Language Models

Advanced Machine Learning for NLP

Jordan Boyd-Graber

KNESSER-NEY AND BAYESIAN NONPARAMETRICS

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

Let's remember what a language model is

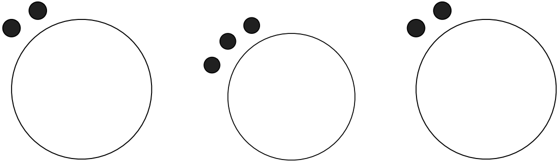
- It is a distribution over the *next word* in a sentence
- Given the previous $n-1$ words

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

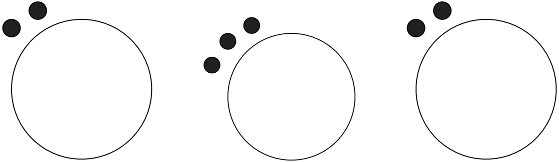
The Chinese Restaurant as a Distribution

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.



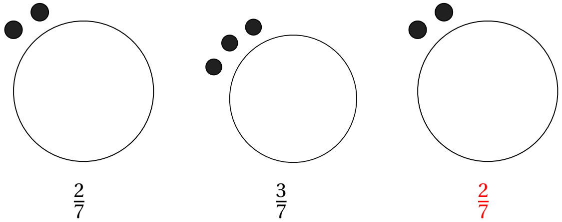
The Chinese Restaurant as a Distribution

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.



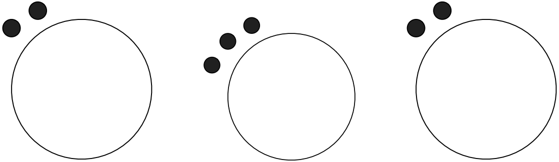
$\frac{2}{7}$ $\frac{3}{7}$ $\frac{2}{7}$

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

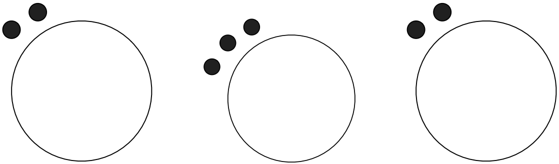
To generate a word, you first sit down at a table. You sit down at a table proportional to the number of people sitting at the table.



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

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.

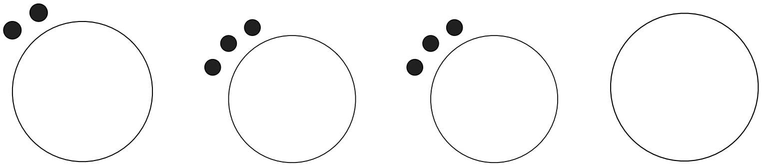


$\frac{2}{7}$ dog | $\frac{3}{7}$ cat | $\frac{2}{7}$ purple

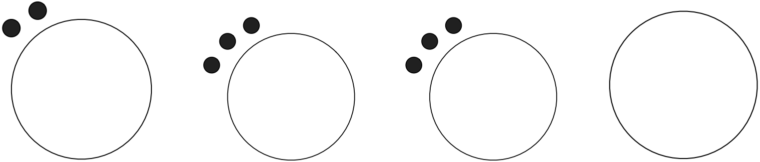
But this is just Maximum Likelihood

Why are we talking about Chinese Restaurants?

Always one more table . . .

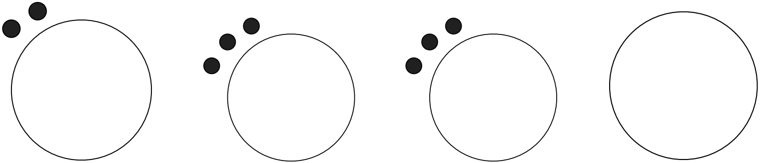


Always one more table . . .



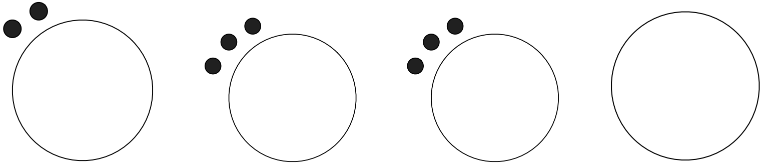
$\frac{2}{7+\alpha}$ $\frac{3}{7+\alpha}$ $\frac{2}{7+\alpha}$ $\frac{\alpha}{7+\alpha}$

Always one more table . . .



$\frac{2}{7+\alpha}$ dog | $\frac{3}{7+\alpha}$ cat | $\frac{2}{7+\alpha}$ purple | $\frac{\alpha}{7+\alpha}$???

Always one more table . . .



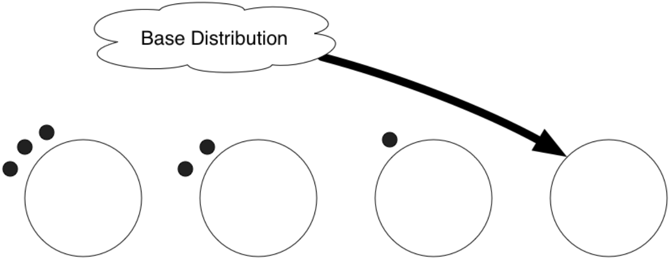
$\frac{2}{7+\alpha}$ dog

$\frac{3}{7+\alpha}$ cat

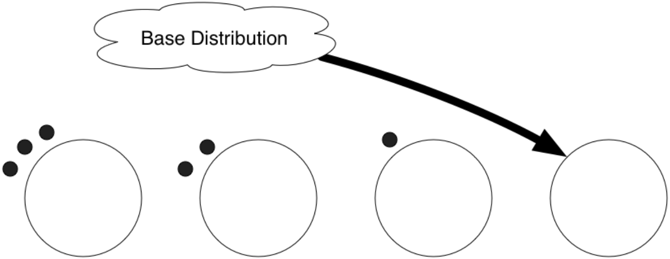
$\frac{2}{7+\alpha}$ purple

$\frac{\alpha}{7+\alpha}$???

What to do with a new table?



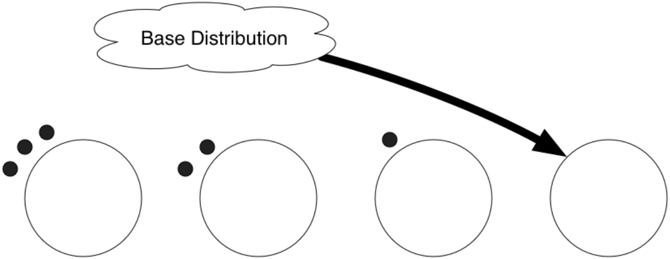
What to do with a new table?



What can be a base distribution?

- Uniform (Dirichlet smoothing)

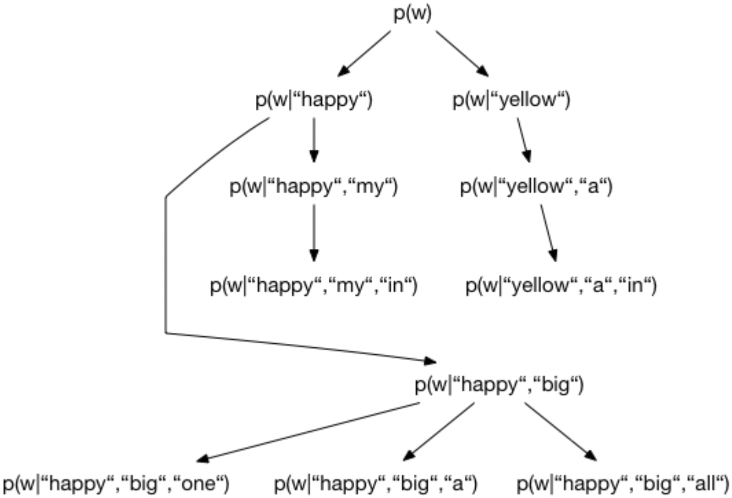
What to do with a new table?



What can be a base distribution?

- Uniform (Dirichlet smoothing)
- Specific contexts → less-specific contexts (backoff)

A hierarchy of Chinese Restaurants



Seating Assignments

Dataset:

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

Seating Assignments

Dataset:

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

Unigram Restaurant

<s> Restaurant

a Restaurant

b Restaurant

c Restaurant

Seating Assignments

Dataset:

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

Unigram Restaurant

<s> Restaurant

$\boxed{*}^{1}$

a Restaurant

b Restaurant

c Restaurant

Seating Assignments

Dataset:

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

Unigram Restaurant

* 1

<s> Restaurant

* 1

a Restaurant

b Restaurant

c Restaurant

Seating Assignments

Dataset:

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

Unigram Restaurant

a 1

<s> Restaurant

a 1

a Restaurant

b Restaurant

c Restaurant

Seating Assignments

Dataset:

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

Unigram Restaurant

a 1

<s> Restaurant

a 1

a Restaurant

* 1

b Restaurant

c Restaurant

Seating Assignments

Dataset:

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

Unigram Restaurant

a [1]

<s> Restaurant

a [1]

a Restaurant

* [1]

b Restaurant

c Restaurant

Seating Assignments

Dataset:

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

Unigram Restaurant

a [2]

<s> Restaurant

a [1]

a Restaurant

a [1]

b Restaurant

c Restaurant

Seating Assignments

Dataset:

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

Unigram Restaurant

$\boxed{\texttt{a}}^{2}$

<s> Restaurant

$\boxed{\texttt{a}}^{1}$

a Restaurant

$\boxed{\texttt{a}}^{1}$

b Restaurant

c Restaurant

Seating Assignments

Dataset:

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

Unigram Restaurant

a 2

<s> Restaurant

a 1

a Restaurant

a 2

b Restaurant

c Restaurant

Seating Assignments

Dataset:

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

Unigram Restaurant

a [2]

<s> Restaurant

a [1]

a Restaurant

a [2] * [1]

b Restaurant

c Restaurant

Seating Assignments

Dataset:

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

Unigram Restaurant

a [2] * [1]

<s> Restaurant

a [1]

a Restaurant

a [2] * [1]

b Restaurant

c Restaurant

Seating Assignments

Dataset:

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

Unigram Restaurant

$\boxed{a}^2$ $\boxed{b}^1$

<s> Restaurant

$\boxed{a}^1$

a Restaurant

$\boxed{a}^2$ $\boxed{*}^1$

b Restaurant

c Restaurant

Seating Assignments

Dataset:

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

Unigram Restaurant

a 2 b 1

<s> Restaurant

a 1

a Restaurant

a 2 b 1

b Restaurant

c Restaurant

Seating Assignments

Dataset:

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

Unigram Restaurant

a[2] b[1]

<s> Restaurant

a[1]

a Restaurant

a[2] b[1]

b Restaurant

c Restaurant

Seating Assignments

Dataset:

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

Unigram Restaurant

a[2] b[1]

<s> Restaurant

a[1]

a Restaurant

a[2] b[1]

b Restaurant

*[1]

c Restaurant

Seating Assignments

Dataset:

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

Unigram Restaurant

a[2] b[1]

<s> Restaurant

a[1]

a Restaurant

a[2] b[1]

b Restaurant

*[1]

c Restaurant

Seating Assignments

Dataset:

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

Unigram Restaurant

a 3 b 1

<s> Restaurant

a 1

a Restaurant

a 2 b 1

b Restaurant

a 1

c Restaurant

Seating Assignments

Dataset:

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

Unigram Restaurant

a 3 b 1

<s> Restaurant

a 1

a Restaurant

a 2 b 1

b Restaurant

a 1

c Restaurant

Seating Assignments

Dataset:

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

Unigram Restaurant

$\boxed{a}^{3}$ $\boxed{b}^{1}$

<s> Restaurant

$\boxed{a}^{1}$

a Restaurant

$\boxed{a}^{2}$ $\boxed{b}^{1}$ $\boxed{*}^{1}$

b Restaurant

$\boxed{a}^{1}$

c Restaurant

Seating Assignments

Dataset:

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

Unigram Restaurant

a[3] b[1] * [1]

<s> Restaurant

a[1]

a Restaurant

a[2] b[1] * [1]

b Restaurant

a[1]

c Restaurant

Seating Assignments

Dataset:

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

Unigram Restaurant

a[3] b[1] c[1]

<s> Restaurant

a[1]

b Restaurant

a[1]

a Restaurant

a[2] b[1] c[1]

c Restaurant

Seating Assignments

Dataset:

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

Unigram Restaurant

$\boxed{\text{a}}^{3}$ $\boxed{\text{b}}^{1}$ $\boxed{\text{c}}^{1}$

<s> Restaurant

$\boxed{\text{a}}^{1}$

b Restaurant

$\boxed{\text{a}}^{1}$

a Restaurant

$\boxed{\text{a}}^{2}$ $\boxed{\text{b}}^{1}$ $\boxed{\text{c}}^{1}$

c Restaurant

Seating Assignments

Dataset:

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

Unigram Restaurant

a	b	c
3	1	1

<s> Restaurant

a
1

a Restaurant

a	b	c
2	1	1

b Restaurant

a
1

c Restaurant

*
1

Seating Assignments

Dataset:

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

Unigram Restaurant

$\mathtt{a}^{3}$ $\mathtt{b}^{1}$ $\mathtt{c}^{1}$ $*^{1}$

<s> Restaurant

$\mathtt{a}^{1}$

a Restaurant

$\mathtt{a}^{2}$ $\mathtt{b}^{1}$ $\mathtt{c}^{1}$

b Restaurant

$\mathtt{a}^{1}$

c Restaurant

$*^{1}$

Seating Assignments

Dataset:

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

Unigram Restaurant

a 3 b 1 c 1 </s> 1

<s> Restaurant

a 1

a Restaurant

a 2 b 1 c 1

b Restaurant

a 1

c Restaurant

</s> 1

Real examples

- San Francisco

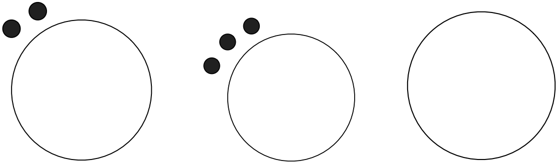
Real examples

- San Francisco
- Star Spangled Banner

Real examples

- San Francisco
- Star Spangled Banner
- Bottom Line: Counts go to the context that explains it best

The rich get richer



$\frac{2}{5+\theta}$ $\quad$ $\frac{3}{5+\theta}$ $\quad$ $\frac{\theta}{5+\theta}$

Computing the Probability of an Observation

$$p(w = x|\vec{s}, \theta, u) = \underbrace{\frac{c_{u,x}}{\theta + c_{u,\cdot}}}_{\text{existing table}} + \underbrace{\frac{\theta}{\theta + c_{u,\cdot}} p(w = x|\vec{s}, \theta, \pi(u))}_{\text{new table}} \quad (1)$$

- Word type x
- Seating assignments $\vec{s}$
- Concentration θ
- Context u
- Number seated at table serving x in restaurant u, $c_{u,x}$
- Number seated at all tables in restaurant u, $c_{u,\cdot}$
- The backoff context $\pi(u)$

Computing the Probability of an Observation

$$p(w=x|\vec{s},\theta,u) = \underbrace{\frac{c_{u,x}}{\theta + c_{u,\cdot}}}_{\text{existing table}} + \underbrace{\frac{\theta}{\theta + c_{u,\cdot}} p(w=x|\vec{s},\theta,\pi(u))}_{\text{new table}} \tag{1}$$

- Word type x
- Seating assignments $\vec{s}$
- Concentration θ
- Context u
- Number seated at table serving x in restaurant u, $c_{u,x}$
- Number seated at all tables in restaurant u, $c_{u,\cdot}$
- The backoff context $\pi(u)$

Computing the Probability of an Observation

$$p(w=x|\vec{s},\theta,u)=\underbrace{\frac{c_{u,x}}{\theta+c_{u,\cdot}}}_{\text{existing table}}+\underbrace{\frac{\theta}{\theta+c_{u,\cdot}}p(w=x|\vec{s},\theta,\pi(u))}_{\text{new table}} \qquad (1)$$

- Word type x
- Seating assignments $\vec{s}$
- Concentration θ
- Context u
- Number seated at table serving x in restaurant u, $c_{u,x}$
- Number seated at all tables in restaurant u, $c_{u,\cdot}$
- The backoff context $\pi(u)$

Computing the Probability of an Observation

$$p(w = x|\vec{s}, \theta, u) = \underbrace{\frac{c_{u,x}}{\theta + c_{u,\cdot}}}_{\text{existing table}} + \underbrace{\frac{\theta}{\theta + c_{u,\cdot}} p(w = x|\vec{s}, \theta, \pi(u))}_{\text{new table}} \quad (1)$$

- Word type x
- Seating assignments $\vec{s}$
- Concentration θ
- Context u
- Number seated at table serving x in restaurant u, $c_{u,x}$
- Number seated at all tables in restaurant u, $c_{u,\cdot}$
- The backoff context $\pi(u)$

Computing the Probability of an Observation

$$p(w=x|\vec{s},\theta,u) = \underbrace{\frac{c_{u,x}}{\theta + c_{u,\cdot}}}_{\text{existing table}} + \underbrace{\frac{\theta}{\theta + c_{u,\cdot}} p(w=x|\vec{s},\theta,\pi(u))}_{\text{new table}} \tag{1}$$

- Word type x
- Seating assignments $\vec{s}$
- Concentration θ
- Context u
- Number seated at table serving x in restaurant u, $c_{u,x}$
- Number seated at all tables in restaurant u, $c_{u,\cdot}$
- The backoff context $\pi(u)$

Computing the Probability of an Observation

$$p(w = x|\vec{s}, \theta, u) = \underbrace{\frac{c_{u,x}}{\theta + c_{u,\cdot}}}_{\text{existing table}} + \underbrace{\frac{\theta}{\theta + c_{u,\cdot}} p(w = x|\vec{s}, \theta, \pi(u))}_{\text{new table}} \quad (1)$$

- Word type x
- Seating assignments $\vec{s}$
- Concentration θ
- Context u
- Number seated at table serving x in restaurant u, $c_{u,x}$
- Number seated at all tables in restaurant u, $c_{u,\cdot}$
- The backoff context $\pi(u)$

Computing the Probability of an Observation

$$p(w = x|\vec{s}, \theta, u) = \underbrace{\frac{c_{u,x}}{\theta + c_{u,\cdot}}}_{\text{existing table}} + \underbrace{\frac{\theta}{\theta + c_{u,\cdot}} p(w = x|\vec{s}, \theta, \pi(u))}_{\text{new table}} \tag{1}$$

- Word type x
- Seating assignments $\vec{s}$
- Concentration θ
- Context u
- Number seated at table serving x in restaurant u, $c_{u,x}$
- Number seated at all tables in restaurant u, $c_{u,\cdot}$
- The backoff context $\pi(u)$

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

Unigram Restaurant

a [3] b [1] c [1] </s> [1]

<s> Restaurant

a [1]

a Restaurant

a [2] b [1] c [1]

b Restaurant

a [1]

c Restaurant

</s> [1]

$$p(w = \mathrm{b} | \ldots) = \frac{c_{\mathrm{a},\mathrm{b}}}{\theta + c_{u,\cdot}} + \frac{\theta}{\theta + c_{u,\cdot}} p(w = x | \vec{s}, \theta, \pi(u)) \tag{2}$$

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

Unigram Restaurant

$\texttt{a}^{3}$ $\texttt{b}^{1}$ $\texttt{c}^{1}$ $\texttt{</s>}^{1}$

<s> Restaurant

$\texttt{a}^{1}$

a Restaurant

$\texttt{a}^{2}$ $\texttt{b}^{1}$ $\texttt{c}^{1}$

b Restaurant

$\texttt{a}^{1}$

c Restaurant

$\texttt{</s>}^{1}$

$$p(w = \mathrm{b} | \ldots) = \frac{c_{\mathrm{a,b}}}{\theta + c_{u,\cdot}} + \frac{\theta}{\theta + c_{u,\cdot}} p(w = x | \vec{s}, \theta, \pi(u)) \quad (2)$$

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

Unigram Restaurant

a [3] b [1] c [1] </s> [1]

<s> Restaurant

a [1]

a Restaurant

a [2] b [1] c [1]

b Restaurant

a [1]

c Restaurant

</s> [1]

$$p(w = \mathrm{b} | \ldots) = \frac{1}{\theta + c_{u,\cdot}} + \frac{\theta}{\theta + c_{u,\cdot}} p(w = x | \vec{s}, \theta, \pi(u)) \tag{2}$$

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

Unigram Restaurant

a^{3} b^{1} c^{1} </s>^{1}

<s> Restaurant

a^{1}

a Restaurant

a^{2} b^{1} c^{1}

b Restaurant

a^{1}

c Restaurant

</s>^{1}

$$p(w = \mathrm{b}|\ldots) = \frac{1}{1.0 + c_{u,\cdot}} + \frac{1.0}{1.0 + c_{u,\cdot}} p(w = x|\vec{s}, \theta, \pi(u)) \tag{2}$$

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

Unigram Restaurant

$\texttt{a}^{3}$ $\texttt{b}^{1}$ $\texttt{c}^{1}$ $\texttt{</s>}^{1}$

<s> Restaurant

$\texttt{a}^{1}$

a Restaurant

$\texttt{a}^{2}$ $\texttt{b}^{1}$ $\texttt{c}^{1}$

b Restaurant

$\texttt{a}^{1}$

c Restaurant

$\texttt{</s>}^{1}$

$$p(w = \mathrm{b} | \ldots) = \frac{1}{1.0 + 4} + \frac{1.0}{1.0 + 4} p(w = x | \vec{s}, \theta, \pi(u)) \quad (2)$$

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

Unigram Restaurant

a [3] b [1] c [1] </s> [1]

<s> Restaurant

a [1]

a Restaurant

a [2] b [1] c [1]

b Restaurant

a [1]

c Restaurant

</s> [1]

$$p(w = \mathrm{b} | \ldots) = \frac{1}{1.0 + 4} + \frac{1.0}{1.0 + 4} p(w = x | \vec{s}, \theta, \pi(u)) \tag{2}$$

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

Unigram Restaurant

a^{3} b^{1} c^{1} </s>^{1}

<s> Restaurant

a^{1}

a Restaurant

a^{2} b^{1} c^{1}

b Restaurant

a^{1}

c Restaurant

</s>^{1}

$$p(w = \mathrm{b} | \ldots) = \frac{1}{1.0 + 4} + \frac{1.0}{1.0 + 4} p(w = x | \vec{s}, \theta, \pi(\emptyset)) \tag{2}$$

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

Unigram Restaurant

a[3] b[1] c[1] </s>[1]

<s> Restaurant

a[1]

a Restaurant

a[2] b[1] c[1]

b Restaurant

a[1]

c Restaurant

</s>[1]

$$p(w = \mathrm{b} | \ldots) = \frac{1}{1.0 + 4} + \frac{1.0}{1.0 + 4} p(w = x | \vec{s}, \theta, \pi(\emptyset)) \tag{2}$$

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

Unigram Restaurant

$\boxed{\texttt{a}}^{3}$ $\boxed{\texttt{b}}^{1}$ $\boxed{\texttt{c}}^{1}$ $\boxed{\texttt{</s>}}^{1}$

<s> Restaurant

$\boxed{\texttt{a}}^{1}$

a Restaurant

$\boxed{\texttt{a}}^{2}$ $\boxed{\texttt{b}}^{1}$ $\boxed{\texttt{c}}^{1}$

b Restaurant

$\boxed{\texttt{a}}^{1}$

c Restaurant

$\boxed{\texttt{</s>}}^{1}$

$$p(w = \mathrm{b} | \ldots) = \frac{1}{5} + \frac{1}{5}\left(\frac{c_{\emptyset,\mathrm{b}}}{c_{\emptyset,\cdot} + \theta} + \frac{\theta}{c_{\emptyset,\cdot} + \theta}\frac{1}{V}\right) \tag{2}$$

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

Unigram Restaurant

a [3] b [1] c [1] </s> [1]

<s> Restaurant

a [1]

a Restaurant

a [2] b [1] c [1]

b Restaurant

a [1]

c Restaurant

</s> [1]

$$p(w = \mathrm{b} | \ldots) = \frac{1}{5} + \frac{1}{5} \left(\frac{c_{\emptyset,\mathrm{b}}}{c_{\emptyset,\cdot} + \theta} + \frac{\theta}{c_{\emptyset,\cdot} + \theta} \frac{1}{5} \right) \tag{2}$$

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

Unigram Restaurant

$\boxed{\texttt{a}}^{3}$ $\boxed{\texttt{b}}^{1}$ $\boxed{\texttt{c}}^{1}$ $\boxed{\texttt{</s>}}^{1}$

<s> Restaurant

$\boxed{\texttt{a}}^{1}$

a Restaurant

$\boxed{\texttt{a}}^{2}$ $\boxed{\texttt{b}}^{1}$ $\boxed{\texttt{c}}^{1}$

b Restaurant

$\boxed{\texttt{a}}^{1}$

c Restaurant

$\boxed{\texttt{</s>}}^{1}$

$$p(w = \mathrm{b} | \ldots) = \frac{1}{5} + \frac{1}{5}\left(\frac{c_{\emptyset,\mathrm{b}}}{c_{\emptyset,\cdot} + 1.0} + \frac{1.0}{c_{\emptyset,\cdot} + 1.0} \frac{1}{5} \right) \quad (2)$$

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

Unigram Restaurant

a [3] b [1] c [1] </s> [1]

<s> Restaurant

a [1]

a Restaurant

a [2] b [1] c [1]

b Restaurant

a [1]

c Restaurant

</s> [1]

$$p(w = \mathrm{b} | \ldots) = \frac{1}{5} + \frac{1}{5}\left(\frac{1}{c_{\emptyset,\cdot} + 1.0} + \frac{1.0}{c_{\emptyset,\cdot} + 1.0}\frac{1}{5}\right) \tag{2}$$

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

Unigram Restaurant

a[3] b[1] c[1] </s>[1]

<s> Restaurant

a[1]

a Restaurant

a[2] b[1] c[1]

b Restaurant

a[1]

c Restaurant

</s>[1]

$$p(w = \mathrm{b} | \ldots) = \frac{1}{5} + \frac{1}{5}\left(\frac{1}{6+1.0} + \frac{1.0}{6+1.0}\frac{1}{5}\right) \quad (2)$$

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

Unigram Restaurant

a [3] b [1] c [1] </s> [1]

<s> Restaurant

a [1]

a Restaurant

a [2] b [1] c [1]

b Restaurant

a [1]

c Restaurant

</s> [1]

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

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 δ 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 δ from every table and give it to the new table (backoff)

$$p(w = x | \vec{s}, \theta, u) = \underbrace{\frac{c_{u,x}}{\theta + c_{u,\cdot}}}_{\text{existing table}} + \underbrace{\frac{\theta}{\theta + c_{u,\cdot}} p(w = x | \vec{s}, \theta, \pi(u))}_{\text{new table}} \quad (3)$$

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 δ from every table and give it to the new table (backoff)

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

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 δ from every table and give it to the new table (backoff)

$$p(w = x|\vec{s}, \theta, u) = \underbrace{\frac{c_{u,x} - \delta}{\theta + c_{u,\cdot}}}_{\text{existing table}} + \underbrace{\frac{\theta + {\color{red}T}\delta}{\theta + c_{u,\cdot}} p(w = x|\vec{s}, \theta, \pi(u))}_{\text{new table}} \tag{3}$$

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 δ from every table and give it to the new table (backoff)

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

Interpolated Kneser-Ney!

More advanced models

- Interpolated Kneser-Ney assumes **one table with a dish (word)** per restaurant (known as **minimal** path assumption)
- Can get slightly better performance by assuming you can have duplicated tables: **Pitman-Yor** language model
- Requires Gibbs Sampling of the seating assignments
    - Initialize seating assignments
    - Remove word from context
    - Add it back in (seating probabilistically)

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>`

A busy night at the restaurant

Unigram Restaurant

a [3] b [1] c [1] </s> [1]

<s> Restaurant

a [1]

a Restaurant

a [2] b [1] c [1]

b Restaurant

a [1]

c Restaurant

</s> [1]

A busy night at the restaurant

Unigram Restaurant

a 3 | b 1 | c 1 | </s> 1

<s> Restaurant

a 1 | b 1

a Restaurant

a 2 | b 1 | c 1

b Restaurant

a 1

c Restaurant

</s> 1

A busy night at the restaurant

Unigram Restaurant

a 3 b 2 c 1 </s> 1

<s> Restaurant

a 1 b 1

a Restaurant

a 2 b 1 c 1

b Restaurant

a 1

c Restaurant

</s> 1

A busy night at the restaurant

Unigram Restaurant

a[3] b[2] c[1] </s>[1]

<s> Restaurant

a[1] b[1]

a Restaurant

a[2] b[1] c[1]

b Restaurant

a[1] b[1]

c Restaurant

</s>[1]

A busy night at the restaurant

Unigram Restaurant

a 3 b 3 c 1 </s> 1

<s> Restaurant

a 1 b 1

a Restaurant

a 2 b 1 c 1

b Restaurant

a 1 b 1

c Restaurant

</s> 1

A busy night at the restaurant

Unigram Restaurant

a^{3} b^{3} c^{1} $\langle/s\rangle^{1}$

<s> Restaurant

a^{1} b^{1}

a Restaurant

a^{2} b^{1} c^{1}

b Restaurant

a^{2} b^{1}

c Restaurant

$\langle/s\rangle^{1}$

A busy night at the restaurant

Unigram Restaurant

a 3, b 3, c 1, </s> 1

<s> Restaurant

a 1, b 1

a Restaurant

a 2, b 1, c 2

b Restaurant

a 2, b 1

c Restaurant

</s> 1

A busy night at the restaurant

Unigram Restaurant

a	b	c	</s>
3	3	1	1

<s> Restaurant

a	b
1	1

a Restaurant

a	b	c
2	1	2

b Restaurant

a	b
2	1

c Restaurant

</s>
2

A busy night at the restaurant

Unigram Restaurant

a^{3} b^{3} c^{1} </s>^{1}

<s> Restaurant

a^{1} b^{1}

a Restaurant

a^{2} b^{1} c^{2}

b Restaurant

a^{2} b^{1}

c Restaurant

</s>^{2}

As you see more data, bottom restaurants do more work.

$$= \frac{1-\delta}{\theta+5} + \frac{\theta+3\delta}{\theta+5} p(\mathtt{b}) \tag{4}$$

$$= \frac{1-\delta}{\theta+5} + \frac{\theta+3\delta}{\theta+5} \left(\frac{3-\delta}{\theta+8} + \frac{\theta+4\delta}{\theta+8} \frac{1}{V} \right) \tag{5}$$

$$\tag{6}$$

$\mathtt{b}$ following $\mathtt{a}$

$$= \frac{1-\delta}{\theta+5} + \frac{\theta+3\delta}{\theta+5} p(\mathtt{b}) \tag{4}$$

$$= \frac{1-\delta}{\theta+5} + \frac{\theta+3\delta}{\theta+5} \left(\frac{3-\delta}{\theta+8} + \frac{\theta+4\delta}{\theta+8} \frac{1}{V} \right) \tag{5}$$

$$\tag{6}$$

$\mathtt{b}$ following $\mathtt{a}$

$$= \frac{1-\delta}{\theta+5} + \frac{\theta+3\delta}{\theta+5} p(\mathtt{b}) \tag{4}$$

$$= \frac{1-\delta}{\theta+5} + \frac{\theta+3\delta}{\theta+5} \left(\frac{3-\delta}{\theta+8} + \frac{\theta+4\delta}{\theta+8} \frac{1}{V} \right) \tag{5}$$

$$\tag{6}$$

0.23

`a` following `b`

$$= \frac{2-\delta}{\theta+3} + \frac{\theta+2\delta}{\theta+3} p(\mathtt{a}) \tag{7}$$

$$= \frac{2-\delta}{\theta+3} + \frac{\theta+2\delta}{\theta+3} \left(\frac{3-\delta}{\theta+8} + \frac{\theta+4\delta}{\theta+8} \frac{1}{V} \right) \tag{8}$$

$$\tag{9}$$

a following b

$$= \frac{2-\delta}{\theta+3} + \frac{\theta+2\delta}{\theta+3} p(\mathtt{a}) \tag{7}$$

$$= \frac{2-\delta}{\theta+3} + \frac{\theta+2\delta}{\theta+3}\left(\frac{3-\delta}{\theta+8} + \frac{\theta+4\delta}{\theta+8}\frac{1}{V}\right) \tag{8}$$

$$\tag{9}$$

a following b

$$
\begin{aligned}
&= \frac{2-\delta}{\theta+3} + \frac{\theta+2\delta}{\theta+3} p(\mathtt{a}) && (7)\\
&= \frac{2-\delta}{\theta+3} + \frac{\theta+2\delta}{\theta+3} \left(\frac{3-\delta}{\theta+8} + \frac{\theta+4\delta}{\theta+8} \frac{1}{V} \right) && (8)\\
& && (9)
\end{aligned}
$$

0.55

</s> following <s>

$$= \frac{\theta + 2\delta}{\theta + 2} p(\text{</s>}) \tag{10}$$

$$= \frac{\theta + 2\delta}{\theta + 2} \left(\frac{1 - \delta}{\theta + 8} + \frac{\theta + 4\delta}{\theta + 8} \frac{1}{V} \right) \tag{11}$$

(12)

</s> following <s>

$$= \frac{\theta + 2\delta}{\theta + 2} p(</\mathrm{s}>) \tag{10}$$

$$= \frac{\theta + 2\delta}{\theta + 2} \left(\frac{1 - \delta}{\theta + 8} + \frac{\theta + 4\delta}{\theta + 8} \frac{1}{V} \right) \tag{11}$$

$$\tag{12}$$

</s> following <s>

$$= \frac{\theta + 2\delta}{\theta + 2} p(\texttt{</s>}) \tag{10}$$

$$= \frac{\theta + 2\delta}{\theta + 2} \left(\frac{1 - \delta}{\theta + 8} + \frac{\theta + 4\delta}{\theta + 8} \frac{1}{V} \right) \tag{11}$$

$$\tag{12}$$

0.08