Chinese Restaurants and Backoff

Natural Language Processing: Jordan Boyd-Graber
University of Colorado Boulder
SEPTEMBER 10, 2014
After this class, you’ll be able to:

- Understand probability distributions through the metaphor of the Chinese Restaurant Process
- Be able to calculate Kneser-Ney smoothing
- Understand the role of contexts in language models
Intuition

- Some words are “sticky”
- “San Francisco” is very common (high ungram)
- But Francisco only appears after one word
Intuition

- Some words are “sticky”
- “San Francisco” is very common (high ungram)
- But Francisco only appears after one word
- Our goal: to tell a statistical story of bay area restaurants to account for this phenomenon
How does a CRP encode a probability distribution?

How do many CRPs encode backoff?

Language Model Probabilities
How does a crp encode a probability distribution?

Let’s remember what a language model is

- It is a distribution over the next word in a sentence
- Given the previous $n-1$ words
Let’s remember what a language model is

- It is a distribution over the next word in a sentence
- Given the previous \( n - 1 \) words
- The challenge: backoff and sparsity
The Chinese Restaurant as a Distribution

To generate a word, you first sit down at a table. You sit down at a table proportional to the number of people sitting at the table.
How does a crp encode a probability distribution?

The Chinese Restaurant as a Distribution

To generate a word, you first sit down at a table. You sit down at a table proportional to the number of people sitting at the table.

2/7

3/7

2/7

Natural Language Processing: Jordan Boyd-Graber | Boulder

Chinese Restaurants and Backoff | 6 of 18
How does a crp encode a probability distribution?

The Chinese Restaurant as a Distribution

To generate a word, you first sit down at a table. You sit down at a table proportional to the number of people sitting at the table.
The Chinese Restaurant as a Distribution

To generate a word, you first sit down at a table. You sit down at a table proportional to the number of people sitting at the table.

\[
\begin{align*}
&\text{dog} & \frac{2}{7} \\
&\text{cat} & \frac{3}{7} \\
&\text{purple} & \frac{2}{7}
\end{align*}
\]
The Chinese Restaurant as a Distribution

To generate a word, you first sit down at a table. You sit down at a table proportional to the number of people sitting at the table.

- $\frac{2}{7}$ dog
- $\frac{3}{7}$ cat
- $\frac{2}{7}$ purple

But this is just Maximum Likelihood

Why are we talking about Chinese Restaurants?
How does a crp encode a probability distribution?

Always one more table . . .
How does a crp encode a probability distribution?

Always one more table . . .

\[
\begin{align*}
\frac{2}{7+\alpha} & \\
\frac{3}{7+\alpha} & \\
\frac{2}{7+\alpha} & \\
\frac{\alpha}{7+\alpha} & 
\end{align*}
\]
How does a crp encode a probability distribution?

**Always one more table . . .**

\[
\begin{align*}
\frac{2}{7+\alpha} & \quad \text{dog} \\
\frac{3}{7+\alpha} & \quad \text{cat} \\
\frac{2}{7+\alpha} & \quad \text{purple} \\
\frac{\alpha}{7+\alpha} & \quad ???
\end{align*}
\]
How does a crp encode a probability distribution?

**Always one more table . . .**

\[
\begin{align*}
\frac{2}{7+\alpha} & \quad \text{dog} \\
\frac{3}{7+\alpha} & \quad \text{cat} \\
\frac{2}{7+\alpha} & \quad \text{purple} \\
\frac{\alpha}{7+\alpha} & \quad ???
\end{align*}
\]
How does a crp encode a probability distribution?

What to do with a new table?

- Uniform (Dirichlet smoothing)
- Specific contexts
  - less-specific contexts (backoff)
How does a crp encode a probability distribution?

What to do with a new table?

What can be a base distribution?

- Uniform (Dirichlet smoothing)
How does a crp encode a probability distribution?

What to do with a new table?

What can be a base distribution?

- Uniform (Dirichlet smoothing)
- Specific contexts $\rightarrow$ less-specific contexts (backoff)
How do many CRPs encode backoff?

Outline

How does a CRP encode a probability distribution?

How do many CRPs encode backoff?

Language Model Probabilities
How do many CRPs encode backoff?

A hierarchy of Chinese Restaurants
Seating Assignments
Dataset:

<s> a a a b a c </s>
How do many crps encode backoff?

Seating Assignments

Dataset:

<\s> a a a b a c </\s>

Unigram Restaurant

<\s> Restaurant

a Restaurant

b Restaurant

c Restaurant
How do many crps encode backoff?

Seating Assignments
Dataset:

<s> a a a b a c </s>

Unigram Restaurant

<s> Restaurant

* 1

b Restaurant

a Restaurant

c Restaurant
How do many crps encode backoff?

Seating Assignments
Dataset:

<s> a a a b a c </s>

Unigram Restaurant

<star> 1

<s> Restaurant

<star> 1

a Restaurant
c Restaurant

b Restaurant
How do many crps encode backoff?

Seating Assignments
Dataset:

```
<s> a a a b a c </s>
```

<table>
<thead>
<tr>
<th>Unigram Restaurant</th>
<th>Restaurant</th>
</tr>
</thead>
<tbody>
<tr>
<td>a 1</td>
<td>a Restaurant</td>
</tr>
<tr>
<td>b Restaurant</td>
<td>c Restaurant</td>
</tr>
</tbody>
</table>

11 of 18
How do many crps encode backoff?

Seating Assignments
Dataset:

<s> a a a b a c </s>

Unigram Restaurant

- a $^{1}$

Unigram Restaurant

- a $^{1}$

- * $^{1}$

b Restaurant
c Restaurant
How do many crps encode backoff?

Seating Assignments
Dataset:

<s> a a a b a c </s>

Unigram Restaurant

<table>
<thead>
<tr>
<th>Restaurant</th>
<th>a</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>1</td>
</tr>
</tbody>
</table>

<s> Restaurant

<table>
<thead>
<tr>
<th>Restaurant</th>
<th>a</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Restaurant</th>
<th>*</th>
</tr>
</thead>
<tbody>
<tr>
<td>*</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Restaurant</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>b</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Restaurant</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>c</td>
<td>1</td>
</tr>
</tbody>
</table>
How do many CRPs encode backoff?

Seating Assignments

Dataset:

\[
<s> \text{a a a b a c} \langle/s> 
\]

<table>
<thead>
<tr>
<th>Unigram Restaurant</th>
<th>Restaurant a</th>
<th>Restaurant b</th>
<th>Restaurant c</th>
</tr>
</thead>
<tbody>
<tr>
<td>a ( ^2 )</td>
<td>a ( ^1 )</td>
<td>a ( ^1 )</td>
<td>c Restaurant</td>
</tr>
</tbody>
</table>
Seating Assignments

Dataset:

<s> a a a b a c </s>

Unigram Restaurant

<table>
<thead>
<tr>
<th>Restaurant</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>2</td>
</tr>
</tbody>
</table>

<s> Restaurant

<table>
<thead>
<tr>
<th>Restaurant</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>1</td>
</tr>
</tbody>
</table>

b Restaurant

c Restaurant
How do many crps encode backoff?

Seating Assignments

Dataset:

\[
\begin{array}{cccccc}
<s> & a & a & a & b & a & c \end{array}
\]

<table>
<thead>
<tr>
<th>Unigram Restaurant</th>
<th>&lt;s&gt; Restaurant</th>
<th>a Restaurant</th>
<th>b Restaurant</th>
<th>c Restaurant</th>
</tr>
</thead>
<tbody>
<tr>
<td>a ( ^2 )</td>
<td>a ( ^1 )</td>
<td>a ( ^2 )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

How do many crps encode backoff?

Seating Assignments

Dataset:

\(<s>\ a\ a\ a\ b\ a\ c\ </s>\)

Unigram Restaurant

- Restaurant
  - a
  - b
  - c

a Restaurant

- a
- *

b Restaurant

c Restaurant
Seating Assignments
Dataset:

<s> a a a b a c </s>

Unigram Restaurant

<table>
<thead>
<tr>
<th>Restaurant</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>2</td>
</tr>
<tr>
<td>*</td>
<td>1</td>
</tr>
</tbody>
</table>

<s> Restaurant

<table>
<thead>
<tr>
<th>Restaurant</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>1</td>
</tr>
</tbody>
</table>

b Restaurant

<table>
<thead>
<tr>
<th>Restaurant</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>2</td>
</tr>
<tr>
<td>*</td>
<td>1</td>
</tr>
</tbody>
</table>

c Restaurant
How do many crps encode backoff?

Seating Assignments
Dataset:

<s> a a a b a c </s>

Unigram Restaurant

a²  b¹

Restaurant

a¹

b Restaurant

a²  *¹

c Restaurant
**Seating Assignments**

**Dataset:**

$$\text{<s> a a a b a c </s>}$$

<table>
<thead>
<tr>
<th>Unigram Restaurant</th>
</tr>
</thead>
<tbody>
<tr>
<td>a (2)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>&lt;s&gt; Restaurant</th>
</tr>
</thead>
<tbody>
<tr>
<td>a (1)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>a Restaurant</th>
</tr>
</thead>
<tbody>
<tr>
<td>a (2)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>b Restaurant</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>c Restaurant</th>
</tr>
</thead>
</table>

How do many CRPs encode backoff?

Seating Assignments

Dataset:

<s> a a a b a c </s>

Unigram Restaurant

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

<s> Restaurant

<table>
<thead>
<tr>
<th>a</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
</tbody>
</table>

a Restaurant

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

b Restaurant

c Restaurant
How do many crps encode backoff?

Seating Assignments

Dataset:

\[ <s> \, a \, a \, a \, b \, a \, c \, </s> \]

Unigram Restaurant

\[
\begin{array}{c}
\text{a}^2 \\
\text{b}^1
\end{array}
\]

<s> Restaurant

\[
\begin{array}{c}
\text{a}^1
\end{array}
\]

b Restaurant

\[
\begin{array}{c}
\ast^1
\end{array}
\]

a Restaurant

\[
\begin{array}{c}
\text{a}^2 \\
\text{b}^1
\end{array}
\]

c Restaurant
How do many crps encode backoff?

Seating Assignments
Dataset:

<s> a a a b a c </s>

Unigram Restaurant

<s> Restaurant

\begin{array}{c}
\text{a}^1 \\
\text{b}^1
\end{array}

\begin{array}{c}
\text{a}^2 \\
\text{b}^1
\end{array}

\begin{array}{c}
\text{a}^1 \\
\text{b}^1
\end{array}

\begin{array}{c}
\text{a}^2 \\
\text{b}^1
\end{array}

\begin{array}{c}
\text{*}^1 \\
\text{c}^1
\end{array}

\end{array}

Natural Language Processing: Jordan Boyd-Graber | Boulder
Chinese Restaurants and Backoff | 11 of 18
Seating Assignments

Dataset:

\[
\langle s \rangle \ a \ a \ a \ b \ a \ c \ \langle /s \rangle
\]

Unigram Restaurant

- \( a \): 3
- \( b \): 1

Restaurant

- \( a \): 1

Chinese Restaurants and Backoff
How do many crps encode backoff?

Seating Assignments

Dataset:

<s> a a a b a c </s>

Unigram Restaurant

| a 3  |
| b 1 |

Restaurant

| a 1 |

Restaurant

| a 2  |
| b 1 |

c Restaurant

| a 1 |

Natural Language Processing: Jordan Boyd-Graber | Boulder
How do many crps encode backoff?

Seating Assignments
Dataset:

<s> a a a b a c </s>

Unigram Restaurant

| a | 3 |
| b | 1 |

Restaurant

| a | 1 |

Restaurant

| a | 2 |
| b | 1 |
| * | 1 |

Restaurant

| a | 1 |

Restaurant

| a | 1 |
Seating Assignments
Dataset:

<s> a a a b a c </s>

Unigram Restaurant

Unigram Restaurant

Unigram Restaurant

Unigram Restaurant

Unigram Restaurant

Unigram Restaurant

Unigram Restaurant

Unigram Restaurant

Unigram Restaurant

Unigram Restaurant

Unigram Restaurant
How do many crps encode backoff?

Seating Assignments
Dataset:

<s> a a a b a c </s>

Unigram Restaurant

```
<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>b</td>
<td>c</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
```

<s> Restaurant

```
<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>
```

b Restaurant

```
<p>| |</p>
<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
</tr>
<tr>
<td>1</td>
</tr>
</tbody>
</table>
```

c Restaurant

```
<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>b</td>
<td>c</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
```
Seating Assignments

Dataset:

<s> a a a b a c </s>

Unigram Restaurant

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Restaurant

<s> Restaurant

<table>
<thead>
<tr>
<th>a</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
</tbody>
</table>

b Restaurant

<table>
<thead>
<tr>
<th>a</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
</tbody>
</table>

a Restaurant

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

c Restaurant
How many crps encode backoff?

Seating Assignments
Dataset:

<s> a a a b a c </s>

Unigram Restaurant

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>b</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>c</td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

<s> Restaurant

<table>
<thead>
<tr>
<th></th>
<th>a</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>1</td>
</tr>
</tbody>
</table>

b Restaurant

<table>
<thead>
<tr>
<th></th>
<th>a</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>1</td>
</tr>
</tbody>
</table>

da Restaurant

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>2</td>
<td>b</td>
<td>1</td>
</tr>
<tr>
<td>c</td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>
c Restaurant

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>*</td>
<td>1</td>
</tr>
</tbody>
</table>
How do many crps encode backoff?

Seating Assignments

Dataset:

<s> a a a b a c </s>

Unigram Restaurant

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>c</th>
<th>*</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Restaurant

<s> Restaurant

<table>
<thead>
<tr>
<th>a</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
</tbody>
</table>

Restaurant

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Restaurant

<table>
<thead>
<tr>
<th>a</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
</tbody>
</table>

Restaurant

<table>
<thead>
<tr>
<th>*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
</tbody>
</table>
How do many crps encode backoff?

Seating Assignments
Dataset:

<s> a a a b a c </s>

Unigram Restaurant

Restaurant

b Restaurant

c Restaurant
Outline

How does a CRP encode a probability distribution?

How do many CRPs encode backoff?

Language Model Probabilities
The rich get richer

\[ \frac{2}{5+\theta} \quad \frac{3}{5+\theta} \quad \frac{\theta}{5+\theta} \]
Computing the Probability of an Observation

\[
p(w = x | \tilde{s}, \theta, u) = \frac{c_{u,x}}{\theta + c_{u,x}} \cdot \text{existing table} + \frac{\theta}{\theta + c_{u,x}} p(w = x | \tilde{s}, \theta, \pi(u)) \tag{1}
\]

- Word type \( x \)
- Seating assignments \( \tilde{s} \)
- Concentration \( \theta \)
- Context \( u \)
- Number seated at table serving \( x \) in restaurant \( u, c_{u,x} \)
- Number seated at all tables in restaurant \( u, c_{u,x} \)
- The backoff context \( \pi(u) \)
Computing the Probability of an Observation

\[
p(w = x|\vec{s}, \theta, u) = \frac{c_{u,x}}{\theta + c_{u,\cdot}} + \frac{\theta}{\theta + c_{u,\cdot}} p(w = x|\vec{s}, \theta, \pi(u))
\]

1. Word type \(x\)
2. Seating assignments \(\vec{s}\)
3. Concentration \(\theta\)
4. Context \(u\)
5. Number seated at table serving \(x\) in restaurant \(u\), \(c_{u,x}\)
6. Number seated at all tables in restaurant \(u\), \(c_{u,\cdot}\)
7. The backoff context \(\pi(u)\)
Computing the Probability of an Observation

\[
p(w = x|\tilde{s}, \theta, u) = \frac{c_{u,x}}{\theta + c_{u,.}} + \frac{\theta}{\theta + c_{u,.}} p(w = x|\tilde{s}, \theta, \pi(u)) \tag{1}
\]

- Word type \(x\)
- Seating assignments \(\tilde{s}\)
- Concentration \(\theta\)
- Context \(u\)
- Number seated at table serving \(x\) in restaurant \(u, c_{u,x}\)
- Number seated at all tables in restaurant \(u, c_{u,.}\)
- The backoff context \(\pi(u)\)
Computing the Probability of an Observation

\[ p(w = x|\tilde{s}, \theta, u) = \frac{c_{u,x}}{\theta + c_{u,u}} + \frac{\theta}{\theta + c_{u,u}} p(w = x|\tilde{s}, \theta, \pi(u)) \]  \hspace{1cm} (1)

- Word type $x$
- Seating assignments $\tilde{s}$
- Concentration $\theta$
- Context $u$
- Number seated at table serving $x$ in restaurant $u$, $c_{u,x}$
- Number seated at all tables in restaurant $u$, $c_{u,u}$
- The backoff context $\pi(u)$
Computing the Probability of an Observation

\[
p(w = x|\vec{s}, \theta, u) = \frac{c_{u,x}}{\theta + c_u} + \frac{\theta}{\theta + c_u} p(w = x|\vec{s}, \theta, \pi(u)) \quad (1)
\]

- Word type \(x\)
- Seating assignments \(\vec{s}\)
- Concentration \(\theta\)
- Context \(u\)
- Number seated at table serving \(x\) in restaurant \(u\), \(c_{u,x}\)
- Number seated at all tables in restaurant \(u\), \(c_u\)
- The backoff context \(\pi(u)\)
Computing the Probability of an Observation

\[ p(w = x|\tilde{s}, \theta, u) = \frac{c_{u,x}}{\theta + c_{u,x}} \cdot \text{existing table} + \frac{\theta}{\theta + c_{u,x}} \cdot p(w = x|\tilde{s}, \theta, \pi(u)) \]  

\[ (1) \]

- Word type \( x \)
- Seating assignments \( \tilde{s} \)
- Concentration \( \theta \)
- Context \( u \)
- Number seated at table serving \( x \) in restaurant \( u, c_{u,x} \)
- Number seated at all tables in restaurant \( u, c_{u,x} \)
- The backoff context \( \pi(u) \)
Computing the Probability of an Observation

\[ p(w = x|\tilde{s}, \theta, u) = \frac{c_{u,x}}{\theta + c_{u,x}} + \frac{\theta}{\theta + c_{u,x}} p(w = x|\tilde{s}, \theta, \pi(u)) \] (1)

- Word type \( x \)
- Seating assignments \( \tilde{s} \)
- Concentration \( \theta \)
- Context \( u \)
- Number seated at table serving \( x \) in restaurant \( u, c_{u,x} \)
- Number seated at all tables in restaurant \( u, c_{u,.} \)
- The backoff context \( \pi(u) \)
**Example:** \( p(w = b|\vec{s}, \theta = 1.0, u = a) \)

\[ p(w = b|\ldots) = \frac{c_{a,b}}{\theta + c_{u}} + \frac{\theta}{\theta + c_{u}} p(w = x|\vec{s}, \theta, \pi(u)) \] (2)
Example: \( p(w = b|\vec{s}, \theta = 1.0, u = a) \)

Unigram Restaurant

\[
\begin{array}{cccc}
  a & b & c & </s> \\
  3 & 1 & 1 & 1 \\
\end{array}
\]

\(<s>\) Restaurant

\[
\begin{array}{c}
  a \\
  1 \\
\end{array}
\]

\(b\) Restaurant

\[
\begin{array}{c}
  a \\
  1 \\
\end{array}
\]

\(c\) Restaurant

\[
\begin{array}{c}
  </s> \\
  1 \\
\end{array}
\]

\[
p(w = b|\ldots) = \frac{c_{a,b}}{\theta + c_{u,1}} + \frac{\theta}{\theta + c_{u,1}} p(w = x|\vec{s}, \theta, \pi(u))
\]  

(2)
Example: \( p(w = b|\vec{s}, \theta = 1.0, u = a) \)

Unigram Restaurant

\[
\begin{array}{c}
\text{a}^3 \text{ b}^1 \text{ c}^1 \langle/s\rangle^1
\end{array}
\]

<\text{s}> Restaurant

\[
\begin{array}{c}
\text{a}^1
\end{array}
\]

c Restaurant

\[
\begin{array}{c}
\text{a}^1
\end{array}
\]

\[
p(w = b|\ldots) = \frac{1}{\theta + c_u,} + \frac{\theta}{\theta + c_u,} p(w = x|\vec{s}, \theta, \pi(u)) \quad (2)
\]
**Example:**  \( p(w = b | \vec{s}, \theta = 1.0, u = a) \)

<table>
<thead>
<tr>
<th>Representative Sentence</th>
<th>Unigram Restaurant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Restaurant a</td>
<td>[ a^3 b^1 c^1 ]</td>
</tr>
<tr>
<td>Restaurant b</td>
<td>[ a^1 ]</td>
</tr>
<tr>
<td>Restaurant c</td>
<td>[ a^1 ]</td>
</tr>
</tbody>
</table>

\[
p(w = b | \ldots ) = \frac{1}{1.0 + c_u} + \frac{1.0}{1.0 + c_u} p(w = x | \vec{s}, \theta, \pi(u)) \quad (2)
\]
Example: \( p(w = b | \vec{s}, \theta = 1.0, u = a) \)

Unigram Restaurant

\[
\begin{array}{cccc}
\text{a} & 3 & \text{b} & 1 \\
\text{c} & 1 & \text{}</s> & 1 \\
\end{array}
\]

Unigram Restaurant

\[
\begin{array}{c}
\text{a} & 1 \\
\end{array}
\]

Unigram Restaurant

\[
\begin{array}{c}
\text{a} & 1 \\
\end{array}
\]

Unigram Restaurant

\[
\begin{array}{c}
\text{}</s> & 1 \\
\end{array}
\]

Unigram Restaurant

\[
\begin{array}{ccc}
a & 2 & \text{b} & 1 \\
\text{c} & 1 \\
\end{array}
\]

Unigram Restaurant

\[
\begin{array}{c}
a & 1 \\
\end{array}
\]

Unigram Restaurant

\[
\begin{array}{c}
\text{}</s> & 1 \\
\end{array}
\]

\[
p(w = b | \ldots) = \frac{1}{1.0 + 4} + \frac{1.0}{1.0 + 4} p(w = x | \vec{s}, \theta, \pi(u)) \tag{2}
\]
Example: \( p(w = b | \vec{s}, \theta = 1.0, u = a) \)

Unigram Restaurant

\[
\begin{array}{ccc}
\text{a}^3 & \text{b}^1 & \text{c}^1 & \langle /s \rangle^1 \\
\end{array}
\]

<\text{s}> Restaurant

\[
\begin{array}{c}
\text{a}^1 \\
\end{array}
\]

a Restaurant

\[
\begin{array}{ccc}
\text{a}^2 & \text{b}^1 & \text{c}^1 \\
\end{array}
\]

b Restaurant

\[
\begin{array}{c}
\text{a}^1 \\
\end{array}
\]

c Restaurant

\[
\begin{array}{c}
\langle /s \rangle^1 \\
\end{array}
\]

\[
p(w = b | \ldots ) = \frac{1}{1.0 + 4} + \frac{1.0}{1.0 + 4} p(w = x | \vec{s}, \theta, \pi(u)) \tag{2}
\]
**Example:** \( p(w = b | \vec{s}, \theta = 1.0, u = a) \)

<table>
<thead>
<tr>
<th>Unigram Restaurant</th>
<th>Unigram Restaurant</th>
</tr>
</thead>
<tbody>
<tr>
<td>a(^3) b(^1) c(^1) (&lt;!!!/&gt;!)(^1)</td>
<td>a(^2) b(^1) c(^1)</td>
</tr>
</tbody>
</table>

\[
p(w = b | \ldots) = \frac{1}{1.0 + 4} + \frac{1.0}{1.0 + 4} p(w = x | \vec{s}, \theta, \pi(\emptyset)) \quad (2)
\]
Example: \( p(w = b|\vec{s}, \theta = 1.0, u = a) \)

Unigram Restaurant

\[
\begin{array}{ccc}
a & b & c \\
3 & 1 & 1 \\
\end{array}
\]

\( <s> \) Restaurant

\[
\begin{array}{c}
a \\
1 \\
\end{array}
\]

\( a \) Restaurant

\[
\begin{array}{ccc}
a & b & c \\
2 & 1 & 1 \\
\end{array}
\]

\( b \) Restaurant

\[
\begin{array}{c}
a \\
1 \\
\end{array}
\]

\( c \) Restaurant

\[
\begin{array}{c}
<s> \\
1 \\
\end{array}
\]

\[
p(w = b|\ldots) = \frac{1}{1.0 + 4} + \frac{1.0}{1.0 + 4} p(w = x|\vec{s}, \theta, \pi(\emptyset))
\] (2)
### Example: $p(w = b|\vec{s}, \theta = 1.0, u = a)$

#### Unigram Restaurant

<table>
<thead>
<tr>
<th>a</th>
<th>³</th>
<th>b</th>
<th>¹</th>
<th>c</th>
<th>¹</th>
<th>$&lt;$s$&gt;$</th>
<th>¹</th>
</tr>
</thead>
</table>

#### <s> Restaurant

<table>
<thead>
<tr>
<th>a</th>
<th>¹</th>
</tr>
</thead>
</table>

#### a Restaurant

<table>
<thead>
<tr>
<th>a</th>
<th>²</th>
<th>b</th>
<th>¹</th>
<th>c</th>
<th>¹</th>
</tr>
</thead>
</table>

#### b Restaurant

<table>
<thead>
<tr>
<th>a</th>
<th>¹</th>
</tr>
</thead>
</table>

#### c Restaurant

| $<$s$>$ | ¹ |

\[
p(w = b|\ldots) = \frac{1}{5} + \frac{1}{5} \left( \frac{c_{\theta,b}}{c_{\theta,.} + \theta} + \frac{\theta}{c_{\theta,.} + \theta} \frac{1}{V} \right)\] (2)
Example: \( p(w = b | \vec{s}, \theta = 1.0, u = a) \)

Unigram Restaurant

\[
\begin{array}{ccc}
\text{a}^3 & \text{b}^1 & \text{c}^1 \ \langle /s \rangle^1 \\
\end{array}
\]

\(<s> \) Restaurant

\[
\begin{array}{c}
\text{a}^1 \\
\end{array}
\]

\(\text{b} \) Restaurant

\[
\begin{array}{c}
\text{a}^1 \\
\end{array}
\]

\(\text{c} \) Restaurant

\[
\begin{array}{c}
\langle /s \rangle^1 \\
\end{array}
\]

\[
p(w = b | \ldots) = \frac{1}{5} + \frac{1}{5} \left( \frac{c_{\emptyset, b}}{c_{\emptyset, .} + \theta} + \frac{\theta}{c_{\emptyset, .} + \theta} \right) \frac{1}{5} \tag{2}
\]
Example: \( p(w = b | \vec{s}, \theta = 1.0, u = a) \)

\[
p(w = b | \vec{s}, \theta = 1.0, u = a) = 1 \frac{1}{5} + \frac{1}{5} \left( \frac{c_{\emptyset,b}}{c_{\emptyset,.} + 1.0} + \frac{1.0}{c_{\emptyset,.} + 1.0} \frac{1}{5} \right)
\]
Example: \( p(w = b|\vec{s}, \theta = 1.0, u = a) \)

Unigram Restaurant

\[
\begin{array}{c}
\text{a}^3 \\
\text{b}^1 \\
\text{c}^1 \\
\langle/s\rangle^1
\end{array}
\]

\[
p(w = b|\ldots) = \frac{1}{5} + \frac{1}{5} \left( \frac{1}{c_{\emptyset, .} + 1.0} + \frac{1.0}{c_{\emptyset, .} + 1.0} \right)
\]
Example: \( p(w = b | \tilde{s}, \theta = 1.0, u = a) \)

\[
\begin{align*}
\text{Unigram Restaurant} & \quad \begin{array}{cccc}
a \quad 3 & b \quad 1 & c \quad 1 & \langle/s\rangle \quad 1 \\
\end{array} \\
\text{<s> Restaurant} & \quad \begin{array}{c}
a \quad 1 \\
\end{array} \\
\text{b Restaurant} & \quad \begin{array}{c}
a \quad 1 \\
\end{array} \\
\text{c Restaurant} & \quad \begin{array}{c}
\langle/s\rangle \quad 1 \\
\end{array}
\end{align*}
\]

\[
p(w = b | \ldots) = \frac{1}{5} + \frac{1}{5} \left( \frac{1}{6 + 1.0} + \frac{1.0}{6 + 1.0} \right)
\]  
(2)
Example: $p(w = b|s, \theta = 1.0, u = a)$

Unigram Restaurant

$$\begin{align*}
a^3 & \quad b^1 & \quad c^1 & \quad <s>^1
\end{align*}$$

$s$ Restaurant

$$\begin{align*}
a^1
\end{align*}$$

$a$ Restaurant

$$\begin{align*}
a^2 & \quad b^1 & \quad c^1
\end{align*}$$

$b$ Restaurant

$$\begin{align*}
a^1
\end{align*}$$

$c$ Restaurant

$$\begin{align*}
<s>^1
\end{align*}$$

$p(w = b|\ldots) = \frac{1}{5} + \frac{1}{5} \left( \frac{1}{7} + \frac{1}{75} \right) = 0.24$ (2)
Discounting

- Empirically, it helps favor the backoff if you have more tables
- Otherwise, it gets too close to maximum likelihood
- Idea is called *discounting*
- Steal a little bit of probability mass $\delta$ from every table and give it to the new table (backoff)
Discounting

- Empirically, it helps favor the backoff if you have more tables
- Otherwise, it gets too close to maximum likelihood
- Idea is called *discounting*
- Steal a little bit of probability mass $\delta$ from every table and give it to the new table (backoff)

$$p(w = x|\vec{s}, \theta, u) = \frac{c_{u,x}}{\theta + c_{u,x}} \cdot \left[ \frac{\theta}{\theta + c_{u,x}} p(w = x|\vec{s}, \theta, \pi(u)) \right]$$

(3)
Discounting

- Empirically, it helps favor the backoff if you have more tables
- Otherwise, it gets too close to maximum likelihood
- Idea is called *discounting*
- Steal a little bit of probability mass $\delta$ from every table and give it to the new table (backoff)

$$p(w = x|\tilde{s}, \theta, u) = \frac{c_{u,x} - \delta}{\theta + c_{u,x}}_{\text{existing table}} + \frac{\theta + T\delta}{\theta + c_{u,x}}_{\text{new table}} p(w = x|\tilde{s}, \theta, \pi(u)) \quad (3)$$
Discounting

- Empirically, it helps favor the backoff if you have more tables
- Otherwise, it gets too close to maximum likelihood
- Idea is called *discounting*
- Steal a little bit of probability mass $\delta$ from every table and give it to the new table (backoff)

\[
p(w = x | \vec{s}, \theta, u) = \frac{c_{u,x} - \delta}{\theta + c_{u,x}} \cdot \text{existing table} + \frac{\theta + T \delta}{\theta + c_{u,x}} \cdot p(w = x | \vec{s}, \theta, \pi(u)) \quad (3)
\]
Discounting

- Empirically, it helps favor the backoff if you have more tables
- Otherwise, it gets too close to maximum likelihood
- Idea is called *discounting*
- Steal a little bit of probability mass $\delta$ from every table and give it to the new table (backoff)

$$p(w = x|\tilde{s}, \theta, u) = \frac{c_{u,x} - \delta}{\theta + c_{u,x}} \underbrace{\frac{\theta + T \delta}{\theta + c_{u,x}} p(w = x|\tilde{s}, \theta, \pi(u))}_{\text{new table}}$$

*Interpolated Kneser-Ney!*
More advanced models

- Interpolated Kneser-Ney assumes one table with a dish (word) per restaurant
- Can get slightly better performance by assuming you can have duplicated tables: Pitman-Yor language model
- Requires Gibbs Sampling of the seating assignments (GS, later, but not for language models)
Exercise

- Start with restaurant we had before
- Assume you see <s> b b a c </s>; add those counts to tables
- Compute probability of b following a ($\theta = 1.0, \delta = 0.5$)
- Compute the probability of a following b
- Compute probability of </s> following <s>