Statistical Natural Language Processing - Assignment 01

Rayyan Mohammad Minhaj (7074982/rami00002@stud.uni-saarland.de)
Abdullah Abdul Wahid (7075730/abyy00002@stud.uni-saarland.de)

April 20, 2025

Exercise 1.1 - Probability Basics

(a) All missing bigram probabilities

Consider the full set of letters S = {a,b,c}, all possible bigrams (ordered pairs) are:

$$SxS = \{(a,a),(a,b),(a,c),(b,a),(b,b),(b,c),(c,a),(c,b),(c,c)$$

There are 9 total bigrams, we have been provided the values for:

$$\begin{aligned} p(a,a) &= 0.25 \\ p(a,c) &= 0.25 \\ p(b,a) &= 0.125 \\ p(b,b) &= 0 \\ p(c,c) &= 0.25 \end{aligned}$$

Adding these up:

$$0.25 + 0.25 + 0.125 + 0 + 0.25 = 0.875$$

That means the remaining probability needs to be spread over the remaining bigrams.

$$1 - 0.875 = 0.125$$

The remaining bigrams are:

$$\begin{gathered} p(a,b) \\ p(b,c) \\ p(c,a) \\ p(c,b) \end{gathered}$$

From additional information we know that pL(a) = 0.5, which means the total of all bigrams where a occurs at the left is equal to 0.5.

$$\begin{aligned} p(a,a) + p(a,b) + p(a,c) &= 0.5 \\ 0.25 + p(a,b) + 0.25 &= 0.5 \\ p(a,b) &= 0 \end{aligned}$$

From additional information we know that pR(b) = 0.125, which means the total of all bigrams where b occurs at the right is equal to 0.125.

$$\begin{aligned} p(a,b)+p(b,b)+p(c,b) &= 0.125 \\ 0+0+p(c,b) &= 0.125 \\ p(c,b) &= 0.125 \end{aligned}$$

Summing the newfound values now gives us:

$$\begin{aligned} p(a,b)+p(c,b)+0.875 &= 1.0 \\ 0+0.125+0.875 &= 1.0 \end{aligned}$$

Which implies that p(b,c) and p(c,a) must be equal to 0, hence the final bigram possibilities are:

$$\begin{aligned} p(a,a) &= 0.25 \\ p(a,c) &= 0.25 \\ p(b,a) &= 0.125 \\ p(b,b) &= 0 \\ p(c,c) &= 0.25 \\ p(a,b) &= 0 \\ p(b,c) &= 0 \\ p(c,a) &= 0 \\ p(c,b) &= 0.125 \end{aligned}$$

(b) Determining whether any pairs of consecutive events (x, y) are independent (i.e., p(x, y) = pL(x) · pR(y)).

We know the following unigram possibilities:

$$\begin{aligned} pL(a) &= 0.5 \\ pR(b) &= 0.125 \end{aligned}$$

Using the previously established bigram possibilities, we can find the rest of the unigram possibilities:

$$\begin{aligned} pL(a) &= 0.5 \\ pL(b) &= 0.125 \\ pL(c) &= 0.375 \\ \\ pR(a) &= 0.375 \\ pR(b) &= 0.125 \\ pR(c) &= 0.5 \end{aligned}$$

Now we can check which of the pairs of consecutive ends are independent.

$$\begin{aligned}
p(a,b) &= pL(a) * pR(b) = 0.5 \neq 0.0625 \\
p(a,c) &= pL(a) * pR(c) = 0.25 = 0.25 \\
p(b,a) &= pL(b) * pR(a) = 0.125 \neq 0.0468 \\
p(b,c) &= pL(b) * pR(c) = 0 \neq 0.0625 \\
p(c,a) &= pL(c) * pR(a) = 0 \neq 0.1406 \\
p(c,b) &= pL(c) * pR(b) = 0.125 \neq 0.0468
\end{aligned}$$

Therefore, only the bigram (a,c) satisfies the independence condition.

(c) Is it enough to compute $p(b \mid c)$ (i.e., the probability of seeing b if we already know that the preceding event generated c)? Justify your answer.

The conditional probability of $p(b \mid c)$ can be represented mathematically as:

$$\begin{aligned}
p(b|c) &= \frac{p(c,b)}{pL(c)} \\
p(b|c) &= \frac{0.125}{0.375} = \frac{1}{3} = 0.333
\end{aligned}$$

However, $p(b \mid c)$ tells us only one specific transition: from c to b, but there can be multiple possible transitions after c. For ex,

$$\begin{gathered}
p(c,c) \\
p(a,c)
\end{gathered}$$

So, to fully model the behavior after seeing c, we will need all the conditional probabilities given c:

$$\begin{gathered}
p(a,c) \\
p(b,c) \\
p(c,c)
\end{gathered}$$

Only then can we understand all possible next steps after c.

Exercise 1.2 - Zipf's Law

(a) What is Zipf's Law?

Zipf's Law states that in any large collection of words, the most common word appears about twice as often as the second most common word, three times as often as the third most common word, and so on. In general, a word's frequency is inversely proportional to its rank in the frequency table.

(b) Does every kind of language (natural, man-made, programming) follow Zipf's Law?

Zipf's Law generally applies to natural languages, where certain words (like "the," "and," etc.) appear much more frequently than others. However, it does not always work for man-made languages like constructed languages or programming languages, where word frequency is more controlled and less variable. In programming languages, for example, certain keywords or symbols may appear equally, making Zipf's Law less applicable.

(c) What are the limitations of Zipf's Law?

Zipf's Law often fails to account for semantic structures (because it assumes word frequency will always follow a simple pattern), which becomes obvious in specialized or technical vocabularies. Additionally, in cases where word frequencies are more evenly distributed like formal languages or controlled vocabularies, Zipf's Law becomes limited.

Exercise 1.3 - Python Basics

Completed the attached Jupyter Notebook.

Exercise 1.4 - Zipf 's and Mandelbrot's Law

Completed the attached Jupyter Notebook.