VIETNAM NATIONAL UNIVERSITY, HO CHI MINH CITY
UNIVERSITY OF TECHNOLOGY
FACULTY OF COMPUTER SCIENCE AND ENGINEERING

Natural Language Processing - Exercise (CO3086)

Lab 3

Math Exercises

Semester 2, Academic Year 2023 - 2024

Teacher: Bui Khanh Vinh
Students: Nguyen Quang Phu - 2252621

HO CHI MINH CITY, FEBRUARY 2025

Contents

Chapter 1

Problem Description and Solution

1.1 Problem 1

1.1.1 Description

Write out the equation for trigram probability estimation. Now write out all the **non-zero trigram probabilities** for the *I am Sam* corpus from:

$$< s > \text{ I am Sam } < /s >$$

$$< s > \text{ Sam I am } < /s >$$

$$< s > \text{ I do not like green eggs and Sam } < /s >$$

1.1.2 Solution

Non-Zero Trigram Probabilities for the Corpus

The *I am Sam* corpus contains the following sentences:

1. $< s > \text{ I am Sam } < /s >$
2. $< s > \text{ Sam I am } < /s >$
3. $< s > \text{ I do not like green eggs and Sam } < /s >$

We calculate all **non-zero trigram probabilities** from the corpus.

Step 1: Trigrams from Each Sentence

- From $< s > \text{ I am Sam } < /s >$:

$$(< s >, \text{I}, \text{am}),\ (\text{I}, \text{am}, \text{Sam}),\ (\text{am}, \text{Sam}, < /s >)$$

- From $< s >$ Sam I am $< /s >$:

$$(< s >, \text{Sam}, \text{I}),\ (\text{Sam}, \text{I}, \text{am}),\ (\text{I}, \text{am}, < /s >)$$

- From $< s >$ I do not like green eggs and Sam $< /s >$:

$$(< s >, \text{I}, \text{do}),\ (\text{I}, \text{do}, \text{not}),\ (\text{do}, \text{not}, \text{like}),\ (\text{not}, \text{like}, \text{green}),$$

$$(\text{like}, \text{green}, \text{eggs}),\ (\text{green}, \text{eggs}, \text{and}),\ (\text{eggs}, \text{and}, \text{Sam}),\ (\text{and}, \text{Sam}, < /s >)$$

Step 2: Trigram Counts and Probabilities

Using the trigram probability formula:

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

we compute the probabilities for all non-zero trigrams. Below are the results:

- $P(\text{am}| < s >, \text{I}) = \frac{1}{2}$
- $P(\text{Sam}|\text{I}, \text{am}) = \frac{1}{2}$
- $P(< /s > |\text{am}, \text{Sam}) = \frac{1}{1} = 1.0$

1.2 Problem 2

1.2.1 Description

Given two tables:

Table 1: Bigram probabilities for eight words

	i	want	to	eat	chinese	food	lunch	spend
i	0.002	0.33	0	0.0036	0	0	0	0.00079
want	0.0022	0	0.66	0.0011	0.0065	0.0065	0.0054	0.0011
to	0.00083	0	0.0017	0.28	0.00083	0	0.0025	0.087
eat	0	0	0.0027	0	0.021	0.0027	0.056	0
chinese	0.0063	0	0	0	0	0.52	0.0063	0
food	0.014	0	0.014	0	0.00092	0.0037	0	0
lunch	0.0059	0	0	0	0	0.0029	0	0
spend	0.0036	0	0.0036	0	0	0	0	0

Table 2: Add-one smoothed bigram probabilities for eight of the word

	i	want	to	eat	chinese	food	lunch	spend
i	0.0015	0.21	0.00025	0.0025	0.00025	0.00025	0.00025	0.00075
want	0.0013	0.00042	0.26	0.00084	0.0029	0.0029	0.0025	0.00084
to	0.00078	0.00026	0.0013	0.18	0.00078	0.00026	0.0018	0.055
eat	0.00046	0.00046	0.0014	0.00046	0.0078	0.0014	0.02	0.00046
chinese	0.0012	0.00062	0.00062	0.00062	0.00062	0.52	0.0012	0.00062
food	0.0063	0.00039	0.0063	0.00039	0.00079	0.002	0.00039	0.00039
lunch	0.0017	0.00056	0.00056	0.00056	0.00056	0.0011	0.00056	0.00056
spend	0.0012	0.00058	0.0012	0.00058	0.00058	0.00058	0.00058	0.00058

Figure 1.1: Bigram probabilities and problem statement.

Assume the additional Laplace smoothed probabilities $P(i \mid < s >) = 0.19$ and $P(< /s > \mid \text{food}) = 0.40$. Calculate the **probability of the sentence** *i want chinese food*. ($< s >$ and $< /s >$ are not smoothed.)

Solution

The probability of the sentence $< s >$ i want chinese food $< /s >$ is given by:

$$P(< s > \text{ i want chinese food } < /s >)$$

=

$$P(i \mid < s >) \cdot P(\text{want} \mid \text{i}) \cdot P(\text{chinese} \mid \text{want}) \cdot P(\text{food} \mid \text{chinese}) \cdot P(< /s > \mid \text{food})$$

We are given the following bigram probabilities:

Without Add-One Smoothing

$$\begin{aligned} P(i \mid < s >) &= 0.19 \\ P(\text{want} \mid \text{i}) &= 0.33 \\ P(\text{chinese} \mid \text{want}) &= 0.0065 \quad \text{(from Table 1)} \\ P(\text{food} \mid \text{chinese}) &= 0.52 \quad \text{(from Table 1)} \\ P(< /s > \mid \text{food}) &= 0.40 \end{aligned}$$

Using these values, the probability of the sentence "<s> I want Chinese food </s>" is:

$$P(< s > \text{ i want chinese food } < /s >) = 0.19 \times 0.33 \times 0.0065 \times 0.52 \times 0.40$$

With Add-One Smoothing

$$\begin{aligned} P(i \mid < s >) &= 0.19 \\ P(\text{want} \mid \text{i}) &= 0.33 \\ P(\text{chinese} \mid \text{want}) &= 0.0065 \quad \text{(from Table 2)} \\ P(\text{food} \mid \text{chinese}) &= 0.052 \quad \text{(from Table 2)} \\ P(< /s > \mid \text{food}) &= 0.40 \end{aligned}$$

Note: From the textbook, for the add-one smoothed bigram probabilities, the value of $P(\text{food} \mid \text{chinese})$ is 0.052 but in the table you give us it is 0.52 (which is the same as the value when without add-one smoothed). So i decided to still use the value of 0.052 here.

Thus, the probability of the same sentence without add-one smoothing is:

$$\begin{aligned} P(< s > \text{ i want chinese food } < /s >) &= 0.19 \times 0.33 \times 0.0065 \times 0.52 \times 0.40 \\ &= 0.0000847704 \end{aligned}$$

And the probability of the same sentence under add-one smoothing is:

$$\begin{aligned} P(< s > \text{ i want chinese food } < /s >) &= 0.19 \times 0.21 \times 0.0029 \times 0.052 \times 0.40 \\ &= 0.000002406768 \end{aligned}$$

1.3 Problem 3

1.3.1 Description

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

1.3.2 Solution

The probability computed without add-one smoothing is higher than the probability computed with smoothing.

This happens because add-one smoothing significantly alters the original word counts. When we apply smoothing, we redistribute probability mass by assigning small probabilities to previously unseen word pairs. As a result, the probabilities of observed bigrams are reduced to compensate for this redistribution.

Looking at the two tables, we notice that probabilities that were originally nonzero before smoothing become smaller after smoothing. This occurs because add-one smoothing assigns probability mass to previously unseen word pairs (those with zero probability in the original table), effectively decreasing the probabilities of the originally observed word pairs.

Relating this back to problem 2, in the unsmoothed case, all the bigram probabilities come directly from the observed data, leading to a relatively higher final probability. However, when smoothing is applied, each individual bigram probability decreases, and since the final sentence probability is the product of these probabilities, the overall sentence probability is also lower.

Therefore, smoothing decreases the computed probability of the given sentence.

1.4 Problem 4

1.4.1 Description

We are given the following corpus:

```
< 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 $P(\text{Sam} \mid \text{am})$? Include $\langle s \rangle$ and $\langle /s \rangle$ in your counts just like any other token.

1.4.2 Solution

The probability $P(\text{Sam} \mid \text{am})$ can be computed using the formula for a bigram model with add-one smoothing:

$$P(\text{Sam} \mid \text{am}) = \frac{C(\text{am}, \text{Sam}) + 1}{C(\text{am}) + V}$$

Where:

- $C(\text{am}, \text{Sam})$ is the count of the bigram (am, Sam) in the corpus.
- $C(\text{am})$ is the total count of the unigram (am).
- V is the size of the vocabulary, including $\langle s \rangle$ and $\langle /s \rangle$.

	$\langle s \rangle$	$\langle /s \rangle$	I	am	Sam	do	not	like	green	eggs	and	Total
$\langle s \rangle$	0	0	3	0	1	0	0	0	0	0	0	4
$\langle /s \rangle$	0	0	0	0	0	0	0	0	0	0	0	0
I	0	0	0	3	0	1	0	0	0	0	0	4
am	0	1	0	0	2	0	0	0	0	0	0	3
Sam	0	3	1	0	0	0	0	0	0	0	0	4
do	0	0	0	0	0	0	1	0	0	0	0	1
not	0	0	0	0	0	0	0	1	0	0	0	1
like	0	0	0	0	0	0	0	0	1	0	0	1
green	0	0	0	0	0	0	0	0	0	1	0	1
eggs	0	0	0	0	0	0	0	0	0	0	1	1
and	0	0	0	0	0	0	0	0	0	0	1	1

Table 1.1: Bigram Count Matrix for the Corpus

	$\langle s \rangle$	$\langle /s \rangle$	I	am	Sam	do	not	like	green	eggs	and	Total
$\langle s \rangle$	1	1	4	1	2	1	1	1	1	1	1	15
$\langle /s \rangle$	1	1	1	1	1	1	1	1	1	1	1	11
I	1	1	1	4	1	2	1	1	1	1	1	15
am	1	2	1	1	3	1	1	1	1	1	1	14
Sam	1	4	2	1	1	1	1	1	1	1	1	15
do	1	1	1	1	1	1	2	1	1	1	1	12
not	1	1	1	1	1	1	1	2	1	1	1	12
like	1	1	1	1	1	1	1	1	2	1	1	12
green	1	1	1	1	1	1	1	1	1	2	1	12
eggs	1	1	1	1	1	1	1	1	1	1	2	12
and	1	1	1	1	1	1	1	1	1	1	2	12

Table 1.2: Bigram Count Matrix with Add-One Smoothing Applied

Based on the corpus:

- $C(\text{am}, \text{Sam}) = 2$
- $C(\text{am}) = 3$
- $V = 11$ (calculated as the total number of unique tokens in the corpus).

Substituting into the formula:

$$P(\text{Sam} \mid \text{am}) = \frac{2+1}{3+11} = \frac{3}{14} \approx 0.21428571428571427$$

1.5 Problem 5

1.5.1 Description

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 $\lambda_1 = \frac{1}{2}$ 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.

1.5.2 Solution

The probability $P(\text{Sam} \mid \text{am})$ using linear interpolation smoothing is computed as:

$$P(\text{Sam} \mid \text{am}) = \lambda_1 \cdot P_{\text{bigram}}(\text{Sam} \mid \text{am}) + \lambda_2 \cdot P_{\text{unigram}}(\text{Sam})$$

Where:

- $P_{\text{bigram}}(\text{Sam} \mid \text{am}) = \frac{C(\text{am,Sam})}{C(\text{am})}$
- $P_{\text{unigram}}(\text{Sam}) = \frac{C(\text{Sam})}{N}$
- N is the total number of tokens in the corpus.

From the corpus:

- $C(\text{am}, \text{Sam}) = 2$
- $C(\text{am}) = 4$
- $C(\text{Sam}) = 4$
- $N = 25$ (number of tokens)

Calculating the probabilities:

$$P_{\text{bigram}}(\text{Sam} \mid \text{am}) = \frac{2}{4} = 0.5$$

$$P_{\text{unigram}}(\text{Sam}) = \frac{4}{25} = 0.16$$

Substituting into the interpolation formula with $\lambda_1 = \lambda_2 = \frac{1}{2}$:

$$P(\text{Sam} \mid \text{am}) = \frac{1}{2} \cdot 0.5 + \frac{1}{2} \cdot 0.16 = 0.25 + 0.1 = 0.33$$

Thus, $P(\text{Sam} \mid \text{am}) = 0.33$.

1.6 Problem 6

1.6.1 Description

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?

1.6.2 Solution

In the training set:

- The digit 0 appears 91 times out of 100.
- Each digit from 1 to 9 appears 1 time out of 100.

The unigram probabilities are:

$$P(0) = \frac{91}{100} = 0.91, \quad P(i) = \frac{1}{100} = 0.01 \quad \text{for } i \in \{1, 2, \ldots, 9\}.$$

For the test set 0 0 0 0 0 3 0 0 0 0, the probability of the test sequence is:

$$P(\text{test}) = P(0)^9 \cdot P(3) = (0.91)^9 \cdot 0.01.$$

The unigram perplexity is given by:

$$\text{Perplexity} = \left(\frac{1}{P(\text{test})}\right)^{1/n},$$

where n is the length of the test sequence (here $n = 9$).

Substituting:

$$\text{Perplexity} = \left(\frac{1}{P(\text{test})}\right)^{1/9} \approx 1.8331.$$