ I am Sam } \langle /s\rangle$$
$$\langle s\rangle \text{ I do not like green eggs and Sam } \langle /s\rangle$$

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 $P(\text{Sam} \mid \text{am})$?

Include $\langle s\rangle$ and $\langle /s\rangle$ in your counts just like any other token.

Linear Interpolation Smoothing:

$$P(w_n \mid w_{n-1}) = \lambda_1 P(w_n) + \lambda_2 P(w_n \mid w_{n-1})$$

Using linear interpolation, the probability is computed as:

$$P(\text{Sam} \mid \text{am}) = \lambda_1 P(\text{Sam}) + \lambda_2 P(\text{Sam} \mid \text{am})$$

From the corpus:

- **Unigram counts**: - Sam appears 3 times. - Total number of tokens (including $\langle s\rangle$ and $\langle /s\rangle$): 25.

$$P(\text{Sam}) = \frac{\text{count(Sam)}}{\text{total word count}} = \frac{4}{25}$$

- **Bigram counts**: - (am, Sam) appears 2 times. - am appears 3 times.

$$P(\text{Sam} \mid \text{am}) = \frac{\text{count(am, Sam)}}{\text{count(am)}} = \frac{2}{3}$$

$$P(\text{Sam} \mid \text{am}) = \frac{1}{2} \times \frac{4}{25} + \frac{1}{2} \times \frac{2}{3}$$

$$\approx 0.4133$$

6 Problem 6

You are given a training set of 100 numbers that 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?

A training set of 100 numbers consists of 91 zeros and 1 each of the other digits 1-9. Therefore, the unigram probabilities are:

$$P(0) = 0.91, \quad P(k) = 0.01, \quad \text{for } k = 1, 2, 3, \ldots, 9.$$

The given test set:

$$0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 3 \quad 0 \quad 0 \quad 0 \quad 0$$

The probability of the test sequence is:

$$P(\text{test}) = 0.91^9 \times 0.01$$

The unigram perplexity is computed as:

$$\text{Perplexity} = P(\text{test})^{-\frac{1}{N}}, \quad \text{where } N = 10.$$
$$\text{Perplexity} = (0.91^9 \times 0.01)^{-\frac{1}{10}}$$
$$\text{Perplexity} \approx 1.7253$$