Unsupervised Bayesian Induction of Spatial Relations from Captioned Scenes: A Dissertation Proposal (in progress)

Colin Reimer Dawson

June 4, 2014

1 Overview

The purpose of this dissertation is the development of a probabilistic generative model, and associated inference and learning algorithms, to model the recovery of semantic scene representations from the conjunction of natural language text (captions or narratives) and visual data (images or videos). This enterprise can be seen as lying in the realm of *grounded semantics*: given some text, the goal is to recover a semantic representation which can be connected to the physical environment. However, the connection between textual and visual information is not one-directional: the presence of the visual scene constrains the meaning of the text, but the text also constrains the interpretation of the raw visual signal.

The specific focus of the project is *relations* among *objects*. Starting with an image that is equipped with a human-generated caption, the inferential goals are to determine (1) what entities are in the scene (e.g., people, vehicles, furniture), (2) the features of those entities (size, color, category), (3) what entities are parts of/attached to/supported by/contained in what other entities, and (4) what/who is located where in relation to what/whom? I will adopt the approach that the representations that support answering these what/where questions are neither intrinsically visual nor intrinsically linguistic, but are abstract semantic propositions whose meanings are determined by (a) their (probabilistic) grounding in physical scenes, and (b) the kinds of utterances they (probabilistically) produce. For example, in a scene containing a blue couch and an end table, a possible semantic representation is that

there is a COUCH object that can be represented as a rectangular prism with major and minor axes, by virtue of which it has two END features, and a small table, which is NEAR-TO one of the ENDS of the COUCH.

The formal semantic representation is discussed in 4.1, along with the form of a prior distribution over semantics.

Conditioned on this semantic representation, a 3D ("minds eye") scene representation is generated by sampling physical parameters for the objects (e.g., length, width, height, average RGB value), as well as (x, z) coordinates on the ground plane. The minds-eye representation is projected onto an image plane to produce the 2D image. These representations and their corresponding priors and likelihoods are discussed in Sec. 4.2. Independently, a caption is generated by sampling a syntactic structure, which is then filled in with words. A candidate source of syntactic structure comes from the collapsed typed dependency structure of the Stanford parser (de Marneffe and Manning, 2008). Here, the syntactic representation of a sentence consists of labeled binary relations among words in the sentence. This representation, along with priors and likelihoods, are discussed in Sec. 4.3.

By performing simultaneous inference over several scene-caption pairs, the model will be able to learn a "vocabulary" of relations by parsimoniously explaining patterns of cooccurrence between features of the "minds eye" representation of the scene (e.g., distances between objects, relative orientations with respect to the perspective axis, etc.), and features of the syntactic structures.

2 Related Work

The proposed model is informed by two hitherto largely separate literatures: one on language understanding, and the other on scene understanding. To my knowledge, few published papers exist that attempt to take advantage of images and text as synergistic and complementary sources of information about "meaning", and those that do (Barnard and Forsyth, 2001; Barnard et al., 2003) mostly use individual word tags rather than natural language captions.

On the language understanding side, Tellex et al. (2011) employ a graphical model to understand grounded spatial relationships from text, and Makalic et al. (2008) present a probabilistic model to infer "instantiated concept graphs" (ICGs) from speech by performing ASR, parsing, relation induction, and finally grounding of semantic arguments. However, in both cases, the relational structure is assumed

to be available deterministically from a parse, whereas here, relations are inferred by learning a probabilistic mapping from training sentences that are grounded by images. Moreover, whereas the model in Tellex et al. (2011) is trained on images annotated with relations, the present model will learn from training data that consists only of text and unannotated images.

On the scene understanding side, Del Pero et al. (2011, 2012) develop probabilistic generative models that populate a room with furniture which is modeled with connected parallelepipeds (Schlecht and Barnard, 2009). Given a 2D image, a 3D room and furniture configuration is proposed from the posterior distribution by MCMC sampling.

The work in this dissertation builds on previous work in both computer vision and grounded semantics. Dawson et al. (2013) developed a "toy" model for learning to interpret spatial utterances in the context of a simple tabletop scene on which a small nubmer of objects are arranged. This work was limited, however, by several simplifying assumptions: first, the representation of the scene itself was assumed known, and utterances were assumed to be "about" a single object in view or a single location in space. Second, utterances were assumed to have a single, fixed parse, which was coerced into a specific "relation-landmark" form. Finally, a fixed "vocabulary" of spatial relations was defined *a priori*, with fixed groundings in space.

3 Summary of Novelty

In the present work, I will extend the spatial language model of Dawson et al. (2013) in several ways, and synthesize it with the scene understanding model of Del Pero et al. (2012), resulting in a generative model resembling (at a high level) the one depicted in Fig. 1.

- I will draw on computer vision work by Del Pero et al. (2011, 2012) by allowing uncertainty about what is in view in the first place, and where it is, to be integrated with uncertainty about the meaning of the language. Together the uncertainty should be lower than that obtained using either side alone.
- I will relax the previous assumption of Dawson et al. (2013) that sentences come pre-equipped with a parse tree, and instead incorporate the parse as a random variable in the generative model. In the model in 1, the probability of a parse is determined by the semantic representation of the scene, as well as a scene-independent grammatical model such as that used by "off-the-shelf" parsers

trained on large annotated corpora that probabilistically parse "ungrounded" utterances.

- I will allow the model to learn sparse, but not one-to-one, dependencies between semantic attributes (e.g., the existence of a couch) and syntactic/lexical features (e.g. the `subject` argument of the verb `is`). In generative terms, given a semantic representation, Ψ, a syntactic representation, Υ, is generated, where $P(\Upsilon|\Psi)$ factors into terms corresponding to the nodes and edges of typed dependency trees (Fig. 5). A number of conditional independence properties will be assumed in this factorization. First, conditioned on Ψ, subtrees of the dependency tree are assumed to be conditionally independent when all common ancestors are known. Second, the distribution for a subtree is assumed to depend on only a small number of semantic properties. Which properties are needed, however, is not assumed to be known *a priori*, but will be learned by clustering. The details of this factorization are given in Sec. 4.3.
- Finally, I will attempt to allow the model to learn a language-specific relational ontology in an unsupervised manner by attempting to parsimoniously explain regularities in the relationship between visual and text data. A simple example might be the distinction between tight and loose containment, which are distinguished lexically in Korean as 'kkita' and 'nehta', but which are both expressed using the preposition 'in' in English. McDonough et al. (2003) found that English-speaking adults did not systematically distinguish visual examples of tight vs. loose containment in a categorization task, whereas Korean-speaking adults distinguished both. A model that attempts to explain the words used to describe a scene should infer the existence of two relational categories when the distinction is expressed in the language, but should prefer one category if no linguistic distinction is made (assuming the visual distributions are the same in either case).

4 Model and Representations

4.1 (Amodal) Semantic Representation

A sketch of a generative model for scene-caption pairs is shown in Fig. 1. A scene has a "topology", Φ, which contains a set of **objects** with associated ontological categories (tables, chairs, people, clothing, cars, etc.), associated unary object **methods**

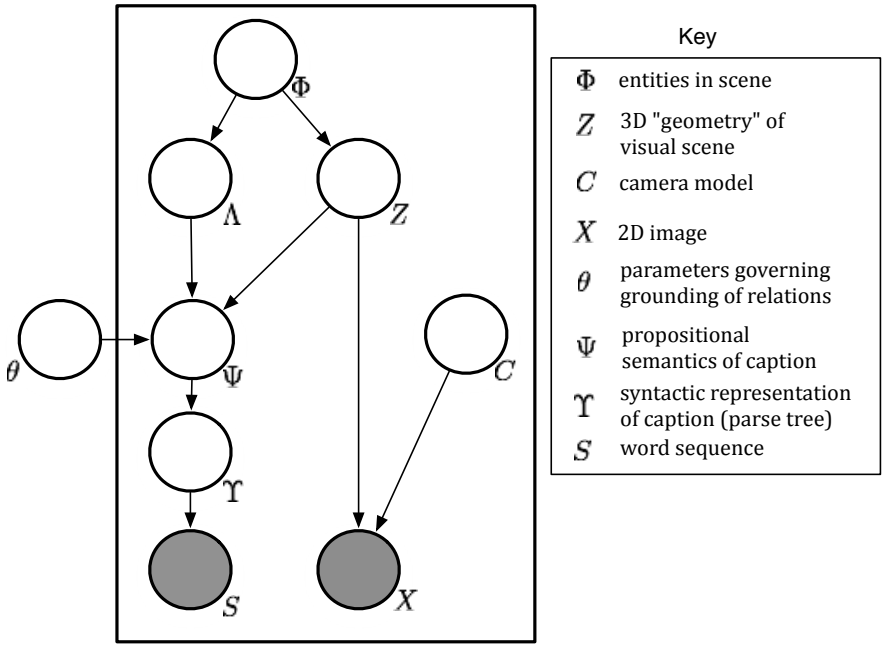
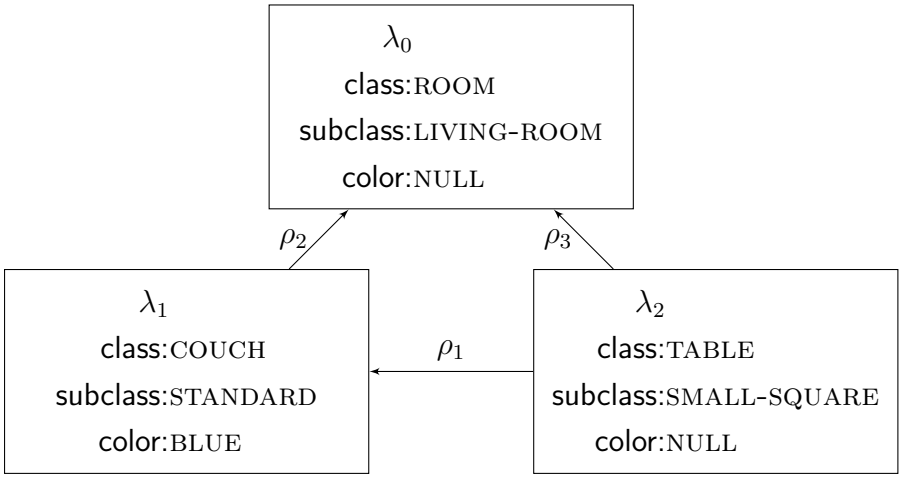


Figure 1: A sketch of a generative model, expressed as a Bayes net, giving rise to scene-caption pairs. Nodes inside the box correspond to scene-caption-specific variables, and are objects of inference from a particular scene-caption pair. Nodes outside the box represent "general knowledge" about the language and/or scene domain, and are either specified *a priori* or are learned over many scene-caption pairs.

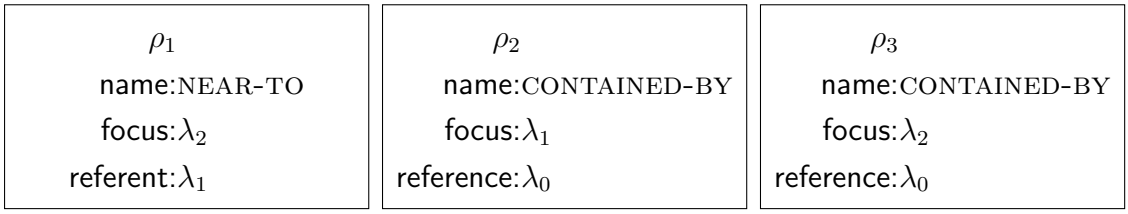
that return **attributes** (e.g., category, subcategory, color, parent object), and binary propositional **relations** between objects (e.g., proximity, relative orientation, containment). Instances of binary location relations are LEFT-OF, BEHIND, NEAR, and IN, though this vocabulary could be built up from primitive features. Each relation has a focal object and a reference entity (in the case that a set of objects is the reference, the set is instantiated as a separate object which is the parent of its members).

For the example scene containing a room with an end table near a blue couch, the semantic representation might be as in Fig. 4.

Constructing a prior over such representations is fairly straightforward. I will attempt to define here a general model that assumes no prior knowledge about object or scene categories, but in practice the set of objects and scenes will most likely be assumed known for purposes of easy compatibility with the existing room model of Del Pero



(a) Example of object and attribute representation



(a) Relation representation

Figure 4: Example semantic representation

et al. (2012).

First, each scene is populated with an initial object, λ_0, constituting the scene itself, and it is assigned a context class (in the example, this is a room). To incorporate the ability to discover new scene types, the distribution $\boldsymbol{\pi}^{ctx}$ over context classes is represented by the stick-breaking process (Sethuraman, 1994). We write $\boldsymbol{\pi}^{ctx} \sim \text{Stick}(\alpha^{ctx})$, where the concentration parameter α^{ctx} represents the tendency to lump or split contexts. If the set of contexts is instead pre-defined, a symmetric Dirichlet prior can be used. Given the context class, c, a subclass c_i may be sampled from another stick-breaking process, where again the concentration parameter represents the tendency to make fine-grained distinctions; however, I will henceforth assume that there is only one context level.

Then, given the scene context, a number of objects, $\lambda_1, \ldots, \lambda_n$ is generated, where n

is drawn from a context-sensitive Poisson distribution with a parameter μ_c^{obj} drawn from a Gamma prior whose parameters are fixed. (Alternatively, if the set of contexts is fixed, the μ_c^{obj} could simply be calibrated from annotated data)

Each object is then assigned a category. I will define two models of object category: one where the set of objects is unknown, the other where the set is known in advance.

If the object categories are unknown, then the context-dependent distribution $\boldsymbol{\pi}_c^{cat}$ over object categories is a draw from a hierarchical Dirichlet Process (HDP) (Teh et al., 2006) with concentration α^{cat} and base measure $\boldsymbol{\pi}^{cat}$. In an HDP, $\boldsymbol{\pi}^{cat}$ is itself a draw from a DP with concentration α_0^{cat} and base measure $\boldsymbol{\pi}_0^{cat}$. In this case, $\boldsymbol{\pi}_0^{cat}$ is drawn from a stick breaking process, Stick(α_0^{cat}). This model has the desirable property that the number of object categories need not be defined *a priori*, but different contexts tend to include certain categories. Here, α_0^{cat} governs the degree of lumping or splitting of object categories, whereas α^{cat} governs the extent to which the same objects appear across scene types: if α^{cat} is small, different contexts will tend to have distinct sets of objects; if it is large, contexts have similar object distributions.

Although the above description in principle allows for object types to be discovered, in practice we will, at least at first, assume a fixed set of K object categories. In this case, the hierarchy of Dirichlet Processes can be replaced by a hierarchy of Dirichlet distributions (i.e., $\boldsymbol{\pi}_c^{cat} \overset{i.i.d.}{\sim} \text{Dir}(\alpha^{cat}\boldsymbol{\pi}^{cat})$ and $\boldsymbol{\pi}^{cat} \sim \text{Dir}(\alpha_0^{cat}\mathbf{1}/K)$), but the qualitative properties are similar: α_0^{cat} governs the overall variability in object types, whereas α^{cat} represents the tendency of scene types to share object distributions.

For each object of category o, its attributes (subcategory, color label, parent object, etc.) are sampled from a category-specific distribution (e.g., $\lambda.\mathsf{color} \sim \boldsymbol{\pi}_o^{col}$). In the case of unknown object categories, these distributions (other than subcategory) are sampled from a parent measure, e.g., $\boldsymbol{\pi}^{col}$, which is drawn from a stick-breaking process, Stick(α^{col}); for known categories, the distributions might have independent *a priori* Dirichlet priors.

Finally, for each pair of objects, a location relation is generated by sampling from a distribution over relations, $\boldsymbol{\pi}^{rel}$, which in the case of a fixed vocabulary of relations has a Dirichlet prior, and in the case of unspecified relations is a sample from another stick-breaking process, Stick(α^{rel}).

4.2 Visual Representation

Given the scene topology, Φ, a 3D "geometric" representation, Z, of the objects in the scene is generated. This includes specific numeric dimensions and locations of objects, as well as properties such as "intrinsic" color values in numeric color space. This step can be thought of as the "mind's eye" representation of the scene. Object- and relation-specific "grounded semantics" are represented by θ, which contains information about, for example, typical dimensions of tables; how near is *near* when the arguments are cups on a table vs. furniture in a room, where are the continuous boundaries between "tight" and "loose containment" (if such a distinction is made at all), etc.

Some features of Z are plausibly modeled as independent given Φ and θ, such as the dimensions of individual tables, but the joint distribution over locations of objects in the scene is likely better handled by a Markov Random Field (MRF) than a directed representation, as multiple soft constraints must be satisfied simultaneously. This seems more conducive to an "energy minimization" approach than to sequential generation. If generation is needed, a stochastic simulator whose dynamics reflect the energy functions in the MRF could be used. However, at least at first, much of this component of the model will be drawn directly from Del Pero et al. (2012).

Given the camera, C, the 3D representation is projected onto the image plane, together with perturbations and noise arising from lighting, etc., to produce an image (or video) X. In principle, aspects of the camera perspective (e.g. occlusion), and specifically visual features (e.g. lighting) are likely to influence the choice of semantic representation, Ψ, to be expressed in the caption; however, as a first approximation, ψ and Z are assumed to be conditionally independent given Φ.

```
blue
```

4.3 Linguistic Representation

Given a scene topology Φ, a subset, $\Psi \subset \Phi$, of the available features are chosen to be expressed. First a set of relations is chosen, and then for each relation argument, a subset of the features of each object argument is selected. The distribution $p(\Psi|\Phi)$ is governed by considerations of salience (which arises, for example, by departures from prior expectations, as well as the behavioral significance of a given relation), descriptiveness (how well could the full scene be recovered from Ψ), and sparsity

(a desire for descriptions to be concise). More will be said about this in a future iteration.

Each sentence is assumed to be associated with a single semantic relation, represented by ψ_0. A semantic feature-vector, $(\psi_0, \psi_1, \ldots, \psi_M)$, can be constructed for the sentence by systematically stepping through the attributes of the arguments of ψ_0, some of which have the value `omitted`, and others of which have the value `null`. By including these gaps, the semantic feature vector has fixed length, and for each fixed m, ψ_m is the same feature across sentences[1]. In the example, ψ_0 has the value NEAR-TO. The next two features are the categories of the focus and referent; in this case $\psi_1 =$ TABLE and $\psi_2 =$ COUCH. The next two features are the subcategory and color of the table; etc.

Given a semantic feature vector (henceforth labeled Ψ by a slight abuse of notation), the speaker generates a syntactic tree, Υ. The tree containins a hierarchically organized set of syntactic relations among lexical content. In the Stanford typed dependency system (de Marneffe and Manning, 2008), syntactic relations roughly correspond to syntactic roles (subject-of, direct object-of) and function words (prepositions, conjunctions), whereas their arguments are "content words" (nouns, verbs, adjectives and adverbs). For example, one parse of the sentence "A small end table is near a blue couch" contains the syntactic relations:

- `root(ROOT-0, is-5)`
- `nsubj(is-5, table-4)`
- `prep_near(is-5, couch-9)`
- `amod(table-4, small-2)`
- `nn(table-4, end-3)`
- `amod(couch-13, blue-12)`

which can be represented as a tree as in Fig. 5.

A tree with J nodes can be decomposed into *arc* features, $\Upsilon^a = (\upsilon^a_{1,1}, \ldots, \upsilon^a_{1,n_1}, \ldots \upsilon^a_{J,1}, \ldots, \upsilon^a_{J,n_j})$, corresponding to the dependency labels between words, and *text* features $\Upsilon^t =$

[1]Eventually recursive semantics should be represented, where the argument of a relation can be another relation. In this case the number of semantic features associated with a given sentence is potentially unbounded, but systematic ordering of the features should still be possible by, for example, leaving a series of gaps whenever an argument is not an object.

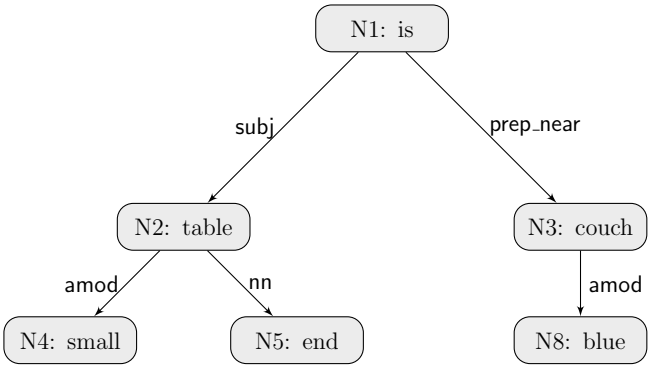


Figure 5: Tree representation of collapsed typed dependencies induced by the Stanford parse of the sentence "A small end table is near the left end of a blue couch."

$(\upsilon_1^t, \ldots, \upsilon_J^t)$ whose value is the word associated with each node. In the example, we have

- $\upsilon_{1,\cdot}^a = (\mathsf{subj}, \mathsf{prep\_near})$
- $\upsilon_1^t = \text{is}$
- $\upsilon_{2,\cdot}^a = (\mathsf{amod}, \mathsf{nn})$
- $\upsilon_2^t = \text{table}$

  ...

For each dependency sequence, one dependency is identified as the *head*, an adaptation of Collins (1997, 2003) to dependency trees. This arc is generated first. Then, adjacent dependency arcs are generated outward in both directions until a value of `stop` is reached. The *syntactic history* for an arc is its parent word (`ROOT` for the root), whether it is the head (0), left of the head (-1) or right of the head (1), and its preceding (toward the head) sibling arc. The syntactic history for a text feature is its incoming arc, its parent word, the adjacent siblings of its parent arc, and its sibling word in the direction of the head (equal to `NULL` if it is in head position). The syntactic history for the text feature associated with the jth node is denoted by $\mathbf{h}_j^t$; similarly the syntactic history of the ℓth arc descending from the jth node is denoted by $\mathbf{h}_{j\ell}^a$.

In terms of the generative model, the distribution $P(\Upsilon_i|\Psi_i)$, corresponding to the ith observation, factors into the probability of each feature $\upsilon_{ij\ell}^a$ and υ_{ij}^t conditioned on the *semantic context*, Ψ, and the *syntactic history*, $h_{ij\ell}^a$ and h_{ij}^t, respectively. That

is,

$$P(\Upsilon_i|\Psi_i) = \prod_{j=1}^{J} \pi^t(\upsilon_{ij}^t|h_{ij}^t, \Psi_i) \prod_{\ell=1}^{n_j} \pi^a(\upsilon_{ij\ell}^a|h_{ij\ell}^a, \Psi_i) \tag{1}$$

Due to the high dimensionality of both semantic context and syntactic history, there will be very little data available for any particular combination. This is a familiar problem for parsing models that attempt to condition production probabilities on more than very local context (even very local context can result in data sparsity problems when individual words are involved). A standard approach 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 $\mathbf{c} = (c_1, c_2, \ldots, c_n)$, then back-off smoothing estimates the probability $p(f|c_1, \ldots, c_n)$ 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

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

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, $\lambda_1, \ldots, \lambda_{n-1}$ 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).

The back-off interpolation method of Collins (2003) has many desireable properties, and could be easily "made Bayesian" through the use of a multilevel HDP model. Let $p_{\mathbf{c}_k} \overset{i.i.d.}{\sim} DP(\alpha_{\mathbf{c}_{k-1}}, p_{\mathbf{c}_{k-1}})$ for $k = 1, \ldots, n$, where $p_{\mathbf{c}_n}$ represents the fully contextualized distribution, $p_{\mathbf{c}_{n-1}}$ represents the distribution with the last context feature deleted, $p_{\mathbf{c}_0} \sim \text{Stick}(\alpha')$, and the concentrations have Gamma hyperpriors. In this model, context distributions are pulled toward a parent distribution with one fewer context feature, with the degree of smoothing at each level governed by a context-dependent concentration parameter. The αs play an analogous role to $1-\lambda$ in Collins (2003), and by placing Gamma priors on their values, contexts with high diversity of productions yield higher posterior αs, and hence more smoothing.

However, a problem with this model is that it requires context features to be ordered *a priori*, which is not easily done when the context includes a mix of semantic and syntactic features. Moreover, the precedence order of semantic features is different depending on the particular syntactic history. For example, at the root of the tree, relation features matter most, and color features matter hardly at all; whereas the reverse might be true following an `amod` arc. Hence, if relation features take precedence, there is no mechanism to enforce similarity of productions that share landmark features if their relations differ; and vice-versa. Ultimately, backing off of semantic features in a fixed order is not reasonable.

Instead of a hierarchical context model, I will assume that each syntactic feature is generated from an infinite mixture of discrete measures, where the mixing weights are informed by the context. Let π^f be the distribution over syntactic feature f (where $f \in \{a, t\}$), and for generality, index instances by $j \in 1, \ldots, J$, and define the *covariate vector*, $x_j = (h_j, \Psi_j)$. Then assume

$$\boldsymbol{\pi}^f | x_j = \sum_{k=1}^{\infty} w_k(x_j) \pi_k^f \tag{3}$$

where the π_k^f are multinomial distributions over the values of syntactic feature f, and $w(x_j)) = (w_1(x_j), w_2(x_j), \ldots)$ represents the covariate-dependent mixing weights.

By defining

$$\pi_k^f \sim Dir(\alpha^f \pi_0^f) \tag{4}$$
$$\pi_0^f \sim Dir(\gamma^f W^f) \tag{5}$$
$$\alpha^f \sim \mathcal{G}(a_\alpha^f, b_\alpha^f) \tag{6}$$
$$\gamma^f \sim \mathcal{G}(a_\gamma^f, b_\gamma^f) \tag{7}$$
$$\tag{8}$$

where α^f is a concentration parameter governing the similarity of the π_k to each other (small values result in peaked components), π_0^f represents the marginal distribution of syntactic features (aggregating across mixture components), and γ^f represents the prior degree of uniformity in the distribution of feature f, and W_f is a prior distribution over values of feature f. The a and b hyperparameters are fixed.

By introducing indicator variables, $\mathbf{z} = \{z_j \in \mathbb{N}\}$, one for each feature instance the weight vector w can be thought of as a covariate-dependent prior on partitions of

the js in the data set. The goal is to define the prior on partitions in such a way that similar covariate vectors are grouped together with high probability; that is, if x_j and $x_{j'}$ are "close", then $w(x_j)$ and $w(x_{j'})$ should be "close" (as distributions) as well.

One attractive possibility comes from the PPMx model of Müller et al. (2011), who introduce a probability model for covariates conditional on cluster membership. The distribution of partitions conditioned on the covariates is then proportional to the product of a prior *cohesion function*, which depends only on cluster sizes, via a Dirichlet Process, and a *similarity function*, which gives a likelihood over partitions given the within-cluster distribution of covariates. Quintana et al. (2012) introduce a method for variable-selection within this model, allowing different clusters to depend on different covariates.

While this model works well when the covariates are observed, in the present case some of them are latent. This presents a problem for the PPMx model, because the normalization factor for the covariate-conditioned partition distribution is no longer constant, as it depends on the overall distribution of covariates in the entire sample. Moreover, although the normalization is a finite sum, it is a sum over all possible partitions of the data, the number of which for n observations is given by the nth Bell number, which grows combinatorially in n.

Hence, it is desirable to model $p(\mathbf{z}|\mathbf{x})$ directly, rather than as the inversion of $p(\mathbf{x}|\mathbf{z})$. One possibility comes from Dahl (2008). Like the PPMx model, his approach begins with a Dirichlet Process prior over partitions, which can be modeled as the stationary distribution of the Polya urn model, in which observations are successively assigned to clusters in proportion to the number already there, and to a new cluster in proportion to a concentration parameter, α. Dahl's approach modifies the Polya urn to take covariates into account by incorporating a *distance function* between observations in covariate space. Let $d_{ii'} = d(x_i, x_{i'})$ represent a distance between two covariate vectors, and let $d^{\max} = \max_{i'} d_{ii'}$. The similarity measure between an observation and a cluster is then given by $g_i(R_k) = c_i \sum_{j \in R_k} (d^{\max}_{i,k} - d_{ii'})$, with the constant chosen so that $\sum_k g(x_i, R_k)$ equals the number of observations.

For a given concentration parameter α, this normalization renders the probability of assigning an observation to a new cluster independent of the covariates. This is undesirable if a new observation's dissimilarity on covariates suggests *a priori* that it likely belongs to a new cluster. Blei and Frazier (2011) relax the normalization assumption and instead define the similarity measure on an absolute scale relative to α, giving rise to the Distance-Dependent Chinese Restaurant Process ($ddCRP$).

Unlike in the traditional CRP, where customers are assigned to tables, in the ddCRP, customers are linked to other customers (in proportion to their similarity), resulting in indirect assignment of tables by identifying connected components of the customer graph.

Here I use a hybrid between these two approaches, retaining the "table-assignment" structure of Dahl (2008)'s model, but using the absolute similarity scale of Blei and Frazier (2011).

Consider first the case of a single categorical covariate, and let $d_{ii'} = 1 - \delta_{x_i,x_{i'}}$, where δ is the Kronecker delta function. Then, $g_i(R_k) = \sum_{i' \in R_k} \left(1 - (1 - \delta_{x_i,x_{i'}})\right) = \sum_{i' \in R_k} \delta_{x_i,x_{i'}}$. That is, g_i is proportional to the number of observations already in cluster k such that $x_{i'} = x_i$.

This is easily extended to multiple categorical covariates, $x_i = (x_{i1}, \ldots, x_{iM})$, by letting $d_{ii'}$ be a function of the L^p norm of the binary vector of position-wise differences. For example, if $d_{ii'} = \frac{1}{M} \sum_{m=1}^{M} (1 - \delta_{x_{im},x_{i'm}})$, then the distance is the proportion of the covariate features that differ, and $g_i(R_k) = \frac{1}{M} \sum_{i' \in R_k} \sum_{m=1}^{M} \delta_{x_{im},x_{i'm}}$ is the total number of matching feature values in a cluster normalized by the number of features per observation. This corresponds to the L_0 (Hamming) distance between categorical vectors, normalized by dimension.

The above is easily generalized to other types of covariates by replacing δ with an arbitrary similarity measure ranging from 0 to 1 (for example, an exponentially decaying function of an arbitrary distance metric). In the present application, it is important that different clusters be allowed to depend on different covariates, as discussed above. Hence it is necessary that the similarity measure be allowed to depend on cluster-specific parameters. This can be accomplished by computing coordinate-wise similarities as desired (ranging from 0 to 1), and defining overall similarity as a weighted average of the coordinate-wise similarities, where the weights are allowed to vary across clusters. Note that the assumption that similarities range from 0 to 1 is not restrictive, since any desired rescaling can be incorporated into the weights.

Let $\pi_k^{sim} = (\pi_{k1}^{sim}, \ldots, \pi_{km}^{sim})$ be a probability vector, and define

$$g_{ik}(R_k) = \sum_{i' \in R_k} \sum_{m=1}^{M} w_{km}^{sim} \delta_{x_{im},x_{i'm}} \tag{9}$$

Then, for cluster k, similarity of the mth covariate contributes to the overall similarity in proportion to π_{km}^{sim}. The π_k^{sim} have a common Dirichlet prior with concentration

α^{sim} and base measure w_0^{sim}, which in turn has a Dirichlet prior with concentration γ^{sim} and symmetric base measure[2]. The interpretation of w_0^{sim} is as the vector of mean weights for each covariate across clusters, with α^{sim} determining the similarity of weights across clusters and γ^{sim} determining the similarity of mean weights across features.

To summarize, given a configuration with K existing components, we have

$$p(z_{ij}|z_{-ij}, x_i, \mathbf{w}) \propto \begin{cases} \sum_{i' \in R_k} \sum_{m=1}^{M} w_{km}^{sim} \delta_{x_{im}, x_{i'm}} & z_{ij} = k \\ \alpha^{cl} & z_{ij} = K+1 \end{cases} \tag{10}$$

$$\pi_k^{sim}|\alpha^{sim}, w_0^{sim} \sim Dir(\alpha^{sim} w_0^{sim}) \tag{11}$$

$$w_0^{sim}|\gamma^{sim} \sim Dir(\alpha^{sim} \mathbf{1}/M) \tag{12}$$

$$\alpha^{sim} \sim \mathcal{G}(a_{\alpha^{sim}}, b_{\alpha^{sim}}) \tag{13}$$

$$\gamma^{sim} \sim \mathcal{G}(a_{\gamma^{sim}}, b_{\gamma^{sim}}) \tag{14}$$

In order to encourage sparsity of the weight vectors, α^{sim} should be less than 1 with high prior probability. This yields weight vectors with mass concentrated in a few components. The prior on γ^{sim} is determined by the extent to which covariates are believed to be equally important *a priori*.

Letting $v_k^{f*} = \{v_j^f : j \in R_k\}$, where $f \in \{a, t\}$, letting $s = 1, \ldots, S^f$ index distinct values of syntactic feature f, and writing $n_{k,s}^f$ for the number of times feature f takes the value s in cluster k, we have the likelihood

$$p(\boldsymbol{v}^f|\mathbf{z}, \pi_0^f, \ldots, \pi_K^f) = \prod_{k=1}^{K} p(v_k^*|\pi_k^f) \tag{15}$$

$$= \prod_{k=1}^{K} \prod_{i \in R_k} \pi_k^f(v_{ki}^*) \tag{16}$$

$$= \prod_{k=1}^{K} \prod_{s=1}^{S^f} (\pi_{k,s}^f)^{n_{k,s}^f} \tag{17}$$

$$\pi_k^f \sim Dir(\alpha^f \pi_0^f) \tag{18}$$

$$\pi_0^f \sim Dir(\gamma \mathbf{1}/S^f) \tag{19}$$

[2]There is no additional difficulty in using a non-symmetric base measure here if there is prior information about the relative importance of the covariates

5 Posterior Inference

Given a scene-caption pair, (S, X) , the goal of inference is to recover the scene topology, Φ, and perhaps additionally the "highlighted" aspects of the scene, Ψ. Since the model yields a posterior distribution over these variables and not a single value, it is necessary to define an objective function, an optimum of which is the desired output of "perception". A sensible solution is to define a loss function, $L(\hat{\Phi}, \hat{\Psi}, \Phi, \Psi)$, where estimates are penalized according to their propositional deviations (some version of "edit distance") from the truth: for example, there would be some cost associated with failing to represent an object, another cost associated with hallucinating an object, and other (presumably smaller) costs associated with misrepresenting the features or (propositional) locations of objects. Misrepresenting objects and relations highlighted by Ψ could have a higher cost, depending on the inference objective.[3]

The optimal solution would then minimize the posterior risk,

$$(\Phi^*, \Psi^*) = \arg\min_{(\hat{\Phi}, \hat{\Psi})} r(\hat{\Phi}, \hat{\psi}|S, X, Y) = \mathbb{E}_{\hat{\phi}, \hat{\psi}}[L(\hat{\phi}, \hat{\psi}, \phi, \psi)]$$

where the expectation is taken with respect to the posterior distribution $P(\Phi, \Psi|S, X, Y)$, itself approximated with a set of representative hypotheses sampled using Markov Chain Monte Carlo.

The overall MCMC algorithm is Gibbs sampling, with MH acceptance-rejection steps in some blocks. Apart from an additional likelihood term corresponding to the prior on relations given the mind's eye, the sampling scheme from Del Pero et al. (2012) can be used largely unchanged.

Next I give a Gibbs sampling algorithm for the caption/semantics side of the model.

5.1 Sampling Cluster Indicators

Recall that we model the distribution of syntactic feature υ_{ij} in the context of semantic representation $\Psi_i = (\psi_{i1}, \ldots, \psi_{iM})$ and syntactic history $h_{ij} = (h_{i,j,1}, \ldots, h_{i,j,H_f})$

[3] As an alternative to this "edit distance"-based loss function, if the semantic interpretation of the scene leads to some action (perhaps based on "upstream" higher-level inference), then the specification of the loss function could be deferred to a more abstract variable on which a decision is based.

as an infinite mixture of multinomials:

$$p(\upsilon_{ij}^f|\Psi_i, h_{ij}) = \sum_{k=1}^{\infty} w_k(\Psi_i, h_{ij})\pi_k(\upsilon_{ij}) \tag{20}$$

Inference is simplified by introducing an indicator variable z_{ij}^{cl} to represent which mixture component generated υ_{ij}, yielding

$$\pi(\upsilon_{ij}^f|\Psi_i, h_{ij}) = \sum_{k=1}^{\infty} p(z_{ij}^{cl} = k|\Psi_i, h_{ij})\pi_k(\upsilon_{ij}) \tag{21}$$

By renumbering, let j range over all i, j combinations, so that z_{ij}^{cl} becomes z_j^{cl}. Let $x_j = (x_{j,1}, \ldots, x_{j,M})$ represent the combined context (semantic context and syntactic history features together). Then the assignment of observations to mixture components ("clusters") is modeled using a modified Polya urn:

$$p(z_j^{cl}|z_{-j}^{cl}, x_j, \mathbf{w}) \propto \begin{cases} \sum_{j' \in R_k} \sum_{m=1}^{M} w_{km}^{sim} \delta_{x_{jm}, x_{j'm}} & z_{ij}^{cl} = k \\ \alpha^{cl} & z_{ij}^{cl} = K+1 \end{cases} \tag{22}$$

$$\tag{23}$$

where δ is a similarity measure between $x_{i,m}$ and $x_{i',m}$ assumed to range between 0 and 1 (e.g., the Kronecker delta), and π_k^{sim} is the weight vector governing which context features matter for determining membership in cluster k. The π_k^{sim} have a two-level hierarchical Dirichlet prior with a fixed, symmetric base measure at the top level, and base measure π_0^{sim} at the lower level, and respective concentrations α^{sim} and γ^{sim}. By introducing an additional set of indicator variables, $\mathbf{z}^{sim}$ ranging over features $1, \ldots, M$ with $z_j^{sim}|z_j^{cl} \sim \pi_{z_j}^{sim}$, the above simplifies to

$$p(z_j^{cl}|z_{-j}^{cl}, \mathbf{z}_{-j}^{sim}, x_j) \propto \begin{cases} \sum_{j' \in R_k} \delta_{x_{j z_{j'}^{sim}}, x_{j' z_{j'}^{sim}}} & z_j^{cl} = k \\ \alpha^{cl} & z_j^{cl} = K+1 \end{cases} \tag{24}$$

The distribution of a single z_j^{sim} given only a cluster assignment z_j^{cl} and the other $z_{j'}^{sim}$ within the cluster, marginalizing out π_k^{sim} and π_0^{sim}, can be computed using the Dirichlet integral. First, integrating out π_k^{sim}, writing $n_{mk}^{sim} = \sum_{j' in R_k} I(z_{j'}^{sim} = m)$ for the number of times feature m is selected in cluster k, $n_{\cdot k}^{sim} = \sum_{m=1}^{M} n_{mk}^{sim}$ and $n_{m\cdot}^{sim} = \sum_{k=1}^{K} n_{mk}^{sim}$,

$$p(z_j^{sim} = m|z_j^{cl}, \mathbf{z}_{-j}^{sim}, x_j, \alpha^{sim}, \pi_0^{sim}) = \frac{n_{m z_j^{cl}}^{sim} + \alpha^{sim}\pi_{0,m}^{sim}}{n_{\cdot z_j^{cl}}^{sim} + \alpha^{sim}}. \tag{25}$$

This expression can be understood as the sum of two components using the Chinese Restaurant Franchise metaphor Teh et al. (2006). Here, observations are "customers", z^{sim} values are "dishes", and mixture components are "restaurants". Given that there are n_{mk}^{sim} customers in restaurant k already assigned to tables eating dish m, the next customer sits at one of those tables with probability proportional to n_{mk}^{sim}, and at a new table with probability proportional to α^{sim}. If the customer starts a new table, she orders a dish from the global menu with probabilities drawn from π_0, and orders dish m with probability $\pi_{0,m}$. Hence there are two ways to get dish m: either by joining an existing table eating that dish, or by ordering it anew from the global menu. Combining these two probabilities gives the expression in 27.

To integrate out π_0^{sim}, it is necessary to distinguish which "path" was taken to each dish: that is, to keep track of how many distinct tables (and not just customers) there are eating each dish. To see why, note that π_0^{sim} only generates a new observation when a new table is created, and so only tables are evidence for mass at a particular location of π_0^{sim} (see Wallach et al. (2009) for a detailed derivation).

Hence, we introduce another set of indicators, τ_j^{sim}, indexing the table number of customer j, where tables are indexed globally by $t = 1, \dots T$. Given $\mathbf{z}_{-j}^{sim}$, z_j^{cl}, and $\boldsymbol{\tau}_{-j}^{sim}$, τ^{sim} is assigned to an existing value, t, in cluster z_j^{cl} in proportion to the number of observations in that cluster already assigned to that value, and to a new value with probability in proportion to α^{sim}. That is,

$$p(\tau_j^{sim}|z_j^{cl}, \mathbf{z}_{-j}^{sim}, \boldsymbol{\tau}_{-j}^{sim}, \alpha^{sim}) = \begin{cases} \frac{\#(\tau^{sim}=t \cap z_j^{cl}=k)}{n_{\cdot z_j^{cl}}^{sim}+\alpha^{sim}} & t \leq T \\ \frac{\alpha^{sim}}{n_{\cdot z_j^{cl}}^{sim}+\alpha^{sim}} & t = T+1 \end{cases} \tag{26}$$

It is then possible to integrate out π_0^{sim}. Let $\hat{n}_{km}^{sim}$ be the number of tables in cluster k with value m, let $\hat{n}_m^{sim} = \sum_{k=1}^K \hat{n}_{km}^{sim}$ be the global number of tables with value m, and let $\hat{n}_\cdot^{sim} = \sum_{t=1}^T \hat{n}_m^{sim}$. Then we have (see Wallach et al. (2009))

$$p(z_j^{sim} = m|z_j^{cl}, \mathbf{z}_{-j}^{sim}, \boldsymbol{\tau}_{-j}^{sim}, x_j, \alpha^{sim}, \gamma^{sim}) = \frac{n_{m z_j^{cl}}^{sim} + \alpha^{sim} \left(\frac{\hat{n}_m^{sim} + \frac{\gamma^{sim}}{M}}{\hat{n}_\cdot^{sim} + \gamma^{sim}} \right)}{n_{\cdot z_j^{cl}}^{sim} + \alpha_0^{sim}} \tag{27}$$

Finally, we need to take the observed υ into account. The distribution of υ does not depend on z^{sim}, and so sampling z^{sim} is just as in (27). The conditional posterior of the z^{cl} depends both on the prior in (27) and the likelihood

$$p(\upsilon_j|z_j^{cl}) = \pi_{z_j^{cl}}(\upsilon_j) \tag{28}$$

Hence, conditioned on the π_k, we have

$$p(z_j^{cl}|\upsilon_j, z_{-j}^{cl}, \mathbf{z}_{-j}^{sim}, x_j, \pi_0, \ldots, \pi_K) \propto \begin{cases} \pi_k(\upsilon_j) \sum_{j' \in R_k} \delta_{x_{jz_{j'}^{sim}}, x_{j'z_{j'}^{sim}}} & z_j^{cl} = k \\ \pi_0(\upsilon_j)\alpha^{cl} & z_j^{cl} = K+1 \end{cases} \tag{29}$$

However, as before, we can integrate out the random measures. Let $n_{sk}^f = \sum_{j \in R_k} I(\upsilon_j = s)$ be the number of times that feature f takes the value s in cluster k. Let $n_{\cdot k}^f = \sum_{s=1}^{S^f} n_{sk}^f$, let $n_{s\cdot}^f = \sum_{k=1}^K n_{sk}^f$. As before, let $\hat{n}_{ks}^f$ be the number of tables in cluster k with value s, let $\hat{n}_{s}^f = \sum_{k=1}^K \hat{n}_{ks}^f$ be the global number of tables with value m, and let $\hat{n}_{\cdot}^f = \sum_{t=1}^T \hat{n}_{s}^f$. Then

$$p(\tau_j^f|z_j^{cl}, \mathbf{z}_{-j}^f, \boldsymbol{\tau}_{-j}^f, \alpha^f) = \begin{cases} \frac{\#(\tau^f = t \cap z_j^{cl} = k)}{n_{\cdot z_j^{cl}}^f + \alpha^f} & t \leq T \\ \frac{\alpha^f}{n_{\cdot z_j^{cl}}^f + \alpha^f} & t = T+1 \end{cases} \tag{30}$$

and

$$p(z_j^{cl}|\upsilon_j, z_{-j}^{cl}, \boldsymbol{\tau}^f, \mathbf{z}_{-j}^{sim}, x_j) \propto \begin{cases} \frac{n_{\upsilon_j, z_j^{cl}}^f + \alpha^f \frac{\hat{n}_{\cdot,\upsilon_j}^f + \frac{\gamma^f}{S^f}}{\hat{n}_{\cdot,\cdot}^f + \gamma^f}}{n_{\cdot, z_j^{cl}}^f + \alpha^f} \sum_{j' \in R_{z_j^{cl}}} \delta_{x_{jz_{j'}^{sim}}, x_{j'z_{j'}^{sim}}} & z_j^{cl} \leq K \\ \frac{n_{\upsilon_j,\cdot}^f + \frac{\gamma^f}{S^f}}{n_{\cdot,\cdot}^f + \gamma^f} \alpha^{cov} & z_j^{cl} = K+1 \end{cases} \tag{31}$$

5.2 Sampling Concentration Parameters

There are several parameters representing concentrations of Dirichlet distributions or Dirichlet Processes. The conditional posteriors for these parameters have simple forms, but are not exponential family distributions, and so exact sampling is not

possible. The following factors involve concentration parameters

$$\pi_k^f|\pi_0^f, \alpha^f \sim Dir(\alpha^f \pi_0^f) \tag{32}$$
$$\alpha^f \sim \mathcal{G}(a_{\alpha^f}, b_{\alpha^f}) \tag{33}$$
$$\pi_0^f|\gamma^f \sim Dir(\gamma^f \mathbf{u}) \tag{34}$$
$$\gamma^f \sim \mathcal{G}(a_{\gamma^f}, b_{\gamma^f}) \tag{35}$$
$$\pi_k^{sim}|\pi_0^{sim}, \alpha^{sim} \sim Dir(\alpha^{sim} \pi_0^{sim}) \tag{36}$$
$$\alpha^{sim} \sim \mathcal{G}(a_{\alpha^{sim}}, b_{\alpha^{sim}}) \tag{37}$$
$$\pi_0^{sim}|\gamma^{sim} \sim Dir(\gamma^{sim} \mathbf{u}) \tag{38}$$
$$\gamma^{sim} \sim \mathcal{G}(a_{\gamma^{sim}}, b_{\gamma^{sim}}) \tag{39}$$
$$\mathbf{z}^{cl}|\mathbf{x} \sim CRP(\alpha^{cl}) \tag{40}$$

5.3 Sampling Ψ and Auxiliary Parameters

The propositional semantics represented by Ψ has in its Markov blanket (1) the parameters, θ governing the grounding of relations, (2) the mind's eye representation, Z, (3) the parse tree, Υ and (4) the set of auxiliary parameters governing the PPMx model. Z together with θ provides the prior on Ψ, while Υ and the auxiliary parameters (such as the indicators various αs and their hyperparameters) determine the likelihood. Let

$$\Psi = \begin{pmatrix} \psi_{10} & \dots & \psi_{1,M} \\ \vdots & \ddots & \vdots \\ \psi_{I,1} & \dots & \psi_{I,M} \end{pmatrix}$$

be the matrix of semantic features, where i indexes sentences and m indexes features (with $m = 0$ denoting the relation label, $m = 1$ and 2 denoting the category and color of the focus object, $m = 3$ and $m = 4$ denoting the category and color of the reference object). Let

$$\Upsilon_i^t = \begin{pmatrix} \upsilon_{i,1}^t & h_{i,1,1}^t & \dots & h_{i,1,H_t}^t \\ \vdots & \vdots & \ddots & \vdots \\ \upsilon_{i,J_i}^t & h_{i,J_i,1}^t & \dots & h_{i,J_i,H_t}^t \end{pmatrix}$$

be the representation of the words features of sentence i, where $\nu_{i,j}^t$ is the jth word, and $h_{i,j,s}^t$ is the sth history feature of that word. Let $\mathbf{z}_i^t = \left(z_{i,1}^t, \dots, z_{i,J}^t\right)$ be the vector

of cluster indicators associated the word features of sentence i. Similarly, let

$$\Upsilon_i^a = \begin{pmatrix} \upsilon_{i,1,1}^a & h_{i,1,1,1}^a & \cdots & h_{i,1,1,H_a}^a \\ \vdots & \vdots & \ddots & \vdots \\ \upsilon_{i,1,L_{i,1}}^a & h_{i,1,L_{i,1},1}^a & \cdots & h_{i,1,L_{i,1},H_a}^a \\ \vdots & \vdots & \ddots & \vdots \\ \upsilon_{i,J_i,1}^a & h_{i,J_i,1,1}^a & \cdots & h_{i,J_i,1,H_a}^a \\ \vdots & \vdots & \ddots & \vdots \\ \upsilon_{i,J_i,L_{i,J_i}}^a & h_{i,J_i,L_{i,J_i},1}^a & \cdots & h_{i,J_i,L_{i,J_i},H_a}^a \end{pmatrix}$$

be the representation of the arc features of sentence i, where $\nu_{i,j,\ell}^t$ is the ℓth arc emitted by the jth word, and $h_{i,j,\ell,s}^a$ is the sth history feature for that arc. Finally, let $\mathbf{z}_i^a = \left(z_{i,1,1}^a, \ldots, z_{i,1,n_{i,1}}^a, \ldots, z_{i,J_i,1}^a, \ldots, z_{i,J_i,L_{i,J_i}}^a\right)$ be the vector of cluster indicators associated with the arc features of sentence i.

Then, $p(\psi_{i,m}|\Psi_{-(i,m)}, \Upsilon, \mathbf{z}, \theta, Z_i)$ is proportional to

$$p(z_{i\cdot}^t|\psi_{i,m}, \Psi_{-(i,m)}, \mathbf{z}_{-i}^t)p(z_{i\cdot}^a|\psi_{i,m}, \Psi_{-(i,m)}, \mathbf{z}_{-i}^a)p(\psi_{i,m}|\theta, Z_i) \tag{41}$$

References

Barnard, K., Duygulu, P., Forsyth, D., De Freitas, N., Blei, D. M., and Jordan, M. I. (2003). Matching words and pictures. *The Journal of Machine Learning Research*, 3:1107–1135.

Barnard, K. and Forsyth, D. (2001). Learning the semantics of words and pictures. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 2, pages 408–415. IEEE.

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.

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.*

Dawson, C. R., Wright, J., Rebguns, A., Valenzuela Escárcega, M. A., Fried, D., and Cohen, P. R. (2013). A generative probabilistic model for learning spatial language. Submitted to 2013 International Conference of Development and Learning.

de Marneffe, M.-C. and Manning, C. D. (2008). The Stanford typed dependencies representation. In *COLING Workshop on Cross-framework and Cross-domain Parser Evaluation.*

Del Pero, L., Bowdish, J., Fried, D., Kermgard, B., Hartley, E., and Barnard, K. (2012). Bayesian geometric modeling of indoor scenes. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2719–2726. IEEE.

Del Pero, L., Guan, J., Brau, E., Schlecht, J., and Barnard, K. (2011). Sampling bedrooms. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 2009–2016. IEEE.

Makalic, E., Zukerman, I., Niemann, M., and Schmidt, D. (2008). A probabilistic model for understanding composite spoken descriptions. In *PRICAI 2008: Trends in Artificial Intelligence*, pages 750–759. Springer.

McDonough, L., Choi, S., and Mandler, J. M. (2003). Understanding spatial relations: Flexible infants, lexical adults. *Cognitive psychology*, 46(3):229–259.

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).

Quintana, F. A., Müller, P., and Papoila, A. L. (2012). Cluster-specific variable selection for product partition models.

Schlecht, J. and Barnard, K. (2009). Learning models of object structure. In *NIPS*, volume 2, page 3.

Sethuraman, J. (1994). A constructive definition of Dirichlet processes. *Statistica Sinica*, 4:639–650.

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).

Tellex, S. A., Kollar, T., Dickerson, S. R., Walter, M. R., Banerjee, A., Teller, S., and Roy, N. (2011). Approaching the symbol grounding problem with probabilistic graphical models. *AI Magazine*, 32(5).

Wallach, H., Mimno, D., and McCallum, A. (2009). Rethinking lda: Why priors matter. *Advances in Neural Information Processing Systems*, 22:1973–1981.