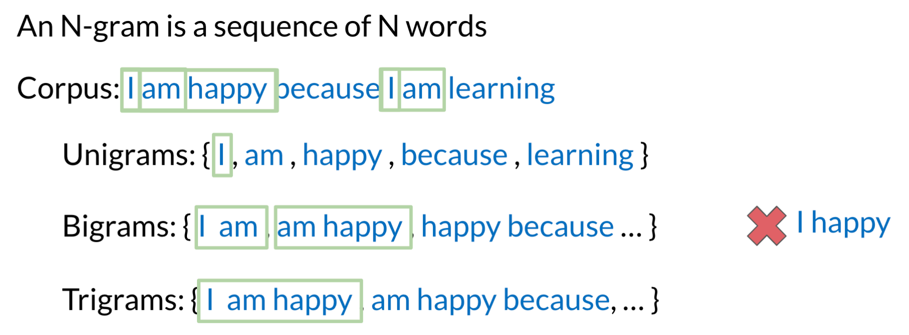


An N-gram is a sequence of N words

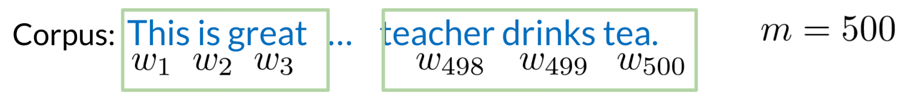
Corpus: I am happy because I am learning

Unigrams: { I , am , happy , because , learning }

Bigrams: { I am , am happy , happy because ... }

Trigrams: { I am happy , am happy because, ... }

Now given the those definitions, we can label a sentence as follows:



Corpus: This is great ... teacher drinks tea.

w_1 w_2 w_3 w_{498} w_{499} w_{500}

$m = 500$

In other notation you can write:

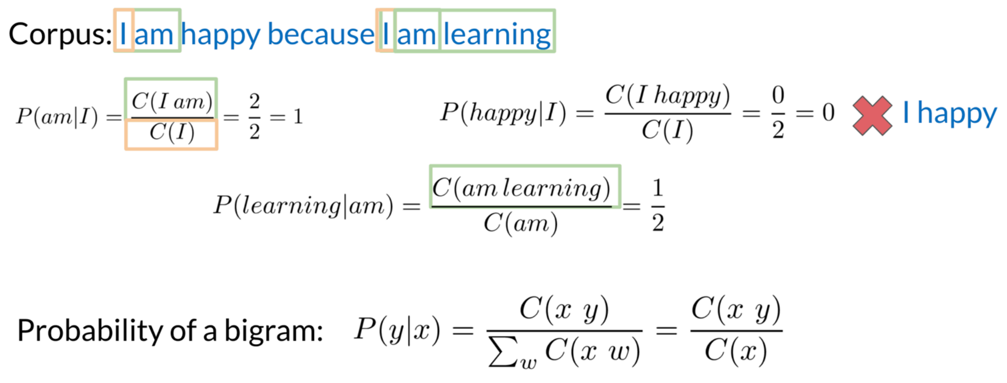
- $\overset{m}{w1} = w1w2w3....wm$
- $\overset{3}{w1} = w1w2w3$
- $\overset{m}{wm}-2 = wm-2wm-1wm$

Given the following corpus: *I am happy because I am learning.*

- Size of corpus m = 7.
- $P(I) = \frac{2}{7}$
- $P(happy) = \frac{1}{7}$

To generalize, the probability of a unigram is $P(w) = \frac{C(w)}{m}$

Bigram Probability:



Corpus: I am happy because I am learning

$$P(am|I) = \frac{C(I\ am)}{C(I)} = \frac{2}{2} = 1$$

$$P(happy|I) = \frac{C(I\ happy)}{C(I)} = \frac{0}{2} = 0$$

$$P(learning|am) = \frac{C(am\ learning)}{C(am)} = \frac{1}{2}$$

Probability of a bigram: $P(y|x) = \frac{C(x\ y)}{\sum_w C(x\ w)} = \frac{C(x\ y)}{C(x)}$

Trigram Probability:

To compute the probability of a trigram:

- $P\left(w3 \mid \overset{2}{w1}\right) = \frac{C(\overset{2}{w1}w3)}{C(\overset{2}{w1})}$
- $C\left(\overset{2}{w1}w3\right) = C\left(w1w2w3\right) = C\left(\overset{3}{w1}\right)$

N-gram Probability: