Language model unit:

- Probability and statistics background (last week Wednesday)
- Statistics about language (last week Friday)
- Language models themselves (Monday)
- Smoothing language models (Friday)
- Interpolation among language models (Friday)
- (Finish and apply language models next week)

Today:

- Review of language models and perplexity
- The need for smoothing
- Laplace smoothing
- Good-Turing smoothing

Basic symbols

V: Vocabulary (set of types) or its size

N: size of the training text (number of tokens)

$C(w)$: count (frequency) of $w \in V$

K: size of test text

Sanity checks:

$\sum_{w \in V} C(w) = N$

$\sum_{w \in V} P(w) = 1$

Perplexity:

$$\left(\prod_{i=1}^{K} P(w_i \mid h)\right)^{\frac{-1}{K}} = \sqrt[K]{\frac{1}{\prod_{i=1}^{K} P(w_i \mid h)}} = 2^{-\frac{1}{K}\sum_{i=1}^{K} \lg P(w_i \mid h)} = B^{-\frac{1}{K}\sum_{i=1}^{K} \log_B P(w_i \mid h)}$$

A perplexity of k means that you are as surprised on average as you would have been if you had had to guess between [sic] k equiprobable choices at each step.

Manning and Shütze, Foundations of Statistical Natural Language Processing, pg 78

Maximum Likelihood Fallacy:

The soldier with the green whiskers led them through the streets of the Emerald City until they reached the room where the Guardian of the Gates lived. This officer unlocked their spectacles to put them back in his great box, and then he politely opened the gate for our friends.

"Which road leads to the Wicked Witch of the West?" asked Dorothy.

"There is no road," answered the Guardian of the Gates. "No one ever wishes to go that way."

"How, then, are we to find her?" inquired the girl.

"That will be easy," replied the man, "for when she knows you are in the country of the Winkies she will find you, and make you all her slaves."

"Perhaps not," said the Scarecrow, "for we mean to destroy her."

"Oh, that is different," said the Guardian of the Gates. "No one has ever destroyed her before, so I naturally thought she would make slaves of you, as she has of the rest. "

F Baum, The Wonderful Wizard of Oz

Maximum Likelihood Fallacy:

I know of no way of judging the future but by the past.

Attributed to Patrick Henry

Prediction is very difficult, especially if it's about the future.

Attributed alternately to Niels Bohr or Yogi Bera

Model families:

Maximum likelihood estimation

$$P_{RF}(w) = \frac{C(w)}{N}$$

Lapace (add-one) smoothing

$$P_L(w) = \frac{C(w)+1}{N+V} = \frac{C_L^*(w)}{N}$$

$$C_L^*(w) = (C(w)+1) \cdot \frac{N}{N+V}$$

Constant

$$P_{\text{const}}(w) = \frac{1}{V} = \frac{C_{\text{const}}^*(w)}{N}$$

$$C_{\text{const}}^*(w) = \frac{N}{V}.$$