moothing

The three main concepts covered here are dealing with missing n-grams, smoothing, and Backoff and interpolation.

$$P\left(w_n \mid w_{n-N+1}^{n-1}\right) = \frac{C\left(w_{n-N+1}^{n-1}, w_n\right)}{C\left(w_{n-N+1}^{n-1}\right)} \text{ can be } 0$$

Hence we can add-1 smoothing as follows to fix that problem:

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

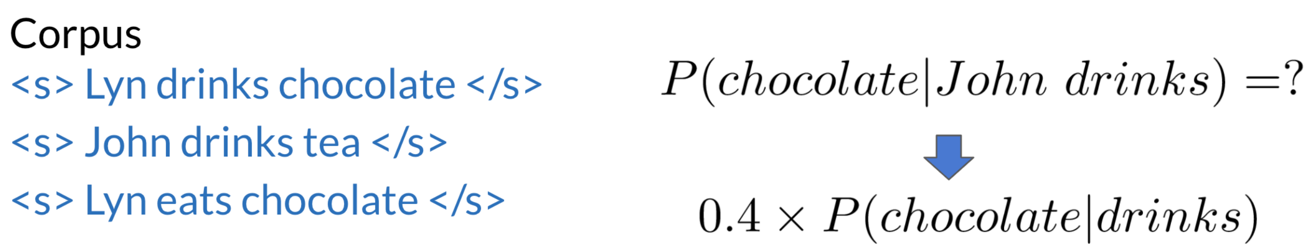
Add-k smoothing is very similar:

$$P\left(w_n \mid w_{n-1}\right) = \frac{C(w_{n-1}, w_n)+k}{\sum_{w \in V}\left(C(w_{n-1}, w)+k\right)} = \frac{C(w_{n-1}, w_n)+k}{C(w_{n-1})+k*V}$$

When using back-off:

- If N-gram missing => use (N-1)-gram, …: Using the lower level N-grams (i.e. (N-1)-gram, (N-2)-gram, down to unigram) distorts the probability distribution. Especially for smaller corpora, some probability needs to be discounted from higher level N-grams to use it for lower level N-grams.
- Probability discounting e.g. Katz backoff: makes use of discounting.
- “Stupid” backoff: If the higher order N-gram probability is missing, the lower order N-gram probability is used, just multiplied by a constant. A constant of about 0.4 was experimentally shown to work well.

Here is a visualization:



You can also use interpolation when computing probabilities as follows:

$$\hat{P}\left(w_n \mid w_{n-2} w_{n-1}\right) = \lambda_1 \times P\left(w_n \mid w_{n-2} w_{n-1}\right) + \lambda_2 \times P\left(w_n \mid w_{n-1}\right) + \lambda_3 \times P\left(w_n\right)$$

Where

$$\sum_i \lambda_i = 1$$