Smoothing.
What if we see new n-grams?

Zero probabilities for test data

Toy train corpus:
This is the house that Jack built.

Toy test corpus:
This is *Jack.*

What's the perplexity of the Bigram LM?

$$p(Jack \mid is) = \frac{c(is\ Jack)}{c(is)} = 0$$

$$p(\mathbf{w}_{\text{test}}) = 0$$

$$\mathcal{P} = \inf$$

Laplacian smoothing

Idea:

- Pull some probability from frequent bigrams to infrequent ones
- Just add 1 to the counts (add-one smoothing):

$$\hat{p}(w_i|w_{i-n+1}^{i-1}) = \frac{c(w_{i-n+1}^{i}) + 1}{c(w_{i-n+1}^{i-1}) + V}$$

- Or tune a parameter (add-k smoothing):

$$\hat{p}(w_i|w_{i-n+1}^{i-1}) = \frac{c(w_{i-n+1}^{i}) + k}{c(w_{i-n+1}^{i-1}) + Vk}$$

Katz backoff

Problem:

- Longer n-grams are better, but data is not always enough

Idea:

- Try a longer n-gram and back off to shorter if needed

$$\hat{p}(w_i|w_{i-n+1}^{i-1}) = \begin{cases} \tilde{p}(w_i|w_{i-n+1}^{i-1}), & \text{if } c(w_{i-n+1}^{i}) > 0 \\ \alpha(w_{i-n+1}^{i-1})\,\hat{p}(w_i|w_{i-n+2}^{i-1}), & \text{otherwise} \end{cases}$$

Katz backoff

Problem:

- Longer n-grams are better, but data is not always enough

Idea:

- Try a longer n-gram and back off to shorter if needed

$$\hat{p}(w_i|w_{i-n+1}^{i-1}) = \begin{cases} \tilde{p}(w_i|w_{i-n+1}^{i-1}), & \text{if } c(w_{i-n+1}^{i}) > 0 \\ \alpha(w_{i-n+1}^{i-1})\,\hat{p}(w_i|w_{i-n+2}^{i-1}), & \text{otherwise} \end{cases}$$

where $\tilde{p}$ and α are chosen to ensure normalization.

Interpolation smoothing

Idea:

- Let us have a mixture of several n-gram models
- Example for a trigram model:

$$\hat{p}(w_i|w_{i-2}w_{i-1}) = \lambda_1 p(w_i|w_{i-2}w_{i-1}) + \lambda_2 p(w_i|w_{i-1}) + \lambda_3 p(w_i)$$

$$\lambda_1 + \lambda_2 + \lambda_3 = 1$$

- The weights are optimized on a test (dev) set
- Optionally they can also depend on the context

Absolute discounting

Idea:

- Let's compare the counts for bigrams in train and test sets

Experiment (Church and Gale, 1991):

- Subtract 0.75 and get a good estimate for the test count!

Train bigram count	Test bigram count
2	1.25
3	2.24
4	3.23
5	4.21
6	5.23
7	6.21
8	7.21

https://web.stanford.edu/~jurafsky/slp3/4.pdf

Absolute discounting

Idea:

- Let's compare the counts for bigrams in train and test sets

Experiment (Church and Gale, 1991):

- Subtract 0.75 and get a good estimate for the test count!

$$\hat{p}(w_i|w_{i-1}) = \frac{c(w_{i-1}w_i) - d}{\sum_x c(w_{i-1}x)} + \lambda(w_{i-1})p(w_i)$$

Kneser-Ney smoothing

Idea:

- The unigram distribution captures the word frequency
- We need to capture the diversity of contexts for the word

$$\hat{p}(w) \propto |x : c(x\,w) > 0|$$

malt

This is the …

Kong

- Probably, the most popular smoothing technique

https://web.stanford.edu/~jurafsky/slp3/4.pdf

Resume

Smoothing techniques:

- Add-one (add-k) smoothing
- Katz backoff
- Interpolation smoothing
- Absolute discounting
- Kneser-Ney smoothing

N-gram models + Kneser-Ney smoothing is a strong baseline in Language Modeling!