Smoothing in N-gram LMs

5-gram models work well!

(2) $P(w \mid to)$ (3) $P(w \mid go\ to)$ (5) $P(w \mid want\ to\ go\ to)$ → Austin

(5) $P(w \mid hate\ to\ go\ to)$ → class

$P(Austin \mid to) > 0$ seen in data

$P(Austin \mid want\ to\ go\ to) = 0$ if corpus isn't huge

Absolute Discounting

Reserve mass from seen 5-grams to allocate to unseen 5-grams

$$P_{AD}(\text{Austin} \mid \text{want to go to}) = \frac{\text{count}(w\,t\,g\,t\,A) - K}{\text{count}(w\,T\,g\,T)} + \lambda \overbrace{P_{AD}(A \mid t\,g\,t)}^{4\text{-gram}}$$

$0 < K < 1$

λ set to make this normalize

$K = 0.2$

# of word types seen in this context times K

$$\lambda = \frac{0.6}{4}$$

want to go to → Maui 2 → 1.8

want to go to → Class 1 → 0.8

want to go to → campus 1 → 0.8

count = 4

$$P_{AD}(A \mid t\,g\,t) = \cdots + \lambda' P_{AD}(A \mid g\,t)$$

$$\cdots \quad P(A) > 0$$

Kneser-Ney smoothing