Natural Language Processing

Lecture 7—9/19/2013

Jim Martin

Today

- More Language modeling (N-grams)
  - Smoothing
    - Finish Good-Turing
    - Pretty good smoothing
      - Bayesian prior smoothing
- Word classes
  - Part of speech tagging

Smoothing
Dealing w/ Zero Counts

- Back to Shakespeare
  - Recall that Shakespeare produced 300,000 bigram types out of V^2= 844 million possible bigrams...
  - So, 99.96% of the possible bigrams were never seen (have zero entries in the table)
  - Does that mean that any sentence that contains one of those bigrams should have a probability of 0?
  - For generation (shannon game) it means we' ll never emit those bigrams
  - But for analysis it' s problematic because if we run across a new bigram in the future then we have no choice but to assign it a probability of zero..

Zero Counts

- Some of those zeros are really zeros...
  - Things that really aren' t ever going to happen
    - Fewer of these than you might think
- On the other hand, some of them are just rare events.
  - If the training corpus had been a little bigger they would have had a count
    - What would that count be in all likelihood?

Zero Counts

- Zipf's Law (long tail phenomenon)
  - A small number of events occur with high frequency
  - A large number of events occur with low frequency
  - You can quickly collect statistics on the high frequency events
  - You might have to wait an arbitrarily long time to get good statistics on low frequency events
- Result
  - Our estimates are necessarily sparse! We have no counts at all for the vast number of events we want to estimate.
- Answer
  - Estimate the likelihood of unseen (zero count) N-grams!

Laplace Smoothing

- Also called Add-One smoothing
- Just add one to all the counts!
- Very simple

- MLE estimate:

$$P(w_i) = \frac{c_i}{N}$$

- Laplace estimate:

$$P_{\text{Laplace}}(w_i) = \frac{c_i + 1}{N + V}$$

- Reconstructed counts:

$$c_i^* = (c_i + 1)\frac{N}{N + V}$$