Bayesian modeling of scenes and captions

Colin Dawson, Luca DelPero, Clayton Morrison, Mihai Surdeanu, Gustave Hahn-Powell, Zachary Chapman and Kobus Barnard

April 5, 2013

Overview

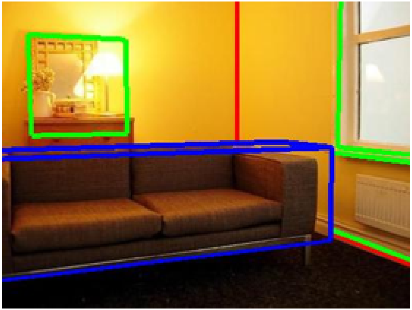


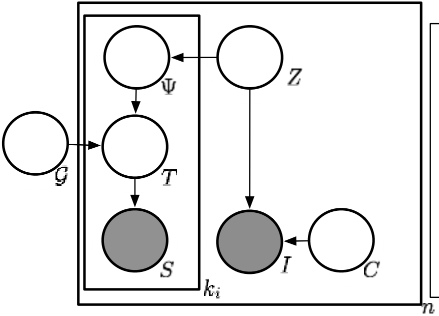
Figure: "There is a brownish couch to the left of the window. It has a dresser behind it with a mirror on top of it that has a white frame."

- Goal: Use 2D image and human-provided caption as joint evidence for the 3D arrangement of objects in the scene.
- Alternatively: Use hypotheses about 3D scene as top-down information about language structure.
- We tackle both goals in one probabilistic model.

Extensions to Related Work

- Delpero et al. [4, 3] developed a model and Bayesian inference methods to infer 3D structure of rooms from 2D images.
- Dawson et al. [2] developed a probabilistic model to learn spatial language in the context of virtual 2D scenes.
- Present model is a synthesis of these two, with the addition of a more sophisticated grammar and parsing model than in [2].
- Sentences and 2D images are assumed to be *jointly* generated from 3D configurations, with a conditional independence assumption to make inference tractable.

Model Overview



Key

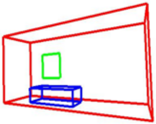
Z	3D "geometry" of visual scene
C	camera model
I	2D image
Ψ	propositional semantics of caption
$\mathcal{G}$	probabilistic grammar
T	parse tree
S	word sequence

Figure: Bayes net representation of the probabilistic model. Each of n scenes, $I_i, \ldots, I_n$, is paired with a caption consisting of k_i sentences, $S_{i1}, \ldots, S_{ik_i}$.

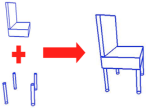
- Scene: What is there? $p(Z)$
- Image: What is seen? $p(I|Z, C)$
- Pragmatics: What is said? $p(\Psi|Z)$
- Syntax: How is it said? $p(T|\mathcal{G}, \Psi)$

Scene and Camera Representation

- Scene Z consists of a room container, r, and m objects, $o_1, \ldots, o_m$.



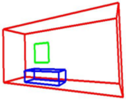
Room is a rectangular box with three location parameters (relative to camera) and three size parameters.



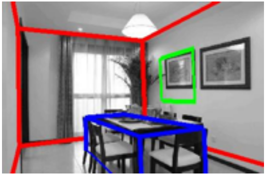
Objects are modeled as a configuration of reusable parts with category-dependent priors on dimensions.

Camera has a 3D orientation and a focal length.

Image Likelihood



$\rightarrow$



$\rightarrow$

- 3D edges are projected onto image plane along with associated edges
- Likelihood $p(I|Z, C)$ measures similarity between detected edges in 2D and hypothesized edges projected from 3D.

Pragmatic Model

- Sentences assumed to be about a *target object* (i.e., the semantic subject), λ. E.g., $\lambda = couch$. Absent other information, prior is uniform over possible objects in the scene.
- Features of λ (e.g. color, size) expressed using a set of discrete symbols (e.g., $\textsc{brown}$). Location of λ expressed using a binary *relation*, ρ (e.g. $\textsc{left-of}$), to a *base object*, β (e.g. $\textsc{window}$).
- Probability of a given elaboration, E, (e.g., $E = \textsc{brown}$, $E = \textsc{left-of}(\textsc{couch}, \textsc{window})$) is proportional to the value of an *applicability function*, $A(E)$, with the probability of no further elaboration ($E = null$) proportional to a constant.
- A new elaboration is sampled for each object until $E = null$ is chosen.

Pragmatic Model

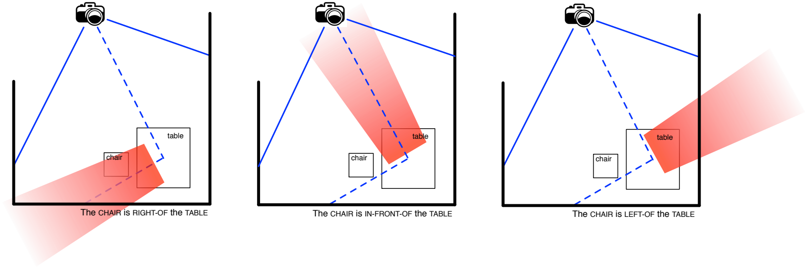


Figure: $A(E)$ can be thought of as a likelihood function. For spatial relations, it is the probability of target position given relation and base; for colors, probability of continuous color statistics given categorical label, etc. In the figure, the "heatmap" represents the degree of applicability.

Semantic Tree Representation

- The result is a tree of predicate, object, attribute, and relation nodes with a head/argument/complement structure.

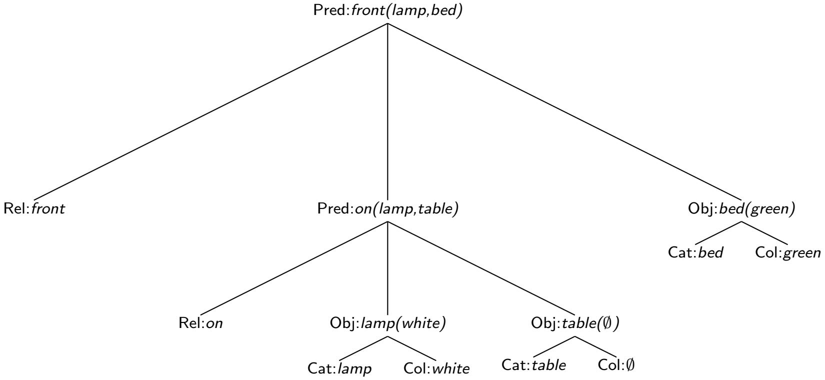
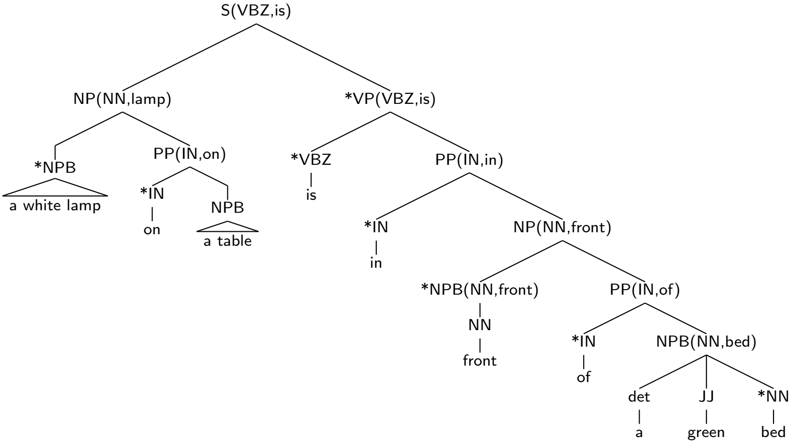


Figure: A possible semantic tree for the sentence "A white lamp on a table is in front of a green bed"

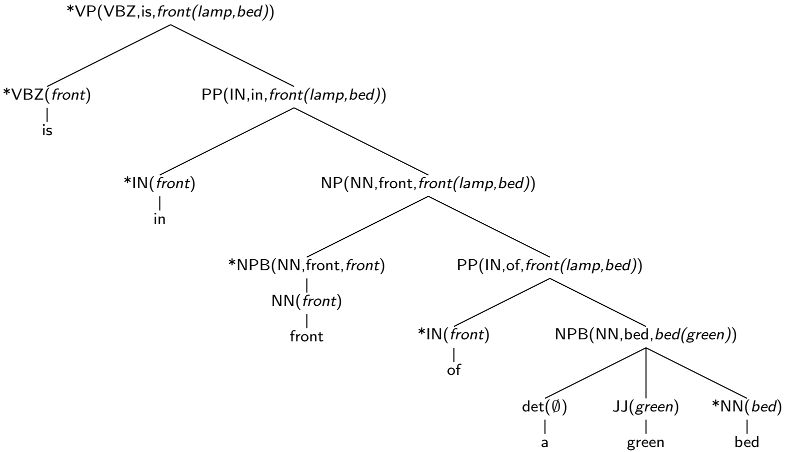
Syntactic Representation

- Syntactic model based on Collins' [1] head-driven dependency parser: constituents keep track of the word and PoS tag associated with their head; modifiers are generated conditioned on head.



Syntactic Representation

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



Training

- Grammar parameters in $\mathcal{G}$ consist of conditional probabilities of two types of syntactic productions (head and modifier), as in Collins [1], as well as new semantic “productions” (type of step taken on the semantic tree)
- All production events are conditioned on syntactic and semantic “history”.
- Parameters estimated from training data using the back-off smoothing method in [1] (extended to include semantic features and history).

Training

- Data consists of photographs of rooms, equipped with captions elicited from human subjects. Subset used for training is annotated with gold constituent parse, and semantic chunk labels, e.g.:

$$\text{A } [\text{white}]_{white} \ [\text{lamp}]_{lamp} \ [\text{on}]_{on} \text{ a } [\text{table}]_{table}$$
$$[\text{is in front of}]_{front} \text{ a } [\text{green}]_{green} \ [\text{bed}]_{bed}$$

- Human involvement only needed to correct first-pass automated annotation.
- Given a parse and a chunked sequence, the semantic tree is determined, and production probabilities can be learned.

Inference

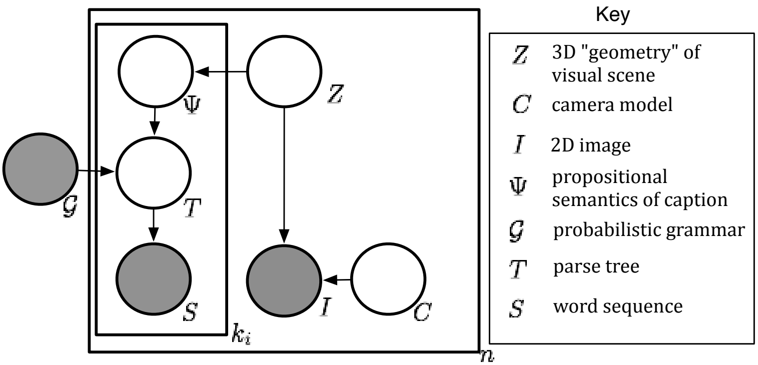


Figure: Bayes net representation of the probabilistic model after training.

- After training, we treat $\mathcal{G}$ as known (for now)
- Goal: Infer posterior, $p(T, \Psi, Z, C | S, I, \mathcal{G}) \propto$ $p(Z)p(C)p(I|Z, C)p(\Psi|Z)p(T|\Psi, \mathcal{G})\mathbb{1}(S \equiv T)$ using MCMC.

Results in progress...

M. Collins.

Head-driven statistical models for natural language parsing.

Computational linguistics, 29(4):589–637, 2003.

C. R. Dawson, J. Wright, A. Rebguns, M. A. Valenzuela Escárcega, D. Fried, and P. R. Cohen.

A generative probabilistic model for learning spatial language.

In *Proceedings of the 2013 International Conference of Development and Learning*, 2013.

L. Del Pero, J. Bowdish, D. Fried, B. Kermgard, E. Hartley, and K. Barnard.

Bayesian geometric modeling of indoor scenes.

In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2719–2726. IEEE, 2012.

L. Del Pero, J. Guan, E. Brau, J. Schlecht, and K. Barnard.

Sampling bedrooms.

In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 2009–2016. IEEE, 2011.