



Department of Computer Science
UNIVERSITY OF COLORADO **BOULDER**



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

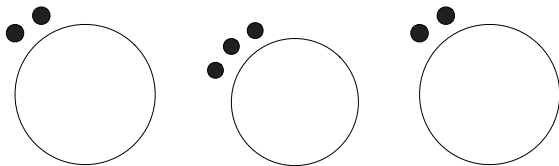
- It is a distribution over the *next word* in a sentence
- Given the previous $n - 1$ words

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

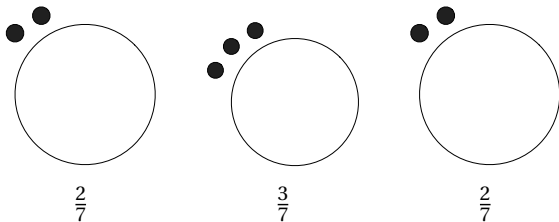
The Chinese Restaurant as a Distribution

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.



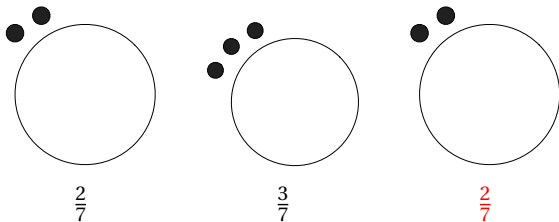
The Chinese Restaurant as a Distribution

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.



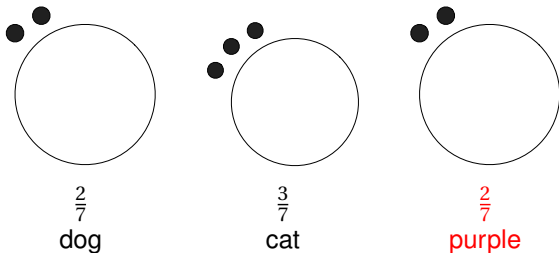
The Chinese Restaurant as a Distribution

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.



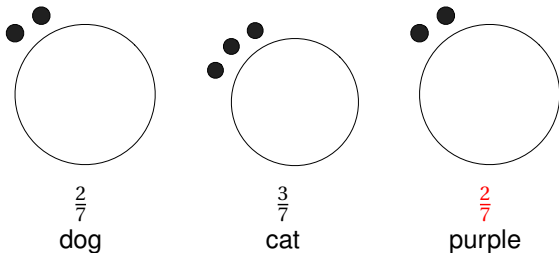
The Chinese Restaurant as a Distribution

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.



The Chinese Restaurant as a Distribution

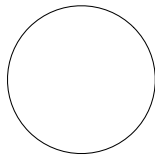
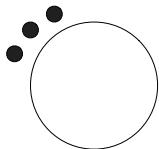
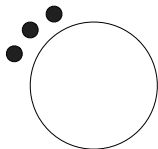
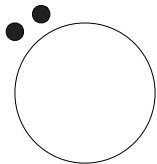
To generate a word, you first sit down at a table. You sit down at a table proportional to the number of people sitting at the table.



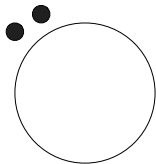
But this is just Maximum Likelihood

Why are we talking about Chinese Restaurants?

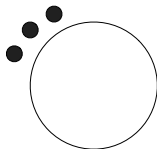
Always one more table ...



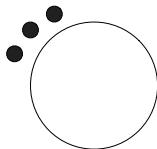
Always one more table ...



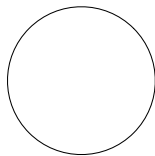
$$\frac{2}{7+\alpha}$$



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

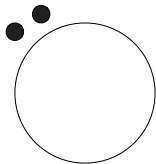


$$\frac{2}{7+\alpha}$$



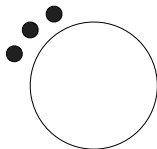
$$\frac{\alpha}{7+\alpha}$$

Always one more table ...



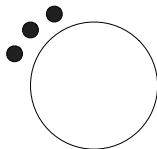
$$\frac{2}{7+\alpha}$$

dog



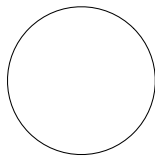
$$\frac{3}{7+\alpha}$$

cat



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

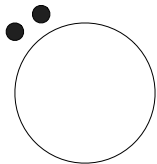
purple



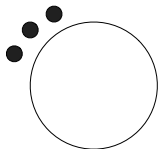
$$\frac{\alpha}{7+\alpha}$$

???

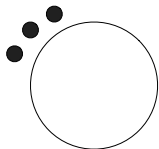
Always one more table ...



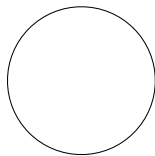
$\frac{2}{7+\alpha}$
dog



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

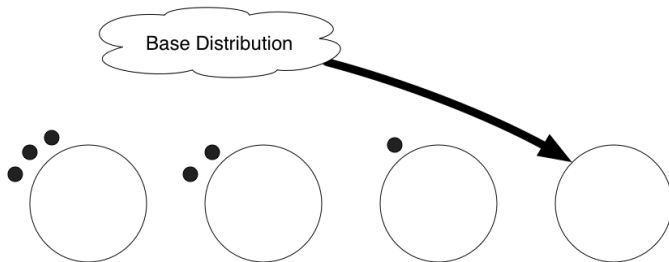


$\frac{2}{7+\alpha}$
purple

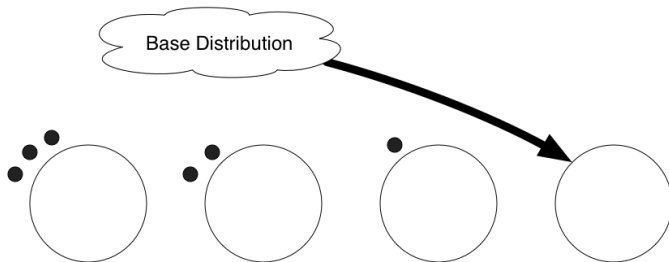


$\frac{\alpha}{7+\alpha}$
???

What to do with a new table?



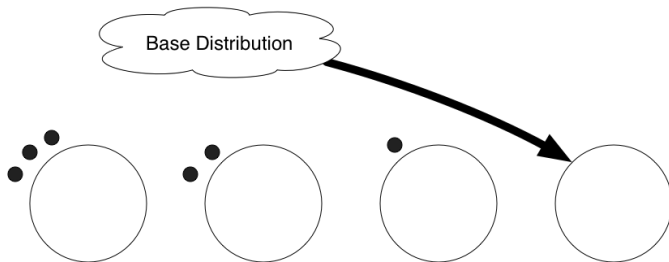
What to do with a new table?



What can be a base distribution?

- Uniform (Dirichlet smoothing)

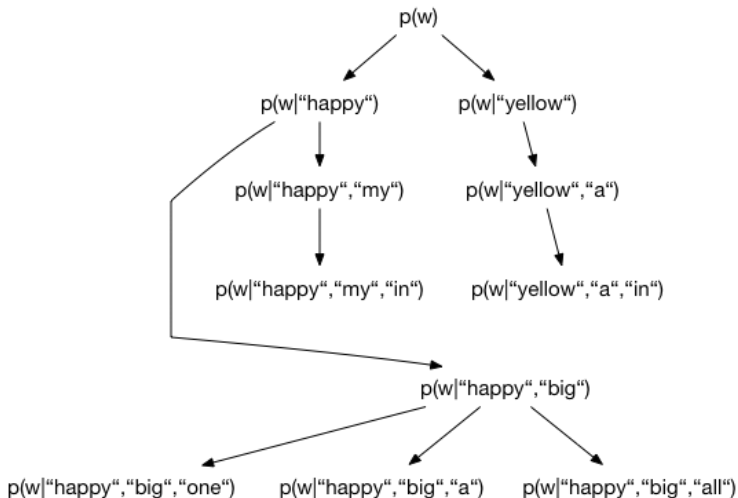
What to do with a new table?



What can be a base distribution?

- Uniform (Dirichlet smoothing)
- Specific contexts \rightarrow 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

*****¹

b Restaurant

a Restaurant

c Restaurant

Seating Assignments

Dataset:

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

Unigram Restaurant

*¹

<s> Restaurant

*¹

b Restaurant

a Restaurant

c Restaurant

Seating Assignments

Dataset:

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

Unigram Restaurant

a¹

<s> Restaurant

a¹

b Restaurant

a Restaurant

c Restaurant

Seating Assignments

Dataset:

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

Unigram Restaurant

a¹

<s> Restaurant

a¹

a Restaurant

*¹

b Restaurant

c Restaurant

Seating Assignments

Dataset:

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

Unigram Restaurant

a¹

<s> Restaurant

a¹

a Restaurant

*¹

b Restaurant

c Restaurant

Seating Assignments

Dataset:

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

Unigram Restaurant

a²

<s> Restaurant

a¹

a Restaurant

a¹

b Restaurant

c Restaurant

Seating Assignments

Dataset:

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

Unigram Restaurant

a²

<s> Restaurant

a¹

a Restaurant

a¹

b Restaurant

c Restaurant

Seating Assignments

Dataset:

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

Unigram Restaurant

a²

<s> Restaurant

a¹

a Restaurant

a²

b Restaurant

c Restaurant

Seating Assignments

Dataset:

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

Unigram Restaurant

a²

<s> Restaurant

a¹

a Restaurant

a² *¹

b Restaurant

c Restaurant

Seating Assignments

Dataset:

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

Unigram Restaurant

a² *¹

<s> Restaurant

a¹

a Restaurant

a² *¹

b Restaurant

c Restaurant

Seating Assignments

Dataset:

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

Unigram Restaurant

a² b¹

<s> Restaurant

a¹

a Restaurant

a² *¹

b Restaurant

c Restaurant

Seating Assignments

Dataset:

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

Unigram Restaurant

a² b¹

<s> Restaurant

a¹

a Restaurant

a² b¹

b Restaurant

c Restaurant

Seating Assignments

Dataset:

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

Unigram Restaurant

a² b¹

<s> Restaurant

a¹

a Restaurant

a² b¹

b Restaurant

c Restaurant

Seating Assignments

Dataset:

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

Unigram Restaurant

a² b¹

<s> Restaurant

a¹

b Restaurant

*¹

a Restaurant

a² b¹

c Restaurant

Seating Assignments

Dataset:

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

Unigram Restaurant

a² b¹

<s> Restaurant

a¹

b Restaurant

*¹

a Restaurant

a² b¹

c Restaurant

Seating Assignments

Dataset:

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

Unigram Restaurant

a³ b¹

<s> Restaurant

a¹

b Restaurant

a¹

a Restaurant

a² b¹

c Restaurant

Seating Assignments

Dataset:

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

Unigram Restaurant

a³ b¹

<s> Restaurant

a¹

b Restaurant

a¹

a Restaurant

a² b¹

c Restaurant

Seating Assignments

Dataset:

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

Unigram Restaurant

a³ b¹

<s> Restaurant

a¹

b Restaurant

a¹

a Restaurant

a² b¹ *¹

c Restaurant

Seating Assignments

Dataset:

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

Unigram Restaurant

a³ b¹ *¹

<s> Restaurant

a¹

b Restaurant

a¹

a Restaurant

a² b¹ *¹

c Restaurant

Seating Assignments

Dataset:

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

Unigram Restaurant

a³ b¹ c¹

<s> Restaurant

a¹

b Restaurant

a¹

a Restaurant

a² b¹ c¹

c Restaurant

Seating Assignments

Dataset:

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

Unigram Restaurant

a³ b¹ c¹

<s> Restaurant

a¹

b Restaurant

a¹

a Restaurant

a² b¹ c¹

c Restaurant

Seating Assignments

Dataset:

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

Unigram Restaurant

a³ b¹ c¹

<s> Restaurant

a¹

a Restaurant

a² b¹ c¹

b Restaurant

a¹

c Restaurant

*¹

Seating Assignments

Dataset:

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

Unigram Restaurant

a³ b¹ c¹ *¹

<s> Restaurant

a¹

a Restaurant

a² b¹ c¹

b Restaurant

a¹

c Restaurant

*¹

Seating Assignments

Dataset:

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

Unigram Restaurant

a³ b¹ c¹ </s>¹

<s> Restaurant

a¹

a Restaurant

a² b¹ c¹

b Restaurant

a¹

c Restaurant

</s>¹

Real examples

- San Francisco

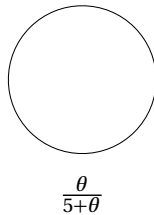
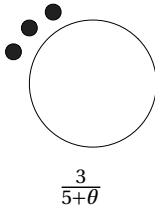
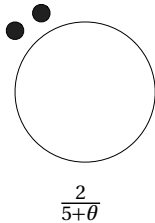
Real examples

- 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



Computing the Probability of an Observation

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

- Word type $\textcolor{red}{x}$
- 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

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

- 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

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

- Word type x
- Seating assignments \vec{s}
- Concentration θ
- Context $\textcolor{red}{u}$
- 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

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

- 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

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

- 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 $\pi(u)$

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

Unigram Restaurant

a³ b¹ c¹ </s>¹

<s> Restaurant

a¹

a Restaurant

a² b¹ c¹

b Restaurant

a¹

c Restaurant

</s>¹

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

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

Unigram Restaurant

a³ b¹ c¹ </s>¹

<s> Restaurant

a¹

a Restaurant

a² b¹ c¹

b Restaurant

a¹

c Restaurant

</s>¹

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

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

Unigram Restaurant

a³ b¹ c¹ </s>¹

<s> Restaurant

a¹

a Restaurant

a² b¹ c¹

b Restaurant

a¹

c Restaurant

</s>¹

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

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

Unigram Restaurant

a³ b¹ c¹ </s>¹

<s> Restaurant

a¹

a Restaurant

a² b¹ c¹

b Restaurant

a¹

c Restaurant

</s>¹

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

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

Unigram Restaurant

a³ b¹ c¹ </s>¹

<s> Restaurant

a¹

a Restaurant

a² b¹ c¹

b Restaurant

a¹

c Restaurant

</s>¹

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

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

Unigram Restaurant

a³ b¹ c¹ </s>¹

<s> Restaurant

a¹

a Restaurant

a² b¹ c¹

b Restaurant

a¹

c Restaurant

</s>¹

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

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

Unigram Restaurant

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

$<s>$ Restaurant

\boxed{a}^1

a Restaurant

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

b Restaurant

\boxed{a}^1

c Restaurant

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

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

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

Unigram Restaurant

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

$<s>$ Restaurant

\boxed{a}^1

a Restaurant

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

b Restaurant

\boxed{a}^1

c Restaurant

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

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

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

Unigram Restaurant

a³ b¹ c¹ </s>¹

<s> Restaurant

a¹

a Restaurant

a² b¹ c¹

b Restaurant

a¹

c Restaurant

</s>¹

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

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

Unigram Restaurant

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

$<s>$ Restaurant

\boxed{a}^1

a Restaurant

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

b Restaurant

\boxed{a}^1

c Restaurant

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

$$p(w = \mathbf{b} | \dots) = \frac{1}{5} + \frac{1}{5} \left(\frac{c_{\emptyset, \mathbf{b}}}{c_{\emptyset, \cdot} + \theta} + \frac{\theta}{c_{\emptyset, \cdot} + \theta} \frac{1}{5} \right) \quad (2)$$

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

Unigram Restaurant

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

$<s>$ Restaurant

\boxed{a}^1

a Restaurant

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

b Restaurant

\boxed{a}^1

c Restaurant

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

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

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

Unigram Restaurant

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

$<s>$ Restaurant

\boxed{a}^1

a Restaurant

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

b Restaurant

\boxed{a}^1

c Restaurant

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

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

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

Unigram Restaurant

a³ b¹ c¹ </s>¹

<s> Restaurant

a¹

a Restaurant

a² b¹ c¹

b Restaurant

a¹

c Restaurant

</s>¹

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

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

Unigram Restaurant

a³ b¹ c¹ </s>¹

<s> Restaurant

a¹

a Restaurant

a² b¹ c¹

b Restaurant

a¹

c Restaurant

</s>¹

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

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)

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

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)

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

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 $\langle s \rangle$ b b a c $\langle /s \rangle$; 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 $\langle /s \rangle$ following $\langle s \rangle$

A busy night at the restaurant

Unigram Restaurant

a³ b¹ c¹ </s>¹

<s> Restaurant

a¹

a Restaurant

a² b¹ c¹

b Restaurant

a¹

c Restaurant

</s>¹

A busy night at the restaurant

Unigram Restaurant

a³ b¹ c¹ </s>¹

<s> Restaurant

a¹ b¹

a Restaurant

a² b¹ c¹

b Restaurant

a¹

c Restaurant

</s>¹

A busy night at the restaurant

Unigram Restaurant

a³ b² c¹ </s>¹

<s> Restaurant

a¹ b¹

a Restaurant

a² b¹ c¹

b Restaurant

a¹

c Restaurant

</s>¹

A busy night at the restaurant

Unigram Restaurant

a³ b² c¹ </s>¹

<s> Restaurant

a¹ b¹

a Restaurant

a² b¹ c¹

b Restaurant

a¹ b¹

c Restaurant

</s>¹

A busy night at the restaurant

Unigram Restaurant

a³ b³ c¹ </s>¹

<s> Restaurant

a¹ b¹

a Restaurant

a² b¹ c¹

b Restaurant

a¹ b¹

c Restaurant

</s>¹

A busy night at the restaurant

Unigram Restaurant

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

$\langle s \rangle$ 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

a³ b³ c¹ </s>¹

<s> Restaurant

a¹ b¹

a Restaurant

a² b¹ c²

b Restaurant

a² b¹

c Restaurant

</s>¹

A busy night at the restaurant

Unigram Restaurant

a³ b³ c¹ </s>¹

<s> Restaurant

a¹ b¹

a Restaurant

a² b¹ c²

b Restaurant

a² b¹

c Restaurant

</s>²

A busy night at the restaurant

Unigram Restaurant

a³ b³ c¹ </s>¹

<s> Restaurant

a¹ b¹

a Restaurant

a² b¹ c²

b Restaurant

a² b¹

c Restaurant

</s>²

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

b following **a**

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

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

$$(6)$$

b following **a**

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

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

$$(6)$$

b following **a**

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

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

(6)

0.23

a following b

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

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

$$(9)$$

a following b

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

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

$$(9)$$

a following b

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

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

(9)

0.55

$\langle /s \rangle$ following $\langle s \rangle$

$$= \frac{\theta + 2\delta}{\theta + 2} p(\langle /s \rangle) \quad (10)$$

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

$$(12)$$

</s> following <s>

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

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

$$(12)$$

</s> following <s>

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

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

$$(12)$$

0.08