Hide menu

Lecture: Autocomplete

- Video: Week Introduction 1 min
- Video: N-Grams: Overview 3 min
- Reading: N-Grams Overview 5 min
- Video: N-grams and Probabilities 7 min
- Reading: N-grams and Probabilities 10 min
- Video: Sequence Probabilities 5 min
- Reading: Sequence Probabilities 6 min
- Video: Starting and Ending Sentences 8 min
- Reading: Starting and Ending Sentences 6 min
- Lab: Lecture notebook: Corpus preprocessing for N-grams 1h
- Video: The N-gram Language Model 6 min
- Reading: The N-gram Language Model 10 min
- Video: Language Model Evaluation 6 min
- Lab: Lecture notebook: Building the language model 1h
- Reading: Language Model Evaluation 10 min
- Video: Out of Vocabulary Words 4 min
- Reading: Out of Vocabulary Words 10 min
- Video: Smoothing 6 min
- Reading: Smoothing 10 min
- Lab: Lecture notebook: Language model generalization 1h
- Video: Week Summary 1 min
- Reading: Week Summary 2 min
- Video: Week Conclusion 46 sec

Lecture Notes (Optional)

The N-gram Language Model

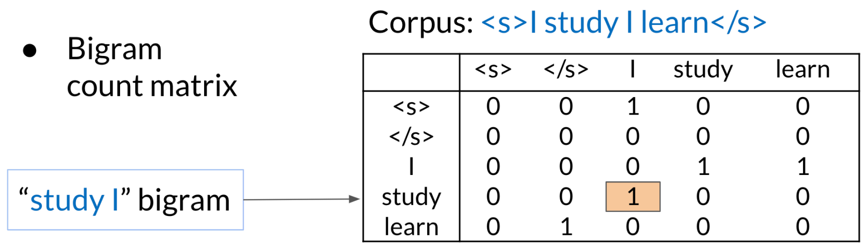
We covered a lot of concepts in the previous video. You have seen:

- 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**.



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

- Bigram count matrix

"study I" bigram

	<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

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

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

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

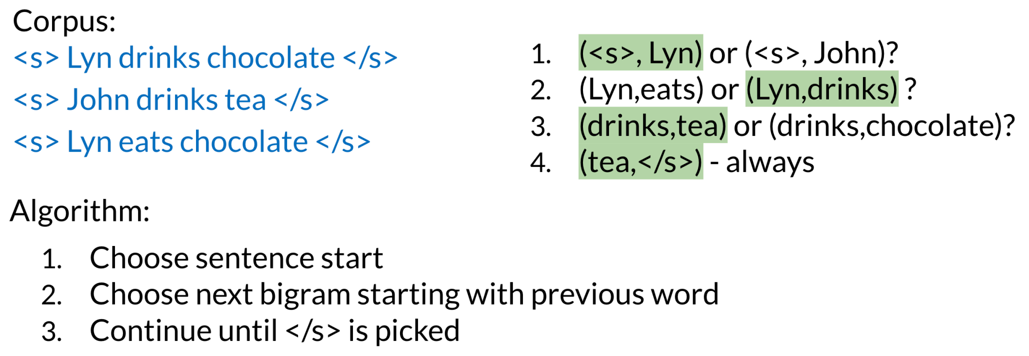
To compute the probability of a sequence, you needed to compute:

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

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