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

June 26, 2015

Contents

Abstract

In a classical mixture modeling, each data point is modeled as arising i.i.d. (typically) from a weighted sum of probability distributions, where both the weights and the parameters of the mixture components are targets of inference. When data arises from different sources that may not give rise to the same mixture distribution, a hierarchical model can allow the source contexts to share components while assigning different weights across them (while perhaps coupling the weights to "borrow strength" across contexts). The Dirichlet Process (DP) Mixture Model (e.g., Rasmussen (2000)) is a Bayesian approach to mixture modeling which models the data as arising from a countably infinite number of components: the Dirichlet Process provides a prior on the mixture weights that guards against overfitting. The Hierarchical Dirichlet Process (HDP) Mixture Model (Teh et al., 2006) employs a separate DP Mixture Model for each context, but couples the weights across contexts by using a common base measure which is itself drawn from a top-level DP. This coupling is critical to ensure that mixture components are reused across contexts. For example, in natural language topic modeling, a common application domain for mixture models, the components represent semantic topics, and the contexts are documents, and it is critical that topics be reused across documents.

These models have been widely adopted in Bayesian statistics and machine learning. However, a limitation of DPs is that the atoms are *a priori* exchangeable, and in the case of HDPs, the component weights are independent conditioned on the top-level measure. This is unrealistic in many applications, including topic modeling, where certain components (e.g., topics) are expected to correlate across contexts (e.g., documents). In the case of topic modeling, the Discrete Infinite Logistic Normal model (DILN; Paisley et al. (2011)) addresses this shortcoming by associating with each mixture component a latent location in an abstract metric, and rescaling each context-specific set of weights, initially drawn from an HDP, by an exponentiated draw from a Gaussian Process (GP), so that components which are nearby in space tend to have their weights be scaled up or down together. However, inference in this model requires the posterior distribution to be approximated by a variational family, as MCMC sampling from the exact posterior was deemed intractable. Thus, one goal of this dissertation is the development of simple MCMC algorithms for HDP models with correlated components.

A second application of HDPs is to time series models, in particular Hidden Markov Models (HMMs), where the HDP can be used as a prior on a doubly infinite transition matrix for the latent Markov chain, giving rise to the HDP-HMM (first developed, as the "Infinite HMM", by Beal et al. (2001), and subsequently shown to be a case of an HDP by Teh et al. (2006)). There, the hierarchy is over rows of the transition matrix, and the distributions across rows are coupled through a top-level Dirichlet Process. The sequential nature of the problem introduces two added wrinkles, namely that: the contexts themselves are random (since the context when generating state t is the state at time $t-1$), and the set of contexts is the same as the set of components. Hence, not only might the components be correlated with each other via locations in some latent space, but we might expect that contexts that correspond to correlated components will overall have similar distributions.

In the first part of the dissertation, I will present a formal overview of Dirichlet Processes and their various representations, as well as associated schemes for tackling the problem of doing approximate inference over an infinitely flexible model with finite computational resources. I will then turn to the Hierarchical Dirichlet Process, and review the literature on modeling correlations between components.

Next, I will present a novel probabilistic model, which I call the Hierarchical Dirichlet Process Hidden Markov Model With Local Transitions, which achieves the goal of simultaneously modeling correlations between contexts and components by assigning each a location in a metric space and promoting transitions between states that are near each other. I present a Gibbs sampling scheme for inference in this model, employing an augmented data representation to simplify the relevant conditional distributions. I give a intuitive interpretation of the augmented representation by casting the discrete time chain as a continuous time chain in which durations are not observed, and in which some jump attempts fail and are never observed. By tying the success probability of a jump between two states to the distance between them, the first successful (and therefore observed) jump is more likely to be to a nearby state. I refer to this representation as a Markov Process With Failed Jump Attempts. I test this model on both synthetic and real data, including a natural language data set drawn from a corpus of biological research articles, in which the goal is inferences about the semantic scope of assertions about biological processes implicated in cancer (to, e.g., species, organ sites, gene variants, etc.). There, the latent states are sets of entities in the scope, and the data is raw text. It is presumed that succesive assertions in a paper apply in similar scopes.

Finally, I present a generative model of natural language phrase structure where the problem is estimation of context-dependent distributions of parse tree symbols, and using that family of distributions to infer, from novel sentences, (1) the best syntactic parse tree, and (2) the semantic context that produced the sentence. There, "context" consists of a combination of the surrounding linguistic elements as well as the semantic context. The chief challenge stems from the huge number of possible contexts, and thus a principled method for tying the contexts based on similarity is needed. I present two approaches, the first based on a multilevel HDP model, where contexts are hierarchically nested based on shared features, and the second based on an adaptation of a generalization of the Dirichlet Process known as the Distance Dependent Chinese Restaurant Process (ddCRP; Blei and Frazier (2011)) to the problem. Here, rather than nesting the contexts, which requires a predetermined order of precedence among context features, I assign to each context an uncountably infinite mixture of principal "topic" components, where the mixing weights are unique per context, but are *a priori* similar across similar contexts. This is achieved by introducing latent cluster assignment variables, where affiliation to a cluster is biased by similarity to its other members. I also present a prior, likelihood, and inference algorithm to learn a sparse, cluster-specific similarity function. I test the model and inference algorithms on a corpus of captioned scenes, where the scenes provide a space of possible semantic contexts for the captions.

Chapter 1

A Hierarchical Dirichlet Process Hidden Markov Model With "Local" Transitions (HDP-HMM-LT)

I describe a generalization of the Hierarchical Dirichlet Process Hidden Markov Model (HDP-HMM; Teh et al. (2006)) which introduces a notion of latent similarity between pairs of hidden states, such that transitions are a priori more likely to occur between states with similar emission distributions. This is achieved by placing a similarity kernel on the space of state parameters, and scaling transition probabilities by the similarity between states. I refer to this model as the Hierarchical Dirichlet Process Hidden Markov Model with Local Transitions (HDP-HMM-LT). Although this achieves the goal of selectively increasing the probability of transitions between similar states, inference is made more complicated since the posterior measure over transition distributions is no longer a Dirichlet Process, due to the heterogenous scale parameters of the Gamma distributed unnormalized weights. I present an alternative representation of this process that facilitates inference by casting the discrete time chain as a continuous time Markov Process in which: (1) some jump attempts fail, (2) the probability of success is proportional to the similarity between the source and destination states, (3) only successful jumps are observed, and (4) the time elapsed between jumps, as well as the number of unsuccessful jump attempts, are latent variables that are sampled during MCMC inference. By introducing these auxiliary latent variables, all conditional distributions in the model are members of an exponential family, admitting exact Gibbs sampling, with the exception of the parameters of the similarity kernel. The choice of similarity kernel is application-specific, but I present results for an exponential (a.k.a. Laplacian) similarity kernel with a single decay parameter whose conditional posterior density is log-concave, and hence admits Adaptive Rejection Sampling (Gilks and Wild, 1992).

The motivating domain for this model is natural language text, in which sentences in a document are arranged in such a way that the sets of relevant entities in successive sentences have a high degree of overlap, even when they are not identical. The goal is to model the entity set in a sentence using a binary vector, indicating which entities are present in the context, and to constrain the dynamics governing latent state transitions so that transitions between similar entity sets are *a priori* more likely, but where the presence or absence of an entity depends on the state of multiple entities in the previous sentence. The latter property makes an ordinary factorial HMM undesireable.

1.1 Transition Dynamics in the HDP-HMM

The conventional HDP-HMM (Teh et al., 2006) is based on a Hierarchical Dirichlet Process defined as follows:

Each of a countably infinite set of states, indexed by j, receives a location θ_j in emission parameter space, Ω, according to base measure H. A top-level weight distribution, $\boldsymbol{\beta}$, is drawn from a stick-breaking process with parameter $\gamma > 0$, so that state j has overall weight β_j, and emission distribution parameterized by θ_j.

$$\theta_j \overset{i.i.d.}{\sim} H \tag{1.1}$$

$$\boldsymbol{\beta} \sim GEM(\gamma) \tag{1.2}$$

The actual transition distribution from state j, denoted by $\boldsymbol{\pi}_j$ is then drawn from a DP with concentration α and base measure $\boldsymbol{\beta}$:

$$\boldsymbol{\pi}_j \overset{i.i.d}{\sim} DP(\alpha\boldsymbol{\beta}) \qquad j = 1, 2, \ldots \tag{1.3}$$

The hidden state sequence is then generated according to the π_j. Let z_t be the index of the chain's state at time t. Then we have

$$z_t \mid z_{t-1}, \boldsymbol{\pi}_{z_{t-1}} \sim \boldsymbol{\pi}_{z_{t-1}} \qquad t = 1, 2, \ldots, T \tag{1.4}$$

where T is the length of the data sequence.

Finally, the emission distribution for state j is a function of θ_j, so that we have

$$y_t \,|\, z_t, \theta_{z_t} \sim F(\theta_{z_t}) \tag{1.5}$$

A shortcoming of this model is that the generative process does not take into account the fact that the set of source states is the same as the set of destination states: that is, the distribution $\boldsymbol{\pi}_j$ has an element which corresponds to state j. Put another way, there is no special treatment of the diagonal of the transition matrix, so that self-transitions are no more likely *a priori* than transitions to any other state. The Sticky HDP-HMM of Fox, et al. (2008) addresses this issue by adding an extra mass of κ at location j to the base measure of the DP that generates $\boldsymbol{\pi}_j$. That is, they replace (1.3) with

$$\boldsymbol{\pi}_j \sim DP(\alpha\boldsymbol{\beta} + \kappa\delta_j). \tag{1.6}$$

An alternative model is presented by Johnson et al. (2013), wherein state duration distributions are modeled separately, and ordinary self-transitions are ruled out. In both of these models, auxiliary latent variables are introduced to simplify conditional posterior distributions and facilitate Gibbs sampling. However, while both of these models have the useful property that self-transitions are treated as "special", they contain no notion of similarity for pairs of states that are not identical: in both cases, when the transition matrix is integrated out, the prior probability of transitioning to state j' depends only on the top-level stick weight associated with state j', and not on the identity or parameters of the previous state j.

1.2 An HDP-HMM With Local Transitions

The goal is to add to the transition model the concept of a transition to a "nearby" state, where nearness of j and j' is possibly a function of θ_j and $\theta_{j'}$. In order to accomplish this, we first consider an alternative construction of the transition distributions, based on the Normalized Gamma Process representation of the Dirichlet Process (Ferguson, 1973).

1.2.1 A Normalized Gamma Process representation of the HDP-HMM

Define a random measure, $\mu = \sum_{j=1}^{\infty} \pi_j \delta_{\theta_j}$, where

$$\pi_j \overset{ind}{\sim} \mathcal{G}(w_j, 1) \tag{1.7}$$

$$T = \sum_{j=1}^{\infty} \pi_j \tag{1.8}$$

$$\tilde{\pi}_j = \frac{\pi_j}{T} \tag{1.9}$$

$$\theta_j \overset{i.i.d}{\sim} H \tag{1.10}$$

and subject to the constraint that $\sum_{j\geq 1} w_j < \infty$, which ensures that $T < \infty$ almost surely. As shown by Paisley et al. (2011), for fixed $\{w_j\}$ and $\{\theta_j\}$, μ is distributed as a Dirichlet Process with base measure $\mathbf{w} = \sum_{j=1}^{\infty} w_j \delta_{\theta_j}$. If we draw $\boldsymbol{\beta}$ from a stick-breaking process and then draw a series $\{\mu_m\}_{m=1}^{M}$ of i.i.d. random measures from the above process, setting $\mathbf{w} = \alpha\boldsymbol{\beta}$ for some $\alpha > 0$, then this defines a Hierarchical Dirichlet Process. If, moreover, there is one μ_m associated with every state j, then we obtain the HDP-HMM.

We can thus write

$$\boldsymbol{\beta} \sim \mathsf{GEM}(\gamma) \tag{1.11}$$

$$\theta_j \overset{i.i.d.}{\sim} H \tag{1.12}$$

$$\pi_{jj'} \overset{ind}{\sim} \mathcal{G}(\alpha\beta_{j'}, 1) \tag{1.13}$$

$$T_j = \sum_{j'=1}^{\infty} \pi_{jj'} \tag{1.14}$$

$$\tilde{\pi}_{jj'} = \frac{\pi_{jj'}}{T_j}, \tag{1.15}$$

where γ and α are prior concentration hyperparameters for the two DP levels, where

$$p(z_t \mid z_{t-1}, \boldsymbol{\pi}) = \tilde{\pi}_{z_{t-1}z_t} \tag{1.16}$$

and the observed data $\{y_t\}_{t\geq 1}$ distributed as

$$y_t \mid z_t \overset{ind}{\sim} F(\theta_{z_t}) \tag{1.17}$$

for some family, F of probability measures indexed by values of θ.

1.2.2 Promoting "Local" Transitions

In the preceding formulation, the θ_j and the $\pi_{jj'}$ are independent conditioned on the top-level measure. Our goal is to relax this assumption, in order to allow for prior knowledge that certain "locations", θ_j, are more likely than others to produce large weights. This can be accomplished by letting the rate parameter in the distribution of the $\pi_{jj'}$ be a function of θ_j and $\theta_{j'}$. Let $\Phi : \Omega \times \Omega \to [0, \infty)$ represent a "similarity function", and define a collection of random variables $\{\phi_{jj'}\}_{j,j' \geq 1}$ according to

$$\phi_{jj'} = \phi(\theta_j, \theta_j') \tag{1.18}$$

We can then generalize (1.11)-(1.15) to

$$\boldsymbol{\beta} \sim \mathrm{GEM}(\gamma) \tag{1.19}$$

$$\theta_j \overset{i.i.d}{\sim} H \tag{1.20}$$

$$\pi_{jj'} \,|\, \boldsymbol{\beta}, \boldsymbol{\theta} \sim \mathcal{G}(\alpha\beta_{j'}, \phi_{jj'}^{-1}) \tag{1.21}$$

$$T_j = \sum_{j'=1}^{\infty} \pi_{jj'} \tag{1.22}$$

$$\tilde{\pi}_{jj'} = \frac{\pi_{jj'}}{T_j} \tag{1.23}$$

so that the expected value of $\pi_{jj'}$ is $\alpha\beta_{j'}\phi_{jj'}$. Since a similarity between one object and another should not exceed the similarity between an object and itself, we will assume that $\phi_{jj'} \leq B < \infty$ for all j and j', with equality holding iff $j = j'$. Moreover, there is no loss of generality by taking $B = 1$, since a constant rescaling of $\phi_{jj'}$ gets absorbed in the normalization.

The above model is equivalent to simply drawing the $\pi_{jj'}$ as in (1.11) and scaling each one by $\phi_{jj'}$ prior to normalization.

Unfortunately, this formulation complicates inference significantly, as the introduction of non-constant rate parameters to the prior on $\boldsymbol{\pi}$ destroys the conjugacy between $\boldsymbol{\pi}$ and $\mathbf{z}$, and worse, the conditional likelihood function for $\boldsymbol{\pi}$ contains an infinite sum of the elements in a row, rendering all entries within a row mutually dependent.

1.2.3 The HDP-HMM-LT as a continuous-time Markov Jump Process with "failed" jumps

We can gain stronger intuition, as well as simplify posterior inference, by re-casting the HDP-HMM-LT described in the last section as a continuous time Markov Jump Process where some of the attempts to jump from one state to another fail, and where the failure probability increases as a function of the "distance" between the states.

Let Φ be defined as in the last section, and let $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$ be defined as in the Normalized Gamma Process representation of the ordinary HDP-HMM. That is,

$$\boldsymbol{\beta} \sim \text{GEM}(\gamma) \tag{1.24}$$

$$\theta_j \overset{i.i.d}{\sim} H \tag{1.25}$$

$$\pi_{jj'} \,|\, \boldsymbol{\beta}, \boldsymbol{\theta} \sim \mathcal{G}(\alpha\beta_{j'}, 1) \tag{1.26}$$

Now suppose that when the process is in state j, jumps to state j' are made at rate $\pi_{jj'}$. This defines a continuous-time Markov Process where the off-diagonal elements of the transition rate matrix are the off diagonal elements of $\boldsymbol{\pi}$. In addition, self-jumps are allowed, and occur with rate π_{jj}. If we only observe the jumps and not the durations between jumps, this is an ordinary Markov chain, whose transition matrix is obtained by appropriately normalizing $\boldsymbol{\pi}$. If we do not observe the jumps themselves, but instead an observation is generated once per jump from a distribution that depends on the state being jumped to, then we have an ordinary HMM.

I modify this process as follows. Suppose that each jump attempt from state j to state j' has a chance of failing, which is an increasing function of the "distance" between the states. In particular, let the success probability be $\phi_{jj'}$ (recall that we assumed above that $0 \leq \phi_{jj'} \leq 1$ for all j, j'). Then, the rate of successful jumps from j to j' is $\pi_{jj'}\phi_{jj'}$, and the corresponding rate of unsuccessful jump attempts is $\pi_{jj'}(1 - \phi_{jj'})$. To see this, denote by $N_{jj'}$ the total number of jump attempts to j' in a unit interval of time spent in state j. Since we are assuming the process is Markovian, the total number of attempts is $\mathcal{P}\text{ois}(\pi_{jj'})$ distributed. Conditioned on $N_{jj'}$, $n_{jj'}$ will be successful, where

$$n_{jj'} \,|\, N_{jj'} \sim \mathcal{B}\text{inom}(N_{jj'}, \phi_{jj'}) \tag{1.27}$$

It is easy to show (and well known) that the marginal distribution of $n_{jj'}$ is $\mathcal{P}\text{ois}(\pi_{jj'}\phi_{jj'})$, and the marginal distribution of $\tilde{q}_{jj'} := N_{jj'} - n_{jj'}$ is $\mathcal{P}\text{ois}(\pi_{jj'}(1 - \phi_{jj'}))$. The rate of successful jumps from state j overall is then $T_j := \sum_{j'} \pi_{jj'}\phi_{jj'}$.

Let t index jumps, so that z_t indicates the tth state visited by the process (couting self-jumps as a new time step). Given that the process is in state j at discretized time $t-1$ (that is, $z_{t-1} = j$), it is a standard property of Markov Processes that the probability that the first successful jump is to state j' (that is, $z_t = j'$) is proportional to the rate of successful attempts to j', which is $\pi_{jj'}\phi_{jj'}$.

Let τ_t indicate the time elapsed between the tth and and $t-1$th successful jump (where we assume that the first observation occurs when the first successful jump from a distinguished initial state is made). We have

$$\tau_t \,|\, z_{t-1} \sim \mathcal{E}\mathrm{xp}(T_{z_{t-1}}) \tag{1.28}$$

where τ_t is independent of z_t.

During this period, there will be $\tilde{q}_{j't}$ unsuccessful attempts to jump to state j', where

$$\tilde{q}_{j't} \,|\, z_{t-1} \sim \mathcal{P}\mathrm{ois}(\tau_t \pi_{z_{t-1}j'}(1 - \phi_{z_{t-1}j'})) \tag{1.29}$$

Define the following additional variables

$$\mathcal{T}_j = \{t \,|\, z_{t-1} = j\} \tag{1.30}$$

$$q_{jj'} = \sum_{t \in \mathcal{T}_j} \tilde{q}_{j't} \tag{1.31}$$

$$u_j = \sum_{t \in \mathcal{T}_j} \tau_t \tag{1.32}$$

and let $\mathbf{Q} = (q_{jj'})_{j,j' \geq 1}$ be the matrix of unsuccessful jump attempt counts, and $\mathbf{u} = (u_j)_{j \geq 1}$ be the vector of the total times spent in each state.

Since each of the τ_t with $t \in \mathcal{T}_j$ are i.i.d. $\mathcal{E}\mathrm{xp}(T_j)$, we get the marginal distribution

$$u_j \,|\, \mathbf{z}, \boldsymbol{\pi\theta} \overset{ind}{\sim} \mathcal{G}(n_{j\cdot}, T_j) \tag{1.33}$$

by the standard property that sums of i.i.d. Exponential distributions has a Gamma distribution with shape equal to the number of variates in the sum, and rate equal to the rate of the individual exponentials. Moreover, since the $\tilde{q}_{j't}$ with $t \in \mathcal{T}_j$ are Poisson distributed, the total number of failed attempts in the total duration u_j is

$$q_{jj'} \overset{ind}{\sim} \mathcal{P}\mathrm{ois}(u_j \pi_{jj'}(1 - \phi_{jj'})). \tag{1.34}$$

Thus if we marginalize out the individual τ_t and $\tilde{q}_{j't}$, we have a joint distribution over $\mathbf{z}$, $\mathbf{u}$, and $\mathbf{Q}$, conditioned on the transition rate matrix $\boldsymbol{\pi}$ and the success probability matrix $\boldsymbol{\phi}$, which is

$$\begin{aligned}
p(\mathbf{z}, \mathbf{u}, \mathbf{Q} \mid \boldsymbol{\pi}, \boldsymbol{\theta}) &= \left(\prod_{t=1}^{T} p(z_t \mid z_{t-1})\right) \prod_j p(u_j \mid \mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\theta}) \prod_{j'} p(q_{jj'} \mid u_j \pi_{jj'}, \phi_{jj'}) && (1.35)\\
&= \left(\prod_t \frac{\pi_{z_{t-1}z_t} \phi_{z_{t-1}z_t}}{T_{z_{t-1}}}\right) \prod_j \frac{T_j^{n_{j\cdot}}}{\Gamma(n_{j\cdot})} u_j^{n_{j\cdot}-1} e^{-T_j u_j} && (1.36)\\
&\qquad \times \prod_{j'} e^{-u_j \pi_{jj'}(1-\phi_{jj'})} u_j^{q_{jj'}} \pi_{jj'}^{q_{jj'}} (1-\phi_{jj'})^{q_{jj'}} (q_{jj'}!)^{-1} && (1.37)\\
&= \prod_j \Gamma(n_{j\cdot})^{-1} u_j^{n_{j\cdot}+q_{j\cdot}-1} && (1.38)\\
&\qquad \times \prod_{j'} \pi_{jj'}^{n_{jj'}+q_{jj'}} \phi_{jj'}^{n_{jj'}} (1-\phi_{jj'})^{q_{jj'}} e^{-\pi_{jj'}\phi_{jj'}u_j} e^{-\pi_{jj'}(1-\phi_{jj'})u_j} (q_{jj'}!)^{-1} && (1.39)\\
&= \prod_j \Gamma(n_{j\cdot})^{-1} u_j^{n_{j\cdot}+q_{j\cdot}-1} \prod_{j'} \pi_{jj'}^{n_{jj'}+q_{jj'}} \phi_{jj'}^{n_{jj'}} (1-\phi_{jj'})^{q_{jj'}} e^{-\pi_{jj'}u_j} (q_{jj'}!)^{-1} && (1.40)
\end{aligned}$$

1.2.4 An HDP-HSMM-LT modification

Note that it is trivial to modify the HDP-HMM-LT to allow the number of observations generated each time a state is visited to have a distribution which is not Geometric, by simply fixing the diagonal elements of $\boldsymbol{\pi}$ to be zero, and allowing D_t observations to be emitted *i.i.d.* $F(\theta_{z_t})$ at jump t, where

$$D_t \mid \mathbf{z} \overset{ind}{\sim} g(\omega_{z_t}) \qquad \omega_j \overset{i.i.d}{\sim} G \tag{1.41}$$

The likelihood then includes the additional term for the D_t, and the only inference step which is affected is that instead of sampling $\mathbf{z}$ alone, we sample $\mathbf{z}$ and the D_t jointly, by defining

$$z_s^* = z_{\max\{T \mid s \leq \sum_{t=1}^{T} D_t\}} \tag{1.42}$$

where s ranges over the number of observations, and associating a $\mathbf{y}_s$ with each z_s^*. Inferences about $\boldsymbol{\phi}$ are not affected, since the diagonal elements are assumed to be 1 anyway.

This is the same construction used in the Hierarchical Dirichlet Process Hidden Semi-Markov Model (HDP-HSMM; Johnson and Willsky (2013)). Unlike in the standard representation of the HDP-HSMM, however, there is no need to introduce additional auxiliary variables as a result of this modification, due to the presence of the (continuous) durations, $\mathbf{u}$, which were already needed to account for the normalization of the $\boldsymbol{\pi}$.

1.2.5 Summary

I have defined the following augmented generative model for the HDP-H(S)MM-LT:

$$\boldsymbol{\beta} \sim \mathrm{GEM}(\gamma) \tag{1.43}$$

$$\theta_j \overset{i.i.d}{\sim} H \tag{1.44}$$

$$\pi_{jj'} \mid \boldsymbol{\beta}, \boldsymbol{\theta} \sim \mathcal{G}(\alpha\beta_{j'}, 1) \tag{1.45}$$

$$z_t \mid z_{t-1}, \boldsymbol{\pi}, \boldsymbol{\theta} \sim \sum_j \left(\frac{\pi_{z_{t-1}j}\phi_{z_{t-1}j}}{\sum_{j'} \pi_{z_{t-1}j'}\phi_{z_{t-1}j'}} \right) \delta_j \tag{1.46}$$

$$u_j \mid \mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\theta} \overset{ind}{\sim} \mathcal{G}(n_{j\cdot}, \sum_{j'} \pi_{jj'}\phi_{jj'}) \tag{1.47}$$

$$q_{jj'} \mid \mathbf{u}, \boldsymbol{\pi}, \boldsymbol{\theta} \overset{ind}{\sim} \mathcal{P}\mathrm{ois}(u_j(1-\phi_{jj'})\pi_{jj'}) \tag{1.48}$$

$$\mathbf{y}_t \mid \mathbf{z}, \boldsymbol{\theta} \sim F(\theta_{z_t}) \tag{1.49}$$

If we are using the HSMM variant, then we simply fix π_{jj} to 0 for each j, draw

$$\omega_j \overset{i.i.d}{\sim} G \tag{1.50}$$

$$D_t \mid \mathbf{z} \overset{ind}{\sim} g(\omega_{z_t}), \tag{1.51}$$

for chosen G and g, set

$$z^*_s = z_{\max\{T \mid s \leq \sum_{t=1}^T D_t\}} \tag{1.52}$$

and replace (1.49) with

$$\mathbf{y}_s \mid \mathbf{z}, \boldsymbol{\theta} \sim F(\theta_{z^*_s}) \tag{1.53}$$

1.3 Inference

I develop a Gibbs sampling algorithm based on the Markov Process with Failed Jumps representation, augmenting the data with the duration variables $\mathbf{u}$, the failed jump attempt count matrix, $\mathbf{Q}$, as well as additional auxiliary variables which we will define below. In this representation the transition matrix is not modeled directly, but is a function of the unscaled transition matrix π and the similarity matrix $\boldsymbol{\phi}$. The full set of variables is partitioned into three blocks: $\{\gamma, \alpha, \boldsymbol{\beta}, \boldsymbol{\pi}\}$, $\{\mathbf{z}, \mathbf{u}, \mathbf{Q}, \Lambda\}$, and $\{\boldsymbol{\theta}\}$, where Λ represents a set of auxiliary variables that will be introduced below. The variables in each block are sampled jointly conditioned on the other two blocks.

Since we are representing the transition matrix of the Markov chain explicitly, we approximate the stick-breaking process that produces $\boldsymbol{\beta}$ using a finite Dirichlet distribution with a number of components larger than we expect to need, forcing the remaining components to have zero weight. Let J indicate the maximum number of states. Then, we approximate (1.24) with

$$\boldsymbol{\beta} \mid \gamma \sim \text{Dirichlet}(\gamma/J, \dots, \gamma/J) \tag{1.54}$$

This distribution converges weakly to the Stick-Breaking Process as $J \to \infty$. In practice, J is large enough when the vast majority of the probability mass in $\boldsymbol{\beta}$ is allocated to a strict subset of components, or when the latent state sequence $\mathbf{z}$ never uses all J available states, indicating that the data is well described by a number of states less than J.

1.3.1 Sampling $\boldsymbol{\pi}$, $\boldsymbol{\beta}$, α and γ

The joint conditional over γ, α, $\boldsymbol{\beta}$ and $\boldsymbol{\pi}$ given $\mathbf{z}$, $\mathbf{u}$, $\mathbf{Q}$, Λ and $\boldsymbol{\theta}$ will factor as

$$p(\gamma, \alpha, \boldsymbol{\beta}, \boldsymbol{\pi} \mid \mathbf{z}, \mathbf{u}, \mathbf{Q}, \Lambda, \boldsymbol{\theta}) = p(\gamma \mid \Lambda) p(\alpha \mid \Lambda) p(\boldsymbol{\beta} \mid \gamma, \Lambda) p(\boldsymbol{\pi} \mid \alpha, \boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{z}) \tag{1.55}$$

I will derive these four factors in reverse order.

Sampling $\boldsymbol{\pi}$

The entries in $\boldsymbol{\pi}$ are conditionally independent given α and $\boldsymbol{\beta}$, so we have the prior

$$p(\boldsymbol{\pi} \mid \boldsymbol{\beta}, \alpha) = \prod_j \prod_{j'} \Gamma(\alpha\beta_{j'})^{-1} \pi_{jj'}^{\alpha\beta_{j'}-1} \exp(-\pi_{jj'}), \tag{1.56}$$

and the likelihood given augmented data $\{\mathbf{z}, \mathbf{u}, \mathbf{Q}\}$ given by (1.40). Combining these, we have

$$p(\boldsymbol{\pi}, \mathbf{z}, \mathbf{u}, \mathbf{Q} \mid \boldsymbol{\beta}, \alpha, \boldsymbol{\theta}) = \prod_j u_j^{n_{j\cdot}+q_{j\cdot}-1} \prod_{j'} \Gamma(\alpha\beta_{j'})^{-1} \pi_{jj'}^{\alpha\beta_{j'}+n_{jj'}+q_{jj'}-1} e^{-(1+u_j)\pi_{jj'}} \phi_{jj'}^{n_{jj'}} (1-\phi_{jj'})^{q_{jj'}} (q_{jj'}!)^{-1} \tag{1.57}$$

Conditioning on everything except $\boldsymbol{\pi}$, we get

$$p(\boldsymbol{\pi} \mid \mathbf{Q}, \mathbf{u}, \mathbf{Z}, \boldsymbol{\beta}, \alpha, \boldsymbol{\theta}) \propto \prod_{j} \prod_{j'} \pi_{jj'}^{\alpha\beta_{j'}+n_{jj'}+q_{jj'}-1} \exp(-(1+u_j)\pi_{jj'}) \tag{1.58}$$

and thus we see that the $\pi_{jj'}$ are conditionally independent given $\mathbf{u}$, $\mathbf{Z}$ and $\mathbf{Q}$, and distributed according to

$$\pi_{jj'} \mid n_{jj'}, q_{jj'}, \beta_{j'}, \alpha \overset{ind}{\sim} \mathcal{G}(\alpha\beta_{j'} + n_{jj'} + q_{jj'}, 1 + u_j) \tag{1.59}$$

Sampling $\boldsymbol{\beta}$

Consider the conditional distribution of $\boldsymbol{\beta}$ having integrated out $\boldsymbol{\pi}$. The prior density of $\boldsymbol{\beta}$ from (1.54) is

$$p(\boldsymbol{\beta} \mid \gamma) = \frac{\Gamma(\gamma)}{\Gamma(\frac{\gamma}{J})^J} \prod_{j} \beta_j^{\frac{\gamma}{J}-1} \tag{1.60}$$

After integrating out $\boldsymbol{\pi}$ in (1.57), we have

$$\begin{aligned} p(\mathbf{z}, \mathbf{u}, \mathbf{Q} \mid \boldsymbol{\beta}, \alpha, \gamma, \boldsymbol{\theta}) = \prod_{j=1}^{J} u_j^{-1} \prod_{j'=1}^{J} & u^{n_{jj'}+q_{jj'}-1}(1+u_j)^{-(\alpha\beta_{j'}+n_{jj'}+q_{jj'})} && (1.61) \\ & \times \frac{\Gamma(\alpha\beta_{j'} + n_{jj'} + q_{jj'})}{\Gamma(\alpha\beta_{j'})} \phi_{jj'}^{n_{jj'}} (1-\phi_{jj'})^{q_{jj'}} (q_{jj'}!)^{-1} && (1.62) \\ = \prod_{j=1}^{J} \Gamma(n_{j\cdot})^{-1} u_j^{-1} (1+u_j)^{-\alpha} & \left(\frac{u_j}{1+u_j}\right)^{n_{j\cdot}+q_{j\cdot}} && (1.63) \\ & \times \prod_{j'=1}^{J} \frac{\Gamma(\alpha\beta_{j'} + n_{jj'} + q_{jj'})}{\Gamma(\alpha\beta_{j'})} \phi_{jj'}^{n_{jj'}} (1-\phi_{jj'})^{q_{jj'}} (q_{jj'}!)^{-1} && (1.64) \end{aligned}$$

where we have used the fact that the β_j sum to 1. Therefore

$$p(\boldsymbol{\beta} \mid \mathbf{z}, \mathbf{u}, \mathbf{Q}, \alpha, \gamma, \boldsymbol{\theta}) \propto \prod_{j=1}^{J} \beta_j^{\frac{\gamma}{J}-1} \prod_{j'=1}^{J} \frac{\Gamma(\alpha\beta_{j'} + n_{jj'} + q_{jj'})}{\Gamma(\alpha\beta_{j'})}. \tag{1.65}$$

Following (Teh et al., 2006), we can write the ratios of Gamma functions as polynomials in β_j, as

$$p(\boldsymbol{\beta} \,|\, \mathbf{z}, \mathbf{u}, \mathbf{Q}, \alpha, \gamma, \boldsymbol{\theta}) \propto \prod_{j=1}^{J} \beta_j^{\frac{\gamma}{J}-1} \prod_{j'=1}^{J} \sum_{m_{jj'}=1}^{n_{jj'}} s(n_{jj'} + q_{jj'}, m_{jj'})(\alpha\beta_{j'})^{m_{jj'}} \tag{1.66}$$

where $s(m, n)$ is an unsigned Stirling number of the first kind. This admits an augmented data representation, where we introduce a random matrix $\mathbf{M} = (m_{jj'})_{1 \leq j,j' \leq J}$, whose entries are conditionally independent given $\boldsymbol{\beta}$, $\mathbf{Q}$ and $\mathbf{z}$, with

$$p(m_{jj'} = m \,|\, \beta_{j'}, \alpha, n_{jj'}, q_{jj'}) = \frac{s(n_{jj'} + q_{jj'}, m)\alpha^m \beta_{j'}^m}{\sum_{m'=0}^{n_{jj'}+q_{jj'}} s(n_{jj'} + q_{jj'}, m')\alpha^{m'} \beta_{j'}^{m'}} \tag{1.67}$$

for integer m ranging between 0 and $n_{jj'} + q_{jj'}$. Note that $s(n, 0) = 0$ if $n > 0$, $s(0, 0) = 1$ and $s(0, m) = 0$ if $m > 0$. Then, we have joint distribution

$$p(\boldsymbol{\beta}, \mathbf{M} \,|\, \mathbf{z}, \mathbf{u}, \mathbf{Q}, \alpha, \gamma, \boldsymbol{\theta}) \propto \prod_{j=1}^{J} \beta_j^{\frac{\gamma}{J}-1} \prod_{j'=1}^{J} s(n_{jj'} + q_{jj'}, m_{jj'})\alpha^{m_{jj'}} \beta_{j'}^{m_{jj'}} \tag{1.68}$$

which yields (1.66) when marginalized over $\mathbf{M}$. Again discarding constants in $\boldsymbol{\beta}$ and regrouping yields

$$p(\boldsymbol{\beta} \,|\, \mathbf{M}, \mathbf{Z}, \mathbf{u}, \boldsymbol{\theta}, \alpha, \gamma) \propto \prod_{j=1}^{J} \beta_j^{\frac{\gamma}{J}+m_{\cdot j}-1} \tag{1.69}$$

which is Dirichlet:

$$\boldsymbol{\beta} \,|\, \mathbf{M}, \gamma \sim \text{Dirichlet}(\frac{\gamma}{J} + m_{\cdot 1}, \ldots, \frac{\gamma}{J} + m_{\cdot J}) \tag{1.70}$$

Sampling α and γ

Assume that α and γ have Gamma priors, with

$$p(\alpha) = \frac{b_\alpha^{a_\alpha}}{\Gamma(a_\alpha)} \alpha^{a_\alpha - 1} \exp(-b_\alpha \alpha) \tag{1.71}$$

$$p(\gamma) = \frac{b_\gamma^{a_\gamma}}{\Gamma(a_\gamma)} \gamma^{a_\gamma - 1} \exp(-b_\gamma \gamma) \tag{1.72}$$

Having integrated out $\boldsymbol{\pi}$, we have

$$p(\boldsymbol{\beta}, \mathbf{z}, \mathbf{u}, \mathbf{Q}, \mathbf{M} \,|\, \alpha, \gamma, \boldsymbol{\theta}) = \frac{\Gamma(\gamma)}{\Gamma(\frac{\gamma}{J})^J} \alpha^{m_{\cdot\cdot}} \prod_{j=1}^{J} \beta_j^{\frac{\gamma}{J}+m_{\cdot j}-1} \Gamma(n_{j\cdot})^{-1} u_j^{-1} (1+u_j)^{-\alpha} \left(\frac{u_j}{1+u_j}\right)^{n_{j\cdot}+q_{j\cdot}} \tag{1.73}$$

$$\times \prod_{j'=1}^{J} s(n_{jj'}+q_{jj'}, m_{jj'}) \phi_{jj'}^{n_{jj'}} (1-\phi_{jj'})^{q_{jj'}} (q_{jj'}!)^{-1} \tag{1.74}$$

We can also integrate out $\boldsymbol{\beta}$, to yield

$$p(\mathbf{z}, \mathbf{u}, \mathbf{Q}, \mathbf{M} \,|\, \alpha, \gamma, \boldsymbol{\theta}) = \alpha^{m_{\cdot\cdot}} e^{-\sum_{j''} \log(1+u_{j''})\alpha} \frac{\Gamma(\gamma)}{\Gamma(\gamma+m_{\cdot\cdot})} \tag{1.75}$$

$$\times \prod_{j} \frac{\Gamma(\frac{\gamma}{J}+m_{\cdot j})}{\Gamma(\frac{\gamma}{J})\Gamma(n_{j\cdot})} u_j^{-1} \left(\frac{u_j}{1+u_j}\right)^{n_{j\cdot}+q_{j\cdot}} \tag{1.76}$$

$$\times \prod_{j'=1}^{J} s(n_{jj'}+q_{jj'}, m_{jj'}) \phi_{jj'}^{n_{jj'}} (1-\phi_{jj'})^{q_{jj'}} (q_{jj'}!)^{-1} \tag{1.77}$$

demonstrating that α and γ are independent given $\boldsymbol{\theta}$ and the augmented data, with

$$p(\alpha \,|\, \mathbf{z}, \mathbf{u}, \mathbf{Q}, \mathbf{M}, \boldsymbol{\theta}) \propto \alpha^{a_\alpha + m_{\cdot\cdot}} \exp(-(b_\alpha + \sum_j \log(1+u_j))\alpha) \tag{1.78}$$

and

$$p(\gamma \,|\, \mathbf{z}, \mathbf{u}, \mathbf{Q}, \mathbf{M}, \boldsymbol{\theta}) \propto \gamma^{a_\gamma - 1} \exp(-b_\gamma \gamma) \frac{\Gamma(\gamma) \prod_{j=1}^{J} \Gamma(\frac{\gamma}{J}+m_{\cdot j})}{\Gamma(\frac{\gamma}{J})^J \Gamma(\gamma+m_{\cdot\cdot})} \tag{1.79}$$

So we see that

$$\alpha \,|\, \mathbf{z}, \mathbf{u}, \mathbf{Q}, \mathbf{M}, \boldsymbol{\theta} \sim \mathcal{G}(a_\alpha + m_{\cdot\cdot}, b_\alpha + \sum_j \log(1+u_j)) \tag{1.80}$$

To sample γ, we introduce a new set of auxiliary variables, $\mathbf{r} = (r_1, \dots, r_j)$ and t with the following distributions:

$$p(r_j = r \,|\, m_{\cdot j}, \gamma) = \frac{\Gamma(\frac{\gamma}{J})}{\Gamma(\frac{\gamma}{J}+m_{\cdot j})} s(m_{\cdot j}, r) \left(\frac{\gamma}{J}\right)^r \qquad r = 1, \dots, m_{\cdot j} \tag{1.81}$$

$$p(t \,|\, m_{\cdot\cdot}\gamma) = \frac{\Gamma(\gamma+m_{\cdot\cdot})}{\Gamma(\gamma)\Gamma(m_{\cdot\cdot})} t^{\gamma-1}(1-t)^{m_{\cdot\cdot}-1} \qquad t \in (0,1) \tag{1.82}$$

so that

$$p(\gamma, \mathbf{r}, t \,|\, \mathbf{M}) \propto \gamma^{a_\gamma - 1} \exp(-b_\gamma \gamma) t^{\gamma - 1} (1-t)^{m_{\cdot\cdot} + q_{\cdot} - 1} \prod_{j=1}^{J} s(m_{\cdot j} + q_j, r_j) \left(\frac{\gamma}{J}\right)^{r_j} \tag{1.83}$$

and

$$p(\gamma \,|\, \mathbf{r}, t) \propto \gamma^{a_\gamma + r_{\cdot} - 1} \exp(-(b_\gamma - \log(t))\gamma), \tag{1.84}$$

which is to say

$$\gamma \,|\, \mathbf{r}, t, \mathbf{z}, \mathbf{u}, \mathbf{Q}, \mathbf{M}, \boldsymbol{\theta} \sim \mathcal{G}(a_\gamma + r_{\cdot}, b_\gamma - \log(t)) \tag{1.85}$$

Summary

I have made the following additional assumptions about the generative model in this section:

$$\gamma \sim \mathcal{G}(a_\gamma, b_\gamma) \qquad \alpha \sim \mathcal{G}(a_\alpha, b_\alpha) \tag{1.86}$$

The joint conditional over γ, α, $\boldsymbol{\beta}$ and $\boldsymbol{\pi}$ given $\mathbf{z}$, $\mathbf{u}$, $\mathbf{Q}$, $\mathbf{M}$, $\mathbf{r}$, t and $\boldsymbol{\theta}$ factors as

$$p(\gamma, \alpha, \boldsymbol{\beta}, \boldsymbol{\pi} \,|\, \mathbf{z}, \mathbf{u}, \mathbf{Q}, \mathbf{r}, t, \boldsymbol{\theta}) = p(\gamma \,|\, \mathbf{r}, t) p(\alpha \,|\, \mathbf{u}, \mathbf{M}) p(\boldsymbol{\beta} \,|\, \gamma, \mathbf{M}) p(\boldsymbol{\pi} \,|\, \alpha, \boldsymbol{\beta}, \mathbf{z}, \mathbf{u}, \mathbf{Q}) \tag{1.87}$$

where

$$\gamma \,|\, \mathbf{r}, t \sim \mathcal{G}(a_\gamma + r_{\cdot}, b_\gamma - \log(t)) \tag{1.88}$$

$$\alpha \,|\, \mathbf{u}, \mathbf{M} \sim \mathcal{G}(a_\alpha + m_{\cdot\cdot}, b_\alpha + \sum_j \log(1 + u_j)) \tag{1.89}$$

$$\boldsymbol{\beta} \,|\, \gamma, \mathbf{M} \sim \text{Dirichlet}(\frac{\gamma}{J} + m_{\cdot 1}, \ldots, \frac{\gamma}{J} + m_{\cdot J}) \tag{1.90}$$

$$\pi_{jj'} \,|\, \alpha, \beta_{j'}, \mathbf{z}, \mathbf{u}, \mathbf{Q} \overset{ind}{\sim} \mathcal{G}(\alpha\beta_{j'} + n_{jj'} + q_{jj'}, 1 + u_j) \tag{1.91}$$

1.3.2 Sampling z and the auxiliary variables

The hidden state sequence, $\mathbf{z}$, is sampled jointly with the auxiliary variables, which consist of $\mathbf{u}$, $\mathbf{M}$, $\mathbf{Q}$, $\mathbf{r}$ and t. The joint conditional distribution of these variables is defined directly by the generative model:

$$p(\mathbf{z}, \mathbf{u}, \mathbf{Q}, \mathbf{M}, \mathbf{r}, t \mid \boldsymbol{\pi}, \boldsymbol{\beta}, \alpha, \gamma, \boldsymbol{\theta}) = p(\mathbf{z} \mid \boldsymbol{\pi}, \boldsymbol{\theta}) p(\mathbf{u} \mid \mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\theta}) p(\mathbf{Q} \mid \mathbf{u}, \boldsymbol{\pi}, \boldsymbol{\theta}) p(\mathbf{M} \mid \mathbf{z}, \mathbf{Q}, \alpha, \boldsymbol{\beta}) \tag{1.92}$$

$$\times\, p(\mathbf{r} \mid \gamma, \mathbf{M}) p(t \mid \gamma, \mathbf{M}) \tag{1.93}$$

Since we are representing the transition matrix explicitly, we can sample the entire sequence $\mathbf{z}$ at once with the forward-backward algorithm, as in an ordinary HMM (or, if we are employing the HSMM variant described in Sec. 1.2.4, then we can use the modified message passing scheme for HSMMs described by Johnson and Willsky (2013). Having done this, we can sample $\mathbf{u}$, $\mathbf{Q}$, $\mathbf{M}$, $\mathbf{r}$ and t from their forward distributions. To summarize, we have

$$u_j \mid \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\theta} \overset{ind}{\sim} \mathcal{G}(n_{j\cdot}, \sum_{j'} \pi_{jj'} \phi_{jj'}) \tag{1.94}$$

$$q_{jj'} \mid u_j, \pi_{jj'}, \phi_{jj'} \overset{ind}{\sim} \mathcal{P}\mathrm{ois}(u_j(1 - \phi_{jj'})\pi_{jj'}) \tag{1.95}$$

$$m_{jj'} \mid n_{jj'}, q_{jj'}, \beta_{j'}, \alpha \overset{ind}{\sim} \frac{\Gamma(\alpha\beta_j)}{\Gamma(\alpha\beta_j + n_{jj'} + q_{jj'})} \sum_{m=1}^{n_{jj'}+q_{jj'}} s(n_{jj'} + q_{jj'}, m) \alpha^m \beta_{j'}^m \delta_m \tag{1.96}$$

$$r_j \mid m_{\cdot j}, \gamma \overset{ind}{\sim} \frac{\Gamma(\frac{\gamma}{J})}{\Gamma(\frac{\gamma}{J} + m_{\cdot j})} \sum_{r=1}^{m_{j\cdot}} s(m_{\cdot j}, r) \left(\frac{\gamma}{J}\right)^r \delta_r \tag{1.97}$$

$$t \mid \gamma, \mathbf{M} \sim \mathcal{B}\mathrm{eta}(\gamma, m_{\cdot\cdot}) \tag{1.98}$$

1.3.3 Sampling state and emission parameters

The state parameters, $\boldsymbol{\theta}$, influence the transition matrix, $\boldsymbol{\pi}$ and the auxiliary vector q through the similarity matrix matrix $\boldsymbol{\phi}$, and also control the emission distributions. We have likelihood factors

$$p(\mathbf{z}, \mathbf{Q} \mid \boldsymbol{\theta}) \propto \prod_j \prod_{j'} \phi_{jj'}^{n_{jj'}} (1 - \phi_{jj'})^{q_{jj'}} \tag{1.99}$$

$$p(\mathbf{Y} \mid \mathbf{z}, \boldsymbol{\theta}) = \prod_{t=1}^{T} f(\mathbf{y}_t; \theta_{z_t}) \tag{1.100}$$

where proportionality is with respect to variation in $\boldsymbol{\theta}$.

The parameter space for the hidden states, the associated prior H on $\boldsymbol{\theta}$, and the similarity function Φ, is application-specific, but we consider here the case where a state, θ_j, consists of a finite-length binary vector, motivated by the application of inferring the set of relevant entities in each sentence of a text document.

Let $\theta_j = (\theta_{j1}, \dots, \theta_{jD})$, with $\theta_{jd} = 1$ indicating presence of feature d in context state j, and $\theta_{jd} = 0$ indicating absence. Of course, in this case, the set of possible states is finite, and so on its face it may seem that a nonparametric model is unnecessary. However, if D is reasonably large, it is likely that most of the 2^D possible states are vanishingly unlikely (and, in fact, the number of observations may well be less than 2^D), and so we would like a model that encourages the selection of a sparse set of states. Moreover, there may be more than one state with the same θ, but with different transition dynamics.

Sampling $\boldsymbol{\theta}$

In principle, H can be any distribution over binary vectors, but we will suppose for simplicity that it can be factored into D independent coordinate-wise Bernoulli variates. Let μ_d be the Bernoulli parameter for the dth coordinate.

We require a similarity function, $\Phi(\theta_j, \theta_{j'})$, which varies between 0 to 1, and is equal to 1 if and only if $\theta_j = \theta_{j'}$. A natural choice in this setting is the Laplacian kernel:

$$\phi_{jj'} = \Phi(\theta_j, \theta_{j'}) = \exp(-\lambda \Delta_{jj'}) \tag{1.101}$$

where $\Delta_{jj'd} = |\theta_{jd} - \theta_{j'd}|$, $\Delta_{jj'} = \sum_{d=1}^{D} \Delta_{jj'}$ is the Hamming distance between θ_j and $\theta_{j'}$, and $\lambda \geq 0$ (if $\lambda = 0$, the $\phi_{jj'}$ are identically 1, and so do not have any influence, reducing the model to an ordinary HDP-HMM).

Let

$$\phi_{jj'-d} = \exp(-\lambda(\Delta_{jj'} - \Delta_{jj'd})) \tag{1.102}$$

so that $\phi_{jj'} = \phi_{jj'-d} e^{-\lambda \Delta_{jj'd}}$.

Since the matrix $\boldsymbol{\phi}$ is assumed to be symmetric, we have

$$\frac{p(\mathbf{z}, \mathbf{Q} \mid \theta_{jd} = 1, \boldsymbol{\theta} \setminus \theta_{jd})}{p(\mathbf{z}, \mathbf{Q} \mid \theta_{jd} = 0, \boldsymbol{\theta} \setminus \theta_{jd})} \propto \prod_{j' \neq j} \frac{e^{-\lambda(n_{jj'} + n_{j'j}) \left|1 - \theta_{j'd}\right|} (1 - \phi_{jj'-d} e^{-\lambda \left|1 - \theta_{j'd}\right|})^{q_{jj'} + q_{j'j}}}{e^{-\lambda(n_{jj'} + n_{j'j}) \left|\theta_{j'd}\right|} (1 - \phi_{jj'-d} e^{-\lambda \left|\theta_{j'd}\right|})^{q_{jj'} + q_{j'j}}} \tag{1.103}$$

$$= e^{-\lambda(c_{jd0} - c_{jd1})} \prod_{j' \neq j} \left(\frac{1 - \phi_{jj'-d} e^{-\lambda}}{1 - \phi_{jj'-d}} \right)^{(-1)^{\theta_{j'd}} (q_{jj'} + q_{j'j})} \tag{1.104}$$

where c_{jd0} and c_{jd1} are the number of successful jumps to or from state j, to or from states with a 0 or 1, respectively, in position d. That is,

$$c_{jd0} = \sum_{\{j' \,|\, \theta_{j'd}=0\}} n_{jj'} + n_{j'j} \qquad c_{jd1} = \sum_{\{j' \,|\, \theta_{j'd}=1\}} n_{jj'} + n_{j'j} \tag{1.105}$$

Therefore, we can Gibbs sample θ_{jd} from its conditional posterior Bernoulli distribution given the rest of $\boldsymbol{\theta}$, where we compute the Bernoulli parameter via the log-odds

$$\log\left(\frac{p(\theta_{jd}=1 \,|\, \mathbf{Y}, \mathbf{z}, \mathbf{Q}, \boldsymbol{\theta} \setminus \theta_{jd})}{p(\theta_{jd}=0 \,|\, \mathbf{Y}, \mathbf{z}, \mathbf{Q}, \boldsymbol{\theta} \setminus \theta_{jd})}\right) = \log\left(\frac{p(\theta_{jd}=1)p(\mathbf{z},\mathbf{Q} \,|\, \theta_{jd}=1, \boldsymbol{\theta} \setminus \theta_{jd})p(\mathbf{Y} \,|\, \mathbf{z}, \theta_{jd}=1, \boldsymbol{\theta} \setminus \theta_{jd})}{p(\theta_{jd}=0)p(\mathbf{z},\mathbf{Q} \,|\, \theta_{jd}=0, \boldsymbol{\theta} \setminus \theta_{jd})p(\mathbf{Y} \,|\, \mathbf{z}, \theta_{jd}=0, \boldsymbol{\theta} \setminus \theta_{jd})}\right) \tag{1.106}$$

$$= \log\left(\frac{\mu_d}{1-\mu_d}\right) + (c_{jd1} - c_{jd0})\lambda + \sum_{j' \neq j} (-1)^{\theta_{j'd}} (q_{jj'} + q_{j'j}) \log\left(\frac{1 - \phi_{jj'}^{(-d)} e^{-\lambda}}{1 - \phi_{jj'}^{(-d)}}\right) \tag{1.107}$$

$$+ \sum_{\{t \,|\, z_t = j\}} \log\left(\frac{f(\mathbf{y}_t; \theta_{jd}=1, \theta_j \setminus \theta_{jd})}{f(\mathbf{y}_t; \theta_{jd}=0, \theta_j \setminus \theta_{jd})}\right) \tag{1.108}$$

Suppose also that the observed data $\mathbf{Y}$ consists of a $T \times K$ matrix, where the tth row $\mathbf{y}_t = (y_{t1}, \ldots, y_{tK})^{\mathsf{T}}$ is a K-dimensional feature vector associated with time t, and let $\mathbf{W}$ be a $D \times K$ weight matrix with kth column $\mathbf{w}_k$, such that

$$f(\mathbf{y}_t; \theta_j) = g(\mathbf{y}_t; \mathbf{W}^{\mathsf{T}}\theta_j) \tag{1.109}$$

for a suitable parametric function g. I will assume for simplicity that g factors as

$$g(\mathbf{y}_t; \mathbf{W}^{\mathsf{T}}\theta_j) = \prod_{k=1}^{K} g_k(y_{tk}; \mathbf{w}_k \cdot \theta_j) \tag{1.110}$$

Define $x_{tk} = \mathbf{w}_k \cdot \theta_{z_t}$, and $x_{tk}^{(-d)} = \mathbf{w}_k^{-d} \cdot \theta_{z_t}^{-d}$, where θ_j^{-d} and $\mathbf{w}_k^{-d}$ are θ_j and $\mathbf{w}_k$, respectively, with the dth coordinate removed. Then

$$\log\left(\frac{f(\mathbf{y}_t; \theta_{jd}=1, \theta_j \setminus \theta_{jd})}{f(\mathbf{y}_t; \theta_{jd}=0, \theta_j \setminus \theta_{jd})}\right) = \sum_{k=1}^{K} \log\left(\frac{g_k(y_{tk}; x_{tk}^{(-d)} + w_{dk})}{g_k(y_{tk}; x_{tk}^{(-d)})}\right) \tag{1.111}$$

If $g_k(y; x)$ is a Normal density with mean x and unit variance, then

$$\log\left(\frac{g_k(y_{tk}; x_{tk}^{(-d)} + w_{dk})}{g_k(y_{tk}; x_{tk}^{(-d)})}\right) = -w_{dk}(y_{tk} - x_{tk}^{(-d)} + \frac{1}{2}w_{dk}) \tag{1.112}$$

Sampling $\boldsymbol{\mu}$

Sampling the μ_d is straightforward with a Beta prior. Suppose

$$\mu_d \overset{ind}{\sim} \mathcal{B}\text{eta}(a_\mu, b_\mu) \tag{1.113}$$

Then, conditioned on $\boldsymbol{\theta}$ the μ_d are independent with

$$\mu_d \,|\, \boldsymbol{\theta} \sim \mathcal{B}\text{eta}(a_\mu + \sum_j \theta_{jd}, b_\mu + \sum_j (1 - \theta_{jd})) \tag{1.114}$$

Sampling λ

The parameter λ governs the connection between $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. Writing (1.99) in terms of λ and the difference matrix $\boldsymbol{\Delta} = (\Delta_{jj'})_{1 \le j, j' \le J}$ gives

$$p(\mathbf{z}, \mathbf{Q} \,|\, \lambda, \boldsymbol{\theta}) \propto \prod_j \prod_{j'} e^{-\lambda \Delta_{jj'} n_{jj'}} (1 - e^{-\lambda \Delta_{jj'}})^{q_{jj'}} \tag{1.115}$$

Put an $\mathcal{E}\text{xp}(b_\lambda)$ prior on λ, so that

$$p(\lambda \,|\, \mathbf{z}, \mathbf{Q}, \boldsymbol{\theta}) \propto e^{-(b_\lambda + \sum_j \sum_{j'} \Delta_{jj'} n_{jj'})\lambda} \prod_j \prod_{j'} (1 - e^{-\lambda \Delta_{jj'}})^{q_{jj'}} \tag{1.116}$$

This density is log-concave, with

$$-\frac{d^2 \log(p(\lambda \,|\, \mathbf{z}, \mathbf{Q}, \boldsymbol{\theta}))}{d\lambda^2} = \sum_{\{(j,j') \,|\, \Delta_{jj'} > 0\}} \frac{\Delta_{jj'}^2 q_{jj'} e^{\lambda \Delta_{jj'}}}{(e^{\lambda \Delta_{jj'}} - 1)^2} > 0 \tag{1.117}$$

and so we can use Adaptive Rejection Sampling (Gilks and Wild, 1992) to sample from it. The relevant h and h', representing the log density and its first derivative, respectively, are

$$h(\lambda) = -(b_\lambda + \sum_{\{(j,j') \,|\, \Delta_{jj'} > 0\}} \Delta_{jj'} n_{jj'})\lambda + \sum_{\{(j,j') \,|\, \Delta_{jj'} > 0\}} q_{jj'} \log(1 - e^{-\lambda \Delta_{jj'}}) \tag{1.118}$$

$$h'(\lambda) = -(b_\lambda + \sum_{\{(j,j') \,|\, \Delta_{jj'} > 0\}} \Delta_{jj'} n_{jj'}) + \sum_{\{(j,j') \,|\, \Delta_{jj'} > 0\}} \frac{q_{jj'} \Delta_{jj'}}{e^{\lambda \Delta_{jj'}} - 1} \tag{1.119}$$

Sampling $\mathbf{W}$

Conditioned on the state matrix $\boldsymbol{\theta}$ and the data matrix $\mathbf{Y}$, the weight matrix $\mathbf{W}$ can be sampled as well using standard methods for Bayesian regression problems. For example, suppose that the weights are *a priori* i.i.d. Normal:

$$p(\mathbf{W}) = \prod_{k=1}^{K} \prod_{d=1}^{D} \mathcal{N}(w_{dk} \,|\, 0, \sigma_0^2) \tag{1.120}$$

and the likelihood is

$$g_k(y; x) = \mathcal{N}(y \,|\, x, 1) \tag{1.121}$$

Then it is a standard result from Bayesian linear modeling that

$$p(\mathbf{W} \,|\, \boldsymbol{\theta}, \mathbf{Y}) = \prod_{k=1}^{K} \mathcal{N}\left(\left(\sigma_0^2 \mathbf{I} + \boldsymbol{\theta}^\top \boldsymbol{\theta}\right)^{-1} \boldsymbol{\theta}^\top \mathbf{y}_k, \sigma_0^2 \mathbf{I} + \boldsymbol{\theta}^\top \boldsymbol{\theta} \right) \tag{1.122}$$

If one or more output features, say $\mathbf{y}_k$, is binary, we can adopt a probit model where we introduce a latent data vector $\mathbf{y}_k^*$ for each such k, and assume

$$p(\mathbf{y}_k^* \,|\, \mathbf{x}_k) = \prod_{t} \mathcal{N}(y_{tk}^* \,|\, x_{tk}, 1) \tag{1.123}$$

and

$$y_{tk} = \begin{cases} 0, & y_{tk}^* \leq 0 \\ 1, & y_{tk}^* > 0 \end{cases} \tag{1.124}$$

And so, after marginalizing over $\mathbf{y}_k^*$

$$p(\mathbf{y}_k \,|\, \mathbf{x}_k) = \prod_{t=1}^{T} F(x_{tk})^{y_{tk}} (1 - F(x_{tk}))^{1 - y_{tk}} \tag{1.125}$$

where F is the standard Normal CDF, since

$$\int_0^\infty dy_{tk}^* \mathcal{N}(y_{tk}^* \,|\, x_{tk}, 1) = \int_{-x_{tk}}^\infty dy_{tk}^* \mathcal{N}(y_{tk}^* \,|\, 0, 1) = 1 - F(-x_{tk}) = F(x_{tk}) \tag{1.126}$$

Then, conditioned on x_{tk} and y_{tk}, we can sample y_{tk}^* from a Normal distribution left- or right-truncated at 0:

$$p(y^*_{tk}\,|\,x_{tk}, y_{tk}) = \begin{cases} \mathcal{N}(x_{tk}, 1) I(y^*_{tk} \leq 0), & y_{tk} = 0 \\ \mathcal{N}(x_{tk}, 1) I(y^*_{tk} > 0), & y_{tk} = 1 \end{cases} \tag{1.127}$$

Conditioned on the y^*_{tk} and $\boldsymbol{\theta}$, the weights are distributed as in (1.122).

Summary

I have made the following assumptions about the representation of the hidden states and observed data in this subsection: (1) $\boldsymbol{\theta}$ consists of D binary features (2) the similarity function Φ is the Laplacian kernel with respect to Hamming distance with decay parameter λ, and (3) $\mathbf{Y}$ consists of K continuous or binary features associated with each time step t. In addition, we make the following distributional assumptions:

$$\mu_d \overset{i.i.d}{\sim} \mathcal{B}\text{eta}(a_\mu, b_\mu) \tag{1.128}$$

$$\lambda \sim \mathcal{E}\text{xp}(b_\lambda) \tag{1.129}$$

$$\theta_{jd}\,|\,\boldsymbol{\mu} \overset{ind}{\sim} \mathcal{B}\text{ern}\left(\mu_d\right) \tag{1.130}$$

$$\mathbf{W} \sim \mathcal{N}(0, \sigma_0^2 \mathbf{I}) y^*_{tk}\,|\,\mathbf{W}, \mathbf{z}, \boldsymbol{\theta} \overset{ind}{\sim} \mathcal{N}(x_{tk}, 1) \tag{1.131}$$

$$y_{tk} = \begin{cases} y^*_{tk}, & \text{if } k \text{ is a continuous feature} \\ \mathbb{I}(y^*_{tk} > 0) & \text{if } k \text{ is a binary feature} \end{cases} \tag{1.132}$$

where we have defined

$$x_{tk} = \mathbf{w}_k \cdot \theta_{z_t} \tag{1.133}$$

I introduce Gibbs blocks corresponding to (1) each θ_{jd} individually, (2) the vector $\boldsymbol{\mu}$, (3) the decay parameter λ, (4) the weight matrix $\mathbf{W}$, and (5) the latent data $\mathbf{Y}^*$ associated with binary features. We have

$$\theta_{jd} \mid \boldsymbol{\theta} \setminus \theta_{jd}, \mathbf{z}, \mathbf{Q}, \boldsymbol{\mu}, \lambda, \mathbf{W}, \mathbf{Y}^* \sim \mathcal{B}\text{ern}\left(\frac{e^{\zeta_{jd}}}{1+e^{\zeta_{jd}}}\right) \tag{1.134}$$

$$\mu_d \mid \boldsymbol{\theta}, \ldots \overset{ind}{\sim} \mathcal{B}\text{eta}(a_\mu + \sum_j \theta_{jd}, b_\mu + \sum_j (1-\theta_{jd})) \tag{1.135}$$

$$p(\lambda \mid \mathbf{z}, \mathbf{Q}, \boldsymbol{\theta}, \ldots) \propto e^{-(b_\lambda + \sum_j \sum_{j'} \Delta_{jj'} n_{jj'})\lambda} \prod_j \prod_{j'} (1 - e^{-\lambda \Delta_{jj'}})^{q_{jj'}} \tag{1.136}$$

$$\mathbf{w}_k \mid \boldsymbol{\theta}, \mathbf{Y}^*, \ldots \overset{ind}{\sim} \mathcal{N}((\sigma_0^2 \mathbf{I} + \boldsymbol{\theta}^\intercal \boldsymbol{\theta})^{-1} \boldsymbol{\theta}^\intercal \mathbf{y}_k^*, \sigma_0^2 \mathbf{I} + \boldsymbol{\theta}^\intercal \boldsymbol{\theta}) \tag{1.137}$$

$$\mathbf{y}_{tk}^* \mid \mathbf{X}, \mathbf{Y}, \ldots \overset{ind}{\sim} \begin{cases} \mathcal{N}(x_{tk}, 1)\mathbb{I}(y_{tk}^* \leq 0), & y_{tk} = 0 \\ \mathcal{N}(x_{tk}, 1)\mathbb{I}(y_{tk}^* > 0), & y_{tk} = 1 \end{cases} \tag{1.138}$$

where $\Delta_{jj'} = \left|\left|\theta_j - \theta_j'\right|\right|_{L_1}$ and

$$\begin{aligned} \zeta_{jd} = \log\left(\frac{\mu_d}{1-\mu_d}\right) + (c_{jd1} - c_{jd0})\lambda + \sum_{j' \neq j} (-1)^{\theta_{j'd}} (q_{jj'} + q_{j'j}) \log\left(\frac{1 - \phi_{jj'}^{(-d)} e^{-\lambda}}{1 - \phi_{jj'}^{(-d)}}\right) \\ - \sum_{\{t \mid z_t = j\}} \sum_{k=1}^{K} w_{dk}(y_{tk}^* - x_{tk}^{(-d)} + \frac{1}{2} w_{dk}) \end{aligned} \tag{1.139}$$

All distributions can be sampled from directly except for λ, which requires Adaptive Rejection Sampling, with the equations

$$h(\lambda) = -(b_\lambda + \sum_{\{(j,j') \mid \Delta_{jj'} > 0\}} \Delta_{jj'} n_{jj'})\lambda + \sum_{\{(j,j') \mid \Delta_{jj'} > 0\}} q_{jj'} \log(1 - e^{-\lambda \Delta_{jj'}}) \tag{1.140}$$

$$h'(\lambda) = -(b_\lambda + \sum_{\{(j,j') \mid \Delta_{jj'} > 0\}} \Delta_{jj'} n_{jj'}) + \sum_{\{(j,j') \mid \Delta_{jj'} > 0\}} \frac{q_{jj'} \Delta_{jj'}}{e^{\lambda \Delta_{jj'}} - 1} \tag{1.141}$$

Chapter 2

Two Dirichlet Process Models of Probabilistic Context-Rich Grammars

In this chapter, I define a probabilistic generative model of syntactic parse trees, in which the probability of generating a particular symbol at a given position in the tree depends on the context in which it occurs. In traditional probabilistic context-free grammars (PCFGs), the context affecting the production probability is limited to the immediate parent node in the tree. Numerous attempts have been made to take into account a richer set of context, including grandparent or sibling nodes. A particularly successful approach in this vein is that of Collins (1997, 2003), in which each node has a special child, called the *head*, which is assumed to define the main content of the phrase, and in which nonterminal nodes carry both the word and part-of-speech tag reached by following the sequence of head children to the leaf level of the tree. These special paths, from phrase nodes to the word nodes, are called *spikes*. The non-head children of a phrase node represent the top of a new spike and are generated conditioned on the enriched ("lexicalized") representation of their immediate parent, as well as their head sibling. The benefit of this representation is that symbol productions can incorporate surrounding words into account into their context, or *syntactic history*, without violating the acyclic tree-structured factorization of the distribution over parse trees.

The cost of taking additional context into account is that there will typically be very few or no occurrences of any particular combination in even very large data sets, due to very rapid combinatorial growth arising from the large set of available symbols. Indeed even very local context can result in data sparsity problems when individual words are part of context. This data sparsity issue is one of the central problems in the field of natural language processing.

A standard solution to the sparsity problem is to start with maximum likelihood estimates (MLEs) of full conditional probabilities and then incorporate some form of smoothing, where probability mass is borrowed from similar contexts (for a review of canonical smoothing methods, see Chen and Goodman (1999)). The parser of Collins (2003) uses successive *back-off interpolation* steps. Abstractly, if f is a syntactic feature to be generated in a context $\mathbf{c} = (c_1, c_2, \dots, c_n)$, then back-off smoothing estimates the probability $p(f|c_1, \dots, c_n)$ by first ordering the context features in perceived decreasing order of importance, and recursively interpolating.

Let $\hat{p}$ the target estimate, let $\tilde{p}$ represent the MLE, and let c_0 be the trivial context. Collins recursively defines

$$\hat{p}(f|c_1, \dots, c_{n-k}) = \lambda_k \tilde{p}(f|c_1, \dots, c_{n-k}) + (1-\lambda_k)\hat{p}(f|c_1, \dots, c_{n-k-1}), \quad k = 0, \dots, n-2 \tag{2.1}$$

where $\hat{p}(f|c_1) = \lambda_{n-1}\tilde{p}(f) + (1 - \lambda_{n-1})\varepsilon$ for a small constant ε. The smoothing weights, $\lambda_1, \dots, \lambda_{n-1}$ are context-dependent, giving greater weight to contexts that have been observed frequently in training, but giving lower weight to contexts in which many values of f were observed with low frequency (intuitively, this is the situation that requires more smoothing, since there are likely many more rare values of f that were not observed).

This approach has some desireable features: first, it allows contexts to "borrow strength" from others that share a prefix, and second, by increasing the degree of borrowing when the symbol distribution in a context has a high degree of diversity, the approach accounts for the asymmetry in the number of observations needed to precisely estimate high- and low-entropy distributions. However, it is nonetheless a heuristic, and does not appear to follow as an optimal estimation strategy in any fully probabilistic model. A second issue with this approach is that it requires a fixed and prespecified ordering of the context features, and hence, for example, two contexts that share almost all of their features will not be coupled at all if the first feature differs. Ideally, contexts would be coupled if they share any features, with the relevance of each feature learned from data.

In the remainder of the dissertation, I have three goals, related to the problem of borrowing strength across related contexts.

1. The first goal does not directly concern borrowing strength, but is motivated by the application of inferring natural language meaning in addition to structure, and places constraints on the space of feasible solutions to the borrowing strength problem. I will augment the context representation to take extralinguistic context into account in

addition to the syntactic history; in particular, the distribution over parse trees will be conditioned on a parallel "semantic tree" containing a propositional representation of the communicative intent of the sentence. The inferences involved in parsing then become a means to the ultimate goal, which is inferences about what the sentence is about. The addition of this added semantic context prohibits solutions to the estimation problem that rely on observations being equipped with a set of known covariates (context features), since even if annotated parse trees are provided in training data, the precise communicative intent will not be given, and thus must be able to be treated as latent, even in the training data.

2. I will show that something close (though not identical) to the smoothing algorithm used by Collins and described above arises by using a multi-level Hierarchical Dirichlet Process as the prior on the conditional symbol distributions, where contexts are nested in the hierarchy according to shared prefixes.

3. I will develop an alternative, non-hierarchical prior on context-conditional symbol distributions which has the properties that (1) two distributions are coupled to the extent that they share context features in general (and not just prefixes) and (2) the importance of particular context features is learned from data. I achieve this by representing each context-specific distribution as a mixture of an uncountably infinite number of "topic" distributions, where (a) latent topic assignment variables are introduced to simplify inference, (b) the probability of assignment to a topic is an increasing function of the total pairwise similarity to existing instances of that topic, (c) the pairwise similarity function is a weighted function of the binary vector indicating which positions overlap, and (d) the weights associated with each position may differ by topic, and are themselves a target of inference.

I turn next to addressing these three goals in turn, beginning by defining the syntactic and semantic representations used by the two probabilistic models, and then defining each model and associated inference algorithms.

2.1 Representation of Captions

The models defined in this chapter are motivated by the application of recovering semantic representations from the conjunction of natural language *captions* and a representation of a physical *scene*, about which visual images provide independent evidnece. Given some text, the goal is to recover a semantic representation which can be connected to the physical environment.

Captions have two layers of latent structure that explain the observed word sequence, S. The first is a propositional, or semantic, *elaboration tree*, Ψ, in which each node represents a semantic entity or relation to be expressed in the text. The second representation is a lexicalized constituency tree, Λ, adapted from the representation of Collins (2003). I describe these two representations in detail next.

2.1.1 Elaboration Trees: Ψ

To constrain the semantic domain, I will assume that the communicative goal of the speaker producing a caption is to describe to a listener a three-dimensional room scene containing furniture and wall hangings, such that the listener may pick out the same objects and relations in the scene that are the speaker's intended referents of the terms in the caption.

Denote the set of objects in the scene representation, including the room itself, by $\mathcal{O}$. Each object in $\mathcal{O}$ is a candidate to be the focus of a description. A focal, or *target* object, $t \in \mathcal{O}$ can be elaborated by including information about its intrinsic features, such as color and size, or about its spatial relationship to other objects in the scene. For each new object introduced into the description via a spatial relation, it may in turn be elaborated. These elaborations define a set of recursive grammatical rewrite rules which instantiate an *elaboration tree*, which we denote by Ψ.

Each elaboration tree is assumed to have a unary predicate, $\textsc{Focus}(t)$, at its root. This predicate has two children: the *relation*, $\rho = \textsc{Focus}$, and the *target*, t, which indicates the object type. The node t may be rewritten with either a spatial relation subtree, or an attribute subtree. A *spatial relation* elaboration is a rewrite rule that transforms the target object, t, into a predicate node, $\rho(t, b_1, \ldots, b_k)$, with $k+2$ children: one for the relation itself, one for the target object, and one for each of the reference objects used in the description An *attribute* elaboration is a rewrite rule that replaces t with a unary predicate node, $\alpha(t)$, with two children: the attribute, α and t itself.

An example of an elaboration tree that might be associated with a sentence about objects in a room is given in Fig. 2.1.

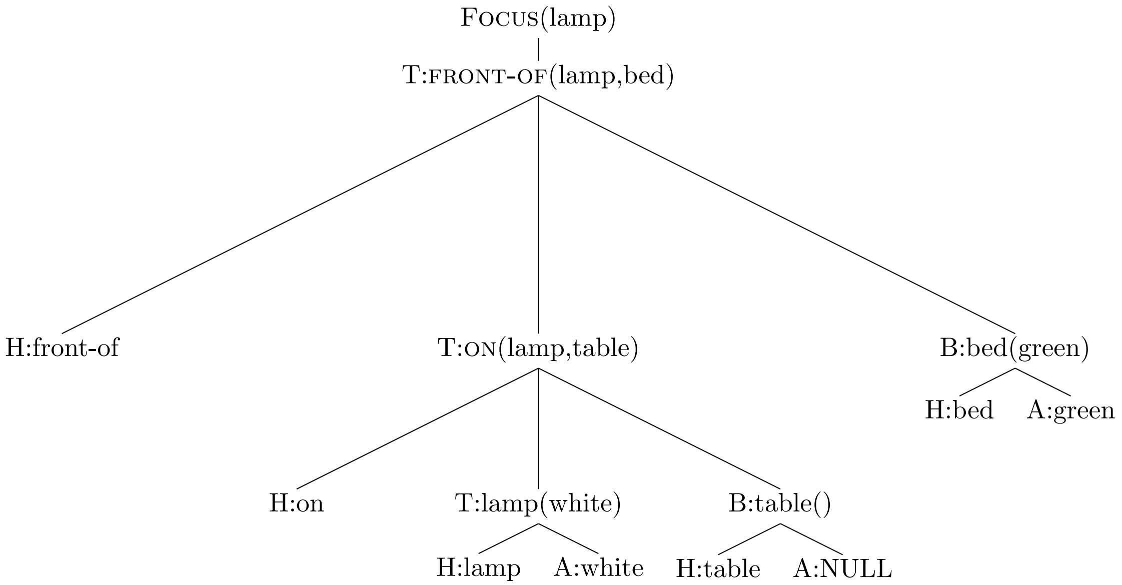

Figure 2.1: A semantic tree that might generate the sentence "A white lamp on a table is in front of a green bed.". Nodes prefixed by "H" are the "head" child of the parent; nodes prefixed by "T" are the "target" (the primary argument), and nodes prefixed by "B" are the "base" (the secondary argument). Nodes prefixed by "A" are attributes.

2.1.2 Syntactic Trees: Λ

The nodes in the elaboration tree consist solely of semantic concepts and predicates. All linguistic information is contained in a *syntactic tree*, Λ. I first describe the representation used by Collins (2003), and then describe how this representation is augmented to incorporate information from the elaboration tree.

The trees in Collins (2003) and related work consist of nodes representing syntactic constituents, labeled by Penn Treebank symbols (Marcus et al., 1993). Each node has a special child, called the *head*, which is assumed to define the main content of the phrase. For example, verb phrases (VPs) typically have a verb as their head; noun phrases have nouns or perhaps other noun phrases; etc. A path from any constituent to the word level, called a *spike*, can always be defined by iteratively descending to the head child until a single word is reached. Each node in the tree carries with it the identities of the word and part-of-speech tag at the end of its respective spike. The non-head children of a phrase node are called *modifiers*, and represent the top of a new spike. Proceeding from the root to the leaves, there are three types of productions, or *events*: (1) *root events*, in which the top level constituent label, along with its associated head word and tag, is generated, (2) *unary events*, in which a head child acquires a label (inheriting the word and tag from the parent), and (3) *modifier* *events*, in which children are generated to the left and right of the head, receiving a label, tag and word (the latter two of which are passed to its own head child). Eventually, each spike of unary events generates a part-of-speech tag as the label, which terminates the spike. An example syntactic tree is shown in Fig. 2.2.

This basic model is extended by adding semantic content from the elaboration tree to each node, in addition to its label, tag, word triple. We assume each node is associated with zero or one nodes from the elaboration tree; that is, that there is a function from the set of nodes in Λ to the set of nodes in Ψ plus a *null elaboration* node. Further, we make a continuity assumption, which requires that every subtree in Λ maps to a subtree in Ψ (or to the null elaboration), and we assume that the root of Λ maps to the root of Ψ. As a result of these assumptions, Λ together with the set of associations can be generated by specifying, for each child node in Λ, whether its associated node in Ψ will be (1) the null node, (2) the same as its parent, or (3) a child of the semantic node associated with its parent. Moves to children are further categorized as moves to the "head", the "target", the "base", or an "attribute" (see Fig. 2.1. We call these decisions *semantic step events*, and distinguish two event types: *head semantic steps*, which determine the semantic association of head children, and *modifier semantic steps*, which determine the association of non-head children. Figure 2.3 depicts an augmented representation in which each node in the parse tree in Fig. 2.2 is associated with a node from the elaboration tree in Fig. 2.1.

2.1.3 Representation of Context

A syntactic tree, Λ, is generated via a sequence of contextualized production events. There are five types of productions, three of which — *root events*, *unary events*, and *modifier events* — generate syntactic symbols, and two of which — *head semantic steps* and *modifier semantic steps* — determine associations between syntactic constituents and nodes in the elaboration tree, Ψ.

Following Collins (2003), the probability of each event depends on a set of discrete-valued *history* features, corresponding to the outcomes of a subset of previous events. We add to the conditioning set features from the semantic nodes associated with nearby syntactic constituents. With the exception of the *root* and *modifier* events, each event type produces a single categorical feature. The *root* and *modifier* events each produce a word, w, a part-of-speech tag, t, and a constituent label, l. As in Collins (2003), we employ the factorization

$$P_{\text{root}}(w, t, l \mid H) = P_{\text{root}}(t, l \mid H) P_{\text{root}}(w \mid t, l, H), \tag{2.2}$$

type	**generated**	**conditioned on**			
		Level 0	Level 1	Level 2	Level 3
root1	l, t	s	sa		
root2	w	t, s	l	sa	
head-sem	s, sa	s_p, l_p	t_h, sa_p	w_h	
unary	l	s, l_p	s_p, t_h, sa_h	sa_p	w_p
mod-sem	s, sa	s_p, l_p	s_h, l_h, $dist$	t_h, sa_p, sa_h	w_h
mod1	l, t	l_p, s	l_h, s_p, s_h, $dist$	t_h, sa_m, sa_p, sa_p	w_h
mod2	w	t, s	l, l_p, l_h, s_p, s_h, $dist$	t_h, sa_m, sa_p, sa_h	w_h

Table 2.1: Features included in each event type. The features l, t and w denote constituent label, part-of-speech tag, and word, respectively. The s and sa features represent the head and arguments, respectively, of a semantic elaboration node. The $dist$ feature for modifier events is Collins' distance feature, which specifies (a) whether a modifier is on the left or right of the head, (b) whether the modifier is adjacent to the head, and (c) whether there is a verb contained in any constituent between the modifier and the head. Features without subscripts refer to the constituent currently being generated; features with the subscript p refer to the parent constituent, and features with the subscript h refer to the sister head constituent.

where H represents the relevant set of history features. We employ the analogous factorization of $P_{\text{modifier}}(w, t, l \mid H)$. Collectively, the full set of conditional event probabilities constitutes the parameter set θ_Λ.

For each event type, the set of history features included at each smoothing level in Collins' recursive definition is listed in table 2.1.

2.2 Heuristic Smoothing

In Collins (2003), the conditional event distributions are estimated from training data by successive back-off smoothing of the MLEs (i.e., empirical conditional proportions). Let $p(x|H)$ be a desired conditional probability of generating event x given the full history H. Let β_k, $k = 0, \ldots, K$ be a set of "abstraction functions" such that $\beta_k(H)$ maps a context H to a context equivalence class which is more general (i.e., the set of contexts sharing a prefix with H), and let $\beta_K(H) = H$. Denote the maximum likelihood estimate of $p(x|\beta_k(B))$ by $\hat{p}(x|\beta_k(B))$. Collins recursively sets

$$p(x|\beta_k(H)) = \lambda_k \hat{p}(x|\beta_k(H)) + (1 - \lambda_k) p(x|\beta_{k-1}(H)) \tag{2.3}$$

where the λ_k are smoothing weights, and

$$p(x|\beta_0(H)) = \lambda_0 \hat{p}(x|\beta_0(H)) + (1 - \lambda_0)\varepsilon \tag{2.4}$$

for a small constant ε.

The smoothing weights are context-sensitive, and defined to be

$$\lambda_k(H) = \frac{n_k(H)}{n_k(H) + \gamma u_k(H)} \tag{2.5}$$

where $n_k(H)$ is the number of training instances of context $\beta_k(H)$, $u_k(H)$ is the number of distinct values of x observed following context $\beta(H)$ (called the *diversity* of $\beta(H)$) and γ is a global smoothing parameter controlling the degree to which the final probability estimates differ from their maximum likelihood estimates.

2.3 An HDP Model of Context-Conditional Event Distributions

We wish to define a fully generative model so that we will have a well defined joint posterior distribution over the parameters of the grammar, so that we can justify our estimate in probabilistic terms, and so that we can incorporate the grammar model in a larger model of scenes and captions. Happily, something very close to Collins' smoothing equations can be derived as the predictive distribution under a reasonable probabilistic model, namely a Hierarchical Dirichlet Process (HDP) Teh et al. (2006), and hence can be used with minimal alteration in Gibbs sampling.

Let x as above represent a particular outcome of an event. Let $x \mid \beta_0(H) \sim \pi_{\beta_0(H)}$, where $\pi_{\beta_0(H))}$ is a distribution over all outcomes available for the event in question. If we define a Dirichlet Process (DP) prior on $\pi_{\beta_0(H)}$, with concentration parameter α_0 and base measure π, then having observed $n_0(H)$ events with history $\beta_0(H)$, whose values are represented by $\mathbf{x}_{\beta_0(H)}$, such that $n_0(x, H)$ of them had outcome x, the predictive distribution of the next event (integrating out $\pi_{\beta_0(H)}$) is given by

$$p(x \mid \beta_0(H), \mathbf{x}_{\beta_0(H)}) = \frac{n_0(H)}{n_0(H) + \alpha_0} \frac{n_0(x, H)}{n_0(H)} + \frac{\alpha_0}{n_0(H) + \alpha_0} \pi(x), \tag{2.6}$$

(see Teh et al. (2006) for a derivation). This is the Chinese Restaurant Process (CRP), in which each event ("customer") "sits at the table" associated with a previous customer, with probability proportional to the number people already sitting there, and "starts a new table" with probability proportional to a concentration parameter, α_0. The actual value (the "dish") associated with each table is chosen randomly from the global base measure, π, by the first customer to sit there.

Now considering contexts $\beta_1(H)$ to be restaurants, several of which belong to a broader context, $\beta_0(H)$, we can add a level to the hierarchy, and define $x \mid \beta_1(H) \sim \pi_{\beta_1(H)}$ and $\pi_{\beta_1(H)} \sim \mathsf{DP}(\alpha_1(H), \pi_{\beta_0(H)})$, yielding

$$p(x \mid \beta_1(H), \mathbf{x}_{\beta_1(H)}) = \frac{n_1(H)}{n_1(H) + \alpha_1(H)} \frac{n_1(x, H)}{n_1(H)} + \frac{\alpha_1(H)}{n_1(H) + \alpha_1(H)} \pi_{\beta_0(H)}(x), \tag{2.7}$$

where $\pi_{\beta_1(H)}$ has been integrated out.

By keeping track of the number of "tables" across all "restaurants" in the broader context, $\beta_0(H)$, that ordered each dish, x, which we represent by $\mathbf{m}_0(H) = (m_0(1, H), m_0(2, H), \dots)$, with $m_0(H) = \sum_x m_0(x, H)$, we can also integrate out $\pi_{\beta_0(H)}$, to get

$$\begin{aligned} p(x \mid \beta_1(H), \mathbf{x}_{\beta_1(H)}, \mathbf{m}_0(H)) = {} & \frac{n_1(H)}{n_1(H) + \alpha_1(H)} \frac{n_1(x, H)}{n_1(H)} + \frac{\alpha_1(H)}{n_1(H) + \alpha_1(H)} \times \\ & \left(\frac{m_0(H)}{m_0(H) + \alpha_0(H)} \frac{m_0(x, H)}{m_0(H)} + \frac{\alpha_0(H)}{m_0(H) + \alpha_0(H)} \pi(x) \right) \end{aligned} \tag{2.8}$$

Equation (2.8) together defines the "Chinese Restaurant Franchise" process as described by Teh et al. (2006), where each narrow context, $\beta_1(H)$ represents a "restaurant" that is affiliated with a broader context, $\beta_0(H)$, representing a "restaurant franchise". Each restaurant works as described above, with a restaurant-specific[1] concentration parameter, $\alpha_1(H)$, except that when a new table is instantiated, instead of sampling new values ("dishes") from the global base measure, π, the new dish is selected in proportion to the number of times it occurs ("the number of tables ordering it") in the broader context ("the franchise"), while a completely new value is chosen from π in proportion to the franchise-level concentration parameter, α_0. It is natural to continue adding levels to the hierarchy (perhaps "affiliations of franchises", etc.), introducing a new set of concentration parameters for each new smoothing level. In that case, we would increment the subscripts in (2.8), and replace $\pi(x)$ with $\pi_{\beta_0(H)}$, which would be integrated out by counting the number of tables selecting each dish in the

[1] The use of separate concentration parameters for each restaurant in the franchise is a departure from the model defined in Teh et al. (2006). To get that model, we would remove the dependence on H, and use a common α_1 for all contexts at level 1.

"affiliation of franchises", and when a new dish is called for at the franchise level, it would be chosen in proportion to the number of tables in the affiliation, with another concentration parameter governing the probability of sampling from the global measure.

2.3.1 Heuristic Smoothing as an Approximation to the HDP Prior

The recursively defined predictive distribution in (2.8) has a similar structure to the predictive distribution defined by (2.3), (2.4) and (2.5) (provided the top level π is uniform over ε^{-1} outcomes), since the MLE, $\hat{p}(x|\beta_k(B))$ is simply the empirical proportion, $n_1(x, H)/n_1(H)$. However, there are some important differences. First, instead of a fixed hyperparameter α, Collins allows the back-off strength to depend on the data, not only through the number of training instances, but also through the number of distinct outcomes, or "dishes" (the diversity).

This qualitative behavior is intuitively reasonable under the HDP model: The greater the value of a (restaurant-specific) α, the more likely it is that a new table will be formed there, hence the larger the expected number of occupied tables. Hence, the likelihood favors larger α values as the number of tables increases. The number of occupied tables in the restaurant(s) governed by α, along with an auxiliary variable, w, whose distribution depends on α and the number of customers in the restaurant, are sufficient statistics for α, since the assignment of tables to dishes depends only on the base measure of the DP, and not on the concentration parameter. Let α govern a restaurant (or restaurant franchise, etc.), with n customers occupying T distinct tables. For the bottom level of the hierarchy, Antoniak (1974) showed that

$$p(T \mid \alpha, n) = s(n, T)\alpha^T \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)} \tag{2.9}$$

where $s(n, T)$ is an unsigned Sterling number of the first kind. If we further define

$$w \mid \alpha, n \overset{ind}{\sim} \mathcal{B}\text{eta}(\alpha, n) \tag{2.10}$$

so that

$$\begin{aligned} p(w, T \mid \alpha, n) &= s(n, T)\alpha^T \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)} \frac{\Gamma(\alpha + n)}{\Gamma(\alpha)\Gamma(n)} w^{\alpha-1}(1 - w)^{n-1} && (2.11) \\ &= s(n, T)\alpha^T \Gamma(n)^{-1} w^{\alpha-1}(1 - w)^{n-1} && (2.12) \end{aligned}$$

then, after placing a $\mathcal{G}(a,\gamma)$ prior on α, we obtain posterior density

$$p(\alpha \mid T, w) \propto \alpha^{a+T-1} e^{-(b-\log(w))\alpha} \tag{2.13}$$

which is that of a $\mathcal{G}(a+T, b-\log(w))$ distribution. Note that the mean of this distribution is approximately proportional to T for small values of a and fixed values of w. Under the assumption that no two tables are assigned to the same outcome, then $T = u$, the diversity of the context, and we can interpret Collins' formula for the smoothing parameter λ given in (2.5) as the result of (1) placing a "vague" Gamma prior on α (i.e., $a, b <<$ 1), (2) approximating all w parameters with a single constant across for all contexts, (3) approximating the number of tables in each context by the number of unique dishes, and (4) approximating the predictive distribution in (2.6) using the resulting approximation posterior mean for α.

The likelihood for concentration parameters at higher levels of the hierarchy has a parallel structure, but importantly, the role played at the bottom level by the number of observations, n, is not played by the total number of observations in the collapsed context (or "restaurant franchise"), but rather by the total number of *tables* across all of the lower-level contexts ("restaurants") belonging to the aggregate ("franchise"), since this corresponds to the number of times a dish/outcome is selected from the broader distribution, and thus represents the number of independent observational units.

Denote by β_{ki} the ith context at level k of the hierarchy (where levels are numbered from top to bottom) which comprises J more specific contexts at level $k+1$, denoted by $\beta_{k+1,1}, \ldots, \beta_{k+1,J}$, and which in turn belongs to a larger unit, β_{k-1}. Let α_{ki} be the concentration parameter for context β_{ki}. Denote by m_{ki} the total number of tables occupied across all the $\{\beta_{k+1,j}\}$; that is, the sum of the T variables used above. Of these, some number, T_{ki}, represent distinct draws from the distribution at context β_{k-1}. The distribution of T_{ki} depends on α_{ki} as before:

$$p(T_{ki} \mid \alpha_{ki}, m_{ki}) = s(m_{ki}, T_{ki}) \alpha^{T_{ki}} \frac{\Gamma(\alpha_{ki})}{\Gamma(\alpha_{ki} + m_{ki})} \tag{2.14}$$

and we can introduce

$$w_{ki} \mid \alpha_{ki}, m_{ki} \overset{ind}{\sim} \mathcal{B}\text{eta}(\alpha_{ki}, m_{ki}) \tag{2.15}$$

so that

$$\alpha_{ki} \,|\, T_{ki}, w_{ki} \sim \mathcal{G}(a + T_{ki}, b - \log(w_{ki})) \tag{2.16}$$

Going up another level, across all I contexts at level k belonging to broader context β_{k-1}, we can define $m_{k-1} = \sum_i T_{ki}$, and model the distribution of T_{k-1} conditioned on the higher-level concentration α_{k-1} and the "observation" count, m_{k-1}. And so on.

We can see then that the implicit assumption in the heuristic smoothing equations that the number of independent observational units equals the number of true observations becomes further removed from the HDP model the higher up we go in the hierarchy.

2.3.2 Posterior Estimation of Event Distributions in the HDP Model

We would like a fully Bayesian approach to inference in the HDP model, and so rather than approximating the α parameters as fixed functions of the data, we develop a Gibbs-sampling algorithm to infer them.

Given a training set of fully annotated parse trees (i.e., where the elaboration tree and the associations between the two trees are known), the only unknown parameters are (1) the number of distinct independent observational units, T (IOUs; "tables" at the lowest level), per context and context prefix, (2) the concentration parameters, α, for each context and context prefix, and (3) the auxiliary variables, w associated with each context and context prefix. Since each α depends only on the corresponding T and w for its context, these can be Gibbs sampled independently, conditioned on the other values. Similarly, the w parameters depend only on α and the n (or m) parameters associated with its context.

From (2.15) and (2.13), and indexing indexing contexts at level k with j, we have

$$\alpha_j \,|\, w_j, T_j \overset{ind}{\sim} \mathcal{G}(a + T_j, b - \log(w_j)) \tag{2.17}$$

$$w_j \,|\, \alpha_j, n_j \overset{ind}{\sim} \mathcal{B}\text{eta}(\alpha_j, n_j) \tag{2.18}$$

Since the number of IOUs for a context at level k depends on the total number of IOUs across all narrower contexts at the level below, the T variables cannot be sampled independently across all contexts. Moreoever, since the assignments of observations to outcomes ("dishes") are known given an annotated training set, this information must be taken into account when sampling the T variables as well. However, it turns out that it is possible to handle each outcome value in each context at a fixed level independently, by simply simulating the Chinese Restaurant Process for the customers in that context assigned to that outcome.

To see why, consider the observations in context j that are assigned to outcome x. By exchangeability of observations we may assume that these observations were the last to be generated, and hence we can consider their distribution conditioned on the rest. Suppose that there are n_{j-x} observations not assigned to outcome x, with the remaining n_{jx} left to be assigned. The prior probability that the next observation would join an existing table is $\frac{n_{j-x}}{n_{j-x}+\alpha}$. However, if this happened, the observation would inherit whatever outcome was associated with the table it joined. Since this did not happen, the posterior probability that it started a new table is 1, just as if it had been the first observation in the entire context. Similarly, the prior probability that the next observation would have joined a table with one of the first n_{j-x} customers is still $\frac{n_{j-x}}{n_{j-x}+1+\alpha}$; but this contradicts its observation value, so it is certain that its table choice came from the remaining $\frac{1+\alpha}{n_{j-x}+1+\alpha}$. The prior probability that it would join the previous customer is $\frac{1}{n_{j-x}+\alpha}$; but conditioned on not joining the first n_{j-x} customers, this probability is increased to $\frac{1}{1+\alpha}$, with the remaining $\frac{\alpha}{1+\alpha}$ going to starting a new table; again just as if the first n_{j-x} observations did not exist. The same logic continues through, demonstrating that the number of distinct tables among the n_{jx}. Moreover, since none of this depends on the number of distinct tables associated with any other outcome, the outcome-specific table counts are independent.

In sum, we can start at the bottom level, K, and sample T_{Kjx} for each combination of context j and outcome x by generating n_{jx} sequential draws from a CRP with concentration α_{Kj}. Then, the number of IOUs for outcome x in the parent context, $T_{K-1,x}$, is obtained by computing $m_{K-1,x}$, the sum of the T_{Kjx} over j for all child contexts, and drawing $m_{K-1,x}$ observations from a CRP with concentration α_{K-1}.

2.4 A Similarity-Based Dependency Structure for Context-Conditional Event Distributions

A limitation of the HDP model defined above is that contexts are only coupled if they share an exact prefix. Thus, the choice of order for the context features is critical, as even two near-identical contexts will not have their distributions tied if they differ on one of the features in the first tier. We would like a model of the dependency structure that takes *any* similarity into account.

As before, the goal is estimation of $p(x \mid H)$, the outcome distribution for an event with history H. We dispense with the projections to broader contexts, and define a direct joint model of these conditional distributions. This model will be an adaptation of the Distance-Dependent Chinese Restaurant Process (*ddCRP*; Blei and Frazier (2011)), which I describe next.

Let $\mathcal{H}$ be a space of possible contexts for an event, and let $\phi : \mathcal{H} \times \mathcal{H} \to [0, 1]$ be a similarity kernel, such that $\phi(h, h) = 1$ for all $h \in \mathcal{H}$, and $\phi(h, h') \geq 0$ for all h, h'. For example, if each $h = (h_1, \ldots, h_n)$ is a context consisting of n categorical features, we might define

$$\phi(h, h') = \sum_{n=1}^{N} \exp(-\lambda_n \mathbb{I}(h_n \neq h'_n)) \tag{2.19}$$

where λ_n are weights associated with each feature. Alternatively, we could use a weighted overlap measure, such as

$$\phi(h, h') = \sum_{n=1}^{N} \omega_n \mathbb{I}(h_n = h'_n), \tag{2.20}$$

where $\sum_n \omega_n = 1$.

In both cases, similarities range between 0 and 1, and the more overlap there is between two context sequences, the greater the similarity score. Let $\alpha > 0$ be an innovation parameter, and π be a base distribution over outcomes $x \in \mathcal{X}$. The following prior on partitions is defined: Introduce set of *link variables*, $\{c_i\}$, one corresponding to each observation, where h_i is the context associated with observation i. Set

$$P(c_i = i' \mid \alpha, \phi) \propto \begin{cases} \phi(h_i, h_{i'}) & i \neq i' \\ \alpha & i = i' \end{cases} \tag{2.21}$$

Subsets of contexts that can be reached through a series of links are co-clustered, with $z(\mathbf{c})_i$ representing the cluster indicator for the ith observation. Each cluster is then assigned an outcome, by drawing from the base distribution π. If the base distribution has a symmetric Dirichlet prior, it can be integrated out, yielding the predictive distribution:

$$p(\bar{x}_k = x \mid) \propto m_x + \gamma \tag{2.22}$$

where $\bar{x}$ is the assignment to an outcome of cluster k, x indexes outcomes, m_x is the number of clusters previously associated with outcome x, and γ is the per-outcome prior weight.

2.4.1 Posterior Estimation of Event Distributions in the Similarity-Based Model

Given a fully annotated training set, the unknown parameters are (1) the cluster assignments, (2) the innovation parameter, α, and (3) the parameters of the similarity function.

Sampling cluster assignments is straightforward in the case where members of a cluster all share an outcome. There, links can only occur among observations with identical outcomes, and hence when resampling a link, we only need to consider linking to those other observations that share an outcome with the source. The posterior probability of a new link is

$$p(c_i = i' \mid \alpha, \phi, x_i) \propto \begin{cases} \phi(h_i, h_{i'})(m_{x_i} + \gamma) & i \neq i', \text{ if no change to clusters} \\ \phi(h_i, h_{i'})(m_{x_i} + 1 + \gamma) & i \neq i', \text{ if a cluster is split} \\ \phi(h_i, h_{i'})(m_{x_i} - 1 + \gamma) & i \neq i', \text{ if two clusters are joined} \\ \alpha(m_{x_i} + \gamma) & i = i', \text{ if no change to clusters} \\ \alpha(m_{x_i} + 1 + \gamma) & i = i' \text{ if a cluster is split} \end{cases} \tag{2.23}$$

For ϕ given by (2.19), the conditional distributions of α and $\boldsymbol{\lambda}$ do not have simple forms; however, an efficient method for sampling arbitrary differentiable densities, such as Hamiltonian Monte Carlo can be used.

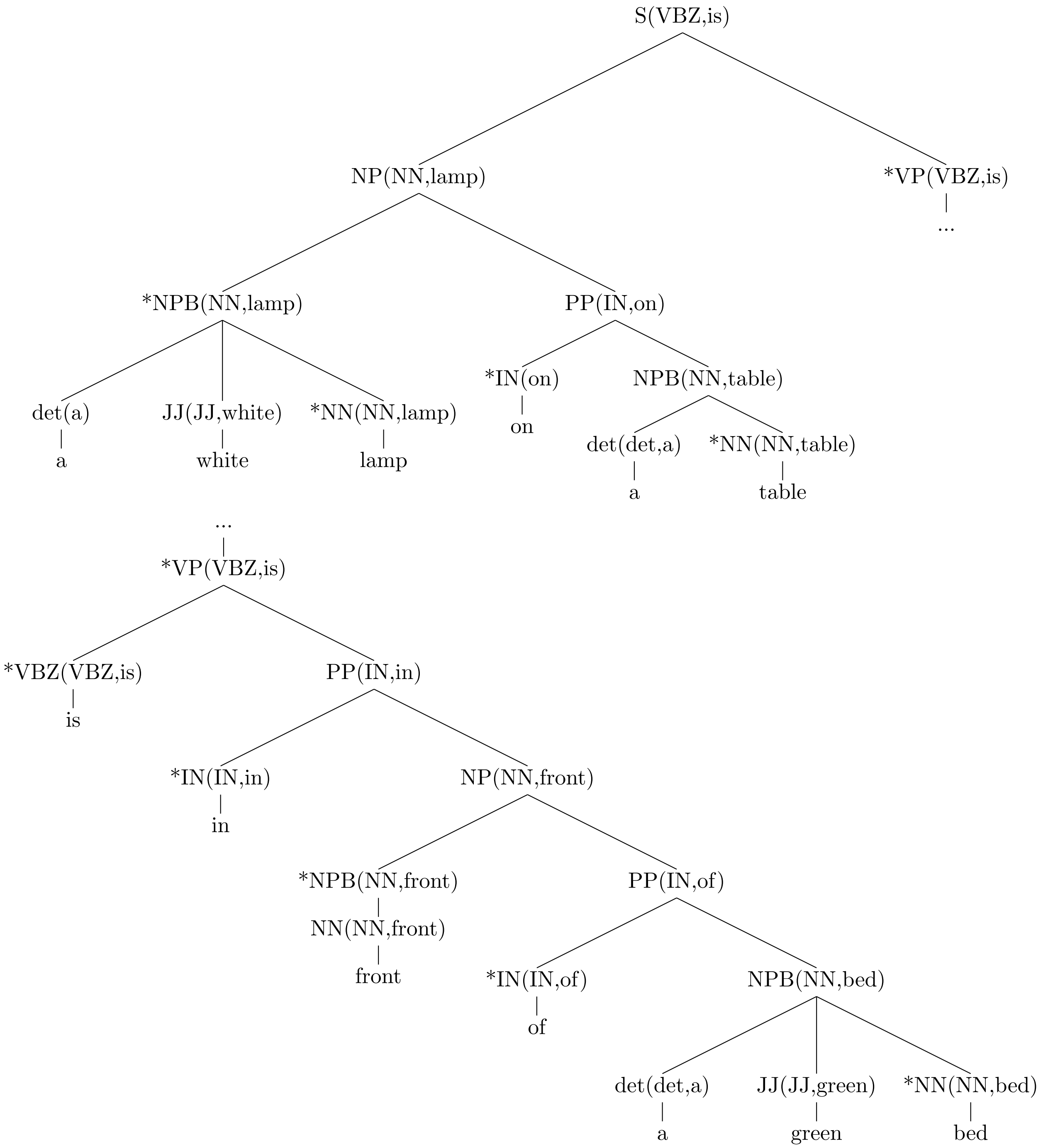

Figure 2.2: A parse tree for the example sentence. Each node has the form LABEL(TAG,Word). Asterisks indicate nodes that are the head child of their parent. All other nodes are "modifiers". Each nonterminal node is the start of a "spike", which is obtained by iteratively descending to the head child.

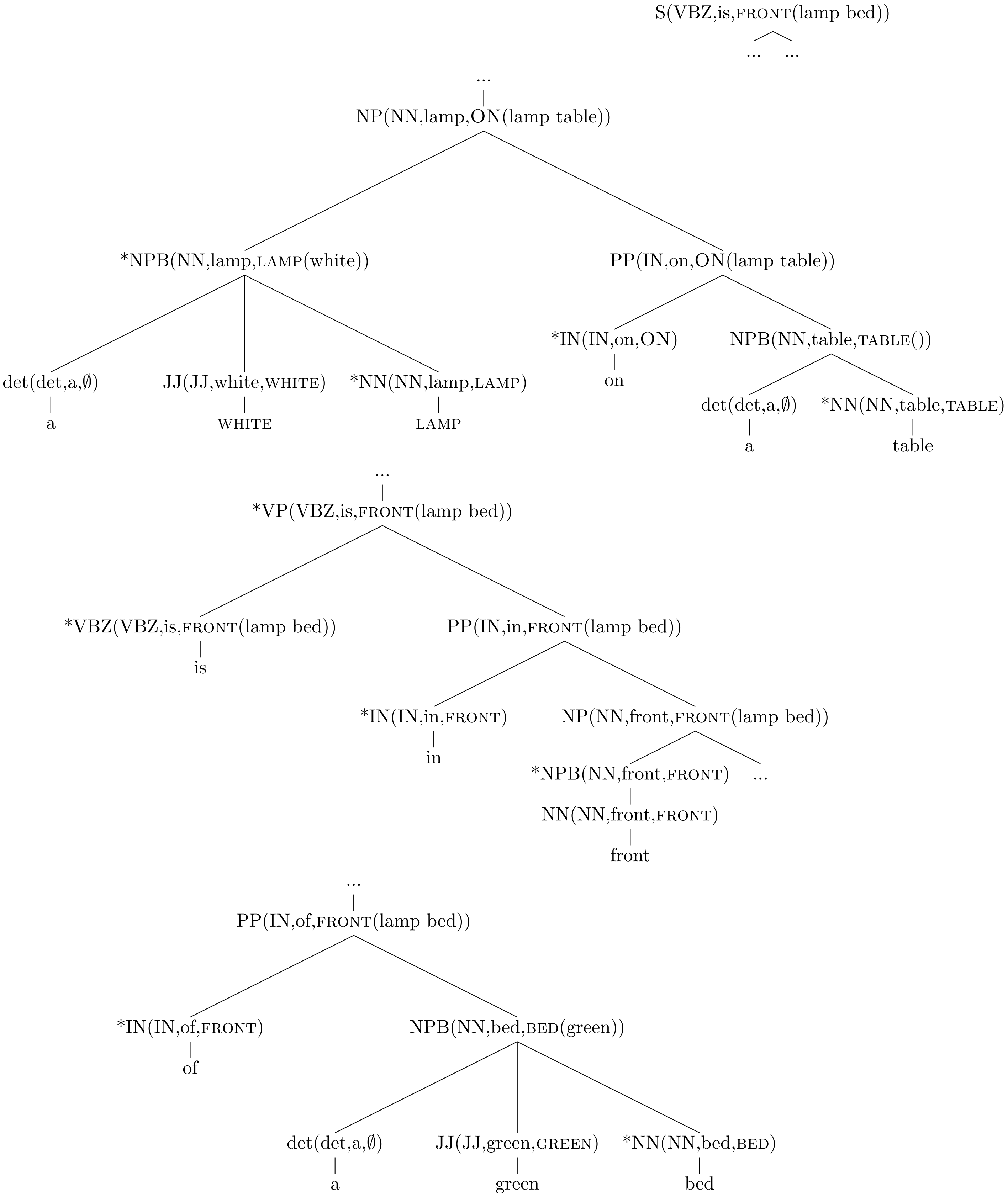

Figure 2.3: A parse tree for the example sentence augmented with semantic information. Nodes have the form LABEL(TAG, WORD, PREDICATE(args))

Bibliography

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.

Chen, S. F. and Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–393.

Collins, M. (1997). Three generative, lexicalised models for statistical parsing. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 16–23. Association for Computational Linguistics.

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

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

Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for gibbs sampling. *Applied Statistics*, pages 337–348.

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

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

Paisley, J., Wang, C., and Blei, D. (2011). The discrete infinite logistic normal distribution. *arXiv preprint arXiv:1103.4789*.

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.

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).