A Note on the Implementation of Hierarchical Dirichlet Processes

Phil Blunsom*, Trevor Cohn*,
Sharon Goldwater* and Mark Johnson†

*School of Informatics, University of Edinburgh
†Department of Cognitive and Linguistic Sciences, Brown University

August 4, 2009

Outline

- GGJ06[1] introduced an approximation for use in hierarchical Dirichlet process (HDP) inference: It's wrong, don't use it.
- We correct that approximation for DP models. However, this doesn't extend to HDPs.
- But that's ok because we'll describe an efficient exact implementation.

[1] S. Goldwater, T. Griffiths, M. Johnson. Contextual dependencies in unsupervised word segmentation. ACL/COLING-06

Outline

- GGJ06[1] introduced an approximation for use in hierarchical Dirichlet process (HDP) inference: **It's wrong, don't use it.**
- We correct that approximation for DP models. **However, this doesn't extend to HDPs.**
- But that's ok because we'll describe an efficient exact implementation.

[1] S. Goldwater, T. Griffiths, M. Johnson. Contextual dependencies in unsupervised word segmentation. ACL/COLING-06

Outline

- GGJ06[1] introduced an approximation for use in hierarchical Dirichlet process (HDP) inference: **It's wrong, don't use it.**
- We correct that approximation for DP models. **However, this doesn't extend to HDPs.**
- But that's ok because we'll describe an efficient exact implementation.

[1] S. Goldwater, T. Griffiths, M. Johnson. Contextual dependencies in unsupervised word segmentation. ACL/COLING-06

Outline

- GGJ06[1] introduced an approximation for use in hierarchical Dirichlet process (HDP) inference: **It's wrong, don't use it.**
- We correct that approximation for DP models. **However, this doesn't extend to HDPs.**
- But that's ok because we'll describe an efficient exact implementation.

[1] S. Goldwater, T. Griffiths, M. Johnson. Contextual dependencies in unsupervised word segmentation. ACL/COLING-06

Outline

- GGJ06[1] introduced an approximation for use in hierarchical Dirichlet process (HDP) inference: **It's wrong, don't use it.**
- We correct that approximation for DP models. **However, this doesn't extend to HDPs.**
- But that's ok because we'll describe an efficient exact implementation.

[1] S. Goldwater, T. Griffiths, M. Johnson. Contextual dependencies in unsupervised word segmentation. ACL/COLING-06

The Chinese Restaurant Process

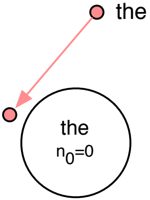
In a Dirichlet Process unigram language model words $w_1 \dots w_n$ are generated as follows:

$$
\begin{aligned}
G|\alpha_0, P_0 &\sim \mathrm{DP}(\alpha_0, P_0) \\
w_i|G &\sim G
\end{aligned}
$$

- G is a distribution over an infinite set of words,
- P_0 is the probability that an word will be in the support of G,
- α_0 determines the variance of G.

One way of understanding the predictions made by the DP model is through the Chinese restaurant process (CRP) ...

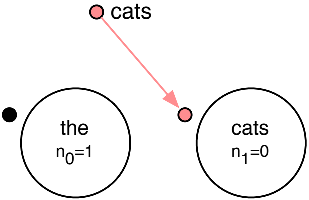
The Chinese Restaurant Process



Customers (words) enter a restaurant and choose a table according to the distribution:

$$P(z_i = k | w_i = w, \mathbf{z}_{-i}) = \begin{cases} \frac{n_k^{\mathbf{z}_{-i}}}{n_w + \alpha_0 P_0(w)}, 0 \leq k < |k| \\ \\ \frac{\alpha_0 P_0(w)}{n_w + \alpha_0 P_0(w)}, k = |k| \end{cases}$$

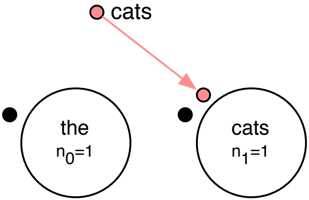
The Chinese Restaurant Process



Customers (words) enter a restaurant and choose a table according to the distribution:

$$P(z_i = k | w_i = w, \mathbf{z}_{-i}) = \begin{cases} \frac{n_k^{\mathbf{z}_{-i}}}{n_w + \alpha_0 P_0(w)}, 0 \leq k < |k| \\ \frac{\alpha_0 P_0(w)}{n_w + \alpha_0 P_0(w)}, k = |k| \end{cases}$$

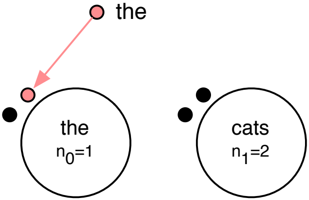
The Chinese Restaurant Process



Customers (words) enter a restaurant and choose a table according to the distribution:

$$P(z_i = k | w_i = w, \mathbf{z}_{-i}) = \begin{cases} \frac{n_k^{\mathbf{z}_{-i}}}{n_w + \alpha_0 P_0(w)}, 0 \leq k < |k| \\ \frac{\alpha_0 P_0(w)}{n_w + \alpha_0 P_0(w)}, k = |k| \end{cases}$$

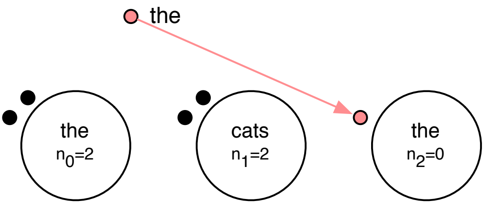
The Chinese Restaurant Process



Customers (words) enter a restaurant and choose a table according to the distribution:

$$P(z_i = k | w_i = w, \mathbf{z}_{-i}) = \begin{cases} \frac{n_k^{\mathbf{z}_{-i}}}{n_w + \alpha_0 P_0(w)}, 0 \leq k < |k| \\ \frac{\alpha_0 P_0(w)}{n_w + \alpha_0 P_0(w)}, k = |k| \end{cases}$$

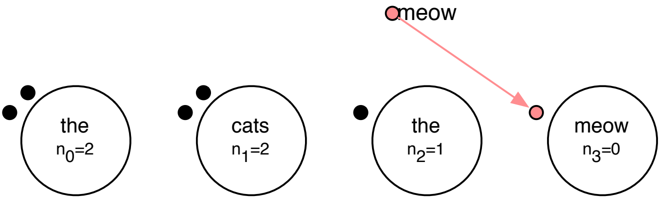
The Chinese Restaurant Process



Customers (words) enter a restaurant and choose a table according to the distribution:

$$P(z_i = k | w_i = w, \mathbf{z}_{-i}) = \begin{cases} \frac{n_k^{\mathbf{z}_{-i}}}{n_w + \alpha_0 P_0(w)}, 0 \leq k < |k| \\ \\ \frac{\alpha_0 P_0(w)}{n_w + \alpha_0 P_0(w)}, k = |k| \end{cases}$$

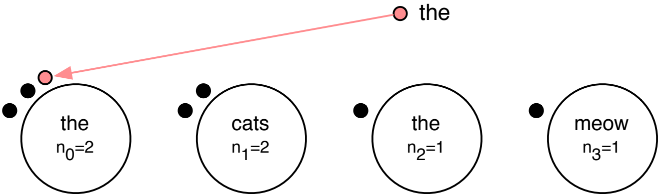
The Chinese Restaurant Process



Customers (words) enter a restaurant and choose a table according to the distribution:

$$P(z_i = k | w_i = w, \mathbf{z}_{-i}) = \begin{cases} \frac{n_k^{\mathbf{z}_{-i}}}{n_w + \alpha_0 P_0(w)}, 0 \leq k < |k| \\ \frac{\alpha_0 P_0(w)}{n_w + \alpha_0 P_0(w)}, k = |k| \end{cases}$$

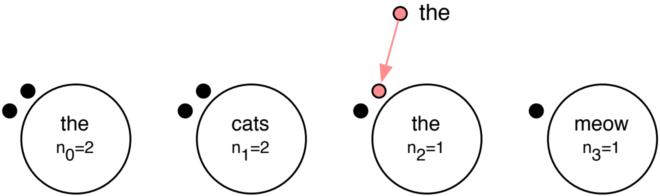
The Chinese Restaurant Process



The 7^{th} customer '*the*' enters the restaurant and choses a table from those already seating '*the*', or opening a new table:

$$P(z_6 = 0 | w_6 = the, \mathbf{z}_{-6}) = \frac{2}{3 + \alpha_0 P_0(the)}$$

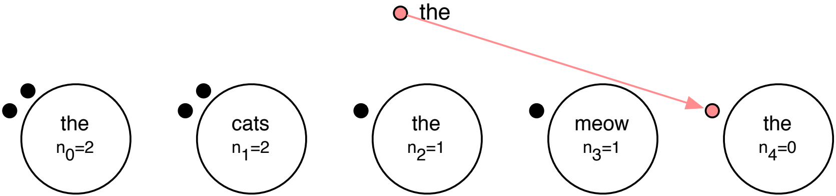
The Chinese Restaurant Process



The 7^{th} customer '*the*' enters the restaurant and choses a table from those already seating '*the*', or opening a new table:

$$P(z_6 = 2 | w_6 = the, \mathbf{z}_{-6}) = \frac{1}{3 + \alpha_0 P_0(the)}$$

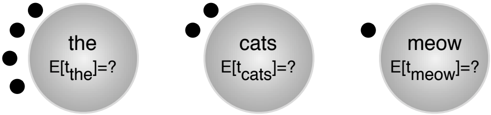
The Chinese Restaurant Process



The 7^{th} customer '*the*' enters the restaurant and choses a table from those already seating '*the*', or opening a new table:

$$P(z_6 = 4 | w_6 = the, \mathbf{z}_{-6}) = \frac{P_0(the)}{3 + \alpha_0 P_0(the)}$$

Approximating the table counts



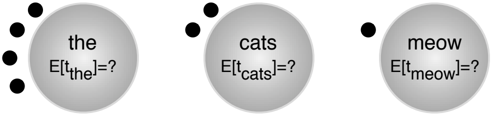
- GGJ06 sought to avoid explicitly tracking tables by reasoning under the expected table counts ($E[t_w]$).
- Antoniak(1974) derives the expected table count as equal to the recurrence:

$$E[t_w] = \alpha_0 P_0(w) \sum_{i=1}^{n_w} \frac{1}{\alpha_0 P_0(w) + i - 1}$$

- Antoniak also suggests an approximation to this expectation which GGJ06 presents as:

$$E[t_w] \approx \alpha_0 \log \frac{n_w + \alpha_0}{\alpha_0}$$

Approximating the table counts



- GGJ06 sought to avoid explicitly tracking tables by reasoning under the expected table counts ($E[t_w]$).
- Antoniak(1974) derives the expected table count as equal to the recurrence:

$$E[t_w] = \alpha_0 P_0(w) \sum_{i=1}^{n_w} \frac{1}{\alpha_0 P_0(w) + i - 1}$$

- Antoniak also suggests an approximation to this expectation which GGJ06 presents as: (corrected)

$$E[t_w] \approx \alpha_0 P_0(w) \log \frac{n_w + \alpha_0 P_0(w)}{\alpha_0 P_0(w)}$$

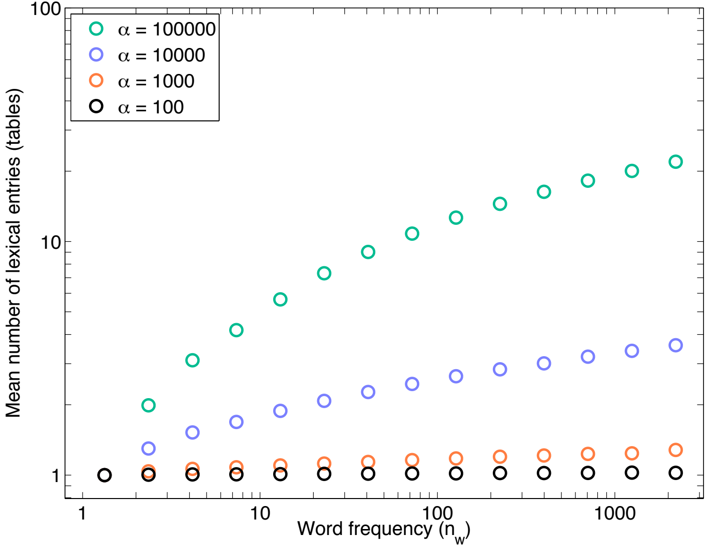
A better table count approximation

- Antoniak's approximation makes two assumptions:
  - α_0 is large, not the predominant situation in recent applications which employ a DP as a sparse prior,
  - $P_0(w)$ is constant, which is not applicable to HDPs.
- In our paper we derive an improved approximation based on a difference of digamma (ψ) functions:

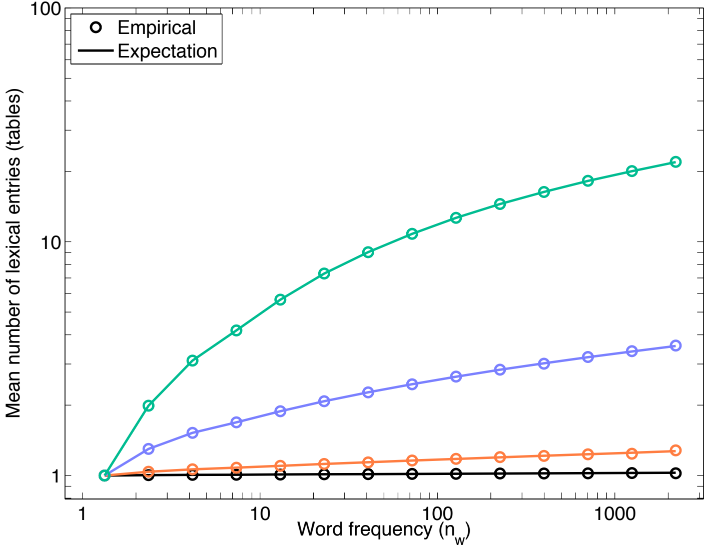
$$E[t_w] = \alpha_0 P_0(w) \cdot \left[\psi\Big(\alpha_0 P_0(w) + n_w\Big) - \psi\Big(\alpha_0 P_0(w)\Big) \right]$$

- However the restriction on $P_0(w)$ being constant remains ...

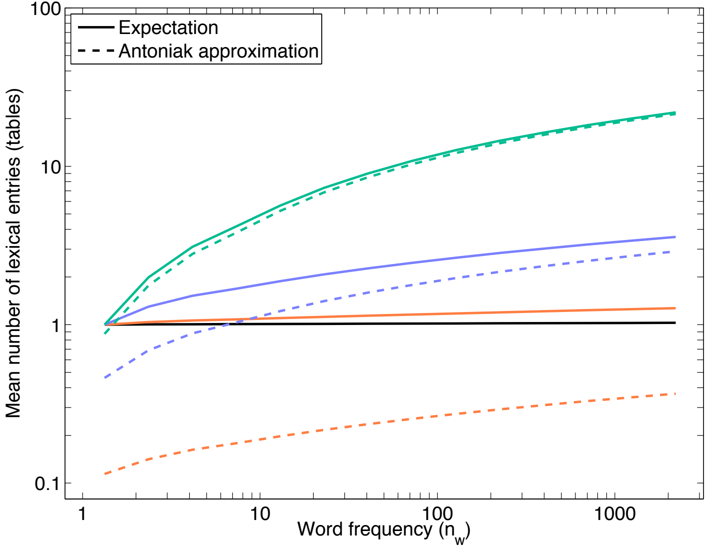
DP performance



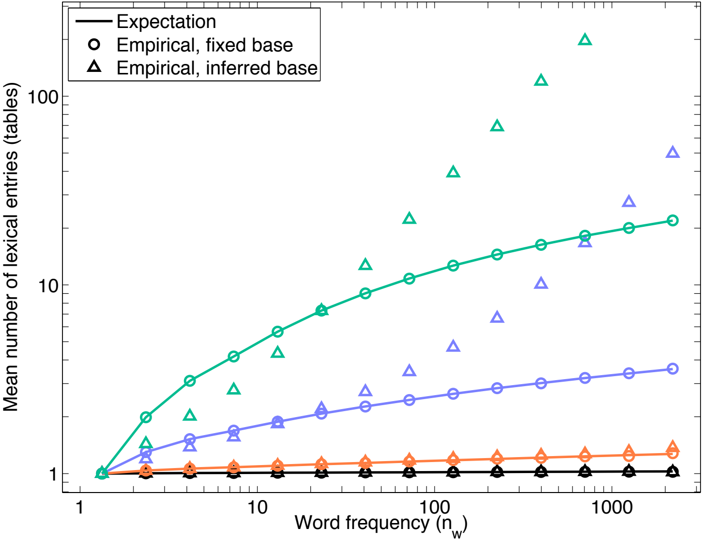
DP performance



DP performance



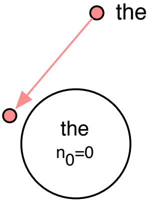
HDP performance

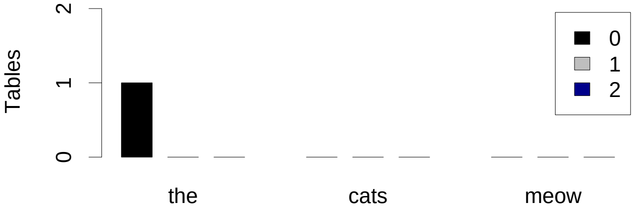


Histogram Method

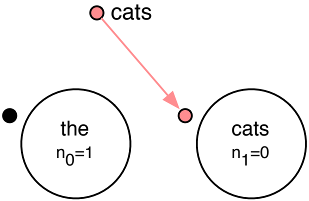
- At this point we don't have a useful approximation of the expected table counts in a HDP model.
- However, we can describe a more compact representation for the state of the restaurant that doesn't require explicit table tracking.
- Instead we maintain a histogram for each dish w_i of the frequency of a table having a particular number of customers.

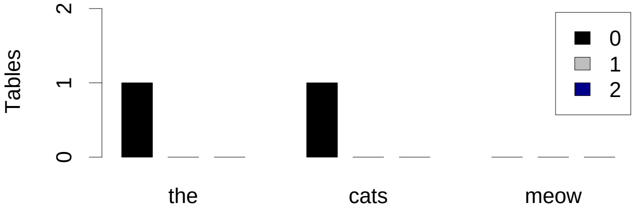
Histogram Method



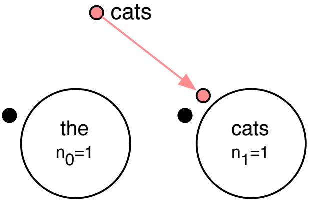


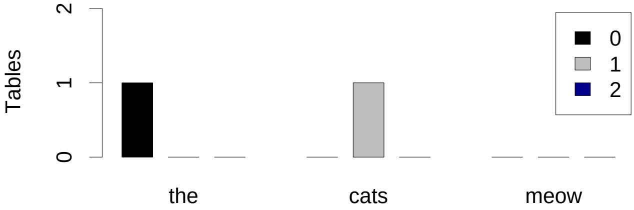
Histogram Method



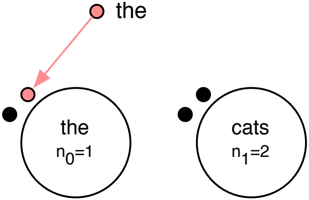


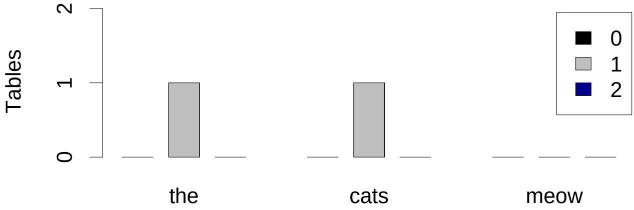
Histogram Method



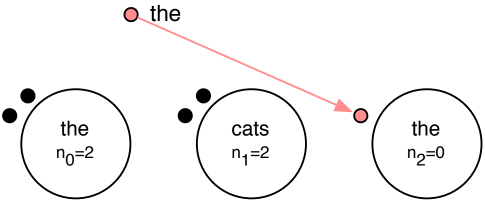


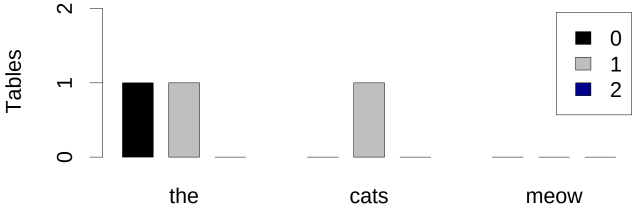
Histogram Method



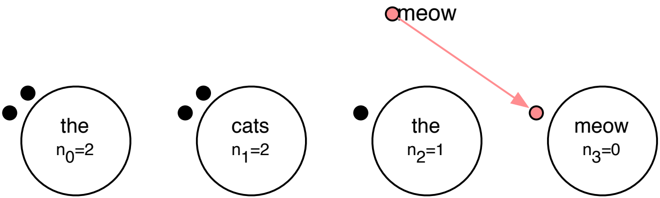


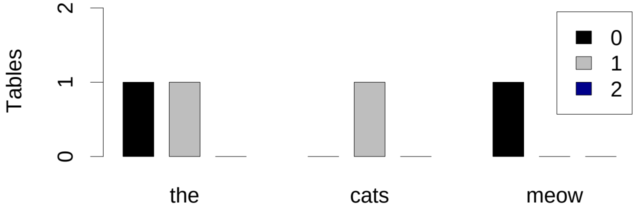
Histogram Method



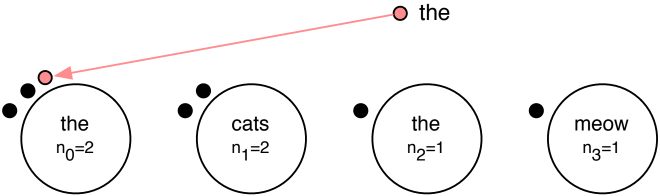


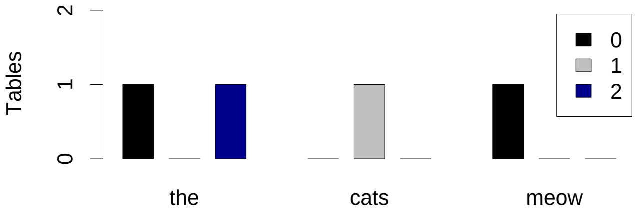
Histogram Method



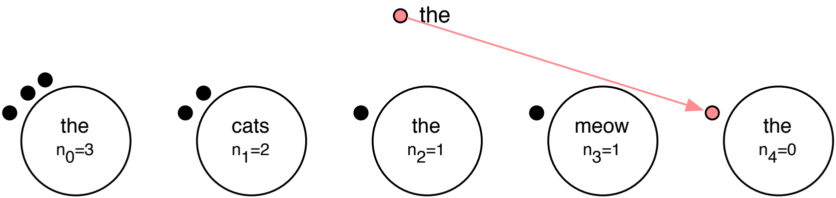


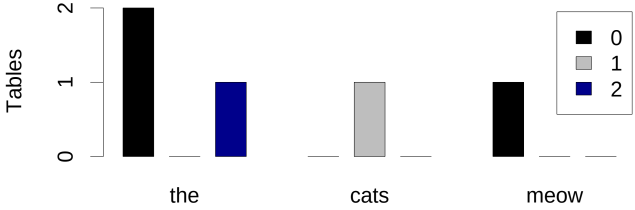
Histogram Method





Histogram Method





Summary

The table count approximation of Goldwater et al. 2006 is broken, don't use it!

Thank you.

References

P. Blunsom, T. Cohn, S. Goldwater and M. Johnson. A note on the implementation of hierarchical Dirichlet processes, *In the Proceedings of ACL-IJCNLP 2009*.

C. E. Antoniak. 1974. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152-1174.

S. Goldwater, T. Griffiths, M. Johnson. Contextual dependencies in unsupervised word segmentation. *In the Proceedings of (COLING/ACL-2006)*.