A Hierarchical Nonparametric Bayesian Approach to Statistical Language Model Domain Adaptation

GATSBY COMPUTATIONAL NEUROSCIENCE UNIT
supported by the Gatsby Charitable Foundation

FRANK WOOD, YEE WHYE TEH

Statistical Natural Language Modelling (SLM)

- Learning distributions over sequences of discrete observations (words)

$$P(w_{1:N}) \qquad w_{1:N} = [w_1 w_2 \ldots w_N]$$

- Useful for
  - Automatic Speech Recognition (ASR)

    $$\text{P(text | speech)} \propto \text{P(speech | text)} \boxed{\text{P(text)}}$$

  - Machine Translation (MT)

    $$\text{P(english | french)} \propto \text{P(french | english)} \boxed{\text{P(english)}}$$

Markov "n–gram" models

Discrete conditional distribution over words that follow given context

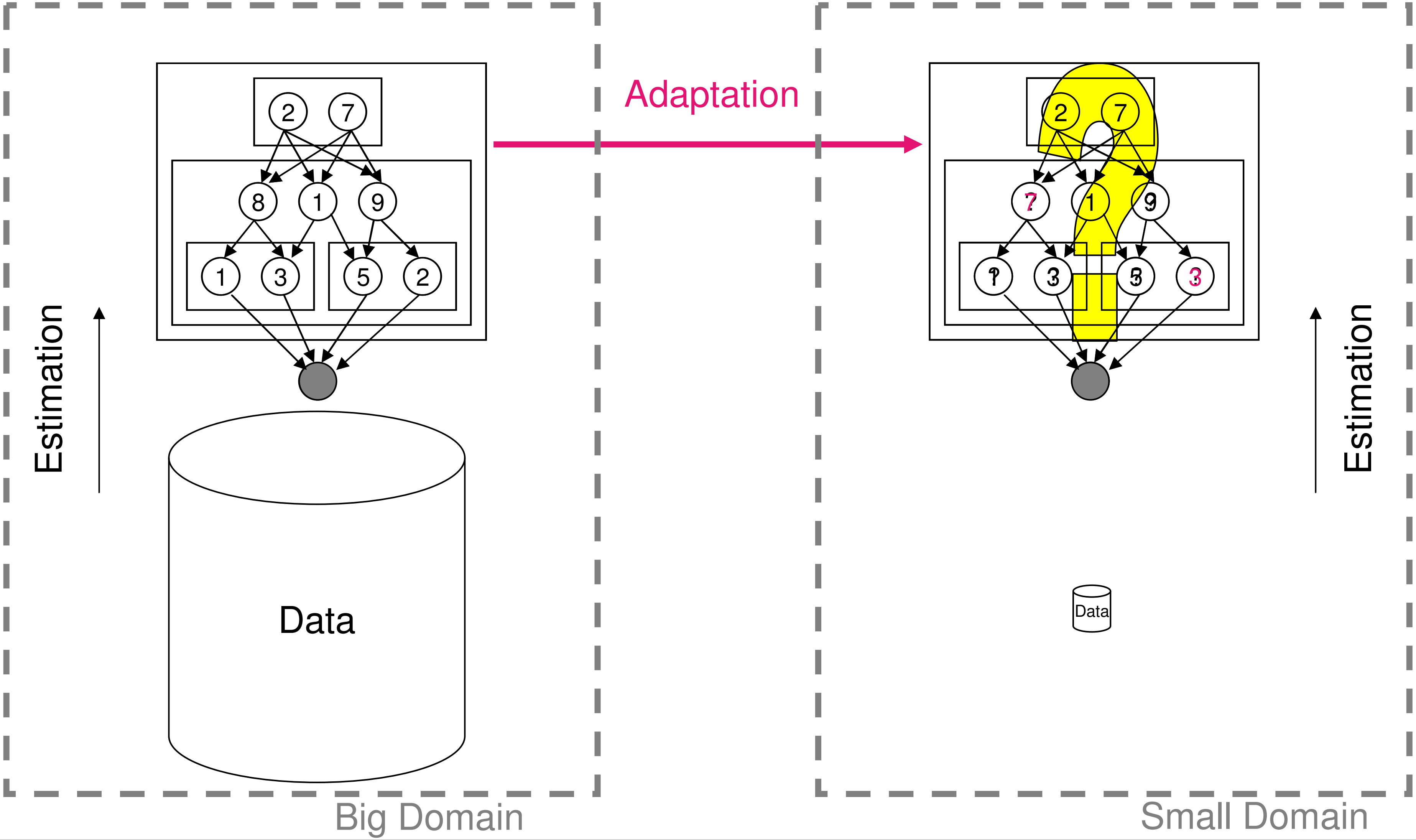
$$P(w_{1:N}) \stackrel{\text{def}}{=} \prod_{n=1}^{N} P(\underbrace{w_n | w_{n-1:n-2}}_{\text{tri-gram}}) = \prod_{n=1}^{N} \boxed{\mathcal{G}}_{\{\boxed{w_{n-1:n-2}}\}}(\boxed{w_n})$$

Context

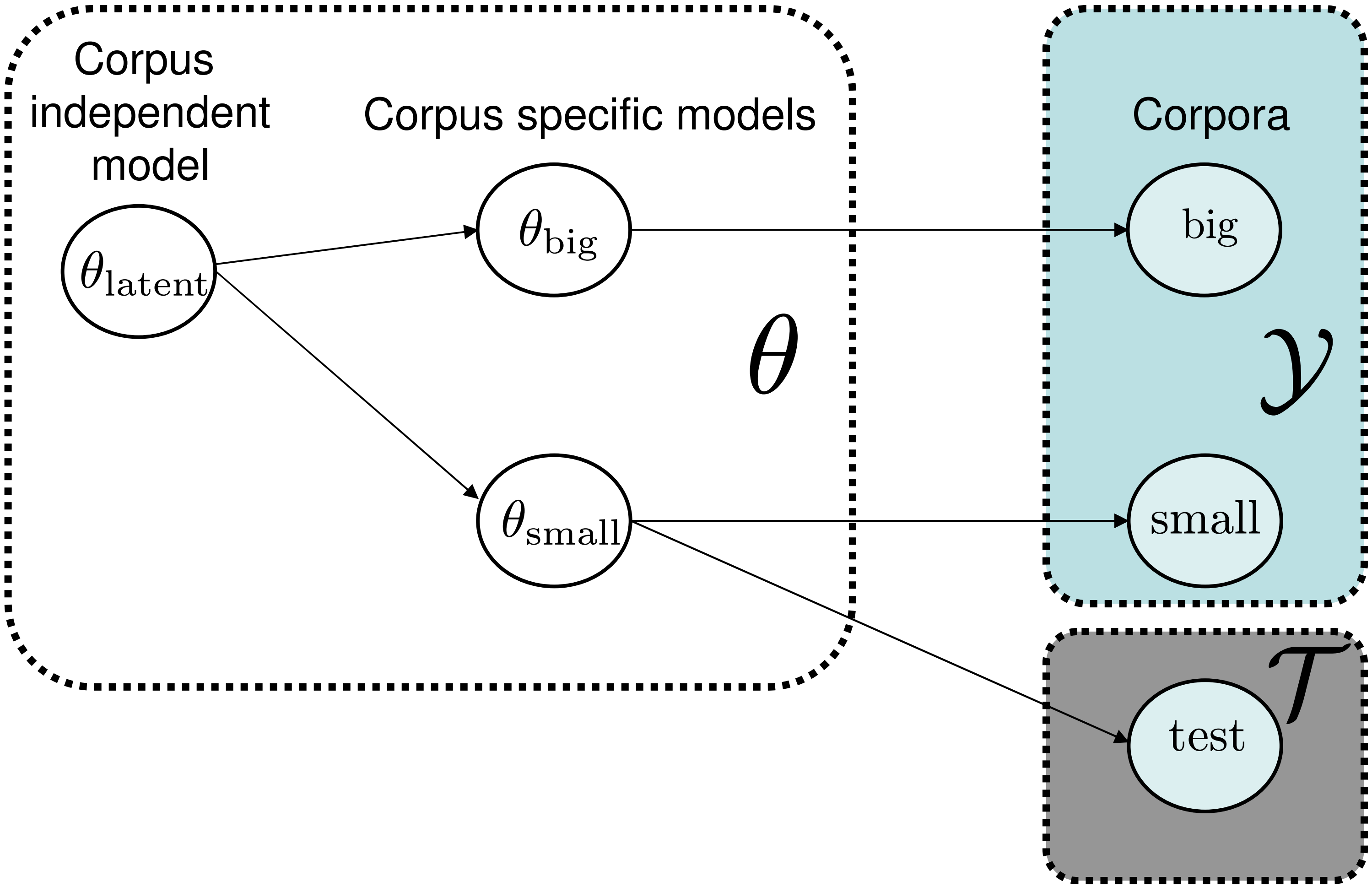
Word following context

e.g. mice | blind, three
league | premier, Barclay's

Domain Adaptation



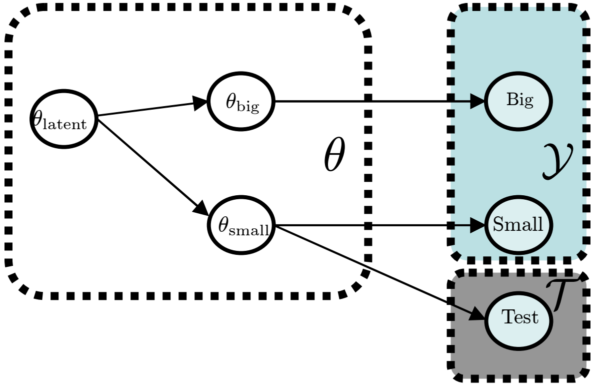
Hierarchical Bayesian Approach



Bayesian Domain Adaptation

- Estimation goal

$$P(\theta|\mathcal{Y}) \propto P(\mathcal{Y}|\theta)P(\theta)$$



- Inference objective

$$P(\mathcal{T}|\mathcal{Y}) = \int P(\mathcal{T}|\theta)P(\theta|\mathcal{Y})d\theta$$

- Simultaneous estimation = automatic domain adaptation

Model Structure : The Likelihood

- One *domain-specific* *n*-gram (Markov) model

Distribution over words that follow given context

Corpus

Word following context

$$P(\boxed{\mathcal{D}}|\boxed{\theta_{\mathcal{D}}}) \stackrel{\text{def}}{=} \prod_{n=1}^{N} \boxed{\mathcal{G}}^{\mathcal{D}}_{\{\boxed{w^{\mathcal{D}}_{n-2:n-1}}\}}(\boxed{w^{\mathcal{D}}_{n}}|\boxed{\theta_{\mathcal{D}}})$$

Corpus specific parameters

Context

for each domain

$$\mathcal{D} \in \{\text{big}, \text{small}\}$$

Model Structure : The Prior

- Doubly hierarchical Pitman-Yor process language model (DHPYLM)

Prior (parameters θ)

$$
\begin{aligned}
\mathcal{G}^{\mathcal{D}}_{\{\}} &\sim \mathrm{PY}(d^{\mathcal{D}}_0, \alpha^{\mathcal{D}}_0, \lambda^{\mathcal{D}}_0 \mathcal{U} + (1-\lambda^{\mathcal{D}}_0) G^{\mathcal{H}}_{\{\}}) \\
\mathcal{G}^{\mathcal{D}}_{\{w_{t-1}\}} &\sim \mathrm{PY}(d^{\mathcal{D}}_1, \alpha^{\mathcal{D}}_1, \lambda^{\mathcal{D}}_1 \mathcal{G}^{\mathcal{D}}_{\{\}} + (1-\lambda^{\mathcal{D}}_1) \mathcal{G}^{\mathcal{H}}_{\{w_{t-1}\}}) \\
&\vdots \\
\mathcal{G}^{\mathcal{D}}_{\{w_{t-j}:w_{t-1}\}} &\sim \mathrm{PY}(d^{\mathcal{D}}_j, \alpha^{\mathcal{D}}_j, \lambda^{\mathcal{D}}_j \mathcal{G}^{\mathcal{D}}_{\{w_{t-j+1}:w_{t-1}\}} + (1-\lambda^{\mathcal{D}}_j) \mathcal{G}^{\mathcal{H}}_{\{w_{t-j}:w_{t-1}\}}) \\
x | w_{t-n+1} : w_{t-1} &\sim \mathcal{G}^{\mathcal{D}}_{\{w_{t-n+1}:w_{t-1}\}}.
\end{aligned}
$$

Likelihood

Every line is conditioned on everything on the right

Novel generalization of back-off

Pitman-Yor Process

$$\begin{aligned} \mathcal{G} &\sim \mathrm{PY}(\underbrace{d}_{\text{discount}}, \underbrace{\alpha}_{\text{concentration}}, \underbrace{\mathcal{G}_0}_{\text{base distribution}}) \\ x &\sim \mathcal{G} \end{aligned}$$

- Distribution over distributions
- Base distribution is the "mean"

$$E[\mathcal{G}(dx)] = \mathcal{G}_0(dx)$$

- Generalization of the Dirichlet Process (d = 0)
  – Different (power-law) "clustering" properties
  – Better for text [Teh, 2006]

DHPYLM

- Focus on a "leaf"
- Imagine recursion

Base distribution

Domain-specific distribution over words following shorter context (by 1)

$$\boxed{\mathcal{G}^{\mathcal{D}}_{\{w_{t-j}:w_{t-1}\}}} \sim \mathrm{PY}\Big(d^{\mathcal{D}}_j, \alpha^{\mathcal{D}}_j, \overbrace{\lambda^{\mathcal{D}}_j \boxed{\mathcal{G}^{\mathcal{D}}_{\{w_{t-j+1}:w_{t-1}\}}} + (1-\lambda^{\mathcal{D}}_j) \boxed{\mathcal{G}^{\mathcal{H}}_{\{w_{t-j}:w_{t-1}\}}}}^{\text{Base distribution}}\Big)$$

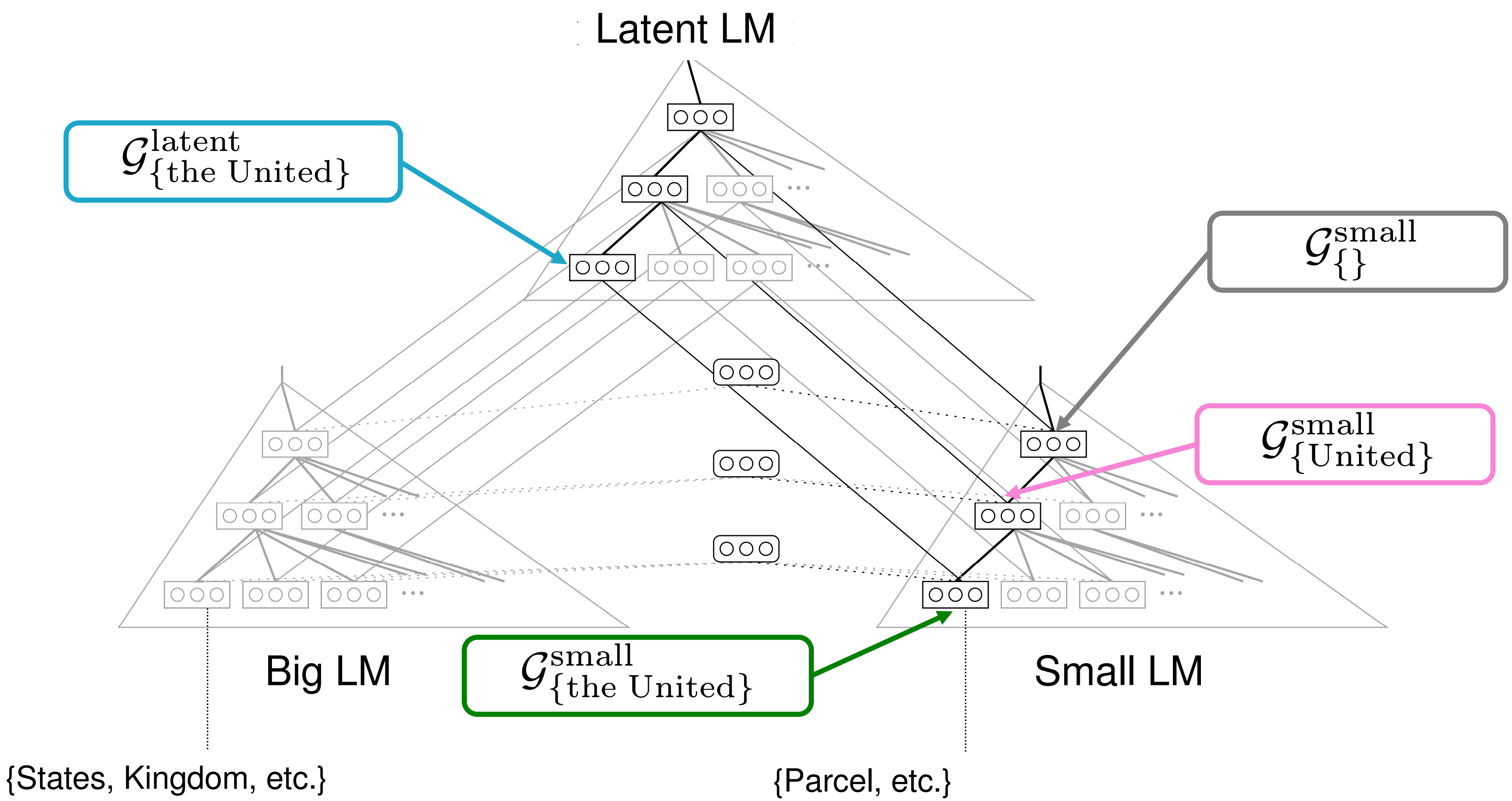
$$\boxed{x | w_{t-n+1} : w_{t-1}} \sim \boxed{\mathcal{G}^{\mathcal{D}}_{\{w_{t-n+1}:w_{t-1}\}}}.$$

Non-specific distribution over words following full context

Observations

Distribution over words following given context

Tree-Shaped Graphical Model



Intuition

- Domain = “The Times of London”
- Context = “the United”
- The distribution over subsequent words *will* look like:

$$
\begin{array}{ll|ll}
\mathcal{G}_{\{\text{the United}\}}(\text{aa}) & = .0000001 & & \vdots \\
\mathcal{G}_{\{\text{the United}\}}(\text{aah}) & = .0000001 & \mathcal{G}_{\{\text{the United}\}}(\text{States}) & = .05 \\
& \vdots & & \vdots \\
\mathcal{G}_{\{\text{the United}\}}(\text{Kingdom}) & = .25 & \mathcal{G}_{\{\text{the United}\}}(\text{zyzzyva}) & = .0000001
\end{array}
$$

Problematic Example

- Domain = "The Times of London"
- Context = "quarterback Joe"
- The distribution over subsequent words *should* look like:

$$
\begin{array}{ll|ll}
\mathcal{G}_{\{\text{quarterback Joe}\}}(\text{aa}) & = .0000001 & & \vdots \\
\mathcal{G}_{\{\text{quarterback Joe}\}}(\text{aah}) & = .0000001 & \mathcal{G}_{\{\text{quarterback Joe}\}}(\text{Namath}) & = .15 \\
 & \vdots & & \vdots \\
\mathcal{G}_{\{\text{quarterback Joe}\}}(\text{Montana}) & = .25 & \mathcal{G}_{\{\text{quarterback Joe}\}}(\text{zyzzyva}) & = .0000001
\end{array}
$$

- But where could this come from?

The Model Estimation Problem

- Domain = "The Times of London"
- Context = "quarterback Joe"
- No counts for American football phrases in UK publications! (hypothetical but nearly true)

$$\mathcal{G}_{\{\text{quarterback Joe}\}}(\text{Montana}) \approx \frac{\#\{\text{quarterback Joe Montana}\}}{\#\{\text{quarterback Joe}\}} = 0$$

Not how we do estimation!

A Solution: In-domain Back-Off

- Regularization $\leftrightarrow$ smoothing $\leftrightarrow$ backing-off
- In-domain back-off
  - [Kneser & Ney, Chen & Goodman, MacKay & Peto, Teh], etc.
  - Use counts from shorter contexts, *roughly*:

$$\begin{aligned}\mathcal{G}_{\{\text{quarterback Joe}\}}(\text{Montana}) &\approx \pi_3 \frac{\#\{\text{quarterback Joe Montana}\}}{\#\{\text{quarterback Joe}\}} + \pi_2 \frac{\#\{\text{Joe Montana}\}}{\#\{\text{Joe}\}} + \pi_1 \frac{\#\{\text{Montana}\}}{N} \\ &\approx \pi_3 \frac{\#\{\text{quarterback Joe Montana}\}}{\#\{\text{quarterback Joe}\}} + \pi_2 \mathcal{G}_{\{\text{Joe}\}}(\text{Montana})\end{aligned}$$

Recursive!

- Due to zero-counts, Times of London training data alone will *still* yield a skewed distribution

Our Approach: Generalized Back-Off

- Use inter-domain shared counts too.
  – *roughly*:

Probability that "Montana" follows "quarterback Joe"

Out-of-domain back-off; same context, no domain-specificity

$$\boxed{\mathcal{G}^{\mathrm{TL}}_{\{\text{quarterback Joe}\}}(\text{Montana})} \approx \pi_3 \underbrace{\frac{\#\{\text{quarterback Joe Montana}\}}{\#\{\text{quarterback Joe}\}}}_{\text{In-domain counts}} + \pi_2(\lambda \boxed{\mathcal{G}^{\mathrm{TL}}_{\{\text{Joe}\}}(\text{Montana})} + (1-\lambda)\boxed{\mathcal{G}_{\{\text{quarterback Joe}\}}(\text{Montana})})$$

In-domain back-off; shorter context

TL: Times of London

DHPYLM generalized back-off

- Desired intuitive form

$$\mathcal{G}^{\mathrm{TL}}_{\{\text{quarterback Joe}\}}(\text{Montana})$$
$$\approx \pi_3 \frac{\#\{\text{quarterback Joe Montana}\}}{\#\{\text{quarterback Joe}\}} + \pi_2 (\lambda \mathcal{G}^{\mathrm{TL}}_{\{\text{Joe}\}}(\text{Montana}) + (1-\lambda)\mathcal{G}_{\{\text{quarterback Joe}\}}(\text{Montana}))$$

- Generic Pitman-Yor single-sample posterior predictive distribution

$$P(x_{n+1}|x_{1:n};\alpha,d) = \frac{\sum_{k=1}^{K}(c_k - d)}{\alpha+N}\delta(\phi_k - x_{n+1}) + \frac{\alpha+dK}{\alpha+N}\mathcal{G}_0(x_{n+1})$$

- Leaf-layer of DHPYLM

$$\mathcal{G}^{\mathcal{D}}_{\{w_{t-j}:w_{t-1}\}} \sim \mathrm{PY}(d^{\mathcal{D}}_j, \alpha^{\mathcal{D}}_j, \lambda^{\mathcal{D}}_j \mathcal{G}^{\mathcal{D}}_{\{w_{t-j+1}:w_{t-1}\}} + (1-\lambda^{\mathcal{D}}_j)\mathcal{G}^{\mathcal{H}}_{\{w_{t-j}:w_{t-1}\}})$$

DHPYLM – Latent Language Model

$$\begin{aligned}
\mathcal{G}^{\mathcal{H}}_{\{\}} &\sim \mathrm{PY}(d^{\mathcal{H}}_0, \alpha^{\mathcal{H}}_0, \mathcal{U}) \\
\mathcal{G}^{\mathcal{H}}_{\{w_{t-1}\}} &\sim \mathrm{PY}(d^{\mathcal{H}}_1, \alpha^{\mathcal{H}}_1, \mathcal{G}^{\mathcal{H}}_{\{\}}) \\
&\vdots \\
\mathcal{G}^{\mathcal{H}}_{\{w_{t-j:t-1}\}} &\sim \mathrm{PY}(d^{\mathcal{H}}_j, \alpha^{\mathcal{H}}_j, \mathcal{G}^{\mathcal{H}}_{\{w_{t-j+1:t-1}\}})
\end{aligned}$$

Generates Latent Language Model

… no associated observations

DHPYLM – Domain Specific Model

$$
\begin{aligned}
\mathcal{G}^{\mathcal{D}}_{\{\}} &\sim \mathrm{PY}(d^{\mathcal{D}}_0, \alpha^{\mathcal{D}}_0, \lambda^{\mathcal{D}}_0 \mathcal{U} + (1-\lambda^{\mathcal{D}}_0) G^{\mathcal{H}}_{\{\}}) \\
\mathcal{G}^{\mathcal{D}}_{\{w_{t-1}\}} &\sim \mathrm{PY}(d^{\mathcal{D}}_1, \alpha^{\mathcal{D}}_1, \lambda^{\mathcal{D}}_1 \mathcal{G}^{\mathcal{D}}_{\{\}} + (1-\lambda^{\mathcal{D}}_1) \mathcal{G}^{\mathcal{H}}_{\{w_{t-1}\}}) \\
&\vdots \\
\mathcal{G}^{\mathcal{D}}_{\{w_{t-j}:w_{t-1}\}} &\sim \mathrm{PY}(d^{\mathcal{D}}_j, \alpha^{\mathcal{D}}_j, \lambda^{\mathcal{D}}_j \mathcal{G}^{\mathcal{D}}_{\{w_{t-j+1}:w_{t-1}\}} + (1-\lambda^{\mathcal{D}}_j) \mathcal{G}^{\mathcal{H}}_{\{w_{t-j}:w_{t-1}\}}) \\
x|w_{t-n+1}:w_{t-1} &\sim \mathcal{G}^{\mathcal{D}}_{\{w_{t-n+1}:w_{t-1}\}}.
\end{aligned}
$$

Generates Domain Specific Language Model

An Generalization

The “Graphical Pitman-Yor Process”

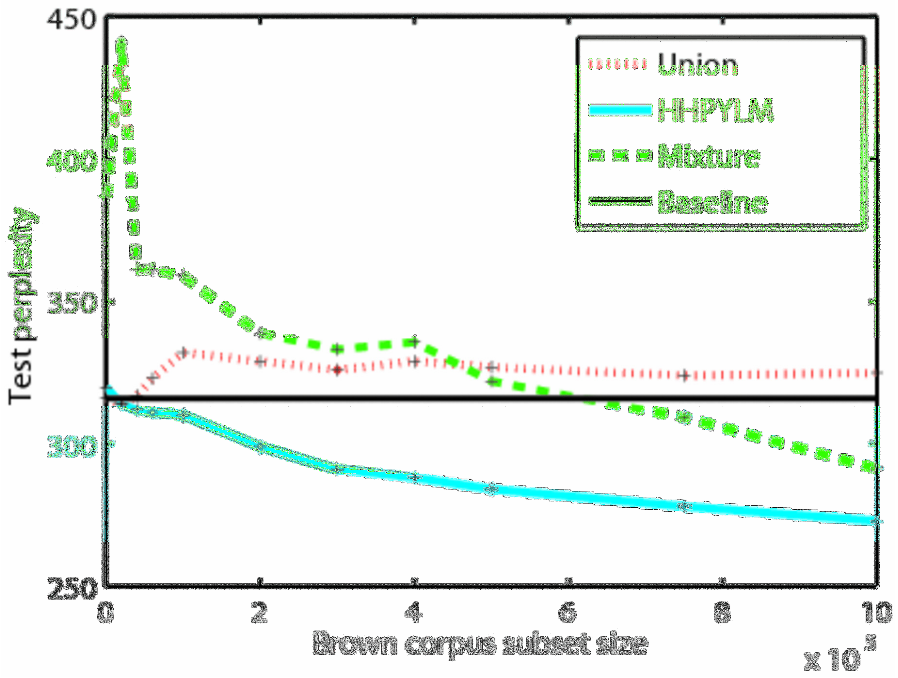
$$\begin{aligned} \Lambda_v &\sim \mathcal{S}_v \\ \mathcal{G}_v | \{G_w : w \in \mathrm{Pa}(v)\} &\sim \mathrm{PY}(d_v, \alpha_v, \sum_{w \in \mathrm{Pa}(v)} \lambda_{w \to v} \mathcal{G}_w) \end{aligned}$$

- Inference in the graphical Pitman-Yor process accomplished via a collapsed Gibbs auxiliary variable sampler

Graphical Pitman-Yor Process Inference

- Auxiliary variable collapsed Gibbs sampler
  - "Chinese restaurant franchise" representation
  - Must keep track of which parent *restaurant* each table comes from
  - "Multi-floor Chinese restaurant franchise"
    - Every *table* has two labels
      - ϕ_k : the parameter (here a word "type")
      - s_k : the parent restaurant from which it came

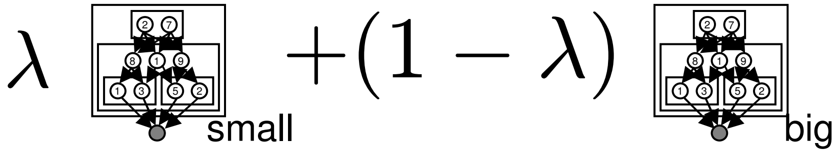
SOU / Brown



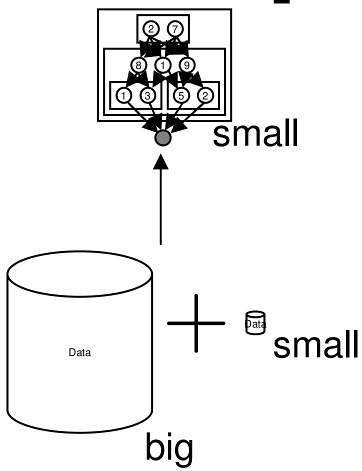
- Training corpora
  - Small
    - State of the Union (SOU)
      - 1945-2006
      - ~ 370,000 words, 13,000 unique
  - Big
    - Brown
      - 1967
      - ~ 1,000,000 words, 50,000 unique
- Test corpus
  - Johnson's SOU Addresses
    - 1963-1969
    - ~ 37,000 words

SLM Domain Adaptation Approaches

- Mixture [Kneser & Steinbiss, 1993]

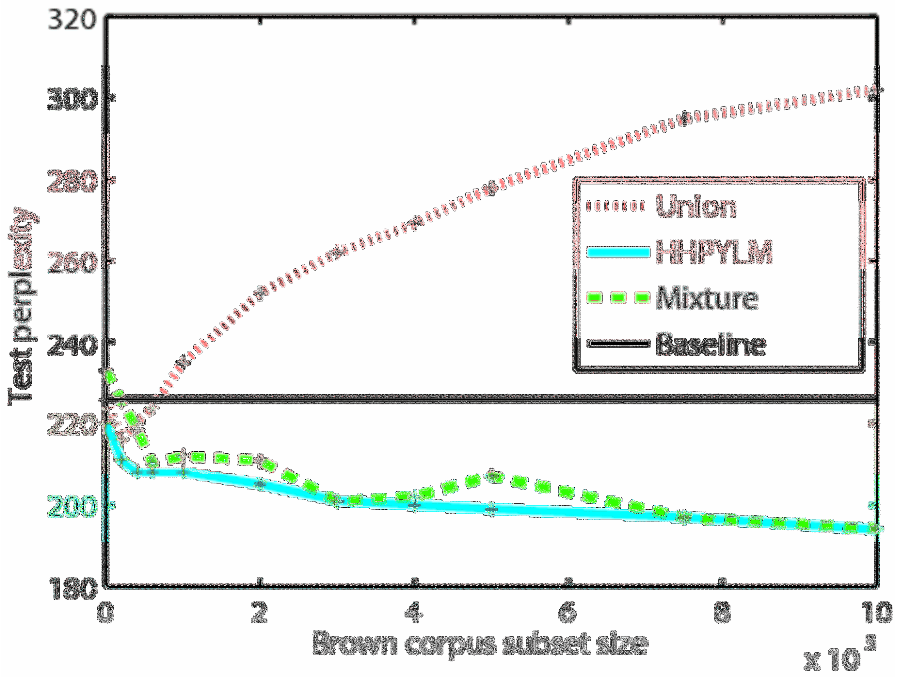


- Union [Bellegarda, 2004]



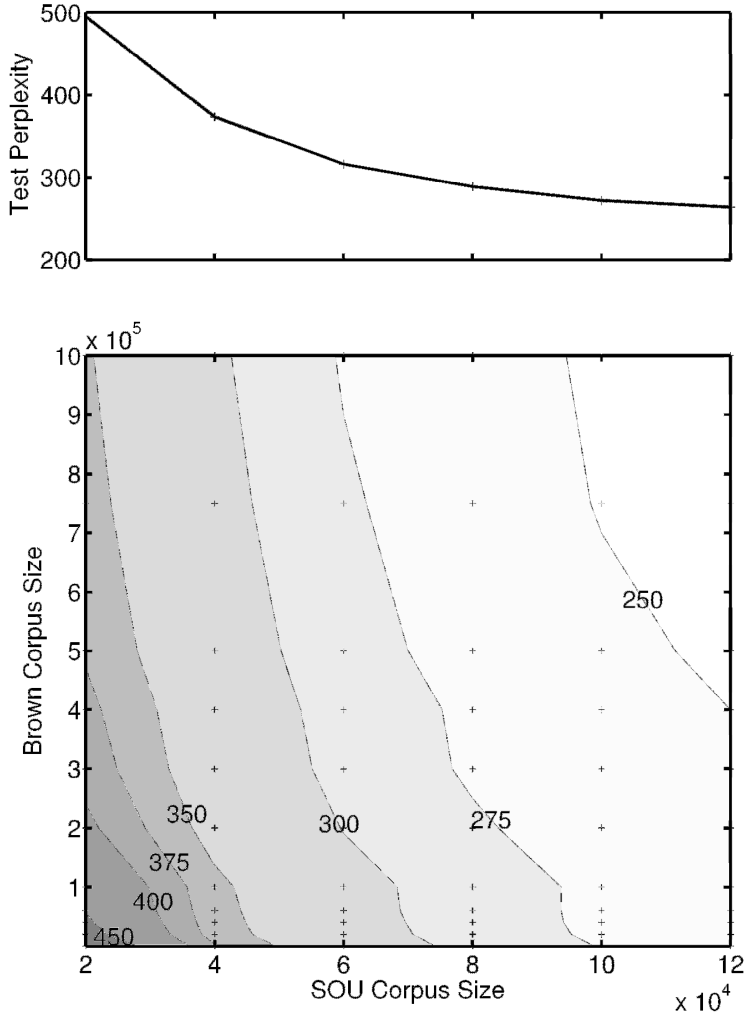
- MAP [Bacchiani, 2006]

AMI / Brown



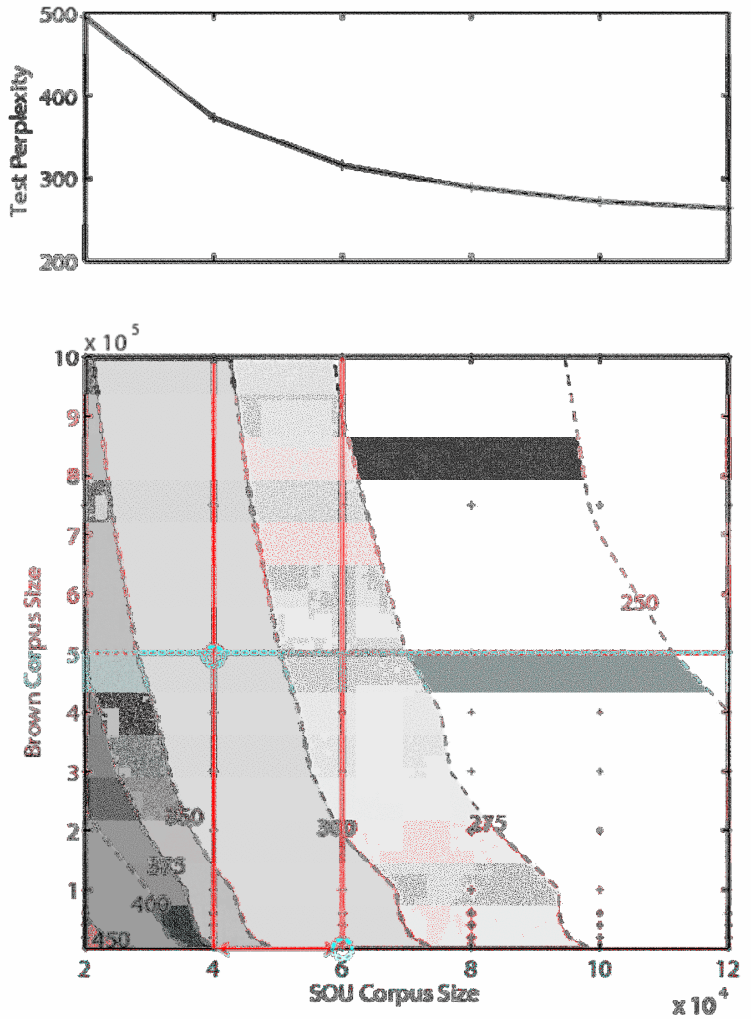
- Training corpora
  - Small
    - Augmented Multi-Party Interaction (AMI)
      - 2007
      - ~ 800,000 words, 8,000 unique
  - Big
    - Brown
      - 1967
      - ~ 1,000,000 words, 50,000 unique
- Test corpus
    - AMI excerpt
      - 2007
      - ~ 60,000 words

Cost / Benefit Analysis



- "Realistic" scenario
  - Computational cost ($) of using more "big" corpus data (Brown corpus)
  - Data preparation cost ($$$) of gathering more "small" corpus data (SOU corpus)
  - Perplexity Goal
- Same test data

Cost / Benefit Analysis



- "Realistic" scenario
  - Computational cost ($) of using more "big" corpus data (Brown corpus)
  - Data preparation cost ($$$) of gathering more "small" corpus data (SOU corpus)
  - Perplexity Goal
- Same test data

Take Home

- Showed how to do SLM domain adaptation through hierarchical Bayesian language modelling
- Introduced a new type of model
  – “Graphical Pitman Yor Process”

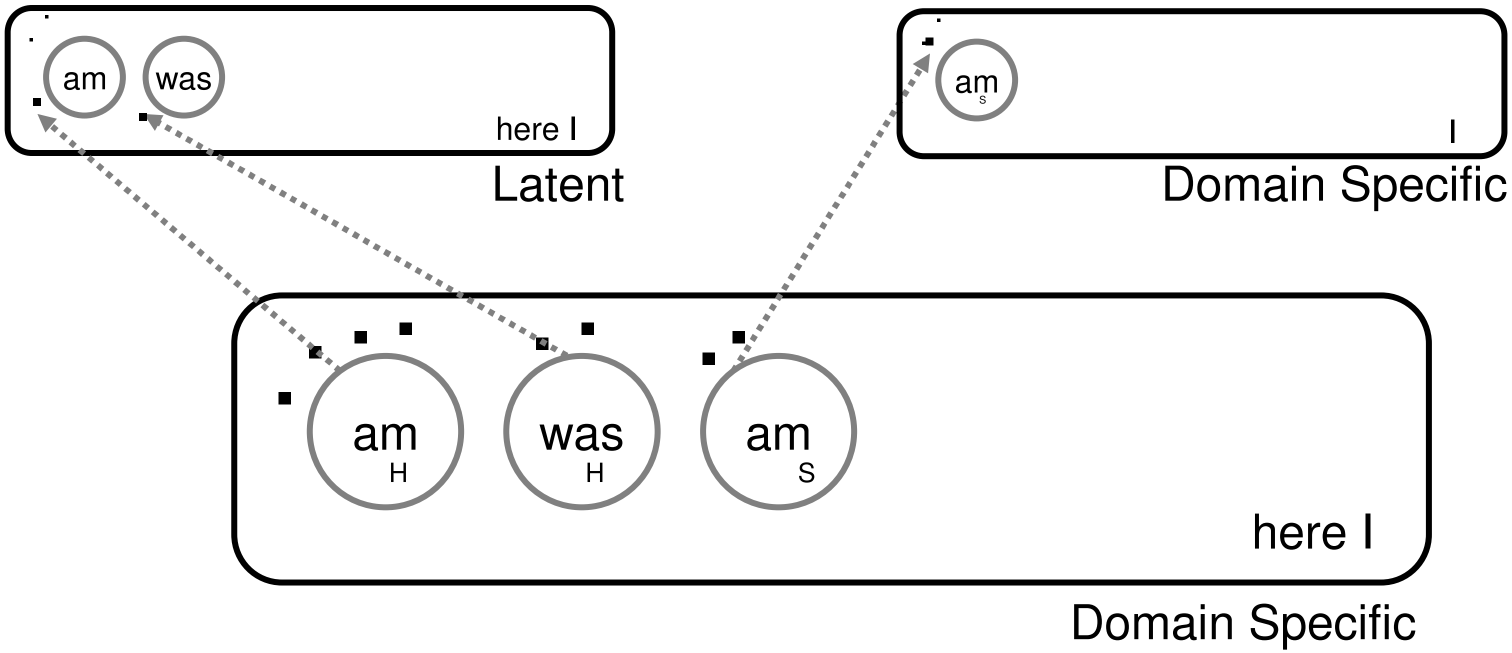
Thank You

Select References

- [1] Bellegarda, J. R. (2004). Statistical language model adaptation: review and perspectives. Speech Communication, 42, 93–108.
- [2] Carletta, J. (2007). Unleashing the killer corpus: experiences in creating the multi-everything AMI meeting corpus. Language Resources and Evaluation Journal, 41, 181–190.
- [3] Daum´e III, H., & Marcu, D. (2006). Domain adaptation for statistical classifiers. Journal of Artificial Intelligence Research, 101–126.
- [4] Goldwater, S., Griffiths, T. L., & Johnson, M. (2007). Interpolating between types and tokens by estimating power law generators. NIPS 19 (pp. 459–466).
- [5] Iyer, R., Ostendorf, M., & Gish, H. (1997). Using out-of-domain data to improve in-domain language models. IEEE Signal processing letters, 4, 221–223.
- [6] Kneser, R., & Steinbiss, V. (1993). On the dynamic adaptation of stochastic language models. IEEE Conference on Acoustics, Speech, and Signal Processing (pp. 586–589).
- [7] Kucera, H., & Francis, W. N. (1967). Computational analysis of present-day American English. Brown University Press.
- [8] Rosenfeld, R. (2000). Two decades of statistical language modeling: where do we go from here? Proceedings of the IEEE (pp. 1270–1278).
- [9] Teh, Y.W. (2006). A hierarchical Bayesian language model based on Pitman-Yor processes. ACL Proceedings (44th) (pp. 985–992).
- [10] Zhu, X., & Rosenfeld, R. (2001). Improving trigram language modeling with the world wide web. IEEE Conference on Acoustics, Speech, and Signal Processing (pp. 533–536).

HHPYLM Inference

- Every table at level n corresponds to a customer in *a* restaurant at level n-1
- All customers in the entire hierarchy are unseated and re-seated in each sampler sweep



HHPYPLM Intuition

- Each G is still a weighted set of (repeated) atoms

$$\begin{aligned}
\mathcal{G}^{\mathcal{D}}_{\{w_{t-j:t-1}\}} &\sim \mathrm{PY}(d_j, \alpha_j, \lambda \mathcal{G}^{\mathcal{D}}_{\{w_{t-j+1:t-1}\}} + (1-\lambda)\mathcal{G}^{\mathcal{L}}_{\{w_{t-j:t-1}\}}) \\
\Rightarrow \mathcal{G}^{\mathcal{D}}_{\{w_{t-j:t-1}\}} &= \sum_{j \in \{\mathcal{D},\mathcal{L}\}} \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k, s_j}
\end{aligned}$$

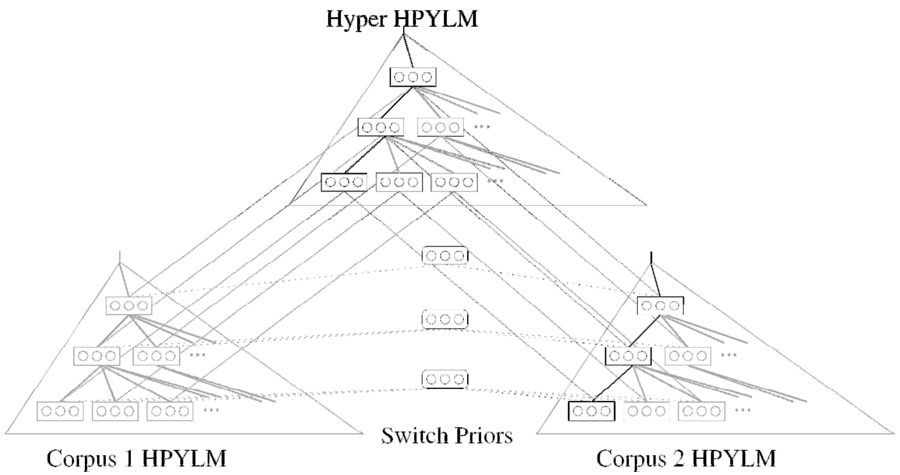
$$\begin{aligned}
\phi_k &\sim \mathcal{G}_{\{w_{t-j+1:t-1}\}} \\
s_j &\sim \{\lambda, 1-\lambda\}
\end{aligned}$$

Sampling the “floor” indicators

- Switch variables drawn from a discrete distribution (mixture weights)

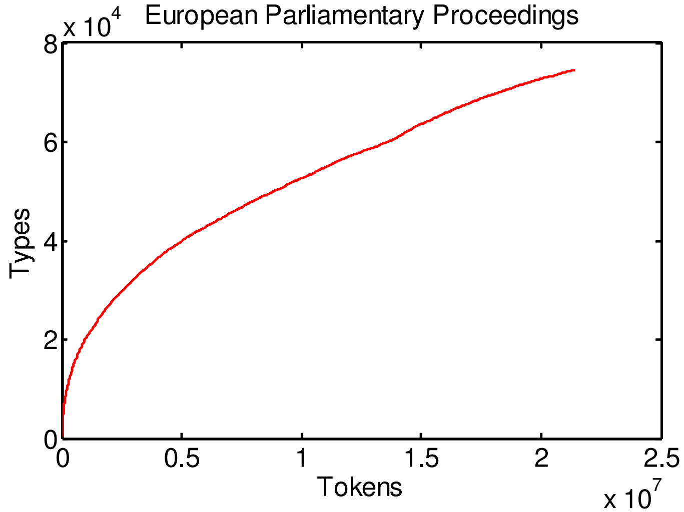
$$\begin{aligned} s_k &\sim \{\lambda, 1-\lambda\} \\ \{\lambda, 1-\lambda\} &\sim PY(d_\lambda, \alpha_\lambda, \mathcal{U}_2) \end{aligned}$$

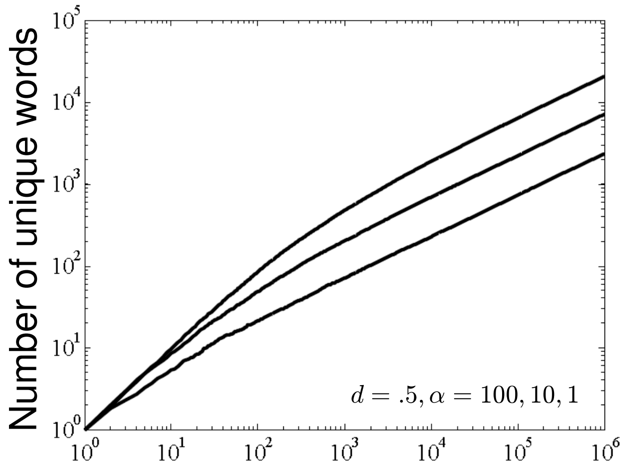
- Integrate out λ’s
- Sample this in the Chinese restaurant representation as well



Why PY vs DP?

- PY power law characteristics match the statistics of natural language well
  - number of unique words in a set of n words follows power law





Number of “words” drawn from Pitman-Yor Process

[Teh 2006]

What does a PY draw look like?

- A draw from a PY process is a discrete distribution with infinite support

$$
\begin{aligned}
\mathcal{G} &\sim \mathrm{PY}(d, \alpha, \mathcal{H}) \\
\Rightarrow \mathcal{G} &= \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}, \sum_{k=1}^{\infty} \pi_k = 1
\end{aligned}
$$

aa (.0001), aah (.00001), aal (.000001), aardvark (.001), …

[Sethurman, 94]

How does one work with such a model?

- Truncation in the stick breaking representation
- Or in a representation where $\mathcal{G}$ is analytically marginalized out

$$P(x_{n+1}|x_{1:n};\alpha,d) = \int P(x_{n+1}|\mathcal{G})P(\mathcal{G}|x_{1:n};\alpha,d)d\mathcal{G}$$

roughly

Posterior PY

- This marginalization is possible and results in the Pitman Yor Polya urn representation.

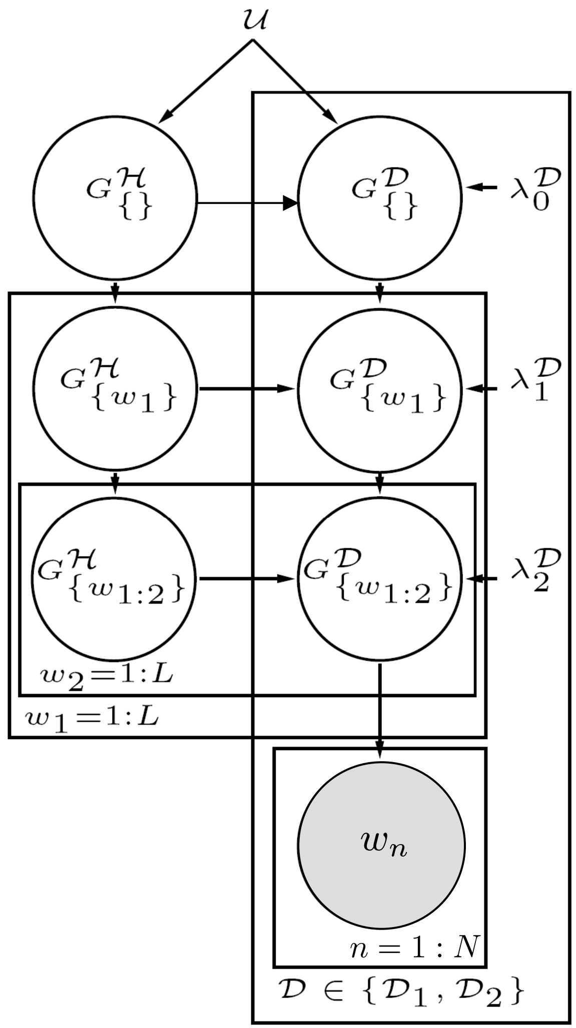
$$P(x_{n+1}|x_{1:n};\alpha,d) = \frac{\sum_{k=1}^{K}(c_k - d)}{\alpha+n}\delta(\phi_k - x_{n+1}) + \frac{\alpha+dK}{\alpha+n}\mathcal{G}_0(x_{n+1})$$

- By invoking exchangeability one can easily construct Gibbs samplers using such representations
  - A generalization of the hierarchical Dirichlet process sampler of Teh [2006].

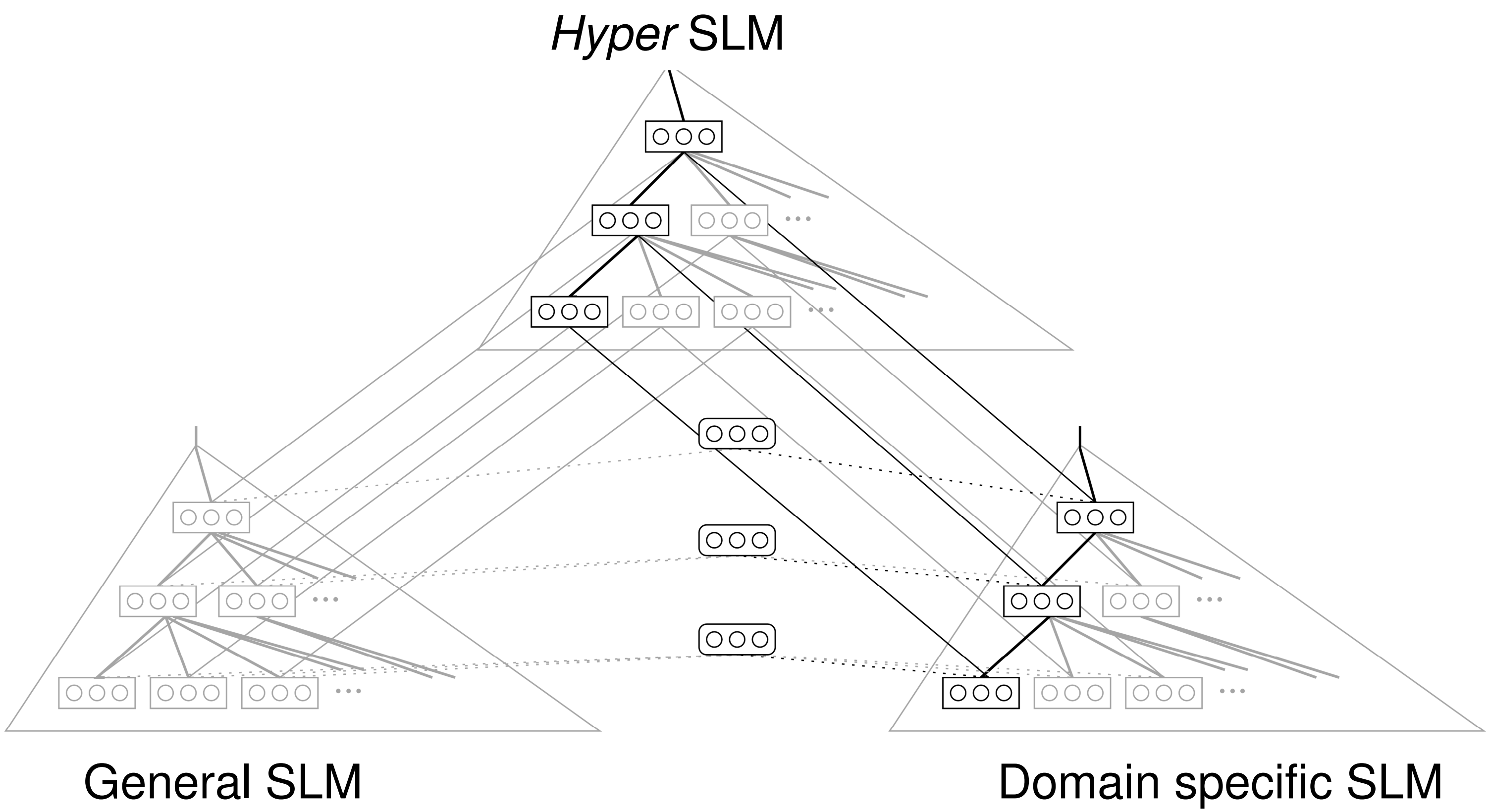
[Pitman, 02]

Doubly-Hierarchical Pitman-Yor Process Language Model

- Finite, fixed vocabularly
- Two domains
  - One big ("general")
  - One small ("specific")
- "Backs-off" to
  - out of domain model with same context
  - or in domain model with one fewer words of context

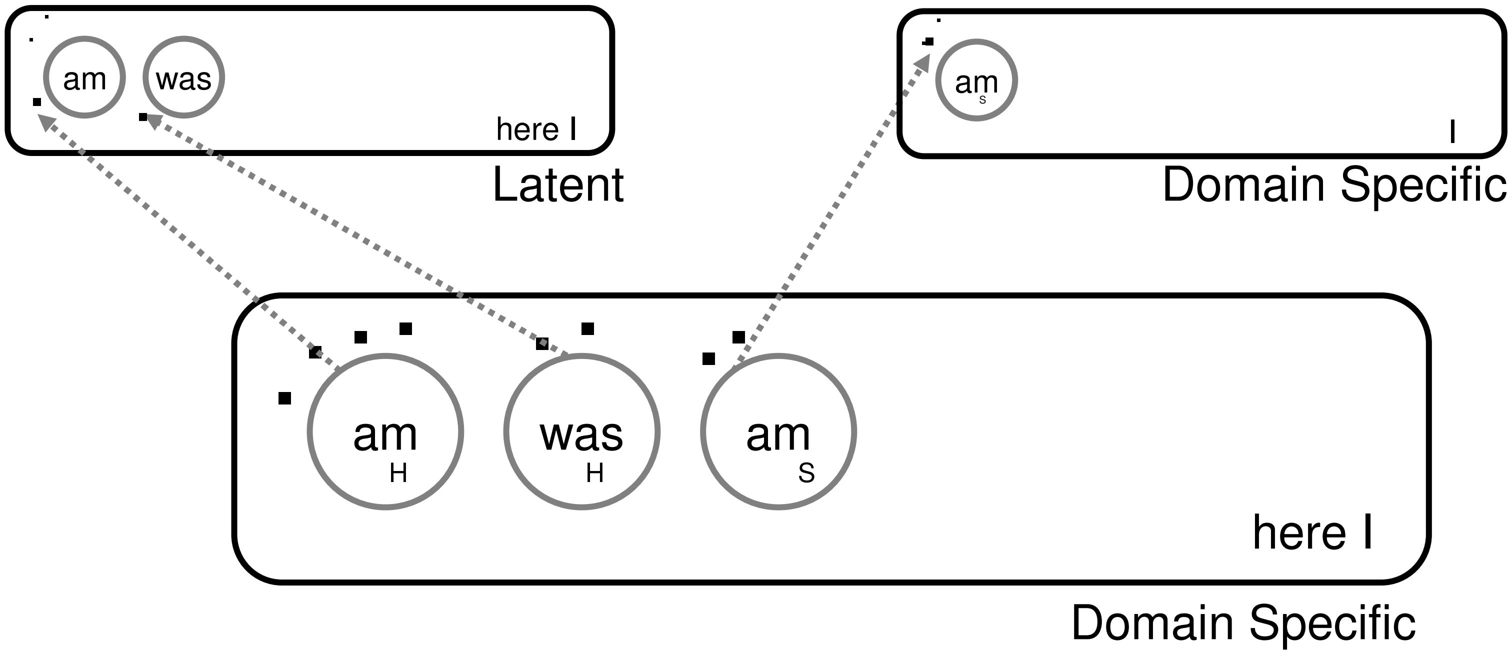


Bayesian SLM Domain Adaptation



HHPYLM Inference

- Every table at level n corresponds to a customer in *a* restaurant at level n-1
- All customers in the entire hierarchy are unseated and re-seated in each sampler sweep



HHPYPLM Intuition

- Each G is still a weighted set of (repeated) atoms

$$
\begin{aligned}
\mathcal{G}^{\mathcal{D}}_{\{w_{t-j:t-1}\}} &\sim \mathrm{PY}(d_j, \alpha_j, \lambda \mathcal{G}^{\mathcal{D}}_{\{w_{t-j+1:t-1}\}} + (1-\lambda)\mathcal{G}^{\mathcal{H}}_{\{w_{t-j:t-1}\}}) \\
\Rightarrow \mathcal{G}^{\mathcal{D}}_{\{w_{t-j:t-1}\}} &= \sum_{j \in \{\mathcal{D},\mathcal{H}\}} \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k, s_j}
\end{aligned}
$$

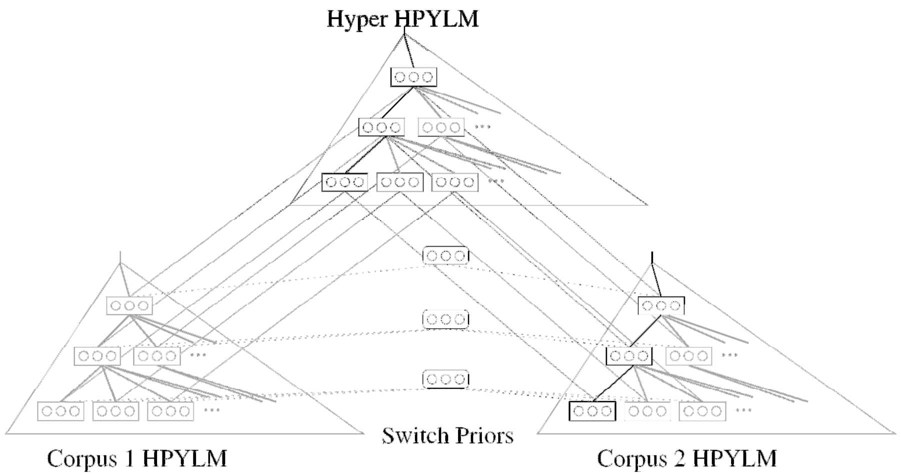
$$
\begin{aligned}
\phi_k &\sim \mathcal{G}_{\{w_{t-j+1:t-1}\}} \\
s_j &\sim \{\lambda, 1-\lambda\}
\end{aligned}
$$

Sampling the "floor" indicators

- Switch variables drawn from a discrete distribution (mixture weights)

$$
\begin{aligned}
s_k &\sim \{\lambda, 1-\lambda\} \\
\{\lambda, 1-\lambda\} &\sim PY(d_\lambda, \alpha_\lambda, \mathcal{U}_2)
\end{aligned}
$$

- Integrate out λ's
- Sample this in the Chinese restaurant representation as well



Language Modelling

- Maximum likelihood
  - Smoothed empirical counts
- Hierarchical Bayes
  - Finite
    - MacKay & Peto, “Hierarchical Dirichlet Language Model,” 1994
  - Nonparametric
    - Teh, “A hierarchical Bayesian language model based on Pitman-Yor processes.” 2006
    - Comparable to the best smoothing *n*-gram model

Questions and Contributions

- How does one do this?
  – Hierarchical Bayesian modelling approach
- What does such a model look like?
  – Novel model architecture
- How does one estimate such a model?
  – Novel auxiliary variable Gibbs sampler
- Given such a model, does inference in the model actually confer application benefits?
  – Positive results

Review: Hierarchical Pitman-Yor Process Language Model

- Finite, fixed vocabulary
- Infinite dimensional parameter space (G's, d's, and α's)

America (.00001), of (.01), States (.00001), the (.01), …

America (.0001), of (.000000001), States (.0001), the (.01), …

America (.8), of (.00000001), States (.0000001), the (. 01), …

$$
\begin{aligned}
\mathcal{G}_{\{\}} &\sim \mathrm{PY}(d_0, \alpha_0, \mathcal{U}) \\
\mathcal{G}_{\{\mathrm{of}\}} &\sim \mathrm{PY}(d_1, \alpha_1, \mathcal{G}_{\{\}}) \\
&\vdots \\
\mathcal{G}_{\{\mathrm{United},\mathrm{States},\mathrm{of}\}} &\sim \mathrm{PY}(d_3, \alpha_3, \mathcal{G}_{\{\mathrm{States},\mathrm{of}\}}) \\
x|\{\mathrm{United}, \mathrm{States}, \mathrm{of}\} &\sim \mathcal{G}_{\{\mathrm{United},\mathrm{States},\mathrm{of}\}}
\end{aligned}
$$

- Forms a suffix tree of depth *n+1* (*n* is from *n*-gram) with one distribution at each node

[Teh, 2006]

Review: HPYPLM General Notation

America (.00001), of (.01), States (.00001), the (.01), …

America (.0001), of (.000000001), States (.0001), the (.01), …

America (.8), of (.00000001), States (.0000001), the (. 01), …

$$
\begin{aligned}
\mathcal{G}_{\{\}} &\sim \mathrm{PY}(d_0, \alpha_0, \mathcal{U}) \\
\mathcal{G}_{\{w_{t-1}\}} &\sim \mathrm{PY}(d_1, \alpha_1, \mathcal{G}_{\{\}}) \\
&\vdots \\
\mathcal{G}_{\{w_{t-j:t-1}\}} &\sim \mathrm{PY}(d_j, \alpha_j, \mathcal{G}_{\{w_{t-j+1:t-1}\}}) \\
x | w_{t-j:t-1} &\sim \mathcal{G}_{\{w_{t-j:t-1}\}}
\end{aligned}
$$

Review: Pitman-Yor / Dirichlet Process Stick-Breaking Representation

- Each G is a weighted set of atoms
- Atoms may repeat
  - Discrete base distribution

$$
\begin{aligned}
\mathcal{G}_{\{w_{t-j:t-1}\}} &\sim \mathrm{PY}(d_j, \alpha_j, \mathcal{G}_{\{w_{t-j+1:t-1}\}}) \\
\Rightarrow \mathcal{G}_{\{w_{t-j:t-1}\}} &= \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k} \qquad \phi_k \sim \mathcal{G}_{\{w_{t-j+1:t-1}\}}
\end{aligned}
$$

aa (.0001), aah (.00001), aal (.000001), aardvark (.001), …

Now you know…

- Statistical natural language modelling
  – Perplexity as a performance measure
- Domain adaptation
- Hierarchical Bayesian natural language models
  – HPYLM
  – HDP Inference

Dirichlet Process (DP) Review

- Notation (G and H are measures over space X)

$$\begin{aligned} \mathcal{G} &\sim \mathrm{DP}(\alpha, \mathcal{H}) \\ \theta_i &\sim \mathcal{G} \end{aligned}$$

- Definition
  - For any fixed partition (A_1, A_2, …, A_K) of X

$$[\mathcal{G}(A_1), \mathcal{G}(A_2), \ldots, \mathcal{G}(A_K)] \sim \mathrm{Dirichlet}(\alpha\mathcal{H}(A_1), \alpha\mathcal{H}(A_2), \ldots, \alpha\mathcal{H}(A_K))$$

Fergusen 1973, Blackwell & MacQueen 1973, Adous 1985, Sethuraman 1994

Inference

- Quantities of interest

$$
\begin{aligned}
P(\mathcal{G}|\theta_1) &= \frac{P(\theta_1|\mathcal{G})P(\mathcal{G})}{P(\theta_1)} \\
P(\theta_1) &= \int P(\theta_1|\mathcal{G})P(\mathcal{G})d\mathcal{G}
\end{aligned}
$$

- Critical identities from multinomial / Dirichlet conjugacy (roughly)

$$
\begin{aligned}
\theta_1 &\sim \mathcal{H} \\
\mathcal{G}|\theta_1 &\sim \mathrm{DP}(\alpha+1, \frac{\alpha\mathcal{H}+\delta_{\theta_1}}{\alpha+1})
\end{aligned}
$$

Posterior updating in one step

- With the following identifications

$$\begin{aligned} \mathcal{H}_i &= \frac{\alpha\mathcal{H} + \sum_{j=1}^{i} \delta_{\theta_j}}{\alpha + i} \\ \alpha_i &= \alpha + i \end{aligned}$$

- then

$$\begin{aligned} \theta_{i+1} | \theta_1, \ldots, \theta_i &\sim \mathcal{H}_i \\ \mathcal{G}_{i+1} | \theta_1, \ldots, \theta_{i+1} &\sim \mathrm{DP}(\alpha_{i+1}, \mathcal{H}_{i+1}) \end{aligned}$$

Don't need G's – "integrated out"

- Many θ's are the same

$$\begin{aligned} \mathcal{H}_i &= \frac{\alpha\mathcal{H} + \sum_{j=1}^{i} \delta_{\theta_j}}{\alpha + i} \\ \theta_{i+1}|\theta_1, \ldots, \theta_i &\sim \mathcal{H}_i \end{aligned}$$

- The "Chinese restaurant process" is a representation (data structure) of the posterior updating scheme that keeps track of only the unique θ's and the number times each was drawn

Review: Hierarchical Dirichlet Process (HDP) Inference

- Starting with training data (a corpus) we estimate the posterior distribution through sampling

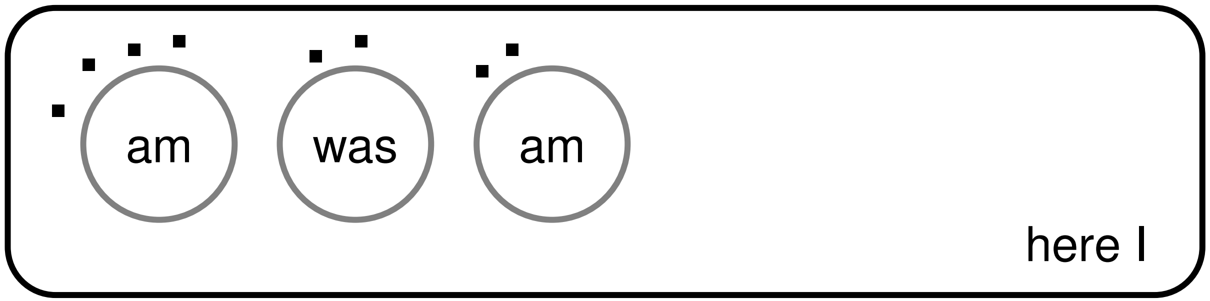
$$P(\Theta|w_1, w_1, \ldots, w_N) \propto P(w_1, w_1, \ldots, w_N|\Theta)P(\Theta)$$

- HPYP Inference is the roughly the same as HDP inference

HDP Inference: Review

- Chinese restaurant franchise [Teh et al, 2006]
  - Seating arrangements of hierarchy of Chinese restaurant processes are the state of the sampler
- Data are the observed tuples/*n*-grams ("here I am" (6), "here I was" (2))

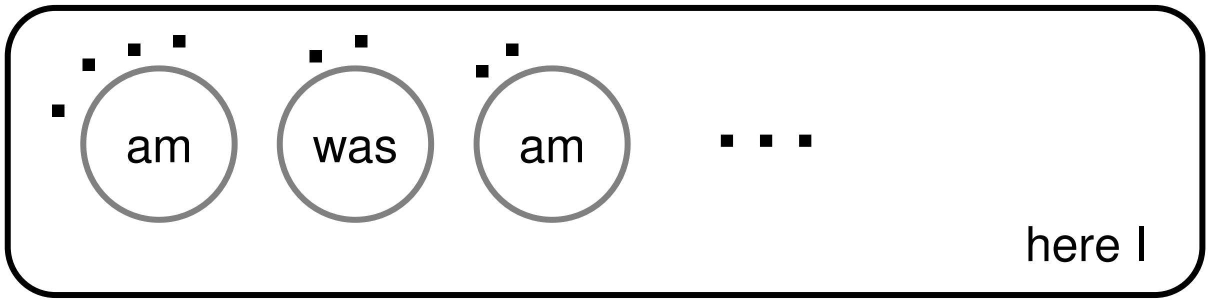
$$
\begin{aligned}
\mathcal{G}_{\{\mathrm{I},\mathrm{here}\}} &\sim \mathrm{PY}(d, \alpha, \mathcal{G}_{\{\mathrm{I}\}}) \\
x | \{\mathrm{I}, \mathrm{here}\} &\sim \mathcal{G}_{\{\mathrm{I},\mathrm{here}\}}
\end{aligned}
$$



HDP Inference: Review

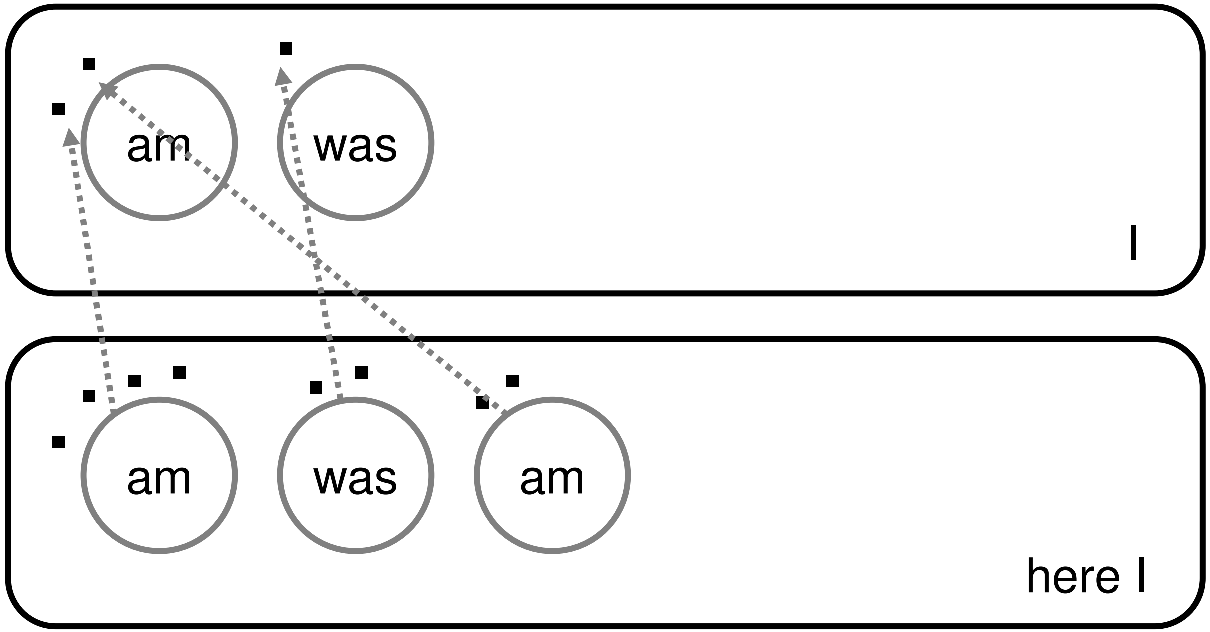
- Induction over base distribution draws -> Chinese restaurant franchise
- Every level is a Chinese restaurant process
- Auxiliary variables indicate from which table each n-gram came
  - Table label is a draw from the base distribution
- State of the sampler: {am, 1}, {am, 3}, {was, 2}, … (in context, “here I”)

$$\begin{aligned} P(z = k \mid \text{everything else}) &\propto \big(c^k - d\big)\, \mathcal{I}(x = \phi^k) \\ P(z = K+1 \mid \text{everything else}) &\propto (\alpha + dK)\mathcal{G}_{\mathrm{I}}(x). \end{aligned}$$



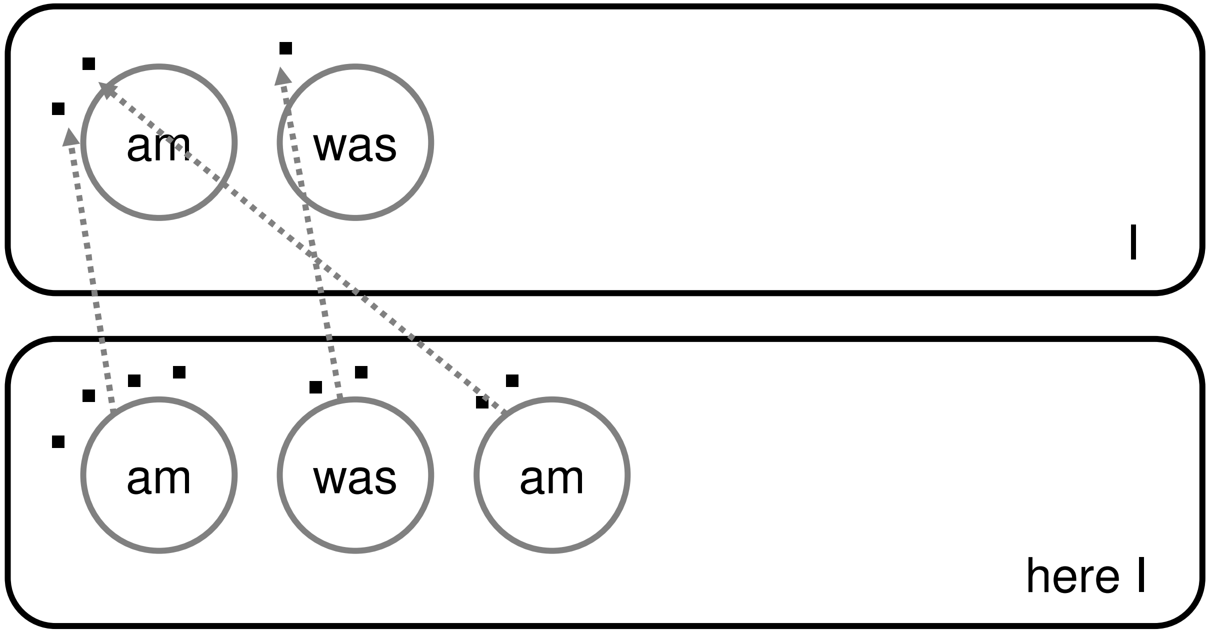
HDP Inference: Review

- Every draw from the base distribution means that the resulting atom must (also) be in the base distribution's set of atoms.



HDP Inference: Review

- Every table at level n corresponds to a customer in a restaurant at level n-1
  - If a customer is seated at a new table, this corresponds to a draw from the base distribution
- All customers in the entire hierarchy are unseated and re-seated in each sweep



Parameter Infestation

- If V is the vocabulary size then a full tri-gram model has V^3 parameters (potentially)

$$P(w_n | w_{n-1}, w_{n-2})$$

$$P(\text{pepper} | \text{and}, \text{salt})$$
$$P(\text{lemon} | \text{and}, \text{salt})$$
$$\vdots$$
$$P(\text{snufalupagus} | \text{and}, \text{salt})$$

Big models

- Assume a 50,000 word vocabulary
  - 1.25×10^{14} parameters
    - Not all must be realized or stored in a practical implementation
  - Maximum number of tri-grams in a billion word corpus, $\sim 1 \times 10^{9}$

Language Modelling

- Zero counts and smoothing

$$P(x|w_{n-1}, w_{n-2}) = \frac{\#\{w_{n-2}, w_{n-1}, x\}}{\#\{w_{n-2}, w_{n-1}\}}$$

- Kneser-Ney

$$P_{KN}(x|w_{n-1}, w_{n-2}) = \frac{\max(0, \#\{w_{n-2}, w_{n-1}, x\} - d)}{\#\{w_{n-2}, w_{n-1}\}} + \frac{dK}{\#\{w_{n-2}, w_{n-1}\}} P_{KN}(x|w_{n-1})$$

Multi-Floor Chinese Restaurant Process

$$
\begin{aligned}
P(z_v^j = k | \mathbf{z}_v \backslash z_v^j, S, X) &\propto \max((c_v^{k-} - d_v), 0)\delta(x_v^j - \phi_v^k) \\
P(z_v^j = K+1, s_v^{K+1} = w | \mathbf{z}_v \backslash z_v^j, S, X) &\propto (\alpha_v + d_v K_v^-)\lambda_{w \to v}\mathcal{G}_w(x_v^j).
\end{aligned}
$$

Domain LM Adaptation Approaches

- MAP [Bacchiani, 2006]

$$
\begin{aligned}
x|w_{n-1}, w_{n-2} &\propto \text{Discrete}(\pi_1, \ldots, \pi_N) \\
\pi_1, \ldots, \pi_N &\propto \text{Dirichlet}(\#\{w_{n-2}, w_{n-1}, x\}_{\mathcal{D}_1} + \#\{w_{n-2}, w_{n-1}, x\}_{\mathcal{D}_2}, \ldots, \#\{w_{n-2}, w_{n-1}, x\}_{\mathcal{D}_1} + \#\{w_{n-2}, w_{n-1}, x\}_{\mathcal{D}_2})
\end{aligned}
$$

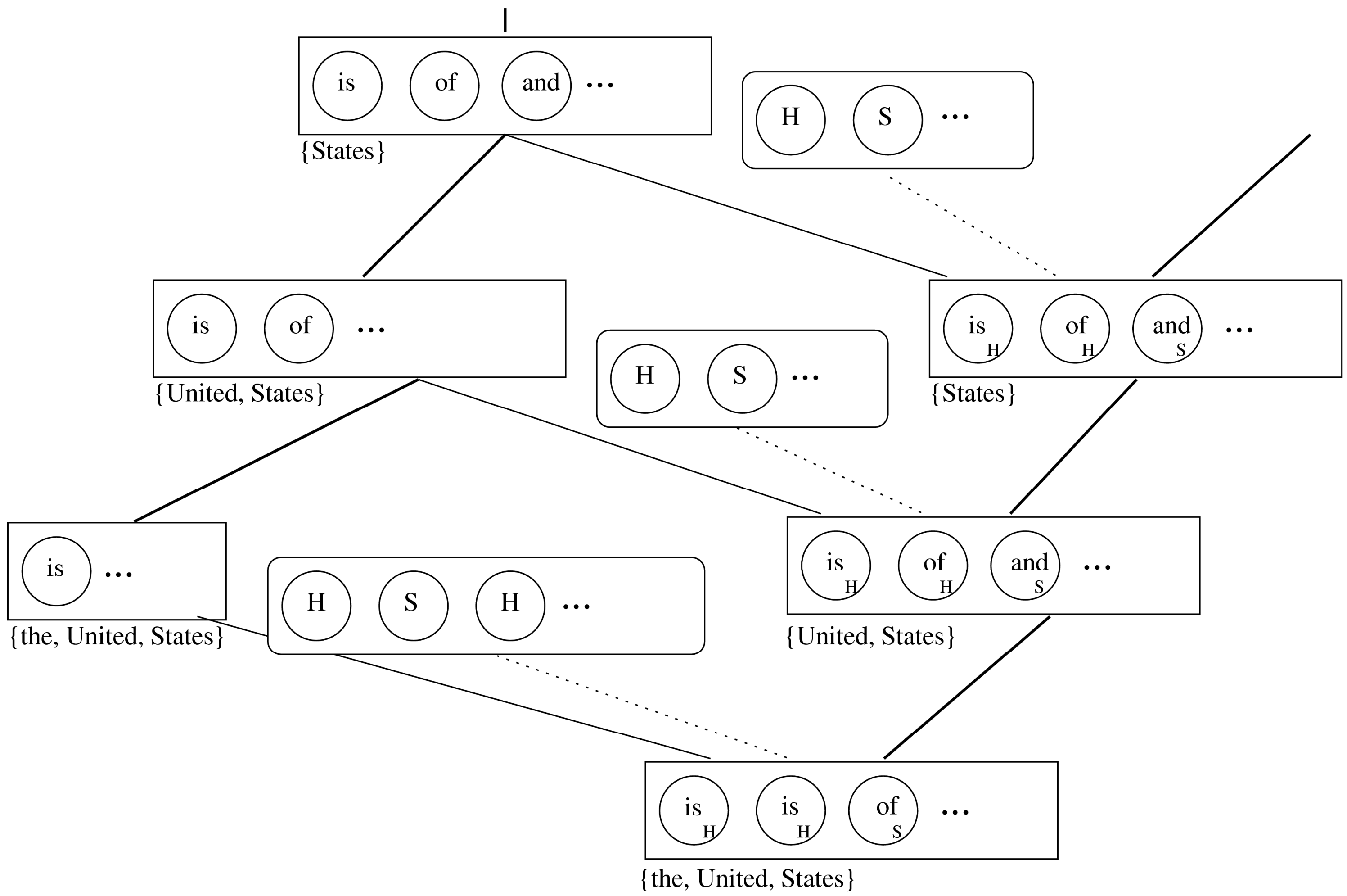
- Mixture [Kneser & Steinbiss, 1993]

$$P(x|w_{n-1}, w_{n-2}) = \lambda P_{\mathcal{D}_1}(x|w_{n-1}, w_{n-2}) + (1-\lambda) P_{\mathcal{D}_2}(x|w_{n-1}, w_{n-2})$$

- Union [Bellegarda, 2004]

$$P(x|w_{n-1}, w_{n-2}) = \frac{\#\{w_{n-2}, w_{n-1}, x\}_{\mathcal{D}_1} + \#\{w_{n-2}, w_{n-1}, x\}_{\mathcal{D}_2}}{\#\{w_{n-2}, w_{n-1}\}_{\mathcal{D}_1} + \#\{w_{n-2}, w_{n-1}\}_{\mathcal{D}_2}}$$

HHPYLM Inference: Intuition



Estimation

- Counting

$$P(x|w_{n-1},w_{n-2}) = \frac{\#\{w_{n-2},w_{n-1},x\}}{\#\{w_{n-2},w_{n-1}\}} = \frac{\#\{w_{n-2},w_{n-1},x\}}{\sum_j \#\{w_{n-2},w_{n-1},j\}}$$

- Zero counts are a problem
- Smoothing (or regularization) is essential
  - Kneser-Ney
  - Hierarchical Dirichlet Language Model (HDLM)
  - Hierarchical Pitman-Yor Process Language Model (HPYLM)
  - etc.

Justification

- Domain specific training data can be costly to collect and process
  - i.e. manual transcription of customer service telephone calls
- Different domains are different!

Why Domain Adaptation?

- *… a language model trained on Dow-Jones newswire text will see its perplexity doubled when applied to the very similar Associated Press newswire text from the same time period*
  - Ronald Rosenfeld, “Two Decades Of Statistical Language Modeling: Where Do We Go From Here”, 2000

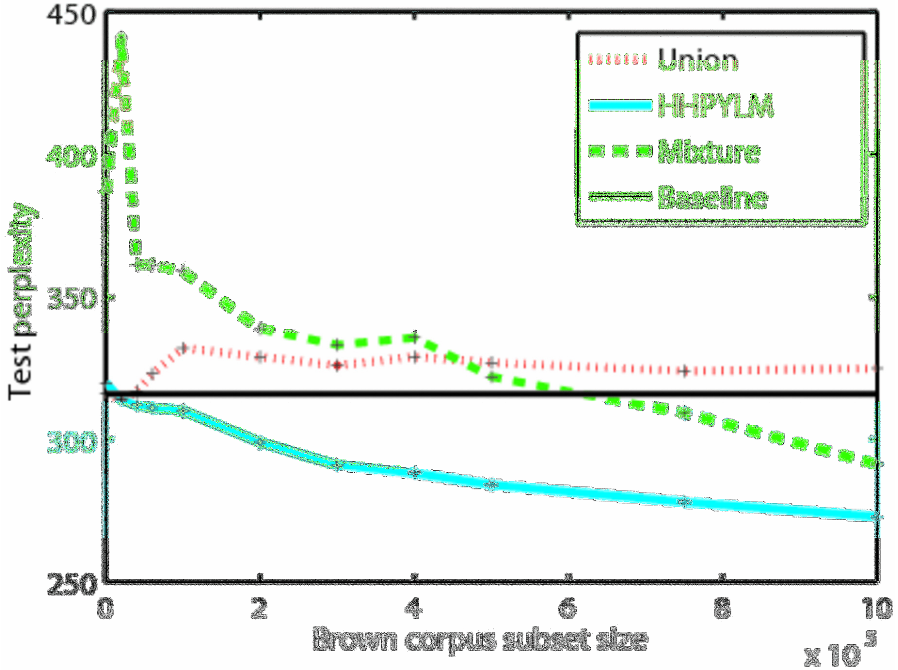
SLM Evaluation

- Per-word perplexity is a model evaluation metric common to the NLP community

$$\text{perplexity}(P_{\text{trained}}, W_{\text{test}}) = 2^{-\frac{1}{N}\sum \log P_{\text{trained}}(w_i^{\text{test}}|w_{i-1}^{\text{test}}, w_{i-2}^{\text{test}})}$$

  - Computed using test data and trained model
  - Related to log likelihood and entropy
  - Language modelling interpretation:
    - Average size of "replacement word" set

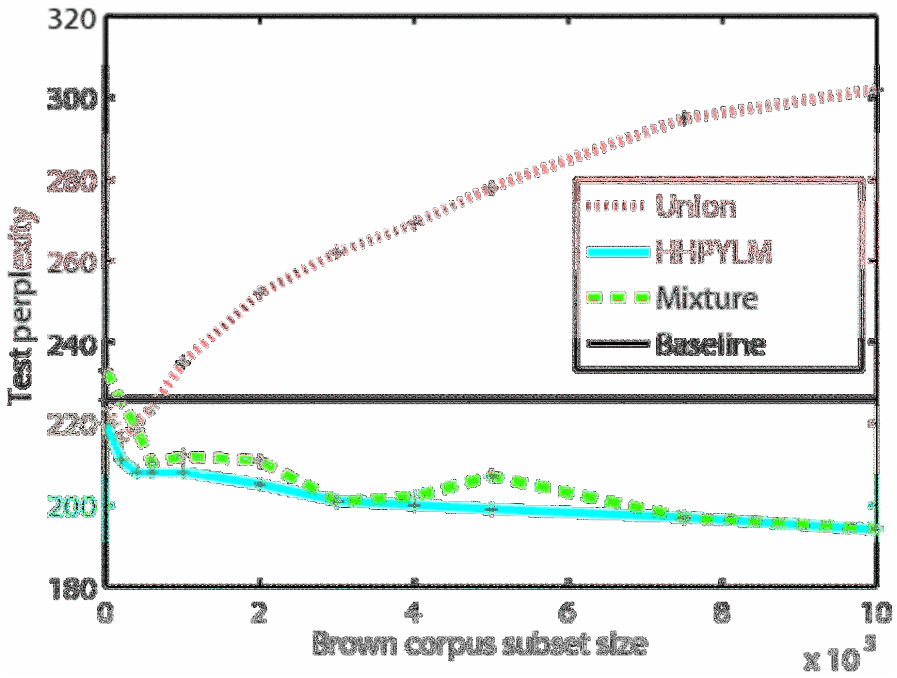
SOU / Brown



- State of the Union Corpus, 1945-2006
  - ~ 370,000 words, 13,000 unique
    - *Today, the entire world is looking to America for enlightened leadership to peace and progress.* Truman, 1945
    - *Today, an estimated 4 out of every 10 students in the 5th grade will not even finish high school - and that is a waste we cannot afford.* Kennedy, 1963
    - *In 1945, there were about two dozen lonely democracies in the world. Today, there are 122. And we're writing a new chapter in the story of self-government -- with women lining up to vote in Afghanistan, and millions of Iraqis marking their liberty with purple ink, and men and women from Lebanon to Egypt debating the rights of individuals and the necessity of freedom.* Bush, 2006
- Brown Corpus, 1967
  - ~ 1,000,000 words, 50,000 unique
    - *During the morning hours, it became clear that the arrest of Spencer was having no sobering effect upon the men of the Somers.*
    - *He certainly didn't want a wife who was fickle as Ann.*
    - *It is being fought, moreover, in fairly close correspondence with the predictions of the soothsayers of the think factories.*
- Test: Johnson's SOU Addresses, 1963-1969
  - ~ 37,000 words
    - *But we will not permit those who fire upon us in Vietnam to win a victory over the desires and the intentions of all the American people.*
    - *This Nation is mighty enough, its society is healthy enough, its people are strong enough, to pursue our goals in the rest of the world while still building a Great Society here at home.*

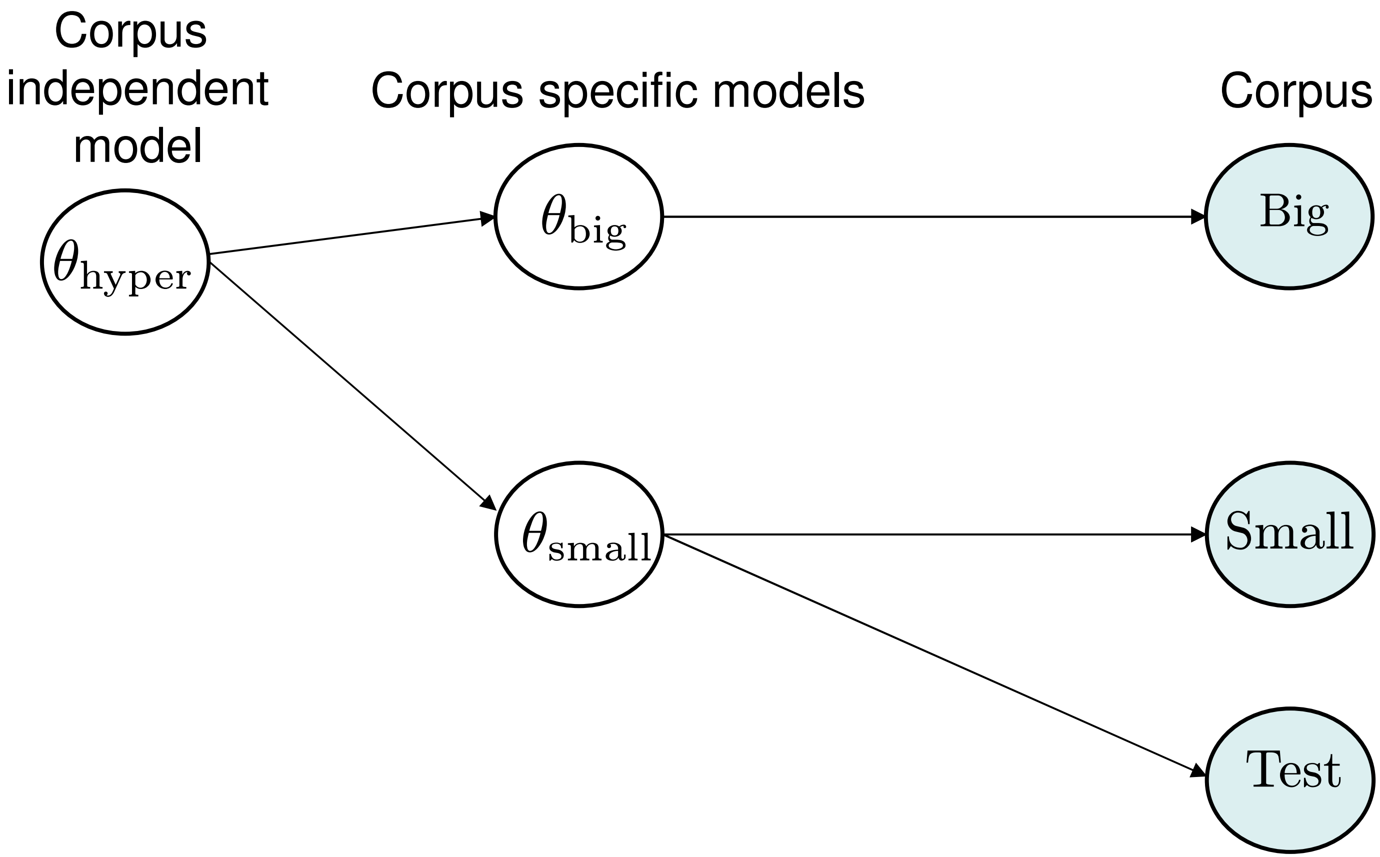
FRANK WOOD, YEE WHYE TEH

AMI / Brown



- AMI, 2007
  - Approx 800,000 tokens, 8,000 types
    - *Yeah yeah for example the l.c.d. you can take it you can put it put it back in or you can use the other one or the speech recognizer with the microphone yeah yeah you want a microphone to pu in the speech recognizer you don't you pay less for the system you see so*

Bayesian Domain Adaptation



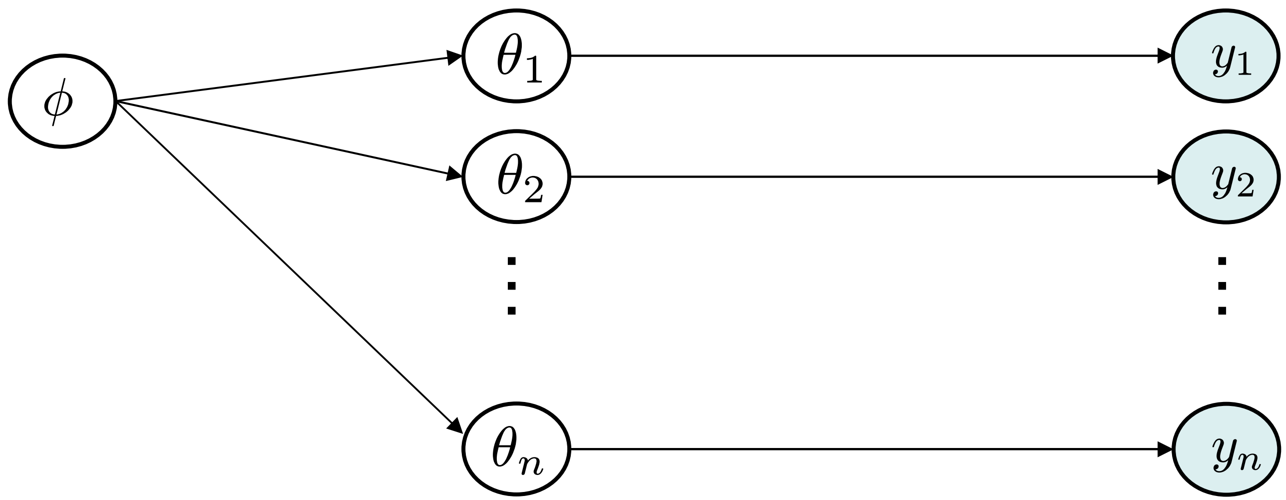
HPYPLM

$$
\begin{aligned}
\mathcal{G}^{\mathcal{D}}_{\{\}} &\sim \mathrm{PY}(d^{\mathcal{D}}_0, \alpha^{\mathcal{D}}_0, \mathcal{U} \qquad\qquad\qquad)\\
\mathcal{G}^{\mathcal{D}}_{\{w_{t-1}\}} &\sim \mathrm{PY}(d^{\mathcal{D}}_1, \alpha^{\mathcal{D}}_1, \mathcal{G}^{\mathcal{D}}_{\{\}} \qquad\qquad\qquad)\\
&\vdots\\
\mathcal{G}^{\mathcal{D}}_{\{w_{t-j}:w_{t-1}\}} &\sim \mathrm{PY}(d^{\mathcal{D}}_j, \alpha^{\mathcal{D}}_j, \mathcal{G}^{\mathcal{D}}_{\{w_{t-j+1}:w_{t-1}\}} \qquad\qquad\qquad)\\
x|w_{t-n+1}:w_{t-1} &\sim \mathcal{G}^{\mathcal{D}}_{\{w_{t-n+1}:w_{t-1}\}}.
\end{aligned}
$$

Bayesian Domain Adaptation

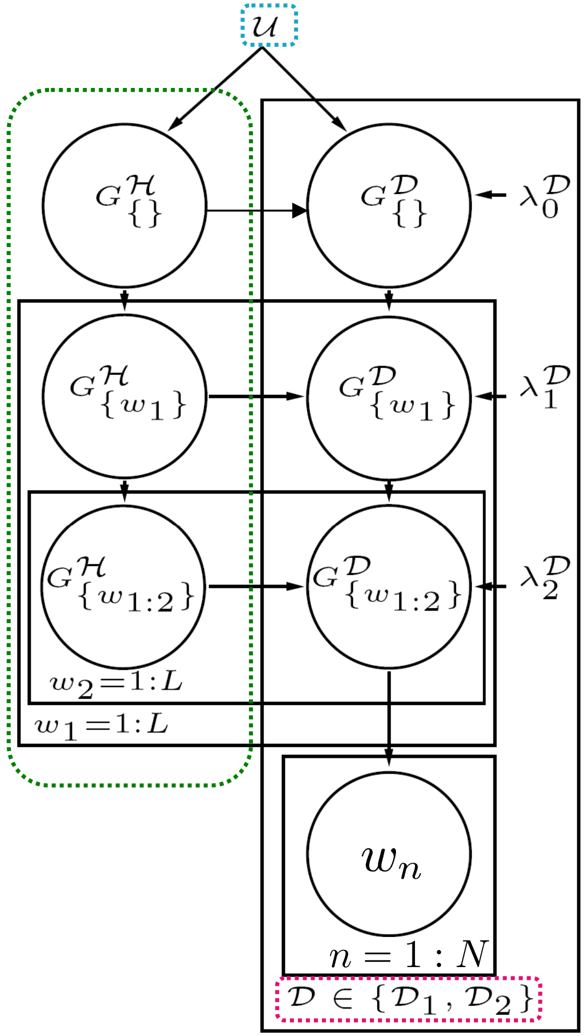
- Domain Adaptation $\leftrightarrow$ Hierarchical Modelling

Hyper model   Domain specific models   Domain specific observations



Doubly hierarchical Pitman-Yor Process Language Model

- Uniform distribution over words
- *Hyper* language model
- Domain



Hierarchical Pitman-Yor Process Language Model

[Teh, Y.W., 2006], [Goldwater, S., Griffiths, T. L., & Johnson, M., 2007]

$$\begin{aligned}
\mathcal{G}_{\{\}} &\sim \mathrm{PY}(d_0, \alpha_0, \mathcal{U}) \\
\mathcal{G}_{\{\mathrm{of}\}} &\sim \mathrm{PY}(d_1, \alpha_1, \mathcal{G}_{\{\}}) \\
&\vdots \\
\mathcal{G}_{\{\mathrm{United},\mathrm{States},\mathrm{of}\}} &\sim \mathrm{PY}(d_3, \alpha_3, \mathcal{G}_{\{\mathrm{States},\mathrm{of}\}}) \\
x|\{\mathrm{United},\mathrm{States},\mathrm{of}\} &\sim \mathcal{G}_{\{\mathrm{United},\mathrm{States},\mathrm{of}\}}
\end{aligned}$$

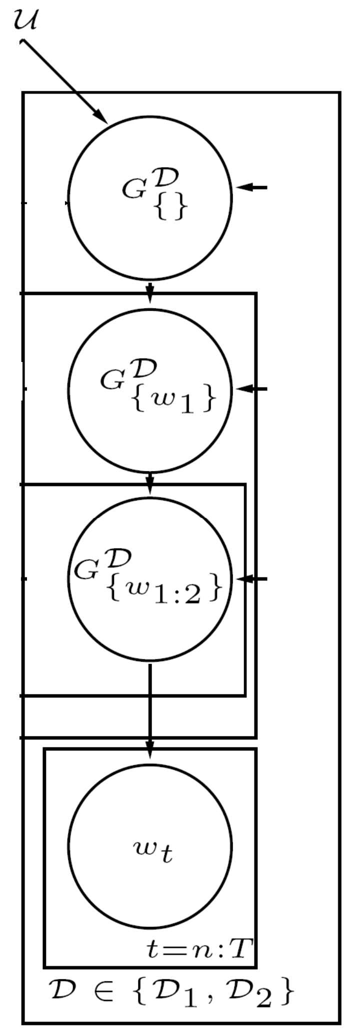
- e.g.

$\mathcal{G}_{\{\}} =$ America (.00001), of (.01), States (.00001), the (.01), …

$\mathcal{G}_{\{\mathrm{of}\}} =$ America (.0001), of (.000000001), States (.0001), the (.01), …

$\vdots$

$\mathcal{G}_{\{\mathrm{United},\mathrm{States},\mathrm{of}\}} =$ America (.8), of (.00000001), States (.0000001), the (. 01), …



Pitman-Yor Process?

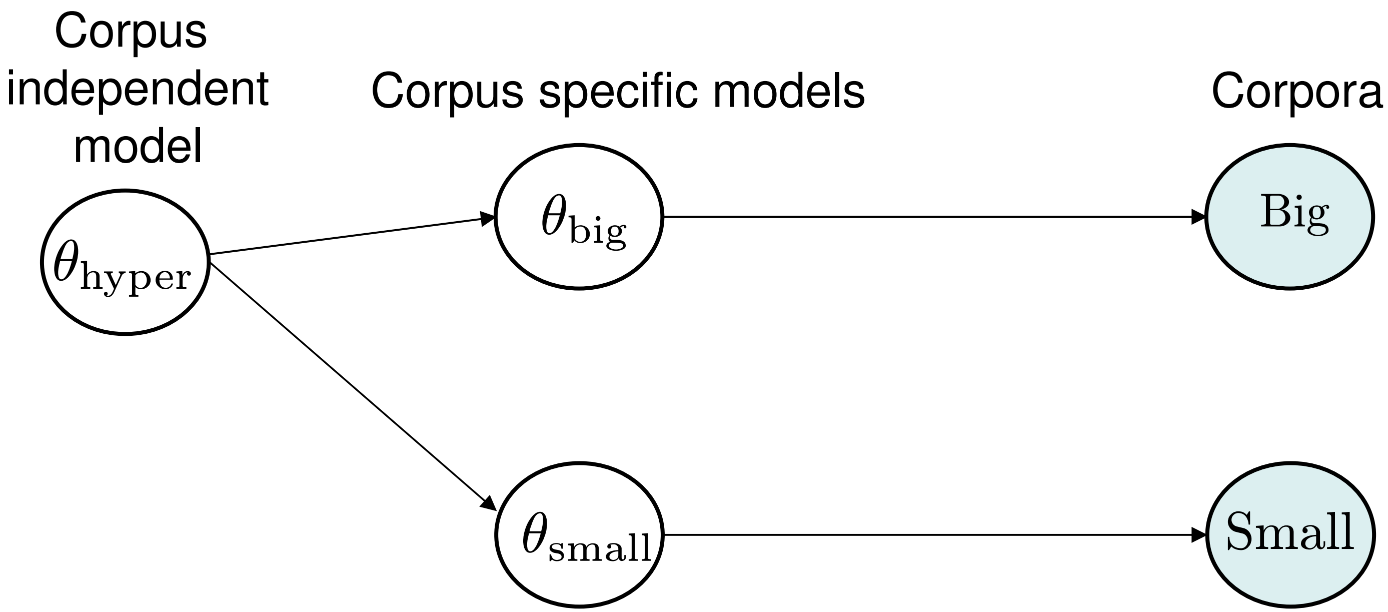
discount concentration

$$\mathcal{G} \sim \mathrm{PY}(\boxed{d}, \boxed{\alpha}, \boxed{\mathcal{G}_0})$$
$$x \sim \mathcal{G}$$

$$\boxed{\mathcal{G}^{\mathcal{D}}_{\{w_{t-j}:w_{t-1}\}}} \sim \mathrm{PY}(\boxed{d^{\mathcal{D}}_j}, \boxed{\alpha^{\mathcal{D}}_j}, \boxed{\lambda^{\mathcal{D}}_j \mathcal{G}^{\mathcal{D}}_{\{w_{t-j+1}:w_{t-1}\}} + (1-\lambda^{\mathcal{D}}_j)\mathcal{G}^{\mathcal{H}}_{\{w_{t-j}:w_{t-1}\}}})$$
$$x | w_{t-n+1} : w_{t-1} \sim \boxed{\mathcal{G}^{\mathcal{D}}_{\{w_{t-n+1}:w_{t-1}\}}}.$$

base distribution

Hierarchical Bayesian Modelling



What does G look like?

- G is a discrete distribution with infinite support

$$
\begin{aligned}
\mathcal{G} &\sim \mathrm{PY}(d, \alpha, \mathcal{H}) \\
\Rightarrow \mathcal{G} &= \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}, \sum_{k=1}^{\infty} \pi_k = 1
\end{aligned}
$$

aa (.0001), aah (.00001), aal (.000001), aardvark (.001), …

[Sethurman, 94]

What does G look like?

- Stick breaking representation (GEM)

$$\pi_k = w_k \prod_{\ell=1}^{k-1}(1-w_\ell), w_k \sim \mathrm{Beta}\boxed{(1-d, \alpha+kd)}$$

$$\phi_k \quad \sim \quad \mathcal{H}$$

The Dirichlet process arises when d = 0

$$\begin{aligned} \mathcal{G} \quad &\sim \quad \mathrm{PY}(d, \alpha, \mathcal{H}) \\ \Rightarrow \mathcal{G} \quad &= \quad \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k} \end{aligned}$$

[Sethurman, 94]