Estimating and Smoothing LM

$$p(w_1) = \frac{\text{count}(w_1)}{\text{total words observed}}$$ Unigram probabilities.

$$p(w_2|w_1) = \frac{\text{count}(w_1 w_2)}{\text{count}(w_1)}$$ Bigram probabilities.

$$p(w_3|w_2, w_1) = \frac{\text{count}(w_1 w_2 w_3)}{\text{count}(w_1 w_2)}$$ Trigram probabilities.

Unseen ngrams ($p(\textit{ngram}) = 0$) are a big problem, invalidate whole sentence: $p_{\text{LM}}(e_1^I) = \cdots 0 \cdots = 0$

$\Rightarrow$ Back-off with shorter ngrams:

$$p_{\text{LM}}(e_1^I) = \prod_{i=1}^{I} \Big(\begin{array}{l} 0.8 \cdot p(e_i|e_{i-1}, e_{i-2}) + \\ 0.15 \cdot p(e_i|e_{i-1}) + \\ 0.049 \cdot p(e_i) + \\ 0.001 \end{array} \Big) \neq 0 \tag{5}$$