A generative probabilistic framework for learning spatial language

Colin R. Dawson, Jeremy Wright, Antons Rebguns, Marco Valenzuela Escárcega, Daniel Fried and Paul R. Cohen
School of Information, The University of Arizona, Tucson, AZ 85721–0077
Email: cdawson@email.arizona.edu

***Abstract*—The language of space and spatial relations is a rich source of abstract semantic structure. We develop a probabilistic model that learns to understand utterances that describe spatial configurations of objects in a tabletop scene by seeking the meaning that best explains the sentence chosen. The inference problem is simplified by assuming that sentences express symbolic representations of (latent) semantic relations between referents and landmarks in space, and that given these symbolic representations, utterances and physical locations are conditionally independent. As such, the inference problem factors into a symbol-grounding component (linking propositions to physical locations) and a symbol-translation component (linking propositions to parse trees). We evaluate the model by eliciting production and comprehension data from human English speakers and find that our system recovers the referent of spatial utterances at a level of proficiency approaching human performance.**

I. Introduction

Imagine that a friend asks you to "Bring me the thing toward the far corner of the table." This simple request requires fairly sophisticated cognitive processing. You must first identify that she is referring to something on a table, in particular, one with corners. Then, you must orient the table to distinguish "far" vs. "near" corners. Finally, if there is more than one object near a "far" corner, but one is also very near the edge, you might tend to favor the other one, reasoning that it would have been easy to ask for "the one near the edge".

We model a version of this problem, with a particular focus on recovering the referent of spatial utterances like "the thing toward the far corner of the table", where the only information available about the intended object is its position relative to a landmark in the scene. Beginning with no knowledge about the meanings of words, but equipped with a small vocabulary of spatial relations (e.g., containment, proximity, ordering in cardinal directions) and abstract representations of objects and their parts (e.g., a table can be represented as a line with ends and a middle, or as a rectangle with corners, quadrants and edges), our model learns probabilistic correspondences between sentences and abstract spatial relations between referents and landmarks by "observing" a teacher repeatedly generating an utterance and pointing to a location in space.

Clearly a method that involved supervised learning based on observed propositional semantics would not be developmentally plausible, as children do not get to observe symbolic meaning directly, and so crucially, the abstract relations are never made overt to our learner. Rather, the locations are probabilistically assigned to abstract landmark-relation pairs using simple prior "applicability functions", along with the assumption that a semantic representation is chosen so as to contrast the intended referent with other potential referents — that is, that the speaker's goal is not simply to say something true but to communicate. These assumptions are intended to reflect the developmental situation facing a social agent learning to communicate in a cooperative environment. We represent these constraints using a generative probabilistic model of sentences conditioned on referents, with propositional relations playing a mediating role. After training, the model infers intended referents associated with novel sentences by computing a posterior distribution over landmark-relation pairs, and then averaging together location distributions for individual pairs to generate a "heatmap" over locations. The fact that the model is generative and probabilistic and the inference Bayesian gives rise to the sort of "analysis-by-synthesis" behavior that integrates pragmatics with semantics without the need to posit separate systems to handle the two.

The defining distinctive features of our approach are threefold. First, semantic representations and syntactic representations are cleanly separated, so that sentences need not have clean syntactic argument structures in order to be understood. Second, it is not necessary to explicitly represent the meanings of individual words in order to represent the meanings of the utterances they occur in. Finally, since the probabilistic model is generative, it naturally carries out counterfactual reasoning: the posterior probability of a meaning is reduced when another phrase would have expressed it more easily.

II. Model

We model utterances that take place in the context of a physical scene, π, observed by both speaker and listener. This scene provides a set of potential representations of objects, their features, parts, and physical locations that the speaker might want to refer to. The listener's end goal is then to recover the referential intent of the speaker. In the case of scenes instantiated in the physical world, the system must construct π using its visual system. We describe the formal representation of π in section II-A.

We assume that the speaker intends to refer to an object or location within π. This *referent* is represented by λ in the graphical model in Fig. 1. For example, λ might be the blue cup at the far end of the table, or it might simply be a set of spatial coordinates. We assume that the uttered sentence, S, does not represent the physical referent λ directly; rather, the connection between λ and S is mediated by a propositional

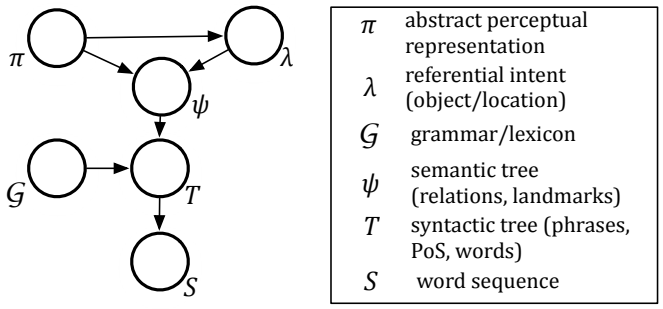


Fig. 1: A graphical model representation of the probabilistic language-generation process

representation, denoted by ψ in the graphical model, as well as a syntactic parse T. For example, the blue cup can be referred to based on its proximity to the far end of the table, or based on being behind another cup. Given ψ, parse trees and locations are assumed conditionally independent. This factorization simplifies the problem of learning mappings from sentences to referents, as it affords generalization of meanings across environments.

We represent ψ as a semantic tree consisting of a *relation*, ψ_{rel} (e.g. NEAR-TO, LEFT-OF) between a *referent*, λ, (e.g. `cup2`) and a *landmark*, ψ_{lmk}, which can be another object, or a part of the table itself (e.g. the far right corner relative to the speaker). The referent and the landmark each get *cast* as a particular abstract mathematical entity, such as a point, line, rectangle or region. A different set of legal relations as well as "child" landmarks is induced depending on the casting. We discuss this casting in section II-A.

We expect the factoring of ψ into referent, relation, and landmark(s) to be sufficient for most spatial relations, with the qualifier that not every part of ψ will necessarily be expressed in language. For instance, the relation "to the north" makes use of the earth as a landmark for absolute orientation, but does not express a landmark phrase. We also do not currently handle relations with multiple landmarks, such as BETWEEN. Doing so would require adapting the present method for selecting landmarks, which currently involves constructing a probability distribution over individual landmarks (see II-B); however, the rest of the method would extend easily.

In order to engage in communicative interaction in a shared physical world, the agent must have representations for (a) the physical world (π), (b) the propositional semantics assumed to be conveyed by an utterance (ψ), (c) the syntactic structure of sentences (T), and (d) links between each of these levels. We describe each of these in detail in the remainder of this section. Finally, recovering the intended referent involves maximizing $P(\lambda|S) \propto P(S|\lambda)P(\lambda)$, by integrating out T and ψ. We discuss a simple inference method in Sec. III.

A. Scenes

We construct an abstract perceptual representation of a table-top scene observed by both speaker and listener instantiated in either virtual or real world environments. In the case of a real world scene, an abstract representation is constructed from camera data processed to create 3D point clouds which are segmented to find objects. Whether we are parsing a virtual or real world scene, two distinct objects are constructed: (1) a *scene*, π, which acts as a container for a collection of *landmarks*, where each landmark is a representation of a corresponding object from the environment, and (2) a *perspective*, which is a representation of the position and orientation of the observer. Landmarks contain information about an ABSTRACTREPRESENTATION chosen for the object, its type or class, its color, unique name and, in the case where it was constructed as a part of another ABSTRACTREPRESENTATION, a parent landmark. We have defined a number of hierarchically organized geometric primitives that can used to represent any object in the scene. For example, objects can be cast as a POINT, a LINE a RECTANGLE, or a SURFACE.

Each representation has a number of landmarks inherent to its shape, regardless of the real world object it represents. For example a LINE will have a *start*, *middle* and *end* landmarks, all of which in turn are instances of POINTs. A more complex example is a RECTANGLE, which also has a number of landmarks inherent to its shape (corners, edges, left/right planes, etc.). Here corners are POINTs, edges are LINEs and left/right planes are SURFACEs. Objects are not restricted to one representation: for example, a long table may be cast as both a RECTANGLE, with corners and sides, and a Line, with two ends. Both representations are available as distinct choices of landmark in the generative model.

B. Spatial Relations

The grounded "meaning" of a relation ψ_{rel} is modeled using a (perspective-dependent) *applicability function*, A_{rel}, which assigns for each referent-landmark pair $(\lambda, \psi_{\text{lmk}})$ an *applicability* between 0 and 1. For instance, if a blue cup is to the left of a red block from the viewer's perspective, then $A_{\text{LEFT-OF}}(\texttt{bluecup}, \texttt{redblock}) = 1$. Relations from any source (such as exploration based-learning) could potentially be used, as long as a suitable applicability function can be constructed. We currently have manually specified three classes of spatial relations: CONTAINMENT, DISTANCE, and ORIENTATION. This set of spatial relations is not meant to cover every relation that speakers could use, but rather some of the most commonly used ones. Although all relations currently in use have arity 2, with only one landmark argument, there is no special difficulty in constructing an applicability function for a relation involving multiple landmarks; the only qualitative change required would be to the prior over landmarks (see II-C below).

Containment Relations Because we represent the scene in only two dimensions, our CONTAINMENT class has only one relation, ON, with $A_{\text{ON}}(\lambda, \psi_{\text{lmk}}) = 1$ for any referent entirely within the boundaries of the landmark, and 0 otherwise. This is meant to account for tabletop scenes where objects are on the table, but not on each other. Allowing intermediate values for referents partially contained by their landmark would be a fairly simple extension, but has so far been unnecessary.

Distance Relations There are two relations in this class, NEAR-TO and FAR-FROM. Each has a *degree* parameter, with three levels. Applicability of distance relations is a sigmoid function (a truncated Gaussian CDF) over Euclidean distances. $A_{\text{FAR-FROM}}(\lambda, \psi_{\text{lmk}})$ increases with the distance between λ and ψ_{lmk}, whereas $A_{\text{NEAR-TO}}(\lambda, \psi_{\text{lmk}})$ decreases. The degree parameter governs the distance at which the sigmoid crosses 0.5.

Orientation Relations This class contains four relations: LEFT-OF RIGHT-OF, FRONT-OF, and BEHIND. To determine their applicabilities, the scene is divided into quadrants with the center of the landmark as the origin and the line from there to the perspective point determining the axes. $A_{\text{LEFT-OF}}(\lambda, \psi_{\text{lmk}})$ is 1 for λ in the two appropriate quadrants and 0 elsewhere; similarly for the other three relations.

Orientation relations have a distance parameter similar to the degree parameter for the FAR-FROM relation, in order to represent ideas such as "far to the left", but in this case distance is the one-dimensional distance from the referent to the appropriate axis.

C. Sampling ψ

We factor $P(\psi|\lambda)$ as $P(\psi_{\text{lmk}}|\lambda)P(\psi_{\text{rel}}|\psi_{\text{lmk}}, \lambda)$. Based on the language usage we have seen in our experiments landmarks near to λ are more likely to be chosen, so we heuristically define $P(\psi_{\text{lmk}}|\lambda) \propto \mathcal{N}(\psi_{\text{lmk}}|\lambda, \mathbf{I}\sigma^2)$ evaluated at the nearest point on the landmark, with σ^2 proportional to the size of the table. $P(\psi_{\text{rel}}|\psi_{\text{lmk}}, \lambda)$ is obtained by normalizing the $A_{\text{rel}}(\lambda, \psi_{\text{lmk}})$ so that for fixed λ and ψ_{lmk} they sum to 1 over ψ_{rel}. That is, we have

$$P(\psi_{\text{rel}}|\psi_{\text{lmk}}, \lambda) = \frac{A_{\text{rel}}(\lambda, \psi_{\text{lmk}})}{\sum_{\text{rel}'} A_{\text{rel}'}(\lambda, \psi_{\text{lmk}})} \tag{1}$$

This produces a "distinctiveness effect" whereby relations are suppressed when there are many others that are applicable to the location (see, e.g., the dark strip immediately to the left of the landmark in Fig. 2, whose edges reflect discontinuities in the applicability IN-FRONT-OF and BEHIND, which creates a discontinuity in the probability of selecting LEFT-OF).

D. Syntax

We use parse trees, T, generated from a probabilistic context-free grammar (PCFG), augmented with a Markovian language model at the word level. Formally, if T contains non-terminal nodes $\{\eta_j\}_{j=1}^{M}$ and lexical nodes $\{w_k\}_{k=1}^{N}$, then $P(T|\psi)$ factors as

$$\prod_{j=1}^{M} P_\psi(\eta_j|\text{Pa}(\eta_j)) \prod_{k=1}^{N} P_\psi(w_k|\text{Fa}(w_k)) \tag{2}$$

where Pa gives the parent of a node and $\text{Fa}(\text{w}_\text{k})$ consists of a word's PoS tag and the preceding word.

$P_\psi(w_k|\text{Fa}(w_k))$ is a weighted average between a context-free parameter, $q_\psi^{\text{Pa}(w_k)}(w_k)$, conditioned only on the PoS tag and semantic features of the sentence, and a context-sensitive parameter, $r_\psi^{\text{Fa}(w_k)}(w_k)$, conditioned on the full $\text{Fa}(w_k)$. If the weights on the r_ψ, denoted below by $\alpha_\psi^{\text{Fa}(w_k)}$, are set to 0, the model reduces to a pure PCFG. In practice the α vary according to the frequency of the preceding word (see Eq. (4) in the next section). The individual q, r and α parameters are fit from training data as described in Sec. III-A, and collectively compose the node labeled $\mathcal{G}$ in the graphical model.

Since at present we are concerned with utterances that convey information about the spatial relation between a landmark and referent, we assume that relational phrases can be chunked into relation and landmark segments, in that order. Each segment is associated with a semantic representation, which determines which conditional probabilities apply in the production of a sentence.

for all $(S_i, \pi_i, \lambda_i) \in \mathcal{T}$ **do**
  Parse S_i to produce the n-best parses $T_{i1}, \ldots T_{in}$ with weights $P(T_{ik}|S)$ normalized over k.
  for all $\psi = (\psi_{\text{rel}}, \psi_{\text{lmk}})$ in the ontology **do**
    Compute $P(\psi_{\text{lmk}}|\lambda_i)$ and $A_{\text{rel}}(\lambda, \psi_{\text{lmk}})$ as in II-B.
  end for
  Compute $P(\psi_{\text{rel}}|\psi_{\text{lmk}}$ from the set of applicabilities using Eq. 1.
  Compute $P(\psi|\lambda) = P(\psi_{\text{lmk}}|\lambda)P(\psi_{\text{rel}}|\psi_{\text{lmk}}, \lambda)$.
  for all $k = 1, \ldots, n$ **do**
    for all CFG productions in T_i **do**
      Add a count with weight $P(\psi|\lambda_i)P(T_{ik}|S)$ to the contingency table entry corresponding to the production.
    end for
  end for
end for

Fig. 3: Algorithm used to learn $\mathcal{G}$. Training computation scales as the product of: the number of training sentences, the number of landmarks available per training scene, the number of relations in the vocabulary and the number of parses considered per sentence. For all results reported here, $n = 1$, i.e., only one parse is considered per sentence.

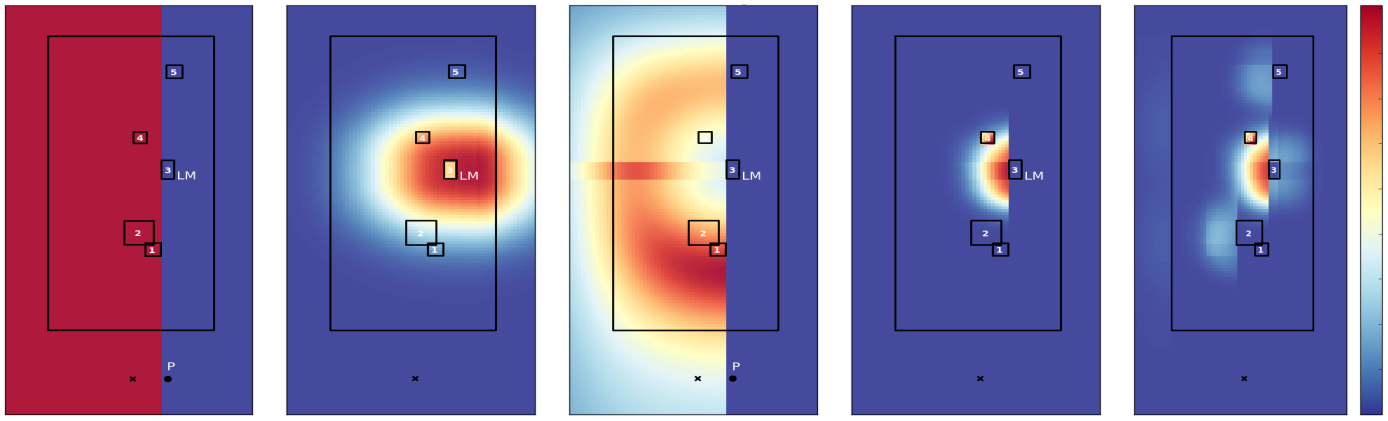
III. TRAINING AND INFERENCE

A. Learning the parameters of $\mathcal{G}$

The parameters governing $P(T|\psi)$ are learned from a training corpus, $\mathcal{T} = \{(S_i, \pi_i, \lambda_i)\}_{i=1}^{N}$, where S_i is an unparsed sequence of word tokens, and λ_i is a deictic referent (either an object or location) within an observable scene, π_i. The full algorithm is given in Fig. 3.

The chief difficulty here is that, while the probabilities in $\mathcal{G}$ relate a symbolic representation, ψ, to a syntactic tree, T, neither ψ nor T is directly available in the training data. Hence, T must be inferred from S (the parsing problem), and ψ from λ (the grounding problem). The parsing process is discussed in Sec. III-B.

Inferring ψ from λ is probabilistic, and proceeds according to the grounded semantics of the referent-relation-landmark tuples discussed in Sec. II-B: landmarks compete for expression based on their distance to the referent, and relations compete



$A_{\text{rel}}(\psi_{\text{lmk}}, \lambda)$ $P(\psi_{\text{lmk}}|\lambda)$ $\text{T}P(\psi_{\text{rel}}|\psi_{\text{lmk}}, \lambda)$ $P(\psi|\lambda)$ $P(S|\lambda)$

Fig. 2: Heat-maps for each stage of calculating $P(S|\lambda)$ for λs sampled across the entire scene, where S is the description "to the left of the orange block". In each case *rel* is LEFT-OF and *lmk* is Object 3, except for the rightmost graph which is a combination of many ψs weighted by the likelihood they give the description.

based on their applicability given the referent/landmark pair, yielding a set of probabilities, $\{P(\psi_{ij}|\lambda_i)\}_{j=1}^{K}$, where j indexes landmark-relation pairs. A weighted observation is added to a contingency table that counts how often each production has been observed in the context of the relevant component of ψ. Since ψ_i is unknown, a fractional observation is added for each $\psi_{ij}, 1 \leq j \leq K$ and each parse T_{ik}, whose weight is $P(\psi_{ij}|\lambda_i)P(T_{ik}|S_i)$, where $P(T_{ik}|S_i)$ is the normalized score of the parse containing the production. In the results reported here we consider only one parse per sentence, so $P(T_{i1}|S_i) = 1$. In addition, a smoothing count of ε is added at each level to represent the possibility of generating an as-yet unobserved grammatical production or lexical item. After training, we then have, for each nonterminal η, each word w, and each landmark-relation pair ψ_j,

$$P_{\psi_j}(\eta|\text{Pa}(\eta)) = \frac{\sum_{\{i,k:\text{Pa}(\eta)\to\eta\in T_{ik}\}} P(\psi_{ij}|\lambda_i)P(T_{ik}|S_i)}{\varepsilon + \sum_{\{i:\text{Pa}(\eta)\in T_{ik}\}} P(\psi_{ij}|\lambda_i)P(T_{ik}|S_i)} \tag{3}$$

The sum of this expression over nonterminal sequences η observed in training is less than 1, with the remaining mass reserved for novel productions. The context-free component, $q_{\psi}^{\text{Pa}(w)}(w)$, of the word probabilities is computed in exactly the same way, replacing ηs with words and $\text{Pa}(\eta)$ with part-of-speech tags, whereas the context-sensitive component $r_{\psi}^{\text{Fa}(w))}(w)$ is computed by restricting the sums in (5) to trees where the preceding word matches as well as the part of speech tag. Similarly, the context-sensitivity weights α are computed as

$$\alpha_{\psi_j}^{\text{Fa}(w)} = \frac{\sum_{\{i:(\text{Fa}(w)\in T_{ik}\}} P(\psi_{ij}|\lambda_i)P(T_{ik}|S_i)}{\sum_{\{i',k':\text{Pa}(w)\in T_{i'k'}\}} P(\psi_{i',k'}|\lambda_{i'})P(T_{i'k'}|S_i)}. \tag{4}$$

The full training algorithm is shown in Fig. 3. In practice, rather than update the q and α parameters for every ψ after every training item according to their probabilities, we sample 5 ψs for each item from the prior $p(\psi|\lambda)$ to reduce computation.

A general grammar would allow the probability of any production in a tree T to depend on the full specification of ψ. At present, we employ syntactic parse trees that have been modified to create a one-to-one correspondence between components of ψ (landmark and relation) and constituents of the parse tree. This alleviates a data sparsity problem in learning the production probabilities, but is admittedly unrealistic, and moreover limits the range of spatial sentences that can be interpreted. and so in ongoing work we are attempting to relax this hard correspondence and learn the extent to which particular production distributions depend on particular semantic features.

The resulting parameters in $\mathcal{G}$ are used to compute $P(T|\psi)$ as described in section II-D, which can then be used both as a production probability to generate sentences, and as a likelihood function to interpret sentences.

B. Parsing

The sentence S is first parsed using the Charniak PCFG parser [1] to obtain the n-best parse trees , which are then manipulated using a sequence of Tsurgeon [2] patterns into a flatter structure with a more transparent correspondence to the semantic features in ψ. The resulting trees (for results reported here, $n = 1$) are structured according to the inheritance structure of the representations discussed in section II-A. The procedure can be loosely described as pulling off the maximal NP starting from the right as the landmark segment. For more details about the syntax of Tsurgeon patterns, see [2].

The productions contained in each T_{ik} are extracted along with the relevant component of ψ. During training, each instance is added to a contingency table, as described in section III-A, which is used to train $\mathcal{G}$. During comprehension, $P(\psi|T_k)$ is computed for each parse of the test sentence,

and used to identify high probability referent locations in the physical scene as described in the next section.

C. Inferring intended referents

Since the semantics of an utterance are partitioned into a predicate component, ψ, and a physical component, λ, understanding can be defined either as correctly recovering the correct referent-relation-landmark tuple, or simply recovering the correct referent. Since ψ is never observed directly, inferences about λ require integrating over the uncertainty associated with inferences about ψ.

Given a new sentence, S, the trained system must first parse it to produce a syntactic tree, T. Given the tree, T, the scene π, and the trained grammar, $\mathcal{G}$, we want to compute $P(\lambda|T, \pi, \mathcal{G})$. For conciseness, we henceforth omit the explicit dependence on π and $\mathcal{G}$ from the notation. This is computed using Bayes' rule as

$$P(\lambda|T) \propto P(T|\lambda)P(\lambda) \tag{5}$$

We assume that $P(\lambda)$ is uniform over all possible referents available in the scene, but this assumption is easily modified if there are discourse-related reasons to suppose that the speaker is more likely to be discussing a particular object or region in space. The marginal likelihood $P(T|\lambda)$ is given by summing out ψ, as $\sum_\psi P(T|\psi)P(\psi|\lambda)$, where $P(T|\psi)$ is estimated from the training database as described in III-A, and $P(\psi|\lambda)$ is calculated from the relational distributions discussed in section II-B. Finally, $P(\lambda|S)$ can be obtained by averaging the $P(\lambda|T)$ over Ts with weights $P(T|S)$; but here we consider only one parse.

The full inference algorithm is given in Fig. 4.

IV. Experiment: Object Identification

To evaluate the model's ability to recover referential intent, we elicited sentences using Amazon Mechanical Turk (AMT). Participants were presented with one of five computer-generated tabletop scenes containing five objects of varying shapes, sizes and colors, each marked with a number (e.g., Fig. 5). On each trial, the participant was asked to give two descriptions of a numbered object: one using its intrinsic properties (e.g., color, shape), and one using only its location. There were no restrictions on the form of the responses. An example prompt is shown in Fig. 5. Each participant completed five trials, and was paid \$0.20. Participants produced a total of 4280 sentences.

After preprocessing (e.g., splitting clauses along conjunctions and commas), 2230 descriptions could not be parsed, and hence were not used. An additional 287 sentences parsed but did not have the required syntactic form. The remaining descriptions were randomly divided into a training set, containing 1818 items, and a test set, containing 295 items.

The training items were formatted as (S, λ, π) triples, where λ is the highlighted object. The production probabilities were learned from this data. After each set of training sentences, the model was presented with the test sentences and their accompanying scenes and chose the most likely object according to the posterior probability distribution, $P(\lambda|S)$ (where, again, locations were substituted for the objects). Performance on the test set was evaluated after each batch of 300 training sentences. A total of 10 training runs were performed, with the training set presented in a different order on each run. Averaged over training runs, the model's maximum a posteriori (MAP) object was the intended object 49.8% of the time (chance: 20%).

For comparison, the test sentences were also presented to AMT workers, who were shown the corresponding scene and asked to choose the most likely referent object. A total of

Parse S to produce the n-best parses $T_1, \ldots T_n$ with normalized scores $P(T_k|S)$.
for all $T_k, k = 1, \ldots, n$ **do**
  for all ψ in the ontology **do**
    for all nonterminal productions $\mathrm{Pa}(\eta) \to \eta \in T_k$ **do**
      Compute $P_\psi(\eta|\mathrm{Pa}(\eta)$ according to (5).
    end for
    for all words $w \in T_k$ and corresponding contexts $\mathrm{Fa}(w)$ **do**
      Compute $q_\psi^{\mathrm{Pa}(w)}(w)$, $r_\psi^{\mathrm{Fa}(w)}(w)$ and $\alpha_\psi^{\mathrm{Fa(w)}}(w)$ using (5) and (4).
      Compute $P_\psi(w|\mathrm{Fa}(w)) = (1 - \alpha_\psi^{\mathrm{Fa(w)}})q_\psi^{\mathrm{Pa}(w)}(w) + \alpha_\psi^{\mathrm{Fa(w)}} r_\psi^{\mathrm{Fa}(w)}(w)$
    end for
    Compute

$$P(T_k|\psi) = \prod_{\eta \in T_k} P_\psi(\eta|\mathrm{Pa}(\eta)) \prod_{w \in T_k} P_\psi(w|\mathrm{Fa}(w))$$

    Compute $P(\psi|\lambda)$ for a dense grid of locations λ as in training (Fig. 3).
  end for
  Compute the marginal likelihood

$$P(T_k|\lambda) = \sum_\psi P(\psi|\lambda)P(T_k|\psi)$$

  for all objects o in the scene π with spatial extent B_o **do**
    Compute

$$P(T_k|o) = \frac{1}{|B_o|} \sum_{\lambda \in B_o} P(T_k|\lambda)$$

    Compute $P(o|T_k) \propto P(T_k|o)P(o)$.
  end for
end for
Compute $P(o|S) = \sum_k P(o|T_k)P(T_k|S)$.
Select $\arg\max_o P(o|S)$.

Fig. 4: General algorithm used to infer the referent λ from the test sentence S. Computation for the exhaustive algorithm scales as the product of the number of possible landmarks in the scene, the number of relations in the ontology and the number of parses considered; plus the product of the number of parses and the number of referents. For the results reported here, only one parse is considered.

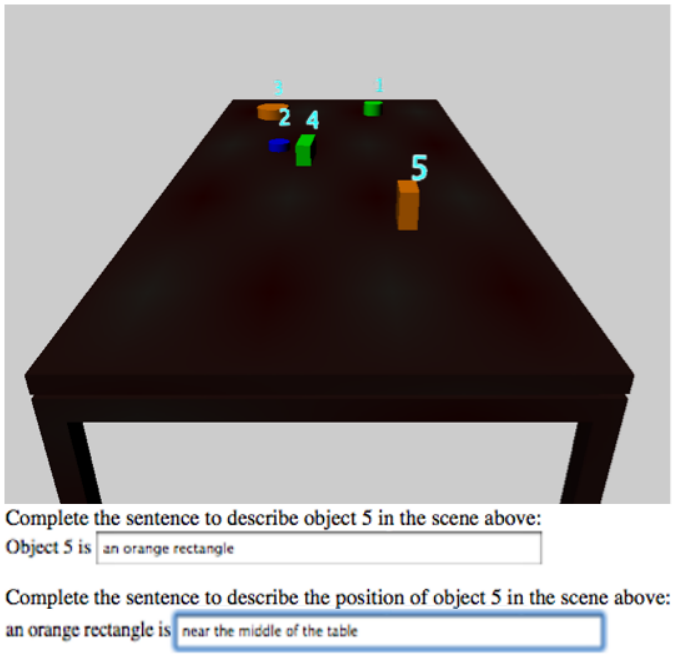


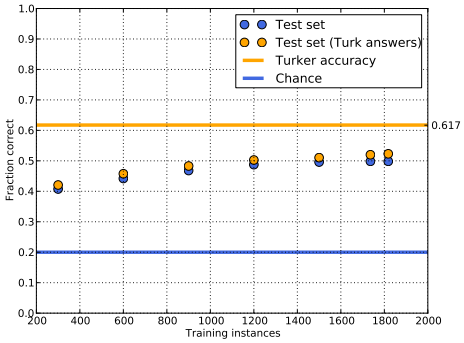
Fig. 5: An example prompt used to elicit descriptions from speakers

2988 responses were collected, with an overall identification accuracy of 61.7%. Since many of the training sentences were ambiguous, we also compared the model's answers to the "consensus object," the one chosen by a majority of raters (whether or not it was the target that produced the sentence). The MAP object corresponded to this consensus 52.4% of the time. The performance increase when using this metric suggests that our learner tends to "misunderstand" sentences in the same way as human raters. These results are shown in Fig. 6.

V. Related Work

Most NLP research has treated preposition meaning as purely lexical. For example, SemEval 2007 contained a task on word-sense disambiguation of prepositions [3]. Evaluations were performed against data from the Preposition Project [4], [5], which models preposition meaning as a fixed inventory of lexically-defined senses. In contrast, our model does not represent lexical semantics at all, but takes the utterance as a whole (or at least phrases within it) to convey a relational meaning. This has at least two advantages. First, if a word is unknown, inference does not break down: semantics can often still be inferred from the remaining context. Second, syntactic and semantic knowledge constrain the meaning of new words.

Baldwin et al [6] summarize linguistic issues in syntactic and semantic accounts of prepositions; such as PP attachment and the semantics of prepositional multi-word phrases. Spatial language and prepositions have been much discussed by cognitive linguists [7]–[11] and by image-schema theorists [12]–[14]. We think of the work we present here as a generative,



(a) Performance as a function of training set size. Accuracy was measured using both the "correct" object (blue points), and the object chosen by a majority of human raters (gold points).

Mod Turk	1	2	3	4	5	Avg	N
1.0	52.4	13.8	10.4	12.4	11.0	2.15	2770
1.5	30.2	27.2	15.0	17.8	09.7	2.49	360
2.0	13.6	22.2	19.9	20.6	23.7	3.19	1250
2.5	21.0	22.1	19.6	17.3	19.9	2.93	1120
≥ 3	10.5	20.9	23.6	22.5	22.5	3.26	9250

(b) Model rank distribution (percentages) conditioned on human rank (determined by the number of raters choosing each item).

	1	2	3	4	5
Model	49.8	16.4	11.4	12.0	10.4
Turk	83.4	8.7	6.4	1.5	0.0

(c) Rank distribution (percentages) for "right" answers.

Fig. 6: Performance on object selection task averaged over 10 iterations (which differ due to random sampling during training).

probabilistic implementation of image-schematic ideas about spatial language.

Previous work on symbol grounding has quantitatively measured the degree of applicability of spatial relations between objects in a physical space [15]. Bateman et al describe a linguistically-derived ontology that includes spatial relations [16]. We employ applicability functions over space, indexed by spatial relations, to induce probability distributions over possible groundings of a symbolic representation. Spranger et al. use a similar approach of selecting relations via competition weighted by applicability functions [17], [18][1].

Gorniak and Roy [19] develop a visually grounded model for understanding spatial language with semantics based on language observed in a human study. They use a compositional parser to combine the meanings of single words into the semantics of complex spatially referring expressions. Their system recovers the correct referent from a large percentage of natural language descriptions. However, no learning occurs; words are associated with their semantics by definition. Piantadosi, et al. [20] learn compositional meanings in a Bayesian fashion, but again assume the existence of grounded lexical

[1]We thank an anonymous reviewer for bringing this work to our attention

semantics.

Tellex et al. [21] employ a graphical model to understand grounded spatial relationships from text, and Makalic et al. [22] 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 we infer relations, learning a probabilistic mapping from grounded training data. spatial relations mapped one-to-one to verbs and prepositions, whereas we separate meaning from text entirely, maintaining more uncertainty about which features of the text inform which parts of the semantic representation. Moreover, whereas the model in [21] is trained on images annotated with correspondences, ours learns from training data that consists only of text and a reference object or location; relations and landmarks are induced.

In research on grounded language learning, our work appears to be first to specifically address learning the meanings and syntax of prepositional phrases. Spranger et al. address learning [17] or co-evoloving [18] language about spatial prepositions, using very similar assumptions to our own about acquiring language, such as the idea that speakers choose semantics that not only apply to an object, but contrast it with other objects. However they use a Fluid Construction Grammar that appears to have mappings between semantic elements and linguistic marker locations predefined rather than learned (although marker morphemes are learned or generated). Additionally, Spranger et al. represent semantics with Incremental Recruitment Language, which uses filtering and early-binding to reduce the set of potential referents, whereas our approach maintains real valued weights on potential referents allowing beliefs to be propagated until a decision must be made.

A related piece of previous work in learning general mappings between visual scenes and sentences is Matuszek et al. [23], who learn to associate object attributes with language through the induction of a grammar of syntax and compositional semantics. In their training data, scenes consist of objects, with both referent and relation explicitly identified. This differs from our work in two important ways: first, we learn to map language directly to object representations and relations; second, we learn probabilistic relationships between sentences and relations without observing the latter directly.

VI. Discussion

After only a few thousand sentences of training without observable semantic representations, our model achieves object-identification performance around 50%, compared to human performance around 62%. The low performance of the human raters is surprising, but is likely attributable to the free-form nature of the description task. If the participants were instead explicitly instructed to give a description that was distinctive for the intended object, human raters would presumably perform better, and presumably the signal-to-noise ratio in the training data would be more favorable for our model as well.

The present work lends itself to several natural extensions. We have already mentioned the possibility of incorporating parsing into the probabilistic process, and employing a less stringent mapping from syntactic constituents to semantic features. The current parsing method unfortunately limits the range of utterances that can be used in training, and hence comprehended after training. For instance, the parser cannot handle utterances that do not explicitly describe a landmark, preventing use of phrases such as "to the north". Nor can it handle relations with more than one landmark, such as BETWEEN or AMONG, or utterances involving a conjunction of relations where the relation phrase is distributed over landmarks ("left of the cup and bowl"). Currently only relation-level conjunctions can be handled ("behind the cup and left of the bowl"), by heuristically splitting utterances on "and", and treating both sub-utterances as describing the same referent. Clearly a more flexible parsing scheme is needed.

A related limitation of the present model is its restricted relational vocabulary. As the complexity of scenes and relational ontologies grows, more sophisticated inferential machinery will be needed, as it will no longer be possible to exhaustively compute posterior distributions over meanings. However, standard Bayesian inference techniques such as Markov Chain Monte Carlo offer promise for approximate inference (indeed we already employ sampling-based approximation during training).

Similarly, the static symbol grounding distributions limit the flexibility of semantic acquisition. It seems obvious, for example, that the grounded meaning of a NEAR-TO relation is not a single function of absolute distance, but is a function of the scale of the landmark and referent involved, as well as the broader context of the scene. We would like to have a model learn such scale-dependencies, and, ultimately, even learn new relations and their associated groundings using both unsupervised and active learning methods such as asking clarifying questions or attempting to produce an utterance and learning from corrections or elaborations received. These forms of interaction with the teacher would let the learner target weak points in its knowledge, mitigating ambiguity in training examples, and allowing it to differentiate between semantics (when a relation is applicable) and pragmatics (when speakers tend to use the relation).

References

[1] E. Charniak, "A maximum-entropy-inspired parser," in *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2000, pp. 132–139. [Online]. Available: http://dx.doi.org/10.3115/979617.979620

[2] R. Levy and G. Andrew, "Tregex and Tsurgeon: tools for querying and manipulating tree data structures," in *Proceedings of the fifth international conference on Language Resources and Evaluation*, 2006.

[3] K. Litkowski and O. Hargraves, "Semeval-2007 task 06: Word-sense disambiguation of prepositions," *SemEval-2007: 4th International Workshop on Semantic Evaluations*, 2007.

[4] ——, "The preposition project," *Proceedings of the Second ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, pp. 171–179, 2005.

[5] ——, "Coverage and inheritance in the preposition project," *Proceedings of the third ACL-SIGSEM workshop on prepositions*, pp. 37–44, 2006.

[6] T. Baldwin, V. Kordoni, and A. Villavicencio, "Prepositions in applications: A survey and introduction to the special issue," *Comput. Linguist.*, vol. 35, no. 2, pp. 119–149, Jun. 2009. [Online]. Available: http://dx.doi.org/10.1162/coli.2009.35.2.119

[7] A. Tyler and V. Evans, *The Semantics of English Prepositions: Spatial Scenes, Embodied Meaning, and Cognition*. Cambridge University Press, 2003. [Online]. Available: http://books.google.com/books?id=OCMCWSt6aQkC

[8] L. Talmy, "The representation of spatial structure in spoken and signed language," in *Perspectives on Classifier Constructions in Sign Language*, K. Emmorey, Ed. Erlbaum, 2003.

[9] ——, "The fundamental system of spatial schemas in language," in *From perception to meaning: Image Schemas in Cognitive Linguistics*, B. Hampe, Ed. Mouton de Gruyter, 2006.

[10] T. Regier, *The Human Semantic Potential: Spatial Language and Constrained Connectionism*. The MIT Press, 1996.

[11] A. Herskovits, *Language and Spatial Cognition: an interdisciplinary study of the prepositions in English*, ser. Studies in Natural Language Processing. London: Cambridge University Press, 1986.

[12] J. Mandler, *The Foundations of Mind: Origins of Conceptual Thought*. Oxford University Press, 2004.

[13] M. Johnson, *The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason*. Chicago, IL: University of Chicago Press, 1987.

[14] G. Lakoff, *Women, Fire and Dangerous Things*. Chicago, IL: University of Chicago Press, 1987.

[15] T. Regier and L. A. Carlson, "Grounding spatial language in perception: an empirical and computational investigation." *Journal of Experimental Psychology: General*, vol. 130, no. 2, p. 273, 2001.

[16] J. A. Bateman, J. Hois, R. Ross, and T. Tenbrink, "A linguistic ontology of space for natural language processing," *Artificial Intelligence*, vol. 174, no. 14, pp. 1027–1071, 2010.

[17] M. Spranger, S. Pauw, and M. Loetzsch, "Open-ended semantics co-evolving with spatial language," in *The Evolution of Language: Proceedings of the 8th International Conference (EVOLANG8)*, 2010, p. 297.

[18] M. Spranger, "The co-evolution of basic spatial terms and categories," in *Experiments in Cultural Language Evolution*, L. Steel, Ed. Amsterdam: John Benjamins Publishing, 2012, pp. 111–141.

[19] P. Gorniak, D. Roy *et al.*, "Grounded semantic composition for visual scenes," *J. Artif. Intell. Res. (JAIR)*, vol. 21, pp. 429–470, 2004.

[20] S. T. Piantadosi, N. D. Goodman, B. A. Ellis, and J. B. Tenenbaum, "A bayesian model of the acquisition of compositional semantics," in *Proceedings of the Thirtieth Annual Conference of the Cognitive Science Society*, 2008.

[21] S. A. Tellex, T. Kollar, S. R. Dickerson, M. R. Walter, A. Banerjee, S. Teller, and N. Roy, "Approaching the symbol grounding problem with probabilistic graphical models," *AI Magazine*, vol. 32, no. 5, Winter 2011.

[22] E. Makalic, I. Zukerman, M. Niemann, and D. Schmidt, "A probabilistic model for understanding composite spoken descriptions," in *PRICAI 2008: Trends in Artificial Intelligence*. Springer, 2008, pp. 750–759.

[23] C. Matuszek, N. FitzGerald, L. Zettlemoyer, L. Bo, and D. Fox, "A joint model of language and perception for grounded attribute learning," *arXiv preprint arXiv:1206.6423*, 2012.