Learning Spatial-Semantic Representations from Natural Language Descriptions and Scene Classifications

Sachithra Hemachandra, Matthew R. Walter, Stefanie Tellex, and Seth Teller

***Abstract*— We describe a semantic mapping algorithm that learns human-centric environment models by interpreting natural language utterances. Underlying the approach is a coupled metric, topological, and semantic representation of the environment that enables the method to fuse information from natural language descriptions with low-level metric and appearance data. We extend earlier work with a novel formulation that incorporates spatial layout into a topological representation of the environment. We also describe a factor graph formulation of the semantic properties that encodes human-centric concepts such as type and colloquial name for each mapped region. The algorithm infers these properties by combining the user's natural language descriptions with image- and laser-based scene classification. We also propose a mechanism to more effectively ground natural language descriptions of distant regions using semantic cues from other modalities. We describe how the algorithm employs this learned semantic information to propose valid topological hypotheses, leading to more accurate topological and metric maps. We demonstrate that integrating language with other sensor data increases the accuracy of the achieved spatial-semantic representation of the environment.**

I. Introduction

A challenge to realizing robots that work productively alongside human partners is the development of efficient command and control mechanisms. Researchers have recently sought to endow robots with the ability to interact more effectively with people through natural language speech [1, 2, 3, 4, 5] and gesture understanding [6]. Efficient interaction is facilitated when robots reason over models that encode high-level semantic properties of the environment. For example, such models could help a micro-aerial vehicle interpret a first responder's command to "fly up the stairway on the right, go down the hall, and observe the kitchen."

Semantic mapping algorithms [7, 8, 9, 10] extend the metric environment models traditionally employed in robotics to include higher-level concepts, including types and colloquial names for regions, and the presence and use of objects in the environment. These methods typically operate by augmenting a standard SLAM metric map with a representation of the environment's topology, and a distinct representation of its semantic properties, the latter of which is populated by interpreting the robot's sensor stream, typically through scene classification. In this layered approach, the underlying metric map induces and embeds the topological and semantic attributes. However, while refinements to the metric map improve the topological and semantic maps, most techniques do not allow knowledge inferred at the semantic and topological levels to influence the other layers. Typical approaches infer semantic information from LIDAR and camera data using pre-trained scene appearance models. Some efforts are capable of integrating *egocentric* descriptions about the robot's current location (e.g., "we are in the kitchen"), but cannot handle *allocentric* descriptions that involve spatial relations between and labels for potentially distant regions in the environment (e.g., "the exit is next to the cafeteria").

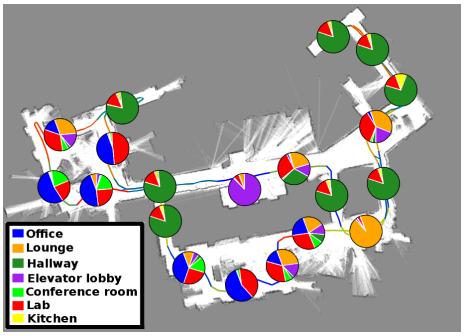


Fig. 1. Maximum likelihood semantic map of the 6th floor of Stata building (pie charts denote the likelihood of different region categories).

We addressed some of these limitations in previous work [11] with an algorithm that maintains a joint distribution over a *Semantic Graph*, a coupled metric, topological, and semantic environment representation learned from user utterances and the robot's low-level sensor data, during a guided tour. Our framework was able to to learn properties of the environment that could not be perceived with typical sensors (e.g. colloquial names for regions, properties of areas outside the robot's sensor range) and use semantic knowledge to influence the rest of the semantic graph, allowing robots to efficiently learn environment models from users.

Our previous algorithm decomposed the environment into a collection of fixed, uniformly-sized regions. This has the potential to result in a topology that is inconsistent with human concepts of space. Consequently, the representation may not model the spatial extent to which the user's descriptions refer, resulting in incorrect language groundings. The semantic information in the framework was limited to user-provided colloquial names and did not provide a means to reason over properties such as region type that can be inferred from LIDARs, cameras, or other onboard sensors. Additionally, due to the absence of other semantic information, the framework required that a landmark location

S. Hemachandra, M.R. Walter, and S. Teller are with the Computer Science and Artificial Intelligence Laboratory at the Massachusetts Institute of Technology, Cambridge, MA 02139 USA {sachih, mwalter, teller}@csail.mit.edu

S. Tellex is with the Computer Science Department at Brown University, Providence, RI 02912 USA stefie10@cs.brown.edu

be labeled by the user before the utterance can be grounded (e.g., processing the phrase "the kitchen is down the hallway" requires the "hallway" to have already been labeled).

This paper describes an extension of our earlier approach to learn richer and more meaningful semantic models of the environment. Whereas our earlier framework reasoned only about the connectivity of deterministically created regions (at fixed intervals), the current approach reasons over the environment's region segmentation as well as its inter-region connectivity. Additionally, we propose a factor graph representation for the semantic model that reasons not only over each region's labels, but also its canonical type. As before, we infer region labels from user-provided descriptions, but we also incorporate scene classification using the robot's onboard sensors, notably camera and laser range-finders, to estimate region types. By modeling the relation between an area's type and its colloquial name, the algorithm can reason over both region type and region label, even in the absence of speech. This enables the method to more effectively ground allocentric user utterances (e.g., when grounding the phrase "the kitchen is down the hallway", we no longer require the user to explicitly label the "hallway" beforehand). We also describe a mechanism by which the algorithm derives a semantically meaningful topology of the environment based upon a factor graph model of the distribution, where edges are proposed using a spatial-semantic prior distribution. We show that the improved topology model then allows the method to better handle ambiguities common in natural language descriptions.

II. Related Work

A number of researchers have focused on the problem of constructing semantic environment models [7, 10, 8, 9]. Most approaches augment lower-level metric maps with higher-level topological and/or semantic information. However, these typically follow a bottom-up approach in which higher-level concepts are constructed from lower-level information, without any information flow back down to lower-level representations. In Walter et al. [11], we addressed this by introducing a framework that uses semantic information derived from natural language descriptions uttered by humans to improve the topological and metric representations. Our proposed approach uses additional semantic cues to evaluate semantic similarity of regions to update the topology.

Several existing approaches [8, 10] have incorporated higher-level semantic concepts such as room type and the presence of objects with the use of appearance models. Pronobis and Jensfelt [10] describe a multi-modal probabilistic framework incorporating semantic information from a wide variety of modalities including object detections, place appearance, and human-provided information. However, their approach is limited to handling egocentric descriptions (e.g., "we are in the living room"). Additionally, they infer topology based on door detections, a heuristic that works well only in certain kinds of environments; they do not maintain a distribution over likely topologies. In [11], we maintained a hypothesis over the distribution of topologies, but the topology was constructed from segments created at fixed spatial intervals, which can be inconsistent with human concepts. In the present work, we address this limitation by proposing a method that maintains multiple hypotheses about region segmentations and about connections among regions.

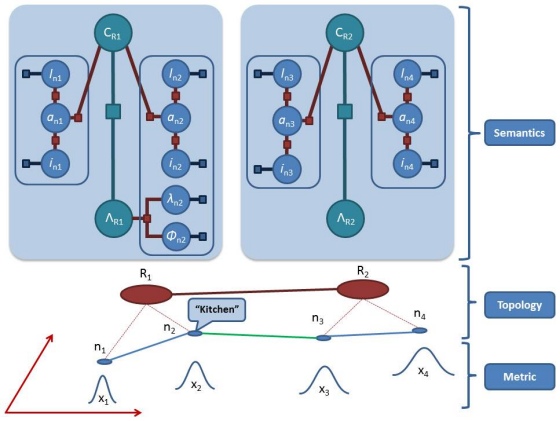


Fig. 2. Example of a semantic graph: Two regions R_1 and $R2$ and their constituent nodes n_i's; distributions over node poses x_i; and the corresponding factor graph.

The problem mapping natural language utterances to their corresponding referents in a robot's world model has been studied by several researchers [3, 4, 2, 12, 13, 14, 15, 16] in the context of command and control. However, these efforts have not focused on constructing semantic representations and, instead, typically assume them to be known a priori. In [11], we augmented our semantic graph with complex language descriptions. However, this could be accomplished only when the user had explicitly labeled a location before describing another in reference to it (e.g., "the gym is down the hallway" requires prior labeling of the hallway). The resulting representation contained only labels obtained through language descriptions; it did not integrate semantic cues from other sources (such as appearance models). The present method improves upon this by integrating information from multiple semantic sources and maintaining a distribution over a larger set of semantic properties, rendering it capable of grounding language even in the absence of pre-labeled landmark locations.

III. Semantic Graph Representation

This section presents our approach to maintaining a distribution over semantic graphs, an environment representation that consists of metric, topological, and semantic maps.

A. Semantic Graphs

We define the semantic graph as a tuple containing topological, metric and semantic representations of the environment. Figure 2 shows an example semantic graph for a trivial environment.

The topology G_t is composed of nodes n_i that denote the robot's trajectory through the environment (sampled at 1 m distances), node connectivity, and node region assignments. We associate with each node a set of observations that include laser scans z_i, semantic appearance observations

a_i based on laser l_i and camera i_i models, and available language observations λ_i. We assign nodes to regions $R_\alpha = \{n_1, .., n_m\}$ that represent spatially coherent areas in the environment compatible with human concepts (e.g., rooms and hallways). Undirected edges exist between node pairs in this graph, denoting traversability. Edges between regions are inferred based on the edges between nodes in the graph. A region edge exists between two regions if at least one graph edge connects a node from one region to a node in the other. The topological layer consists of the nodes, edges, and the region assignments for the nodes.

The pose x_i of each node n_i is represented in a global reference frame. The metric layer is induces by the topology, where edges in the topology also include metric constraints between the corresponding node poses. Metric constraints are calculated by scan-matching the corresponding laser observations of each region. A pose graph representation is employed to maintain the distribution over the pose of each node, conditioned on these constraints. Occupancy maps can be constructed based on the node poses and their corresponding laser observations.

Semantic information is also conditioned on the topology as shown in the Fig. 2. The semantic layer consists of a factor graph with variables that represent the type C_r and labels Λ_r for each region, properties that can be observed at each node (in each region), and factors that denote the joint likelihood of these variables (e.g., the likelihood of observing a label given a particular room type). Observations of these region properties are made using laser- and image-based scene classifiers and by grounding human descriptions of the environment.

Algorithm 1: Semantic Mapping Algorithm

Input: $P_{t-1} = \left\{P_{t-1}^{(i)}\right\}$, and $(u_t, z_t, a_t, \lambda_t)$, where $P_{t-1}^{(i)} = \left\{G_{t-1}^{(i)}, X_{t-1}^{(i)}, S_{t-1}^{(i)}, w_{t-1}^{(i)}\right\}$

Output: $P_t = \left\{P_t^{(i)}\right\}$

for $i = 1$ *to* n **do**

1) Employ proposal distribution to propagate the graph sample based on u_t, λ_t and a_t.
   a) Sample region allocation
   b) Sample region edges
   c) Merge newly connected regions
2) Update the Gaussian distribution over the node poses $X_t^{(i)}$ conditioned on topology.
3) Update the factor graph representing semantic properties for the topology based on appearance observations (l_t and i_t) and language λ_t.
4) Compute the new particle weight $w_t^{(i)}$ based upon the previous weight $w_{t-1}^{(i)}$ and the metric data z_t.

end

Normalize weights and resample if needed.

B. Distribution Over Semantic Graphs

We maintain the joint distribution over the topology G_t, the vector of locations X_t, and the set of semantic properties S_t. Formally, we maintain this distribution over semantic graphs $\{G_t, X_t, S_t\}$ at time t conditioned upon the history of metric exteroceptive sensor data $z^t = \{z_1, z_2, \ldots, z_t\}$, odometry $u^t = \{u_1, u_2, \ldots, u_t\}$, scene appearance observations $a^t = \{a_1, a_2, \ldots, a_t\}$ (where in our implementation $a_t = \{l_t, i_t\}$), and natural language descriptions $\lambda^t = \{\lambda_1, \lambda_2, \ldots, \lambda_t\}$,

$$p(G_t, X_t, S_t | z^t, u^t, a^t, \lambda^t). \tag{1}$$

Each variable λ_i denotes a (possibly null) utterance, such as "This is the kitchen," or "The gym is down the hall." We factor the joint posterior into a distribution over the graphs and a conditional distribution over the node poses and labels,

$$\begin{aligned} p(G_t, X_t, S_t | z^t, a^t, u^t, \lambda^t) &= p(S_t | X_t, G_t, z^t, a^t, u^t, \lambda^t) \\ &\times p(X_t | G_t, z^t, a^t, u^t, \lambda^t) \times p(G_t | z^t, a^t, u^t, \lambda^t) \end{aligned} \tag{2}$$

As in Walter et al. [11], we maintain this factored distribution using a Rao-Blackwellized particle filter, mitigating the hyper-exponential hypothesis space of the topology [11].

We represent the joint distribution over the topology, node locations, and labels as a set of particles

$$\mathcal{P}_t = \{P_t^{(1)}, P_t^{(2)}, \ldots, P_t^{(n)}\}. \tag{3}$$

Each particle $P_t^{(i)} \in \mathcal{P}_t$ consists of the set

$$P_t^{(i)} = \left\{G_t^{(i)}, X_t^{(i)}, S_t^{(i)}, w_t^{(i)}\right\}, \tag{4}$$

where $G_t^{(i)}$ denotes a sample from the space of graphs, $X_t^{(i)}$ is the analytic distribution over locations, $S_t^{(i)}$ is the distribution over semantic properties, and $w_t^{(i)}$ is the weight of particle i.

IV. Semantic Mapping Algorithm

Algorithm 1 outlines the process by which the method recursively updates the distribution over semantic graphs (2) to reflect the latest robot motion, metric sensor data, laser- and image-based scene classifications, and the natural language utterances. The following sections explain each step in detail.

A. The Proposal Distribution

We compute the prior distribution over the semantic graph G_t, given the posterior from the last time step G_{t-1}, by sampling from a proposal distribution. This proposal distribution is a predictive prior over the current graph given the previous graph, sensor data (excluding the current time step), appearance data, odometry, and language,

$$p(G_t | G_{t-1}, z^{t-1}, a^t, u^t, \lambda^t). \tag{5}$$

We augment the topology G_{t-1} to reflect the robot's motion by adding a node n_t to to the topology and an edge to the previous node n_{t-1}, resulting in an intermediate graph G_t^-. This represents the robot's current pose and the connectivity to its previous pose. This yields an updated

vector of poses X_t^- and semantic properties S_t^-. The new node is assigned to the current region.

1) Creation of New Regions: We then probabilistically bisect the current region R_c using the spectral clustering method proposed by Blanco et al. [17]. We construct the similarity matrix using the laser point overlap between each pair of nodes in the region. Equation 6 defines the likelihood of bisecting the region, which is based on the normalized cut value N_c of the graph involving the proposed segments. The likelihood of accepting a proposed segmentation rises as the N_c value decreases, i.e., as the separation of the two segments improves (minimizing the inter-region similarity),

$$P(s/N_{cut}) = \frac{1}{(1 + \alpha N_c^3)}. \tag{6}$$

This results in more spatially distinct areas in the world having a higher likelihood of being distinct regions, leading to more particles modeling these areas as separate regions. If a particle segments the current region, a new region R_i is created that does not include the newly added node.

2) Edge Proposals: When a new region R_i is created, the algorithm proposes edges between this region, and other regions in the topology, excluding the current region R_c.

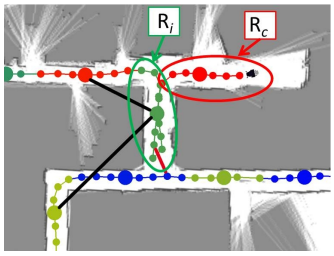


Fig. 3. Example of region edges being proposed (black lines represents rejected edge proposals; the red line represents an accepted edge).

The algorithm samples inter-region edges from a spatial-semantic proposal distribution that incorporates the semantic similarity of regions, as well as the spatial distributions of its constituent nodes. This reflects the notion that regions that are nearby and semantically similar are more likely to be connected. We measure semantic similarity based upon the label distribution associated with each region. The resulting likelihood has the form

$$p_a(G_t|G_t^-, z^{t-1}, u^t, a^t, \lambda^t) = \prod_{j:e_{ij} \notin E^-} p(G_t^{ij}|G_t^-) \tag{7a}$$

$$\propto \prod_{j:e_{ij} \notin E^-} p_x(G_t^{ij}|G_t^-) p_s(G_t^{ij}|G_t^-), \tag{7b}$$

where we have omitted the history of language observations λ^t, metric measurements z^{t-1}, appearance measurements a^t, and odometry u^t for brevity. Equation 7a reflects the assumption that additional edges that express constraints involving the current node $e_{ij} \notin E^-$ are conditionally independent. While $p_x(G_t^{ij}|G_t^-)$ encodes the likelihood of the edge based on the spatial properties of the two regions, $p_s(G_t^{ij}|G_t^-)$ describes the edge likelihood based on the regions' semantic similarity. Equation 7b reflects the assumed conditional independence between the spatial- and the semantic-based edges.

For the spatial distribution prior, we consider the distance d_{ij} between the mean nodes of the two regions, where the mean node is that with its pose closest to the region's average pose

$$p_x(G_t^{ij}|G_t^-) = \int_{X_t^-} p(G_t^{ij}|X_t^-, G_t^-, u_t) p(X_t^-|G_t^-) \tag{8a}$$

$$= \int_{d_{ij}} p(G_t^{ij}|d_{ij}, G_t^-) p(d_{ij}|G_t^-). \tag{8b}$$

The conditional distribution $p(G_t^{ij}|d_{ij}, G_{t-1}, z^{t-1}, u^t)$ expresses the likelihood of adding an edge between regions R_t and R_j based upon the location of their mean nodes. We represent the distribution for a particular edge between regions R_i and R_j with distance $d_{ij} = |\bar{X}_{R_i} - \bar{X}_{R_j}|_2$ as

$$p(G_t^{ij}|d_{ij}, G_t^-, z^{t-1}, u^t) \propto \frac{1}{1 + \gamma d_{ij}^2}, \tag{9}$$

where γ specifies a distance bias. For the evaluations in this paper, we use $\gamma = 0.3$. We approximate the distance prior $p(d_{ij}|G_t^-, z^{t-1}, u^t)$ with a folded Gaussian distribution.

The semantic prior expresses the increased likelihood that edges exist between regions with similar distributions over labels Λ. The label distributions for the regions are modeled in the semantic layer,

$$p_s(G_t^{ij}|G_t^-) = \sum_{S_t^-} p(G_t^{ij}|S_t^-, G_t^-) p(S_t^-|G_t^-) \tag{10a}$$

$$= \sum_{\Lambda_i^-, \Lambda_j^-} p(G_t^{ij}|\Lambda_i^-, \Lambda_j^-, G_t^-) p(\Lambda_i^-, \Lambda_j^-|G_t^-). \tag{10b}$$

Equation 11 expresses the likelihood of an edge existing between two regions, given the value of the regions' respective label values

$$p(G_t^{ij}|\Lambda_i, \Lambda_j) = \begin{cases} \theta_{\Lambda_i} & \text{if } \Lambda_i = \Lambda_j \\ 0 & \text{if } \Lambda_i \neq \Lambda_j \end{cases}, \tag{11}$$

where θ_{Λ_i} denotes the likelihood that edges exist between nodes with the same label. In practice, we assume a uniform saliency prior for each label. Equation 10b measures the cosine similarity between the label distributions.

After a region edge is sampled from the spatial-semantic prior, a scan-match procedure attempts to find the best alignment between the two regions. Upon convergence of the scan-match routine, the edge is accepted and is used to update the topology.

3) Region Merges: After a new region R_i has been created and edges to other regions have been checked and added, the algorithm determines whether it is possible to merge with each region to which it is connected. The newly-created region is merged with an existing (connected) region if the observations associated with the smaller of the two regions can be adequately explained by the larger region. This results in regions being merged when the robot revisits

locations already represented in the graph. This merge process is designed to ensure that the complexity of the topology increases only when the robot explores new areas, leading to more efficient region edge proposals as well as more compact language groundings.

B. Updating the Metric Map Based on New Edges

The algorithm then updates the spatial distribution over the node poses X_t conditioned on the proposed topology,

$$p(X_t|G_t, z^t, u^t, \lambda^t) = \mathcal{N}^{-1}(X_t; \Sigma_t^{-1}, \eta_t), \qquad (12)$$

where we parametrize the Gaussian in the canonical form in terms of the information matrix Σ_t^{-1} and information vector η_t. We use the iSAM algorithm [18], which iteratively solves for the QR factorization of the information matrix.

C. Updating the Semantic Layer

Compared with Walter et al. [11], our updated representation maintains a distribution over a larger set of semantic properties associated with the environment. The distribution over the semantic layer is maintained using a factor graph [19] that is conditioned on the topology for each particle.

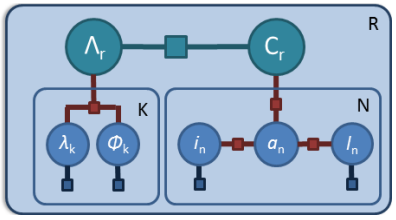


Fig. 4. Semantic Layer (plate representation)

As Fig. 4 shows, the semantic layer maintains two variables associated with each region r, namely the region category C_r (e.g., hallway, conference room, etc.) and the labels associated with the region Λ_r. For example, a conference room can have multiple labels, such as "meeting room" and "conference room." The factor that joins these two variables represents the likelihood of each room category generating a particular label.

At each node n within a region, the robot can observe one or more of these semantic properties. In our current implementation, these are the region *appearance* observed from laser scanners l_n or cameras i_n, and the region *labels* λ_k (and the associated correspondence variables Φ_k). We run belief propagation for each particle at each time step as new variables and factors are added. We subsequently update the category and label distributions for each region.

The following subsections outline how we integrate observations of these node properties into the semantic layer.

1) Integrating Semantic Classification Results: Each node has an appearance variable a_n that is related to its region category. We consider several general appearance classes ("room", "hallway," and "open area") that are then observed using robot sensors. The factor that connects a region category variable C_r to an appearance variable a_n encodes the likelihood of a region category generating an appearance class (e.g., how often does a conference room appear as a room). The category-to-appearance factor was trained using annotated data from several other floors of the Stata building.

We make two appearance observations a_n using the laser and camera observations at node n. These are represented in the factor graph as the laser appearance l_n and the image appearance i_n. We use two pre-trained appearance models for laser observations and camera images. The laser appearance classification model has been trained using laser features similar to those outlined in Mozos et al. [20], while the image appearance model has been trained using CRFH [21]. Laser and camera appearance variables l_n and i_n are connected to the node's appearance a_n using factors built from the confusion matrix for the two trained models. The classification results for the two sensors provide a distribution representing the likelihood of the observations being generated from each appearance category. The classifier outputs are integrated to the factor graph as factors attached to variables l_n and i_n.

2) Integrating Language Observations: The robot can also receive either ego-centric (e.g., "I am at the kitchen") or allocentric (e.g., "The kitchen is down the hall") descriptions of the environment from the user. We use these observations to update the likelihood of observing labels in each region. We maintain the label for each region as another variable Λ_r in our factor graph. The region label is related to the region category, as each category is more likely to generate some labels than others. For example, while a person might describe a "cafeteria" as a "dinning hall," it is unlikely that they will describe it as an "office." For our current experiments, we identified a limited subset of labels associated with each region category in our representation (e.g., the hallway category can generate "hall," "hallway," "corridor," or "walkway"). When building these factors between labels and room categories, we assign higher likelihoods to labels associated with each category and smaller likelihoods to the other labels (capturing the likelihood of generating these labels given a particular room category).

A node can have zero, one, or multiple label observations depending on the way the person describes a location. We represent each label observation with a variable λ_k and a correspondence variable Φ_k, which denotes the likelihood that the label was intended to describe the current location. The correspondence variable Φ_k is a binary-valued variable specifying whether or not the label describes the region. If the label doesn't correspond to that region ($\Phi_k = 0$), the observation λ_k is uninformative about the region's label, and will have equal likelihood for each label value. However, when the correspondence holds ($\Phi_k = 1$), the factor encodes the likely co-occurrences between different labels. For example, if the robot heard the label "conference room" with a high likelihood of $\Phi_k = 1$, it will result in other labels that often co-occur with "conference room" (e.g., "meeting room") as having high likelihoods as well. Currently, high co-occurrence is added for words that are synonyms (e.g., "hallway" and "corridor"). In this way, we use the correspondence variable to handle the ambiguity

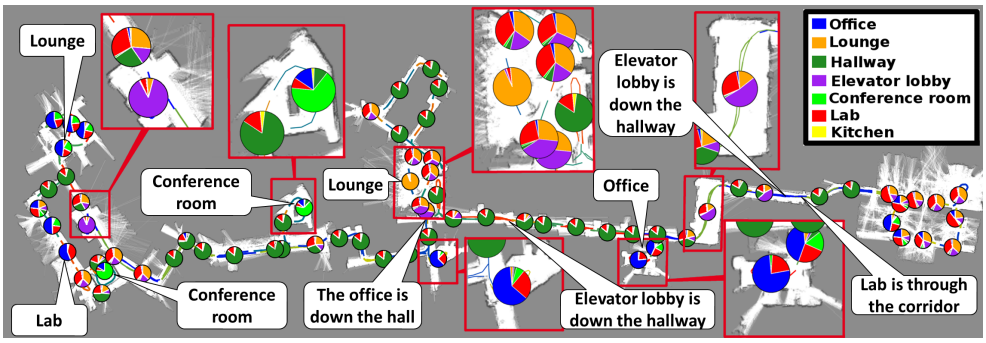


Fig. 5. Maximum likelihood semantic map of a multi-building environment on the MIT campus.

inherent in grounding natural language descriptions. When a label is grounded to a region, we create a label observation λ_k and correspondence variable Φ_k, and connect it to the associated region's label variable Λ_r using a co-occurrence factor. We integrate the correspondence observation Φ_k by attaching a factor encoding this likelihood. We treat the observed label as having no uncertainty, as our current model does not model errors arising from speech recognition.

We derive the factor denoting the observation of Φ_k based on the type of description given by the user. If the user describes the current location (e.g., "I am at the living room"), we have higher correspondence with spatially local nodes. For such descriptions, we allocate a high likelihood of correspondence with the current region, i.e., $p(\Phi_k = 1) = 0.8$. For allocentric descriptions (e.g., "the kitchen is down the hallway", where the user describes the location of the referent "kitchen" with relation to the landmark "hallway"), we use the G^3 framework [4] to calculate the correspondence likelihood given the potential landmarks. We marginalize the landmarks to arrive at correspondence likelihood of each referent region in a manner similar to our previous approach.

In handling allocentric descriptions, the current method improves upon our earlier approach in two ways. Firstly, we no longer require common landmarks to be explicitly described before being able to ground the language correctly. We do this by leveraging the richer semantic representation made possible by integrating additional semantic information to arrive at likely landmarks. Secondly, while some expressions can be ambiguous (e.g., there could be multiple regions down the hall), the presence of other semantic cues increases the accuracy of the final label distribution since incorrect groundings will have less impact if the region's appearance is different from the label observation.

D. Updating the Particle Weights and Resampling

Particle weights are updated and resampling in the same fashion as Walter et al. [11].

V. Results

We evaluate our algorithm through four experiments in which a human gives a robotic wheelchair [9] a narrated guided tour of the Stata Center (S3, S4, S6) as well as a multi-building indoor tour (MF) on the MIT campus. The robot was equipped with a forward-facing LIDAR, a camera, wheel encoders, and a microphone. In these experiments we drove the robot using a joystick, and provided it with textual natural language descriptions at specific salient locations.

We evaluate the resulting semantic maps with regards to their topological accuracy, compactness, segmentation accuracy, and semantic accuracy. All experiments were run with 10 particles. The results show that our framework produces more compact and more accurate semantic graphs than our previous approach. They also demonstrate the improvement in semantic accuracy due to language descriptions. We also show the ability of our framework to ground complex language even in the absence of previous labels for the referent (e.g., it handles the expression "the lobby is down the hall" even when the hall has not been labeled).

A. Topological accuracy

We compare the topological accuracy, conditioned upon the resulting segmentation, by comparing the maximum likelihood map with the ground truth topology. We define a topology as matching ground truth if node pairs that are spatially close (1 m) in a metrically accurate alignment are at most one region hop away. This avoids penalizing occasional regions that do not contain valid edges as long as a nearby region was accurately connected (and possibly merged with the nodes from a prior visit). This can happen when an edge was not sampled or when scan-matching failed to converge.

The percentage of close node pairs that were more than one region hop away from each other for the third, fourth and sixth floor were 2.8%, 3.7%, and 3.8%, respectively. Most region-matching errors occurred in areas with significant clutter, causing scan-matching failures. Metric maps derived from the maximum likelihood particles were accurate for all three floors.

B. Topological Compactness

We compare the allocation of nodes to regions in the current framework to the previous method. In the previous approach, the topology update did not merge regions even when the robot revisited a region; it simply created an edge between the regions. The redundancy of the regions has several negative implications. Firstly, it unnecessarily increases the hypothesis space of possible region edges, reducing

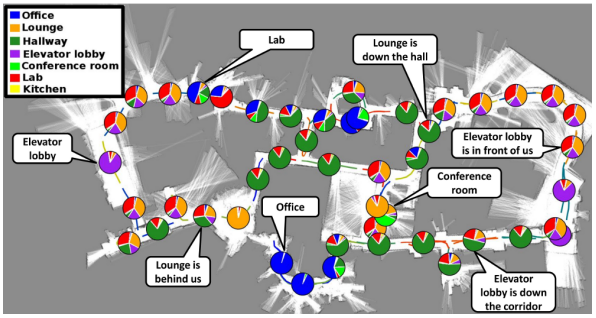


Fig. 6. Maximum likelihood semantic map of the 3rd floor of the Stata building (pie charts denote the likelihood of different region categories).

TABLE I

REGION ALLOCATION EFFICIENCY (S_c)

Floor	Proposed Framework	Old Framework
Stata Floor 3 (S3)	.67	.29
Stata Floor 4 (S4)	.77	.37
Stata Floor 6 (S6)	.69	.52

the likelihood of a sample proposing valid region edges. Secondly, it increases the hypothesis space for grounding language, forcing the framework to consider more region pairs as possible groundings for user descriptions.

We measure the duplicity of the region allocation as

$$S_c = N_s/N_t, \tag{13}$$

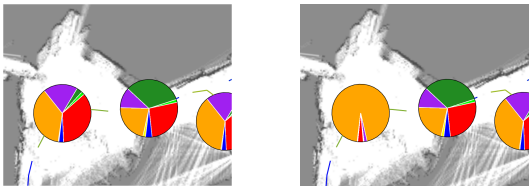
where N_s is the number of close node pairs ($< 1\,m$) assigned to the same region and N_t is the total number of close node pairs. If the topology is efficient at allocating regions, this ratio should be high, as only nodes near region boundaries should belong to different regions. Table I compares these scores three different floors. The new method scores significantly higher in all three experiments. The difference becomes more pronounced as the robot revisits more regions. Since the sixth floor dataset did not have too many revisited regions, the scores for the two approaches are closer.

C. Segmentation Accuracy

Table II outlines the segmentation accuracy for the maximum likelihood particle for two datasets, outlined according to region type. We picked the best matches based on the Jaccard index (number of intersecting nodes divided by the number of union nodes) for each ground truth annotated region and the resulting segmented region. Since our segmentation method depends on the similarity of laser observations, large cluttered region types such as lab spaces and lounges tend to be over-segmented. Additionally long hallways tend to be over-segmented by our method, which is reflected in the lower scores for hallways.

D. Inference of Semantic Properties

Table II also outlines the semantic accuracy for the maximum likelihood particle for two datasets. Semantic accuracy was calculated for each ground truth region by assigning each constituent node with its parent region's category distribution



(a) Appearance only (b) With language

Fig. 7. Region category distribution (a) for a region with only appearance information and (b) and with both appearance and language "the lounge is behind us". (category "lounge": yellow).

TABLE II

REGION SEGMENTATION AND SEMANTIC ACCURACY

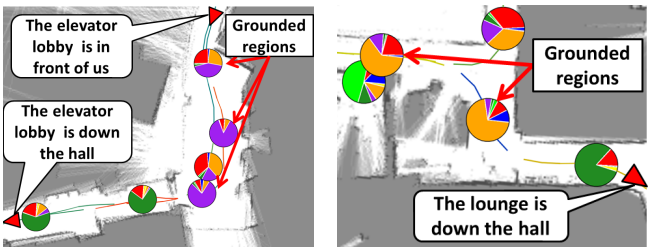
Region Type	Segmentation Accuracy		Semantic Accuracy Without Lang		Semantic Accuracy With Lang	
	S3	MF	S3	MF	S3	MF
Conference room	80.0	81.7	8.8	15.1	48.5	58.7
Elevator lobby	59.7	72.8	18.8	12.8	64.1	46.4
Hallway	49.4	55.7	44.5	58.5	44.4	58.0
Lab	52.8	30.1	11.8	27.2	14.2	30.6
Lounge	42.9	39.4	28.6	36.6	62.0	40.5
Office	62.5	76.1	78.1	45.6	98.6	60.2

and taking the cosine similarity. We observe that the semantic accuracy with language is higher for most region types, with the exception of hallways that show minimal improvement since they were rarely described by users. Some regions such as labs, which were labeled with egocentric descriptions, have low scores because the regions are over-segmented and the language is attributed only to the current region. Figure 7 compares the region category properties with and without language. In the absence of language (Fig. 7(a)), the appearance of the region gives equal likelihood for both "elevator lobby" and "lounge." In Fig. 7(b), the region was grounded with the label "lounge" and the framework inferred a higher likelihood of the region category being a lounge.

E. Grounding Allocentric Language Descriptions

We also tested our framework with allocentric language descriptions. When handling phrases that include a landmark and a referent (e.g., "the gym is down the hall"), our earlier framework required the landmark to have already been labeled before describing the referent location. With our new framework, the robot is able to ground language when the landmark corresponds to locations that may not have been labeled, but can be inferred from other semantic cues (e.g., appearance classification). We tested this situation using several instances in our dataset.

Figure 8 shows instances in which allocentric language utterances were grounded into the semantic graph. As the label distributions for the surrounding regions demonstrate, the framework is able to ground the referent with a high degree of accuracy, even though the landmark was never explicitly labeled. However, since there is more uncertainty about the landmark region, the information derived from the allocentric language has less influence on the semantic properties on the region (since we marginalize the landmark likelihood when calculating the grounding likelihood Φ).



(a) (b)

Fig. 8. Resulting category distribution for complex language phrases (a) "the elevator lobby is down the hall" and (b) "the lounge is down the hall".

VI. Conclusion

We described a framework that enables robots to construct spatial-semantic representations of environments from natural language descriptions and scene classifications. Compared to earlier methods, our approach handles natural language descriptions more efficiently, and produces semantic representations that are richer and more compact.

Currently, our method grounds descriptive language under the assumption that it refers to concepts that exist in the robot's representation, either due to previously interpreted descriptions or to appearance classification. We are exploring a mechanism that reasons about the presence of the referred entities and grounds the utterance only when it is confident about the validity of the grounding. We also plan to integrate additional semantic cues that inform region labels and categories, such as the presence of common objects found indoors. In our current implementation, the factors between region category and labels were constructed using a set of predefined labels and relationships identified by us (using synonyms for region types). In the future we plan to learn these relationships from data.

Our approach maintains multiple hypotheses regarding the region assignments of the topology by using particles to sample region assignments based on a prior that considers the spatial similarity of nodes. However, particle weights are calculated based only on laser data. This can lead to valid segmentation hypotheses being discarded during resampling if the laser observation likelihood is low. We plan to incorporate region hypothesis scores based on appearance. In the current approach, the robot passively accepts human assertions about the environment. We also plan to explore more interactive scenarios where the robot reasons over its hypothesis space and asks questions of the human to resolve ambiguities.

VII. Acknowledgments

We would like to thank S. Shemet and F. Paerhati for their help collecting and annotating datasets as well as in training the appearance models. This work was supported in part by the Robotics Consortium of the U.S Army Research Laboratory under the Collaborative Technology Alliance Program, Cooperative Agreement W911NF-10-2-0016.

References

[1] G. Bugmann, E. Klein, S. Lauria, and T. Kyriacou, "Corpus-based robotics: A route instruction example," *Proc. Intelligent Autonomous Systems*, pp. 96–103, 2004.

[2] J. Dzifcak, M. Scheutz, C. Baral, and P. Schermerhorn, "What to do and how to do it: Translating natural language directives into temporal and dynamic logic representation for goal management and action execution," in *Proc. IEEE Int'l Conf. on Robotics and Automation (ICRA)*, 2009, pp. 4163–4168.

[3] C. Matuszek, D. Fox, and K. Koscher, "Following directions using statistical machine translation," in *Proc. ACM/IEEE Int'l. Conf. on Human-Robot Interaction (HRI)*, 2010, pp. 251–258.

[4] S. Tellex, T. Kollar, S. Dickerson, M. R. Walter, A. G. Banerjee, S. Teller, and N. Roy, "Understanding natural language commands for robotic navigation and mobile manipulation," in *Proc. Nat'l Conf. on Artificial Intelligence (AAAI)*, 2011, pp. 1507–1514.

[5] D. L. Chen and R. J. Mooney, "Learning to interpret natural language navigation instructions from observations," in *Proc. Nat'l Conf. on Artificial Intelligence (AAAI)*, 2011, pp. 859–865.

[6] K. Nickel and R. Stiefelhagen, "Visual recognition of pointing gestures for human-robot interaction," *Image and Vision Computing*, vol. 25, no. 12, pp. 1875–1884, Dec. 2007.

[7] B. Kuipers, "The spatial semantic hierarchy," *Artificial Intelligence*, vol. 119, no. 1, pp. 191–233, 2000.

[8] H. Zender, O. Martínez Mozos, P. Jensfelt, G. Kruijff, and W. Burgard, "Conceptual spatial representations for indoor mobile robots," *Robotics and Autonomous Systems*, vol. 56, no. 6, pp. 493–502, 2008.

[9] S. Hemachandra, T. Kollar, N. Roy, and S. Teller, "Following and interpreting narrated guided tours," in *Proc. IEEE Int'l Conf. on Robotics and Automation (ICRA)*, 2011, pp. 2574–2579.

[10] A. Pronobis and P. Jensfelt, "Large-scale semantic mapping and reasoning with heterogeneous modalities," in *Proc. IEEE Int'l Conf. on Robotics and Automation (ICRA)*, 2012, pp. 3515–3522.

[11] M. R. Walter, S. Hemachandra, B. Homberg, S. Tellex, and S. Teller, "Learning semantic maps from natural language descriptions," in *Proc. Robotics: Science and Systems (RSS)*, 2013.

[12] M. MacMahon, B. Stankiewicz, and B. Kuipers, "Walk the talk: Connecting language, knowledge, and action in route instructions," in *Proc. Nat'l Conf. on Artificial Intelligence (AAAI)*, 2006, pp. 1475–1482.

[13] M. Skubic, D. Perzanowski, S. Blisard, A. Schultz, W. Adams, M. Bugajska, and D. Brock, "Spatial language for human-robot dialogs," *IEEE Trans. on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 34, no. 2, pp. 154–167, 2004.

[14] T. Kollar, S. Tellex, D. Roy, and N. Roy, "Toward understanding natural language directions," in *Proc. ACM/IEEE Int'l. Conf. on Human-Robot Interaction (HRI)*, 2010, pp. 259–266.

[15] S. Tellex, P. Thaker, R. Deits, T. Kollar, and N. Roy, "Toward information theoretic human-robot dialog," in *Proc. Robotics: Science and Systems (RSS)*, 2012.

[16] C. Matuszek, N. FitzGerald, L. Zettlemoyer, L. Bo, and D. Fox, "A joint model of language and perception for grounded attribute learning," in *Proc. Int'l Conf. on Machine Learning (ICML)*, 2012.

[17] J.-L. Blanco, J. Gonzalez, and J. Fernandez-Madrigal, "Consistent observation grouping for generating metric-topological maps that improves robot localization," in *Proc. IEEE Int'l Conf. on Robotics and Automation (ICRA)*, 2006.

[18] M. Kaess, A. Ranganathan, and F. Dellaert, "iSAM: Incremental smoothing and mapping," *Trans. on Robotics*, vol. 24, no. 6, pp. 1365–1378, 2008.

[19] J. M. Mooij, "libDAI: A free and open source C++ library for discrete approximate inference in graphical models," *J. Machine Learning Research*, vol. 11, pp. 2169–2173, Aug. 2010.

[20] O. Mozos, C. Stachniss, and W. Burgard, "Supervised learning of places from range data using adaboost," in *Proc. IEEE Int'l Conf. on Robotics and Automation (ICRA)*, 2005.

[21] A. Pronobis, O. M. Mozos, B. Caputo, and P. Jensfelt, "Multi-modal semantic place classification," *Int'l J. of Robotics Research*, vol. 29, no. 2–3, pp. 298–320, Feb. 2010. [Online]. Available: http://www.pronobis.pro/publications/pronobis2010ijrr