Language Models

Computational Linguistics: Jordan Boyd-Graber
University of Maryland
INTRODUCTION

Slides adapted from Philip Koehn

Roadmap

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 to model this phenomena

Interpolation

- Higher and lower order n-gram models have different strengths and weaknesses
  - high-order n-grams are sensitive to more context, but have sparse counts
  - low-order n-grams consider only very limited context, but have robust counts
- Combine them

$$
\begin{aligned}
p_I(w_3 \mid w_1, w_2) = \quad & \lambda_1 \, p_1(w_3) \\
& + \lambda_2 \, p_2(w_3 | w_2) \\
& + \lambda_3 \, p_3(w_3 | w_1, w_2)
\end{aligned}
$$

Back-Off

- Trust the highest order language model that contains n-gram

$$p_n^{BO}(w_i|w_{i-n+1},...,w_{i-1}) = \begin{cases} \alpha_n(w_i|w_{i-n+1},...,w_{i-1}) & \text{if } \text{count}_n(w_{i-n+1},...,w_i) > 0 \\ d_n(w_{i-n+1},...,w_{i-1})\, p_{n-1}^{BO}(w_i|w_{i-n+2},...,w_{i-1}) & \text{else} \end{cases}$$

- Requires
  - adjusted prediction model $\alpha_n(w_i|w_{i-n+1},...,w_{i-1})$
  - discounting function $d_n(w_1,...,w_{n-1})$

Let's remember what a language model is

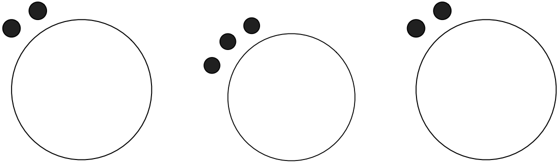
- It is a distribution over the next word in a sentence
- Given the previous $n-1$ words

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

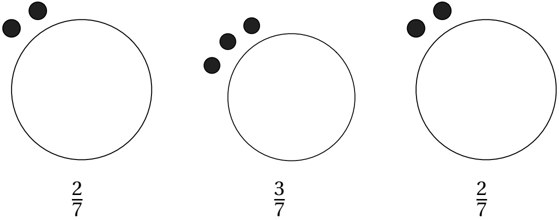
The Chinese Restaurant as a Distribution

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.



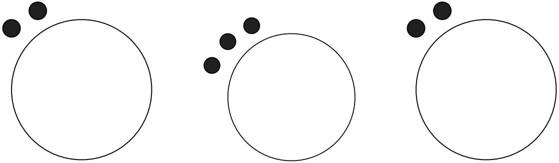
The Chinese Restaurant as a Distribution

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

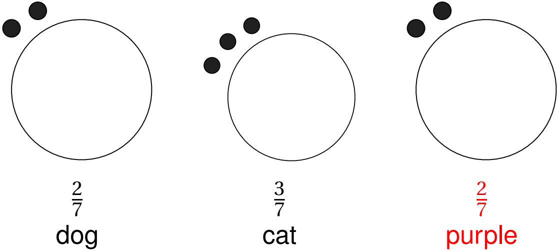
To generate a word, you first sit down at a table. You sit down at a table proportional to the number of people sitting at the table.



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

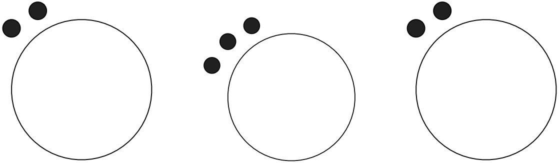
The Chinese Restaurant as a Distribution

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.

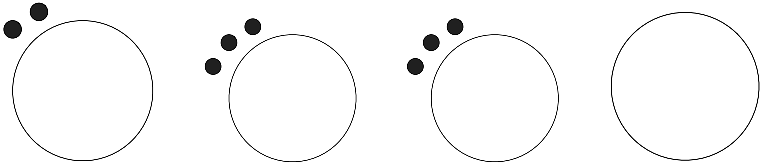


$\frac{2}{7}$ dog $\quad$ $\frac{3}{7}$ cat $\quad$ $\frac{2}{7}$ purple

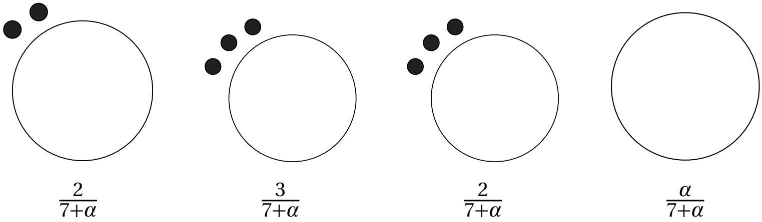
But this is just Maximum Likelihood

Why are we talking about Chinese Restaurants?

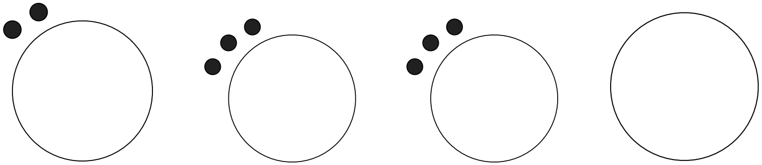
Always one more table . . .



Always one more table . . .



Always one more table . . .



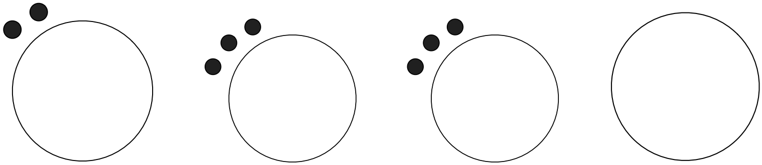
$\frac{2}{7+\alpha}$ dog

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

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

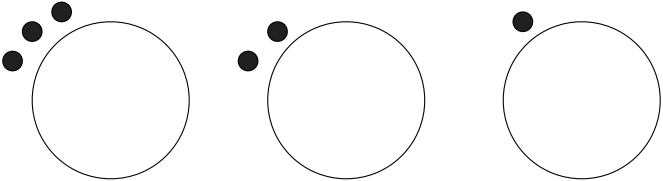
$\frac{\alpha}{7+\alpha}$???

Always one more table . . .

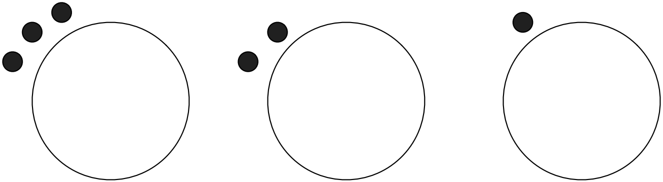


$\frac{2}{7+\alpha}$ dog | $\frac{3}{7+\alpha}$ cat | $\frac{2}{7+\alpha}$ purple | $\frac{\alpha}{7+\alpha}$???

What to do with a new table?



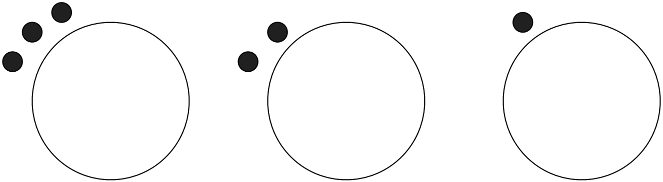
What to do with a new table?



What can be a base distribution?

- Uniform (Dirichlet smoothing)

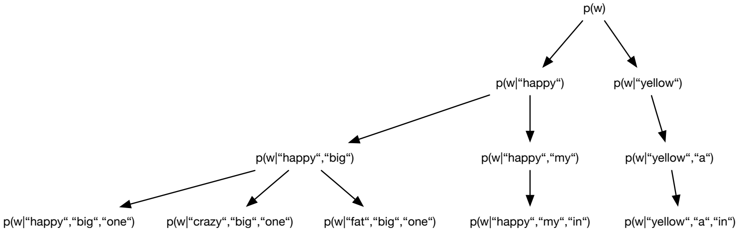
What to do with a new table?



What can be a base distribution?

- Uniform (Dirichlet smoothing)
- Specific contexts $\longrightarrow$ 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]

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

c Restaurant

b Restaurant

Seating Assignments

Dataset:

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

Unigram Restaurant

a [1]

<s> Restaurant

a [1]

b Restaurant

a 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

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

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

a [2] b [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

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)

b Restaurant

* (1)

a Restaurant

a (2) b (1)

c Restaurant

Seating Assignments

Dataset:

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

Unigram Restaurant

a [3] b [1]

<s> Restaurant

a [1]

b Restaurant

a [1]

a Restaurant

a [2] b [1]

c Restaurant

Seating Assignments

Dataset:

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

Unigram Restaurant

a 3 b 1

<s> Restaurant

a 1

b Restaurant

a 1

a Restaurant

a 2 b 1

c Restaurant

Seating Assignments

Dataset:

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

Unigram Restaurant

a [3] b [1]

<s> Restaurant

a [1]

b Restaurant

a [1]

a Restaurant

a [2] b [1] * [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]

b Restaurant

a[1]

a Restaurant

a[2] b[1] * [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

a Restaurant

a 2 b 1 c 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

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

<s> Restaurant

a [1]

a Restaurant

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

b Restaurant

a [1]

c Restaurant

* [1]

Seating Assignments

Dataset:

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

Unigram Restaurant

a ^{3} b ^{1} c ^{1} * ^{1}

<s> Restaurant

a ^{1}

a Restaurant

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

b Restaurant

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

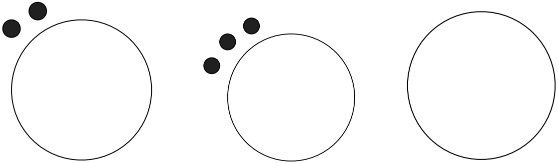
Real examples

- 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



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

Computing the Probability of an Observation

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

- 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

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

- 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 $\pi(u)$

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

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]

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

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

Unigram Restaurant

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]

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

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]

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

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

Unigram Restaurant

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]

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

How do many crps encode backoff?

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

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

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]

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

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]

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

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]

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

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]

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

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]

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

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]

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

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]

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

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]

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

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)

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

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)

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

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)

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

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)

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)
- Neural language models ...