CS474 Natural Language Processing

- Smoothing
  - Add-one
  - Discounting
- Combining estimators
  - Linear interpolation
  - Backoff
- Training issues

Language models: n-grams

- I’d like to make a collect ____
- to make a collect call
- make a collect call
- a collect call

- Markov assumption: only the prior local context --- the last few words --- matters
  \[ P(w_n \mid w_1^{n-1}) \approx P(w_n \mid w_{n-N+1}^{n-1}) \]

Training N-gram models

- N-gram models can be trained by counting and normalizing
  \[ P(\text{call} \mid \text{a collect}) = \frac{\text{Count(}a \text{ collect}\text{ call})}{\text{Count(}a \text{ collect})} \]
- MLE estimates from relative frequencies
- Bigram model
  \[ P(w_n \mid w_1^{n-1}) = \frac{C(w_{n-1}, w_n)}{C(w_{n-1})} \]

Sparse data

- Problem with the maximum likelihood estimate: sparse data

ATIS corpus (~500 sentences, ~400 words):
Show me flights from Boston to Chicago
I need to return on Tuesday
I would like to travel to Westchester
Sparse data (counts)

<table>
<thead>
<tr>
<th></th>
<th>Show</th>
<th>me</th>
<th>flights</th>
<th>from</th>
<th>Boston</th>
<th>to</th>
<th>Chicago</th>
</tr>
</thead>
<tbody>
<tr>
<td>Show</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>me</td>
<td>0</td>
<td>0</td>
<td>14</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>flights</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>from</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Boston</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>to</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Chicago</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
</tbody>
</table>

Sparse data (probabilities)

<table>
<thead>
<tr>
<th></th>
<th>Show</th>
<th>me</th>
<th>flights</th>
<th>from</th>
<th>Boston</th>
<th>to</th>
<th>Chicago</th>
</tr>
</thead>
<tbody>
<tr>
<td>Show</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>me</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>flights</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>from</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Boston</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>to</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Chicago</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Smoothing

- Need better estimators for rare events
- Approach
  - Somewhat decrease the probability of previously seen events, so that there is a little bit of probability mass left over for previously unseen events

Add-one smoothing

- Add one to all of the counts before normalizing into probabilities
- Normal unigram probabilities
  \[ P(w_x) = \frac{C(w_x)}{N} \]
- Smoothed unigram probabilities
  \[ P(w_x) = \frac{C(w_x) + 1}{N + V} \]
+1 Smoothed ATIS (counts)

<table>
<thead>
<tr>
<th>Show</th>
<th>me</th>
<th>flights</th>
<th>from</th>
<th>Boston</th>
<th>to</th>
<th>Chicago</th>
</tr>
</thead>
<tbody>
<tr>
<td>Show</td>
<td>1</td>
<td>93</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>me</td>
<td>1</td>
<td>1</td>
<td>15</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>flights</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>97</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>from</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Boston</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>to</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Chicago</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

+1 Smoothed ATIS (probabilities)

<table>
<thead>
<tr>
<th>Show</th>
<th>me</th>
<th>flights</th>
<th>from</th>
<th>Boston</th>
<th>to</th>
<th>Chicago</th>
</tr>
</thead>
<tbody>
<tr>
<td>Show</td>
<td>0.0019</td>
<td>0.1782</td>
<td>0.0134</td>
<td>0.0019</td>
<td>0.0019</td>
<td>0.0019</td>
</tr>
<tr>
<td>me</td>
<td>0.0019</td>
<td>0.0019</td>
<td>0.0292</td>
<td>0.0019</td>
<td>0.0019</td>
<td>0.0019</td>
</tr>
<tr>
<td>flights</td>
<td>0.0016</td>
<td>0.0016</td>
<td>0.0016</td>
<td>0.1557</td>
<td>0.0016</td>
<td>0.0016</td>
</tr>
<tr>
<td>from</td>
<td>0.0015</td>
<td>0.0015</td>
<td>0.0015</td>
<td>0.0015</td>
<td>0.0073</td>
<td>0.0015</td>
</tr>
<tr>
<td>Boston</td>
<td>0.0024</td>
<td>0.0024</td>
<td>0.0024</td>
<td>0.0024</td>
<td>0.0118</td>
<td>0.0024</td>
</tr>
<tr>
<td>to</td>
<td>0.0014</td>
<td>0.0014</td>
<td>0.0014</td>
<td>0.0014</td>
<td>0.0056</td>
<td>0.0042</td>
</tr>
<tr>
<td>Chicago</td>
<td>0.0024</td>
<td>0.0024</td>
<td>0.0024</td>
<td>0.0024</td>
<td>0.0118</td>
<td>0.0024</td>
</tr>
</tbody>
</table>

Too much probability mass is moved

- Estimated bigram frequencies
- AP data, 44million words
- Church and Gale (1991)
- In general, add-one smoothing is a poor method of smoothing
- Much worse than other methods in predicting the actual probability for unseen bigrams
- Variances of the counts are worse than those from the unsmoothed MLE method

Aside: Methodology

- Cardinal sin: Testing on the training corpus
- Divide data into training set and test set
  - Train the statistical parameters on the training set; use them to compute probabilities on the test set
  - Test set: 5-10% of the total data, but large enough for reliable results
- Divide training into training and validation/held out set
  » Obtain counts from training
  » Tune smoothing parameters on the validation set
- Divide test set into development and final test set
  - Do all algorithm development by testing on the dev set, save the final test set for the very end…
Back to smoothing: solutions

- **Discounting**
  - Better estimates for how much probability to siphon away for unseen words
  - Use higher frequency words to estimate mass of lower frequency words
  - See book: Witten-Bell, Good-Turing (& many more)

- **Combining estimators...**

Combining estimators

- **Discounting methods**
  - Provide the same estimate for all unseen (or rare) n-grams
  - Make use only of the raw frequency of an n-gram

- But there is an additional source of knowledge we can draw on --- the n-gram “hierarchy”
  - If I haven’t seen a collect call, maybe I’ve seen collect call
  - ... collect call call

- For n-gram models, suitably combining various models of different orders is the secret to success.

Simple linear interpolation

- Construct a linear combination of the multiple probability estimates.
  - Weight each contribution so that the result is another probability function.

$$P(w_n | w_{n-1}, w_{n-2}) = \lambda_3 P(w_n | w_{n-1}, w_{n-2}) + \lambda_2 P(w_n | w_{n-1}) + \lambda_1 P(w_n)$$
  - \(\lambda\)s sum to 1.
  - \(\lambda\)s trained on validation set

Backoff (Katz 1987)

- **(this is a lie)**
- Want \(P(\text{call}|a \text{ collect})\)

  - **Seen a collect call?**
    - Yes! Use \(P(\text{call}|\text{a collect})\)
    - No \(\oplus\) Use \(P(\text{call})\)

  - **Seen collect call?**
    - Yes! Use \(P(\text{call}|\text{collect})\)
    - No \(\oplus\) Use \(P(\text{call})\)
Backoff: details

- Ps need to sum to 1!
- Discount each MLE prob (W-B, G-T, ...)
- Apportion the saved mass to lower-orders

\[
\hat{P}(w_n | w_{n-N+1}^{n-1}) = \tilde{P}(w_n | w_{n-N+1}^{n-1}) + \\
\theta (P(w_n | w_{n-N+1}^{n-1})) \alpha(w_{n-N+1}^{n-1}) \hat{P}(w_n | w_{n-N+2}^{n-1})
\]