The N-gram Language Model

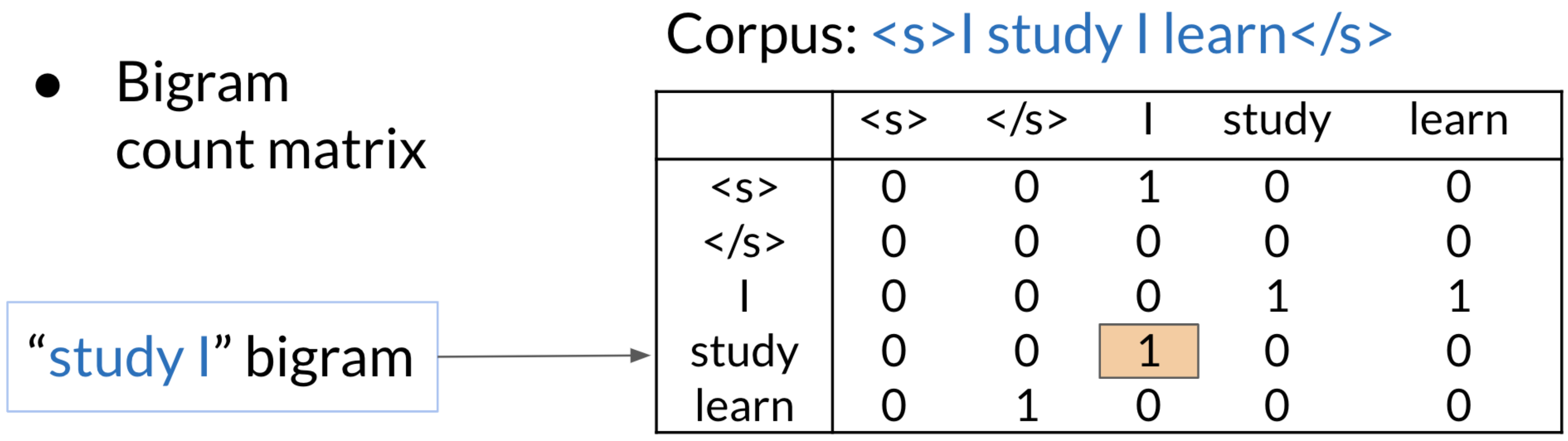
We covered a lot of concepts in the previous video. You have seen:

- Count matrix
- Probability matrix
- Language model
- Log probability to avoid underflow
- Generative language model

In the count matrix:

- Rows correspond to the unique corpus N-1 grams.
- Columns correspond to the unique corpus words.

Here is an example of the count matrix of a **bigram**.



- Bigram count matrix

Corpus: <s>I study I learn</s>

	<s>	</s>	I	study	learn
<s>	0	0	1	0	0
</s>	0	0	0	0	0
I	0	0	0	1	1
study	0	0	1	0	0
learn	0	1	0	0	0

“study I” bigram

To convert it into a probability matrix, you can use the following formula:

- $P\left(w_n \mid w_{n-N+1}^{n-1}\right)=\frac{C\left(w_{n-N+1}^{n-1}, w_n\right)}{C\left(w_{n-N+1}^{n-1}\right)}$
- $\operatorname{sum}(row)=\sum_{w \in V} C\left(w_{n-N+1}^{n-1}, w\right)=C\left(w_{n-N+1}^{n-1}\right)$

Now given the probability matrix, you can generate the language model. You can compute the sentence probability and the next word prediction.

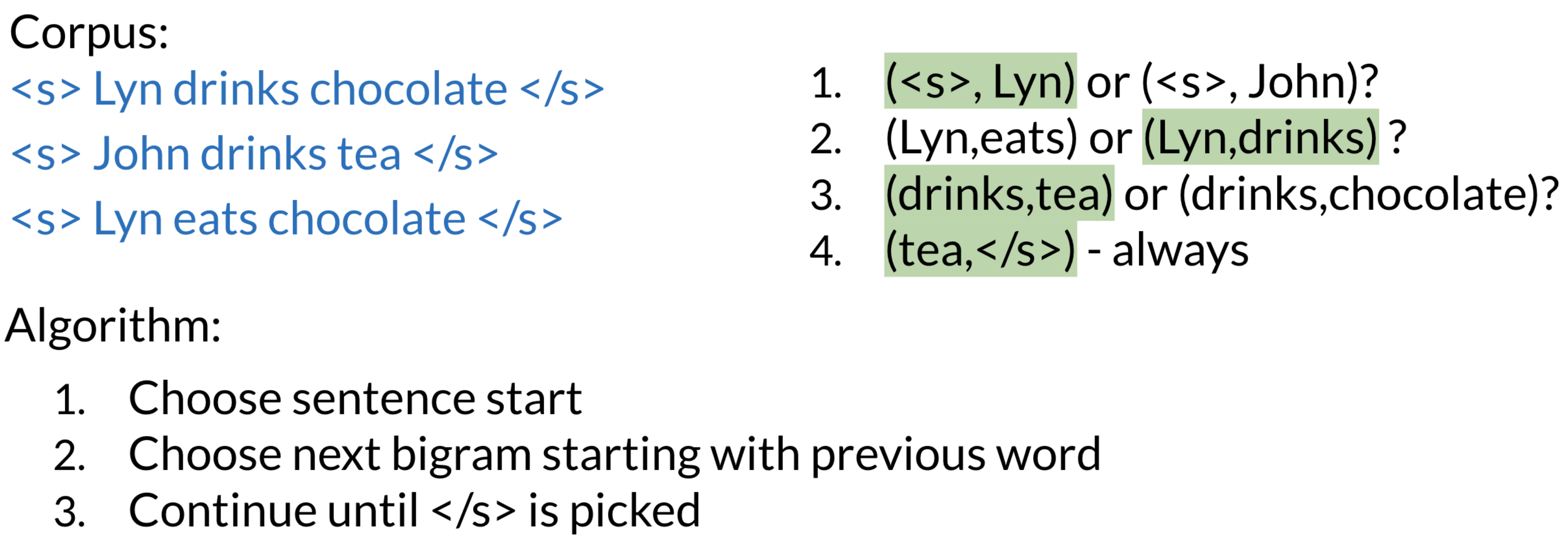
To compute the probability of a sequence, you needed to compute:

$$P\left(w_1^n\right) \approx \prod_{i=1}^n P\left(w_i \mid w_{i-1}\right)$$

To avoid underflow, you can multiply by the log.

$$\log \left(P\left(w_1^n\right)\right) \approx \sum_{i=1}^n \log \left(P\left(w_i \mid w_{i-1}\right)\right)$$

Finally here is a summary to create the generative model:



Corpus:

<s> Lyn drinks chocolate </s>

<s> John drinks tea </s>

<s> Lyn eats chocolate </s>

1. (<s>, Lyn) or (<s>, John)?
2. (Lyn,eats) or (Lyn,drinks) ?
3. (drinks,tea) or (drinks,chocolate)?
4. (tea,</s>) - always

Algorithm:

1. Choose sentence start
2. Choose next bigram starting with previous word
3. Continue until </s> is picked