N-grams and Probabilities

Before we start computing probabilities of certain sequences, we need to first define what is an N-gram language model:

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 ... }

 I happy

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. $m = 500$

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

In other notation you can write:

- $w_1^m = w_1 w_2 w_3 .... w_m$
- $w_1^3 = w_1 w_2 w_3$
- $w_{m-2}^m = w_{m-2} w_{m-1} w_m$

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

$$P(am|I) = \frac{C(I\ am)}{C(I)} = \frac{2}{2} = 1 \qquad P(happy|I) = \frac{C(I\ happy)}{C(I)} = \frac{0}{2} = 0$$ ✖ I happy

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

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:

- $P\left(w_3 \mid w_1^2\right) = \frac{C\left(w_1^2 w_3\right)}{C\left(w_1^2\right)}$
- $C\left(w_1^2 w_3\right) = C\left(w_1 w_2 w_3\right) = C\left(w_1^3\right)$

N-gram Probability:

- $P\left(w_N \mid w_1^{N-1}\right) = \frac{C\left(w_1^{N-1} w_N\right)}{C\left(w_1^{N-1}\right)}$
- $C\left(w_1^{N-1} w_N\right) = C\left(w_1^N\right)$