Modeling and Unsupervised Learning of Structured Similarity Among Source Contexts in Bayesian Hierarchical Infinite Mixture Models

With Two Applications to Modeling Natural Language Semantics

Colin Reimer Dawson

May 18, 2015

Outline

- Background and Problem Formulation
    - Mixture Models
    - Infinite Mixture Models
    - Hierarchical Dirichlet Process Mixtures
    - Non-Exchangeability
- HDP Hidden Markov Model with Local Transitions
    - Hidden Markov Models
    - The Infinite (HDP) Hidden Markov Model
    - Adding the Concept of “Local” Transitions
    - Posterior Inference
- Two DP Models for Context-Sensitive Phrase Structure Grammar
    - Context-Free Grammars
    - An HDP Formulation of Backoff Smoothing
    - Future Work: A Similarity-Based Alternative

Mixture Models

- Goal: Estimate unknown density, f, of the form

$$f(\mathbf{x}) = \sum_k \pi_k f_k(\mathbf{x}) \tag{1}$$

- Traditionally, number of components is fixed and f_k have parametric form:

$$f(\mathbf{x}; \boldsymbol{\pi}, \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f(\mathbf{x}; \theta_k) \tag{2}$$

- Estimate $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$.
  - MLE: can find local optimum using Expectation-Maximization (EM) algorithm.

A Bayesian Approach

- Standard Bayesian version:

$$\pi \sim \text{Dirichlet}(\alpha \mathbf{1}_K) \tag{3}$$

$$\theta_k \overset{i.i.d.}{\sim} f\text{-Conjugate}(\xi) \tag{4}$$

- Straightforward to do Gibbs Sampling

Unbounded number of components

- Having to specify K in advance is limiting. Too high → overfitting. Too low → underfitting.
- We can instead use an *infinite mixture model*, with a prior to guard against overfitting.

Dirichlet Processes

Definition: Dirichlet Process

A **Dirichlet Process** [Ferguson, 1973] with **base probability measure** G_0 and **concentration parameter** $\alpha > 0$ is a random measure, μ on a measure space $(\mathcal{X}, \Sigma)$ with the property that, for any finite partition, $\{A_1, \ldots, A_n\}$ of $\mathcal{X}$, the induced random vector

$$(\mu(A_1), \ldots, \mu(A_n)) \sim \text{Dirichlet}(\alpha G_0(A_1), \ldots, \alpha G_0(A_n)) \tag{5}$$

- Note: μ is atomic $a.s.$.
- Note: If G_0 is atomic with finite support, μ reduces to a Dirichlet distribution over those atoms.

Normalized Gamma Process Representation

The DP is obtained by normalizing a *Gamma Process*:
A **Gamma Process** is a Poisson Process on $\mathbb{R}^+ \times \Theta$ with Lévy intensity measure

$$\nu(d\pi, d\theta) = \alpha \pi^{-1} e^{-\pi} d\pi G_0(d\theta) \tag{6}$$

That is, consider a random collection of point masses $\{\theta_k\}$ on Θ, with respective random masses $\{\pi_k\}$ as points $(\pi_k, \theta_k) \in \mathbb{R}^+ \times \Theta$. The number $n(A)$ of such points in a region $A \subset \mathbb{R}^+ \times \Theta$ is distributed

$$n(A) \sim \mathcal{P}\mathrm{ois}\left(\int_A \nu(d\pi, d\theta)\right) \tag{7}$$

Normalized Gamma Process Representation

The sum $T = \sum_k \pi_k$ is finite almost surely (see, e.g., [Ferguson, 1973]), so we can normalize the set of atoms to form a probability measure on Θ.

This normalized measure is distributed $\mathrm{DP}(\alpha G_0)$.

A Constructive Definition of the DP

The Stick-Breaking Construction [Sethuraman, 1991]

Define

$$\{\pi'_k\}_{k=1}^{\infty} \overset{i.i.d}{\sim} \mathcal{B}\text{eta}(1, \alpha) \tag{8}$$

$$\pi_k = \pi'_k \prod_{k=1}^{k-1} (1 - \pi'_{k-1}) \tag{9}$$

$$\{\theta_k\}_{k=1}^{\infty} \overset{i.i.d.}{\sim} G_0 \tag{10}$$

Then

$$\mu \overset{def}{=} \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k} \tag{11}$$

is distributed as a Dirichlet Process with base measure G_0 and concentration parameter α.

An Infinite Mixture Model

DP Mixture Model

If we let

$$\boldsymbol{\pi} \sim \text{Stick}(\alpha) \tag{12}$$

$$\{\theta_k\} \overset{i.i.d}{\sim} G_0 \tag{13}$$

$$f(\mathbf{x} \,|\, \boldsymbol{\pi}, \boldsymbol{\theta}) = \sum_{k=1}^{\infty} \pi_k f(\mathbf{x} \,|\, \theta_k) \tag{14}$$

then $\mathbf{x}$ are distributed according to a **Dirichlet Process Mixture Model**

Infinite Gaussian Mixture Model

For example, if $f(\mathbf{x} \,|\, \theta_k)$ is a Normal density, we have the **Infinite Gaussian Mixture model** [Rasmussen, 2000].

We could then let G_0 be a Normal Inverse Gamma model for μ and σ^2.

Hierarchical Dirichlet Processes

- If we have data from multiple sources, $j = 1, \ldots, J$, whose generating distributions, $\{G_j\}$ are distinct but related, we can use a hierarchical prior to couple them, e.g.,

$$G_j \overset{i.i.d}{\sim} \mathrm{DP}(\alpha G_0), \tag{15}$$

  where dependence is introduced by putting a hyperprior on G_0 and integrating it out.
- Problem: If G_0 is absolutely continuous, the atoms in the G_j will be at disjoint locations.
- Solution: Let G_0 itself have a DP prior.

Hierarchical Dirichlet Processes

The Hierarchical Dirichlet Process Mixture Model [Teh et al., 2006]

Define

$$G_0 \sim \mathrm{DP}(\gamma H) \tag{16}$$

$$G_j \,|\, G_0 \overset{i.i.d}{\sim} \mathrm{DP}(\alpha G_0) \qquad j = 1, \ldots, J \tag{17}$$

$$\{(\pi_{jk}, \theta_{jk})\}_{k=1}^{\infty} \leftarrow G_j \tag{18}$$

$$\{\mathbf{x}_{jn}\}_{n=1}^{N} \,|\, \{\pi_{jk}\theta_{jk}\} \overset{i.i.d}{\sim} \sum_{k=1}^{\infty} \pi_{jk} f(\mathbf{x} \,|\, \theta_{jk}) \tag{19}$$

This defines a Hierarchical Dirichlet Process (HDP) Mixture. Atoms are shared among contexts by virtue of the discreteness of G_0.

Examples of HDP Mixture Models

- If H is a Normal-Inverse Gamma distribution and f is a Normal density, we have a hierarchical Infinite Gaussian Mixture.
- If H is itself a Dirichlet distribution, and f is a Multinomial mass function, we obtain an **hierarchical infinite topic model**: x are words, and θ_{jk} parameterize multinomial distributions corresponding to “topics” in a document.

Exchangeability vs. Dependence

- A limitation of the HDP is that the components in the bottom level measures are **exchangeable**, and independent given G_0. Can we allow for particular component pairs to have correlated weights across contexts? (E.g., topics cooccur across documents)
- Second, both the *contexts* and the *data points* are exchangeable. Can we couple them through covariates?

Hidden Markov Models

One way to incorporate temporal dependence in a mixture model is the **Hidden Markov Model**.

Hidden Markov Model

Let $\{f_k\}_{k=1}^K$ be a finite family of density functions and $\{\boldsymbol{\pi}_k\}_{k=0}^K$ a family of K-dimensional Multinomial distributions. Define a *latent state sequence*, $\{z_t\}_{t=1}^T$ and an *observation sequence* $\{\mathbf{x}_t\}_{t=1}^T$ such that

$$z_t \,|\, z_{t-1} \sim \pi_{z_{t-1}} \quad t = 1, \ldots, T \tag{20}$$

$$\mathbf{x}_t \,|\, z_t \sim f_{z_t} \quad t = 1, \ldots, T \tag{21}$$

with z_0 defined to be 0. This defines a **Hidden Markov Model** (HMM).

A Bayesian HMM

A standard Bayesian formulation for estimating the emission and transition distributions, $\{f_k\}$ and $\{\boldsymbol{\pi}_k\}$, where the f_k are members of a parametric family, $f_k(\mathbf{x}) = f(\mathbf{x} \mid \theta_k)$ is to use priors

$$\boldsymbol{\pi}_k \overset{i.i.d}{\sim} \text{Dirichlet}(\alpha \mathbf{1}_K) \tag{22}$$

$$\theta_k \overset{i.i.d.}{\sim} f\text{-Conjugate}(\xi) \tag{23}$$

It is then straightforward to do Gibbs sampling over these variables as well as the latent state sequence (which can be sampled jointly using a dynamic programming-based message passing algorithm).

An Infinite State Generalization

We can allow infinitely many states by replacing the Dirichlet prior with a Dirichlet Process prior:

$$(\boldsymbol{\pi}_k, \boldsymbol{\theta}_k) \overset{i.i.d}{\sim} \mathrm{DP}(\alpha G_0) \tag{24}$$

where $\boldsymbol{\theta}_k$ represents the vector of emission parameters for the states reachable from state k.

However, we need G_0 to be atomic, to ensure that the $\boldsymbol{\theta}_k$ contain overlapping values for different k.

Solution: Use a hierarchical prior, with $G_0 \sim \mathrm{DP}(\gamma H)$. This is the Infinite or HDP HMM [Beal et al., 2001, Teh et al., 2006]

Excessive Exchangeability

- Two properties of HDP-HMM not shared with non-temporal HDP:
  1. Contexts (except the first) are random
  2. Set of contexts is identified with set of states
- Self-transitions are not special
- No notion of a state topology: $\pi_{kk'}$ should (perhaps) be similar (but not identical) to $\pi_{k'k}$; states with similar incoming distributions should (perhaps) have similar (but not identical) outgoing distributions.

Making Self-Transitions Special: Two Approaches

- Two solutions to making self-transitions special are:
  1. the Sticky HDP-HMM [Fox et al., 2008]:

     $$\boldsymbol{\pi}_k \sim \mathrm{DP}(\alpha G_0 + \kappa \delta_{\theta_k}) \tag{25}$$

  2. the HDP-HSMM [Johnson and Willsky, 2013]: rule out self-transitions, and model durations separately

Local Transitions: The HDP-HMM-LT

- I incorporate the notion of state similarity by defining an HDP-HMM that favors "local" transitions: The HDP-HMM-LT.
- Key idea: latent states are located an abstract space on which a symmetric similarity kernel, ϕ is defined:

$$0 \leq \phi(\ell_k, \ell_{k'}) = \phi(\ell_{k'}, \ell_k) \leq \phi(\ell_k, \ell_k) \equiv 1 \tag{26}$$

- The transition probabilities generated by the HDP prior are scaled by the corresponding $\phi_{kk'}$.

The HDP-HMM-LT

Definition: HDP-HMM-LT

Assume we have a sequence of location pairs $\{(\theta_k, \ell_k)\}$ from some distribution, and a similarity kernel ϕ. We define

$$\boldsymbol{\beta} \sim \mathrm{Stick}(\gamma) \tag{27}$$

$$\tilde{\boldsymbol{\pi}}_k \sim \mathrm{DP}(\alpha\boldsymbol{\beta}) \tag{28}$$

$$a_{kk'} = \tilde{\pi}_k(k')\phi(\ell_k, \ell_{k'}) \tag{29}$$

$$z_t \mid z_{t-1} \sim \sum_k \frac{a_{z_{t-1}k}}{\sum_{k'} a_{z_{t-1}k'}} \delta_k \tag{30}$$

$$x_t \mid z_t \sim F(\theta_{z_t}) \tag{31}$$

Note that the normalization term is finite and positive almost surely, since it is bounded above by $\sum_{k'} \tilde{\pi}_k(k') = 1$, and below by $\tilde{\pi}_k(k)$.

A Gamma Process representation

By the Gamma Process representation of the DP, we obtain the same model by drawing $\boldsymbol{\beta}$ as above and setting

$$\tilde{\pi}_k(k') = \frac{\pi_{kk'}}{\sum_{k''} \pi_{kk''}} \tag{32}$$

where

$$\{\pi_{kk'}\}_k \overset{i.i.d.}{\sim} \mathcal{G}(\alpha\beta_{k'}, 1) \qquad k' \geq 1 \tag{33}$$

It is known that the Lévy measure underlying the Gamma process meets the sufficient conditions for the normalization constant to be positive and finite almost surely, namely

$$\int_{\mathbb{R}^+} \rho(d\pi) = +\infty \qquad \int_{\mathbb{R}^+} (1 - e^{-\pi})\rho(d\pi) < \infty \tag{34}$$

A Gamma Process representation

Combining the two normalizations yields

HDP-HMM-LT (Gamma Process Representation)

$$\boldsymbol{\beta} \sim \text{Stick}(\gamma) \tag{35}$$

$$\pi_{kk'} \sim \mathcal{G}(\alpha\beta_{k'}, 1) \tag{36}$$

$$a_{kk'} = \pi_{kk'}\phi(\ell_k, \ell_{k'}) \tag{37}$$

$$z_t \mid z_{t-1} \sim \sum_k \frac{a_{z_{t-1}k}}{\sum_{k'} a_{z_{t-1}k'}}\delta_k \tag{38}$$

$$x_t \mid z_t \sim F(\theta_{z_t}) \tag{39}$$

Loss of Conjugacy for $\boldsymbol{\pi}$

A difficulty introduced by the rescaling is that the likelihood for π, fixing the sequence $\mathbf{z}$ is no longer conjugate, due to the normalization. Let $n_{kk'}$ be the number of transitions from k to k' in $\mathbf{z}$. Then

$$p(\mathbf{z} \mid \boldsymbol{\pi}, \boldsymbol{\phi}) = \prod_{k} \left(\sum_{k''} \pi_{kk'} \phi_{kk'} \right)^{-n_{k\cdot}} \prod_{k'} (\pi_{kk'} \phi_{kk'})^{n_{kk'}} \qquad (40)$$

But we can restore conjugacy by introducing auxiliary variables.

The Markov Process With Failed Jumps Representation

A discrete time Markov chain is obtained from a pure jump Markov process by allowing self-jumps, and marginalizing time.

Let A be a rate matrix for such a process. Then time spent in state k is distributed $\mathsf{Exp}(\sum_{k'} a_{kk'})$ and is independent of the state jumped to, which is distributed by the normalized kth row of A.

The Markov Process With Failed Jumps Representation

Given $n_{k\cdot} = \sum_{k'} n_{kk'}$ visits to state k, the chain will spend

$$u_k \sim \mathcal{G}(n_{k\cdot}, \sum_{k'} a_{kk'}) \tag{41}$$

time there.

The augmented likelihood is now

$$\begin{aligned} p(\mathbf{z}, \mathbf{u} \,|\, \boldsymbol{\pi}, \boldsymbol{\phi}) &= \prod_k \Gamma(n_{k\cdot})^{-1} e^{-u_k \sum_{k'} a_{kk'}} \prod_{k'} (\pi_{kk'} \phi_{kk'})^{n_{kk'}} & (42) \\ &= \prod_k \Gamma(n_{k\cdot})^{-1} \prod_{k'} (\pi_{kk'} \phi_{kk'})^{n_{kk'}} e^{-u_k \phi_{kk'} \pi_{kk'}} & (43) \end{aligned}$$

Introducing failed jumps

Suppose also that while in state k, unsuccessful attempts to jump to state k' are made at rate $\pi_{kk'} - a_{kk'} = \pi_{kk'}(1 - \phi_{kk'})$. The total number of these is

$$q_{kk'} \sim \mathcal{P}\text{ois}\left(\pi_{kk'}(1 - \phi_{kk'})u_k\right) \tag{44}$$

Augmented Likelihood for $\boldsymbol{\pi}$

Augmenting the data with $\mathbf{u}$ and $\mathbf{Q}$, the likelihod for $\boldsymbol{\pi}$ is now

$$\mathcal{L}(\boldsymbol{\pi} \mid \mathbf{z}, \mathbf{u}, \mathbf{Q}; \phi) \propto \prod_k \prod_{k'} \pi_{kk'}^{n_{kk'}} e^{-u_k \phi_{kk'} \pi_{kk'}} \tag{45}$$

$$\times e^{-u_k(1-\phi_{kk'})\pi_{kk'}} \frac{(u_k(1-\phi_{kk'})\pi_{kk'})^{q_{kk'}}}{q_{kk'}!} \tag{46}$$

$$\propto \prod_k \prod_{k'} \pi_{kk'}^{n_{kk'}+q_{kk'}} e^{-u_k \pi_{kk'}} \tag{47}$$

which is conjugate to the Gamma prior.

Completing the Gibbs Sampler

- Conditioned on A and the observations, we can sample the state sequence $\mathbf{z}$ jointly with standard HMM message passing.
- We can sample $\boldsymbol{\pi}$ and its hyperparameters ($\boldsymbol{\beta}$, α, and γ) jointly as well by the factorization

  $$p(\gamma, \alpha, \boldsymbol{\beta}, \boldsymbol{\pi} \,|\, \mathcal{D}) = p(\gamma \,|\, \mathcal{D}) p(\alpha \,|\, \mathcal{D}) p(\boldsymbol{\beta} \,|\, \gamma, \mathcal{D}) p(\boldsymbol{\pi} \,|\, \boldsymbol{\beta}, \alpha, \mathcal{D}) \tag{48}$$

  where $\mathcal{D}$ is the augmented "data".
- These terms require marginal likelihoods for $\boldsymbol{\beta}$, α and γ.

Marginal Likelihoods

The marginal likelihood for $\boldsymbol{\beta}$ (integrating out $\boldsymbol{\pi}$) is

$$\mathcal{L}(\boldsymbol{\beta} \,|\, \mathcal{D}) \propto \int \prod_{k} \prod_{k'} \Gamma(\alpha\beta_{k'})^{-1} \pi_{kk'}^{\alpha\beta_{k'}+n_{kk'}+q_{kk'}-1} e^{-(1+u_k)\pi_{kk'}} d\boldsymbol{\pi} \tag{49}$$

$$\propto \prod_{k} \prod_{k'} (1+u_k)^{-\alpha\beta_{k'}} \frac{\Gamma\left(\alpha\beta_{k'}+n_{kk'}+q_{kk'}\right)}{\Gamma(\alpha\beta_{k'})} \tag{50}$$

$$\propto \prod_{k} (1+u_k)^{-\alpha} \prod_{k'} \frac{\Gamma\left(\alpha\beta_{k'}+n_{kk'}+q_{kk'}\right)}{\Gamma(\alpha\beta_{k'})} \tag{51}$$

$$\propto \prod_{k} \prod_{k'} \sum_{m=1}^{n_{kk'}+q_{kk'}} s(n_{kk'}+q_{kk'}, m) \alpha^{m} \beta_{k'}^{m} \tag{52}$$

Marginal Likelihoods

- So we can employ the auxiliary variable method of [Escobar and West, 1995], introducing a collection of random $m_{kk'}$ whose distributions depend on $n_{kk'}$, $q_{kk'}$, α and $\boldsymbol{\beta}$.
- $m_{kk'}$ represents the number "sticks" assigned to component k' in context k, during the DPs generating $\boldsymbol{\pi}$.
- After adding $\mathbf{M}$ to $\mathcal{D}$, the $\boldsymbol{\beta}$ likelihood is conjugate to the DP prior.

Marginal Likelihoods

With the addition of $\mathbf{M}$, the marginal likelihood for α is now simply

$$\mathcal{L}(\alpha \,|\, \mathcal{D}) \propto \prod_k (1 + u_k)^{-\alpha} \prod_{k'} \alpha^{m_{kk'}} \tag{53}$$

$$\propto \alpha^{m_{\cdot\cdot}} e^{-\sum_k \log(1+u_k)\alpha} \tag{54}$$

which is conjugate to a Gamma prior. This is simpler than in the formulation of either [Escobar and West, 1995] or [Teh et al., 2006] due to the presence of $\mathbf{u}$.

Marginal Likelihoods

We can further integrate out $\boldsymbol{\beta}$ to obtain a marginal likelihood for γ. Collapsing all unrepresented components of $\boldsymbol{\beta}$ into a single component β_* yields a degenerate Dirichlet prior, and so

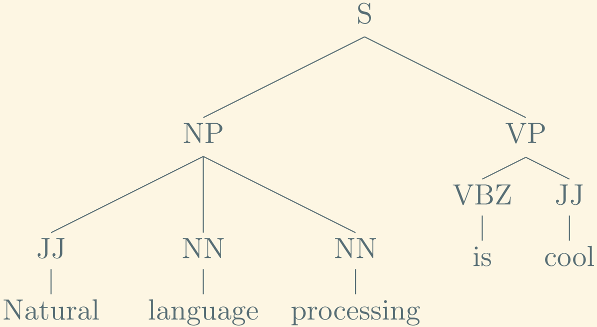
$$\begin{aligned}
\mathcal{L}(\gamma \,|\, \mathcal{D}) &\propto \int \beta_*^{\gamma} \prod_k \beta_k^{m_{\cdot k}} d\boldsymbol{\beta} && (55)\\
&\propto \frac{\Gamma(\gamma) \prod_k \Gamma(m_{\cdot k})}{\Gamma(\gamma + m_{\cdot\cdot})} && (56)\\
&\propto \int_0^1 t^{\gamma-1}(1-t)^{m_{\cdot\cdot}-1} dt && (57)
\end{aligned}$$

and so adding one more auxiliary variable $t \sim \mathcal{B}\text{eta}(\gamma, m_{\cdot\cdot})$ yields a marginal likelihood for γ which is conjugate to a Gamma prior.

Interim Summary of Contributions

- I have introduced a new model, the HDP-HMM-LT, in which there is a notion of topology on the transition state space, such that transitions are more likely between nearby states.
- I have formulated the HDP-HMM-LT as the marginalization of another process, the Markov Process With Failed Jumps.
- By “reinstating” selected functions of the marginalized variables, the model admits a straightforward Gibbs sampler
- The HDP parameters, $\boldsymbol{\pi}$, $\boldsymbol{\beta}$, α and γ can be Gibbs-sampled jointly, conditioned on the augmented data.

Constituency Trees



Context-Free Grammar

Ordinary Context Free Grammar: syntactic constituents (nonterminal phrases) expanded into sequences of smaller phrase nodes and "preterminals" (part of speech tags).

$S \rightarrow NP\ VP$	$NP \rightarrow JJ\ NN\ NN$	$VP \rightarrow VBZ\ JJ$
$JJ \rightarrow$ Natural	$NN \rightarrow$ language	$NN \rightarrow$ processing
$V \rightarrow$ is	$JJ \rightarrow$ cool	

Now with Probabilities

- But the same LHS can expand in more than one way.
- **Probabilistic** Context-Free Grammar associates each production with a conditional probability of the RHS given the LHS.

$S \overset{0.7}{\rightarrow} NP\ VP$	$NP \overset{0.02}{\rightarrow} JJ\ NN\ NN$	$VP \overset{0.2}{\rightarrow} V\ JJ$
$JJ \overset{0.002}{\rightarrow}$ Natural	$NN \overset{0.001}{\rightarrow}$ language	$NN \overset{0.0001}{\rightarrow}$ processing
$V \overset{0.01}{\rightarrow}$ is	$JJ \overset{0.002}{\rightarrow}$ cool	

- Context-free assumption: each sequence is conditionally independent of everything else given the parent.
- Therefore, tree probability is product of production probabilities.

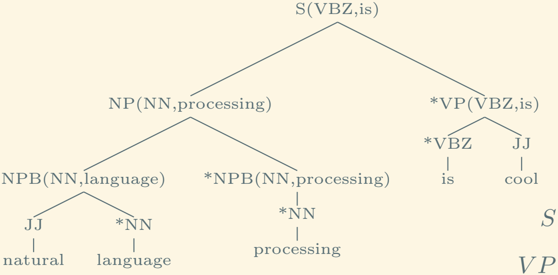
Problems With This Structure

- This grammar is both too flexible and too inflexible.
- Too inflexible: Context-free assumption is unrealistic; grandparents, aunts, etc. matter. Too few parameters!
- Too flexible: Each sequence (say, $JJ\ NN\ NN$) treated as a distinct category, regardless of similarity (to, say, $NN\ NN$). Too many parameters!

Lexicalizing and Factoring the Production Probabilities

- [Collins, 2003] augments representation by labeling constituent nodes with lexical and “head/modifier” information.
- Each constituent has a single head word.
- Multiple production types
- Modifiers generated one at a time (conditionally independent of each other given parent and head).

Head-Driven Lexicalized Grammar



$$ROOT \xrightarrow{\text{head}} S(VBZ,\text{is})$$

$$S(VBZ,\text{is}) \xrightarrow{\text{head}} VP(VBZ,\text{is})$$

$$VP(VBZ,\text{is}) \xrightarrow{\text{head}} VBZ(\text{is})$$

$$VP(VBZ,\text{is}) \xrightarrow{\text{right}} JJ(\text{cool})$$

$$S(VBZ,\text{is}) \xrightarrow{\text{left}} NP(NN,\text{processing})$$

$$NP(NN,\text{processing}) \xrightarrow{\text{head}} NBP(NN,\text{processing})$$

$$NBP(NN,\text{processing}) \xrightarrow{\text{head}} NN(\text{processing})$$

$$NP(NN,\text{processing}) \xrightarrow{\text{left}} NPB(NN,\text{language})$$

$$NPB(NN,\text{language}) \xrightarrow{\text{head}} NN(\text{language})$$

$$NPB(NN,\text{language}) \xrightarrow{\text{left}} JJ(\text{natural})$$

Parameter Estimation

- For modifiers, condition on (sister) head, as well as parent.
- But contexts now contain words — very little training data per context!
- Solution: “smoothing” distributions for similar contexts

$$S(JJ, \mathit{cool}) \overset{\text{head}}{\rightarrow} VP \mid \text{Parent} = \text{Root}$$

$$\begin{aligned}
&P(VP \mid \text{Root}, S, JJ, \text{cool}) \\
&= \lambda_0 \hat{P}(VP \mid \text{Root}, S, JJ, \text{cool}) \\
&\qquad + (1-\lambda_0)\big(\lambda_1 \hat{P}(VP \mid \text{Root}, S, JJ) \\
&\qquad\qquad + (1-\lambda_1)(\lambda_2 \hat{P}(VP \mid \text{Root}) + (1-\lambda_2)\varepsilon)\big)
\end{aligned}$$

How Should We Choose λs?

Collins sets smoothing parameters from the data as

$$\lambda_A = \frac{c_A}{c_A + bu_A}$$

c_A: # of observations of context A
u_A: # of distinct outcomes in context A (the “diversity”)
b: tunable parameter shared over all contexts

Intuition

- More data for context (high c_A) → trust the MLE more
- More distinct outcomes (high u_A) → high entropy → need more data to get a reliable estimate

A Bayesian Interpretation of Smoothing

- With only one smoothing level, i.e.,

  $$P_A = \lambda \hat{P}_A + (1 - \lambda)\varepsilon$$

  this would be the predictive distribution assuming a Dirichlet-multinomial model, with a symmetric prior over $1/\varepsilon$ categories and concentration hyperparameter related to $(1 - \lambda)$. Specifically

  $$\lambda = \frac{c_A}{c_A + \alpha} \tag{58}$$

A Bayesian Interpretation of Smoothing

We can extend an HDP to arbitrary levels by letting $\pi_K(A)$ represent the predictive distribution for context A at level K (where smaller K collapse prefix-equivalent As). Then

$$\pi_K(A) = \lambda_K \hat{\pi}_K(A) + (1 - \lambda_k)\pi_{K-1}(A)$$

where $\hat{\pi}$ is the empirical distribution,

$$\lambda_K = \frac{m_K(A)}{m_K(A) + \alpha_K} \tag{59}$$

and $m_K(A)$ is the total number of distinct masses (“sticks”) with which the observations in context A are associated, and α_K is the concentration parameter of the DP at level K.

Estimating α

The set of $m_K(A)$ for all the contexts at level K is probabilistically dependent on α_K, since higher α_K leads to more a more entropic set of stick weights. [Antoniak, 1974] showed that, for a fixed number n of draws from the distribution, the number of distinct "sticks" represented has distribution given by

$$p(m \mid \alpha, n) = \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)} s(n, m) \alpha^m \tag{60}$$

This is proportional (in α) to

$$\alpha^m \int_0^1 w^{\alpha-1}(1-w)^{n-1} dw \tag{61}$$

so if we draw $w \sim \mathcal{B}\text{eta}(\alpha, n)$, the augmented likelihood is conjugate to a Gamma prior with shape and rate updates m and $-\log(w)$, respectively.

Estimating α

For fixed w, the posterior mean for α is asymptotically proportional to m. Hence if we approximate m by the diversity u and all ws by a fixed constant w_0 (perhaps a function of the sample size), we might estimate α by $w_0 u$, as Collins does:

$$\lambda_A = \frac{c_A}{c_A + b u_A}$$

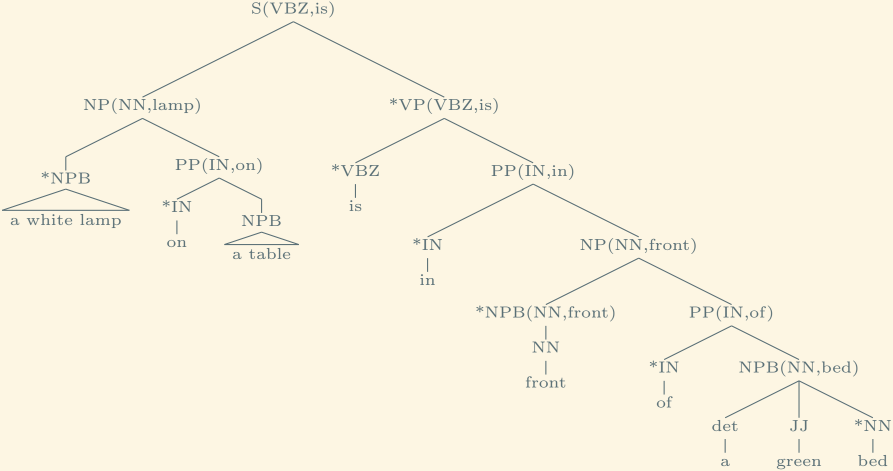
c_A: # of observations of context A
u_A: # of distinct outcomes in context A (the "diversity")
b: tunable parameter shared over all contexts

A true HDP model

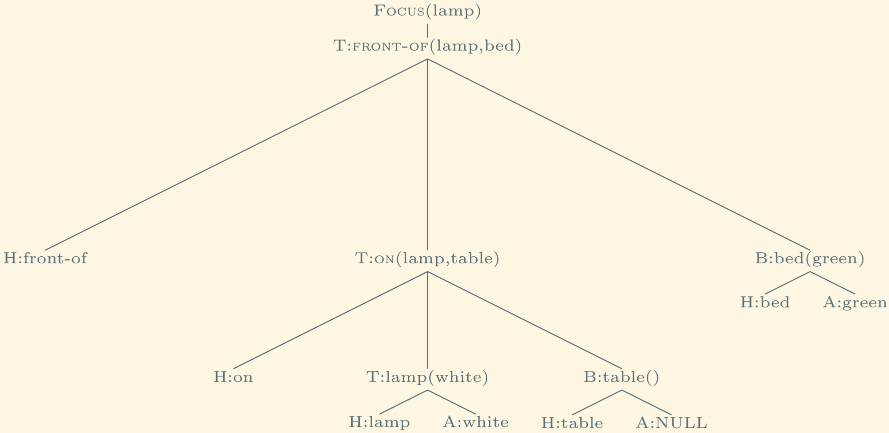
- Instead of these heuristic approximations, I propose Gibbs sampling the m, w and α to derive a Monte Carlo estimate of the predictive probabilities.
- This can perhaps improve parsers trained on annotated corpora (empirical study to come).
- Further, by specifying a full probabilistic model, we can account for missing or incomplete context data in the training bank of parse trees, such as semantic information (informed by paired images).

Conditioning Grammar on Semantics

We want the probability of this...



to depend on this:



Focus(lamp)

T:front-of(lamp,bed)

H:front-of

T:on(lamp,table)

B:bed(green)

H:bed

A:green

H:on

T:lamp(white)

B:table()

H:lamp

A:white

H:table

A:NULL

Sparse Dependency Assumption

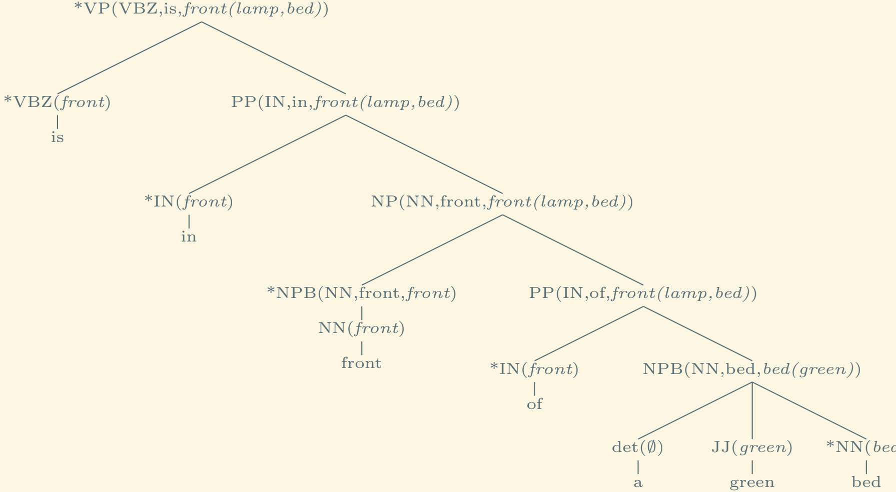
- That's a lot of structure to condition on. Smoothing only helps so much.
- Intuition: A given syntactic subtree expresses a particular "semantic constituent".

$$
\begin{aligned}
front(lamp, bed) &\leftarrow S(VBZ, is) \\
bed(green) &\leftarrow NP(NN, bed)
\end{aligned}
$$

- Moreover, correspondences are largely "continuous" (connected graphs map to connected graphs).
- Hence, add semantic features to the head/modifier features, and define new production types.

Semantic Features and Events in $P(T|\psi, \mathcal{G})$

- Syntactic root is associated with root predicate; each step down syntactic tree associated with (a) no move, (b) move down, or (c) move to null, in semantic tree.



*VP(VBZ,is,*front(lamp,bed)*)

*VBZ(*front*)

is

PP(IN,in,*front(lamp,bed)*)

*IN(*front*)

in

NP(NN,front,*front(lamp,bed)*)

*NPB(NN,front,*front*)

NN(*front*)

front

PP(IN,of,*front(lamp,bed)*)

*IN(*front*)

of

NPB(NN,bed,*bed(green)*)

det(∅)

a

JJ(*green*)

green

*NN(*bed*

bed

Semantic Features and Events in $P(T|\psi, \mathcal{G})$

- Simply adds new production types to the grammar.
- For a fixed mapping, we can employ the same context-sensitive grammar model to estimate the likelihood of a parse.
- Use this as a likelihood to sample (a) mappings, and (b) the semantic trees themselves

A Similarity-Based Alternative

- The HDP model requires context features to be ordered in advance.
- No tying between contexts that differ on a feature in the first "tier".
- Instead: couple context distributions via an overall similarity measure.

A Similarity-Based Alternative

- Treat each context distribution, P_A as an infinite mixture of principal “topic” distributions

$$P_A = \sum_{k=1}^{\infty} w_k(A) P'_k, \qquad \sum_k w_k(A) = 1 \tag{62}$$

- Introduce random topic indicators, $\{z_i\}$ for each observed production, with

$$z_i \,|\, A_i \sim \sum_{k=1}^{\infty} w_k(A_i)\delta_k \tag{63}$$

$$x_i \,|\, z_i \sim P_{z_i} \tag{64}$$

- Key: coupled prior on $\{w(A)\}_A$ so that $w(A)$ and $w(A')$ are similar when A and A' are similar.

The Distance-Dependent Chinese Restaurant Process

- Several approaches exist to place a distribution on partitions that bias groupings based on similarity on covariates: [MacEachern, 2000, Dunson et al., 2007, Dahl, 2008, Müller et al., 2011]
- However, all assume that the covariates are known, which is not the case when the context consists of outcomes above in the tree.
- The Distance-Dependent Chinese Restaurant Process [Blei and Frazier, 2011] assumes this as well, but admits a straightforward modification to resample context.

The Chinese Restaurant Process

- By integrating out the “stick weights”, we get a marginal distribution on partitions of observations to “sticks” called the Polya Urn Process, aka, the **Chinese Restaurant Process**.
- Conditioned on a partition of N other observations, into K clusters (numbered $1, \ldots, K$), with n_k in cluster k, the cluster assignment z_{N+1} of the $N+1$ data point has distribution

$$p(z_{N+1} = k) \propto \begin{cases} n_k & k = 1, \ldots, K \\ \alpha & k = K+1 \end{cases} \tag{65}$$

The Distance-Dependent CRP

- Equivalently, assign observation $N+1$ to the same stick as another observation with probability proportional to N and to a new stick with probability proportional to α. Conditioned on the first case, choose an observation uniformly.
- The **Distance-Dependent CRP** generalizes this uniformity. Given a similarity function $\phi(\mathbf{x}, \mathbf{x}')$, where $\mathbf{x}$ and $\mathbf{x}'$ are covariates, link observation i to observation j according to

$$p(c_i = j) \propto \begin{cases} \phi(\mathbf{x}_i, \mathbf{x}_j) & i \neq j \\ \alpha & i = j \end{cases} \tag{66}$$

- The self-link case corresponds to creating a new cluster.

The Distance-Dependent CRP for Context-Sensitive Grammars

- We can then alternate sampling the per-cluster outcome variables (which in our case, unlike in the original ddCRP, needs to take into account the links, since it will change some contexts and hence the link probabilities), and sampling the links.
- Moreover, we can sample parameters of the similarity function conditional on the links to do relevance-determination.

Antoniak, C. E. (1974).
Mixtures of dirichlet processes with applications to bayesian nonparametric problems.
The annals of statistics, pages 1152–1174.

Beal, M. J., Ghahramani, Z., and Rasmussen, C. E. (2001).
The infinite hidden markov model.
In *Advances in neural information processing systems*, pages 577–584.

Blei, D. M. and Frazier, P. I. (2011).
Distance dependent chinese restaurant processes.
The Journal of Machine Learning Research, 12:2461–2488.

Collins, M. (2003).
Head-driven statistical models for natural language parsing.
Computational linguistics, 29(4):589–637.

Dahl, D. B. (2008).
Distance-based probability distribution for set partitions with applications to bayesian nonparametrics.

JSM Proceedings. Section on Bayesian Statistical Science, American Statistical Association, Alexandria, Va.

Dunson, D. B., Pillai, N., and Park, J.-H. (2007).
Bayesian density regression.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69(2):163–183.

Escobar, M. D. and West, M. (1995).
Bayesian density estimation and inference using mixtures.
Journal of the american statistical association, 90(430):577–588.

Ferguson, T. S. (1973).
A bayesian analysis of some nonparametric problems.
The annals of statistics, pages 209–230.

Fox, E. B., Sudderth, E. B., Jordan, M. I., and Willsky, A. S. (2008).
An hdp-hmm for systems with state persistence.
In *Proceedings of the 25th international conference on Machine learning*, pages 312–319. ACM.

Johnson, M. J. and Willsky, A. S. (2013).
Bayesian nonparametric hidden semi-markov models.
The Journal of Machine Learning Research, 14(1):673–701.

MacEachern, S. N. (2000).
Dependent dirichlet processes.
Unpublished manuscript, Department of Statistics, The Ohio State University.

Müller, P., Quintana, F., and Rosner, G. L. (2011).
A product partition model with regression on covariates.
Journal of Computational and Graphical Statistics, 20(1).

Rasmussen, C. E. (2000).
The infinite gaussian mixture model.
In Solla, S., Leen, T., and Müller, K.-R., editors, *Advances in Neural Information Processing Systems*, volume 12, pages 554–560. MIT Press.

Sethuraman, J. (1991).
A constructive definition of dirichlet priors.

Technical report, DTIC Document.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006).
Hierarchical Dirichlet processes.
Journal of the American Statistical Association, 101(476).