Learning Semantic Maps from Natural Language Descriptions

Matthew R. Walter,[1] Sachithra Hemachandra,[1] Bianca Homberg, Stefanie Tellex, and Seth Teller
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139 USA
{mwalter, sachih, bhomberg, stefie10, teller}@csail.mit.edu

***Abstract*—This paper proposes an algorithm that enables robots to efficiently learn human-centric models of their environment from natural language descriptions. Typical semantic mapping approaches augment metric maps with higher-level properties of the robot's surroundings (e.g., place type, object locations), but do not use this information to improve the metric map. The novelty of our algorithm lies in fusing high-level knowledge, conveyed by speech, with metric information from the robot's low-level sensor streams. Our method jointly estimates a hybrid metric, topological, and semantic representation of the environment. This *semantic graph* provides a common framework in which we integrate concepts from natural language descriptions (e.g., labels and spatial relations) with metric observations from low-level sensors. Our algorithm efficiently maintains a factored distribution over semantic graphs based upon the stream of natural language and low-level sensor information. We evaluate the algorithm's performance and demonstrate that the incorporation of information from natural language increases the metric, topological and semantic accuracy of the recovered environment model.**

Fig. 1. A user giving a tour to a robotic wheelchair designed to assist residents in a long-term care facility.

I. Introduction

Until recently, robots that operated outside the laboratory were limited to controlled, prepared environments that explicitly prevent interaction with humans. There is an increasing demand, however, for robots that operate not as machines used in isolation, but as co-inhabitants that assist people in a range of different activities. If robots are to work effectively as our teammates, they must become able to efficiently and flexibly interpret and carry out our requests. Recognizing this need, there has been increased focus on enabling robots to interpret natural language commands [1, 2, 3, 4, 5]. This capability would, for example, enable a first responder to direct a micro-aerial vehicle by speaking "fly up the stairs, proceed down the hall, and inspect the second room on the right past the kitchen." A fundamental challenge is to correctly associate linguistic elements from the command to a robot's understanding of the external world. We can alleviate this challenge by developing robots that formulate knowledge representations that model the higher-level semantic properties of their environment.

We propose an approach that enables robots to efficiently learn human-centric models of the observed environment from a narrated, guided tour (Fig. 1) by fusing knowledge inferred from natural language descriptions with conventional low-level sensor data. Our method allows people to convey meaningful concepts, including semantic labels and relations for both local and distant regions of the environment, simply by speaking to the robot. The challenge lies in effectively combining these noisy, disparate sources of information. Spoken descriptions convey concepts (e.g., "the second room on the right") that are ambiguous with regard to their metric associations: they may refer to the region that the robot currently occupies, to more distant parts of the environment, or even to aspects of the environment that the robot will never observe. In contrast, the sensors that robots commonly employ for mapping, such as cameras and LIDARs, yield metric observations arising only from the robot's immediate surroundings.

To handle ambiguity, we propose to combine metric, topological, and semantic environment representations into a *semantic graph*. The metric layer takes the form of an occupancy-grid that models local perceived structure. The topological layer consists of a graph in which nodes correspond to reachable regions of the environment, and edges denote pairwise spatial relations. The semantic layer contains the labels with which people refer to regions. This knowledge representation is well-suited to fusing concepts from spoken descriptions with the robot's metric observations of its surroundings.

[1]The first two authors contributed equally to this paper.

We estimate a joint distribution over the semantic, topological and metric maps, conditioned on the language and the metric observations from the robot's proprioceptive and exteroceptive sensors. The space of semantic graphs, however, increases combinatorially with the size of the environment. We efficiently maintain the distribution using a Rao-Blackwellized particle filter [6] to track a factored form of the joint distribution over semantic graphs. Specifically, we approximate the marginal over the space of topologies with a set of particles, and analytically model conditional distributions over metric and semantic maps as Gaussian and Dirichlet, respectively. The algorithm updates these distributions iteratively over time using spoken descriptions and sensor measurements. We model the likelihood of natural language utterances with the Generalized Grounding Graph (G^3) framework [2]. Given a description, the G^3 model induces a distribution over semantic labels for the nodes in the semantic graph that we then use to update the Dirichlet distribution. The algorithm uses the resulting semantic distribution to propose modifications to the graph, allowing semantic information to influence the metric and topological layers.

We demonstrate our algorithm through three "guided tour" experiments within mixed indoor-outdoor environments. The results demonstrate that by effectively integrating knowledge from natural language descriptions, the algorithm efficiently learns semantic environment models and achieves higher accuracy than existing methods.

II. Related Work

Several researchers have augmented lower-level metric maps with higher-level topological and/or semantic information [7, 8, 9, 10, 11]. Zender et al. [9] describe a framework for office environments in which the semantic layer models room categories and their relationship with the labels of objects within rooms. The system can then classify room types based upon user-asserted object labels. Pronobis and Jensfelt [8] describe a multi-modal probabilistic framework incorporating semantic information from a wide variety of modalities including detected objects, place appearance, and human-provided information. These approaches focus on augmenting a metric map with semantic information, rather than jointly estimating the two representations. They do not demonstrate improvement of metric accuracy using semantic information.

The problem of mapping linguistic elements to their corresponding manifestation in the external world is referred to as the symbol grounding problem [12]. In the robotics domain, the grounding problem has been mainly addressed in the context of following natural language commands [1, 2, 3, 13, 14, 15, 16, 17]. Cantrell et al. [18] described an approach that updates the symbolic state, but not the metric state, of the environment.

A contribution of the proposed algorithm is a probabilistic framework that uses learned semantic properties of the environment to efficiently identify loop closures, a fundamental problem in simultaneous localization and mapping (SLAM). Semantic observations, however, are not the only information source useful for place recognition. A number of solutions exist that identify loop closures based upon visual appearance [19, 20] and local metric structure [21], among others.

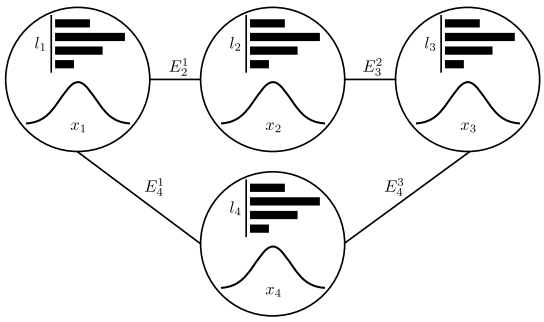


Fig. 2. An example of a semantic graph.

III. Building Semantic Maps with Language

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

A. Semantic Graphs

We model the environment as a set of *places*, regions in the environment a fixed distance apart that the robot has visited. We represent each place by its pose x_i in a global reference frame, and a label l_i (e.g., "gym," "hallway"). More formally, we represent the environment by the tuple $\{G, X, L\}$ that constitutes the semantic graph. The graph $G = (V, E)$ denotes the environment topology with a vertex $V = \{v_1, v_2, \ldots, v_t\}$ for each place that the robot has visited, and undirected edges E that signify observed relations between vertices, based on metric or semantic information. The vector $X = [x_1, x_2, \ldots, x_t]$ encodes the pose associated with each vertex. The set $L = \{l_1, l_2, \ldots, l_t\}$ includes the semantic label l_i associated with each vertex. The semantic graph (Fig. 2) grows as the robot moves through the environment. Our method adds a new vertex v_{t+1} to the topology after the robot travels a specified distance, and augments the vector of poses and collection of labels with the corresponding pose x_{t+1} and labels l_{t+1}, respectively. This model resembles the pose graph representation commonly employed by SLAM solutions [22].

B. Distribution Over Semantic Graphs

We estimate a joint distribution over the topology G_t, the vector of locations X_t, and the set of labels L_t. Formally, we maintain this distribution over semantic graphs $\{G_t, X_t, L_t\}$ at time t conditioned upon the history of metric exteroceptive sensor data $z^t = \{z_1, z_2, \ldots, z_t\}$, odometry $u^t = \{u_1, u_2, \ldots, u_t\}$, and natural language descriptions $\lambda^t = \{\lambda_1, \lambda_2, \ldots, \lambda_t\}$:

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

Each 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:

$$p(G_t, X_t, L_t | z^t, u^t, \lambda^t) = p(L_t | X_t, G_t, z^t, u^t, \lambda^t) \times p(X_t | G_t, z^t, u^t, \lambda^t) \times p(G_t | z^t, u^t, \lambda^t) \quad (2)$$

This factorization explicitly models the dependence of the labels on the topology and place locations, as well as the metric map's dependence on the constraints induced by the topology.

The space of possible graphs for a particular environment is spanned by the allocation of edges between nodes. The number of edges, however, can be exponential in the number of nodes. Hence, maintaining the full distribution over graphs is intractable for all but trivially small environments. To overcome this complexity, we assume as in Ranganathan and Dellaert [23] that the distribution over graphs is dominated by a small subset of topologies while the likelihood associated with the majority of topologies is nearly zero. In general, this assumption holds when the environment structure (e.g., indoor, man-made) or the robot motion (e.g., exploration) limits connectivity. In addition, conditioning the graph on the spoken descriptions further increases the peakedness of the distribution because it decreases the probability of edges when the labels and semantic relations are inconsistent with the language.

The assumption that the distribution is concentrated around a limited set of topologies suggests the use of particle-based methods to represent the posterior over graphs, $p(G_t | z^t, u^t, \lambda^t)$. Inspired by the derivation of Ranganathan and Dellaert [23] for topological SLAM, we employ Rao-Blackwellization to model the factored formulation (2), whereby we accompany the sample-based distribution over graphs with analytic representations for the conditional posteriors over the node locations and labels. Specifically, we represent the posterior over the node poses $p(X_t | G_t, z^t, u^t, \lambda^t)$ by a Gaussian, which we parametrize in the canonical form. We maintain a Dirichlet distribution that models the posterior distribution over the set of node labels $p(L_t | X_t, G_t, z^t, u^t, \lambda^t)$.

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

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

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

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

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

Algorithm 1 outlines the process by which we recursively update the distribution over semantic graphs (2) to reflect the latest robot motion, metric sensor data, and utterances. The following sections explain each step in detail.

Algorithm 1: Semantic Mapping Algorithm

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

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

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

1) Employ proposal distribution $p(G_t | G_{t-1}^{(i)}, z^{t-1}, u^t, \lambda^t)$ to propagate the graph sample $G_{t-1}^{(i)}$ according to odometry u_t and current distributions over labels $L_{t-1}^{(i)}$ and poses $X_{t-1}^{(i)}$.

2) Update the Gaussian distribution over the node poses $X_t^{(i)}$ according to the constraints induced by the newly-added graph edges.

3) Update the Dirichlet distribution over the current and adjacent nodes $L_t^{(i)}$ according to the 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.

C. Augmenting the Graph using the Proposal Distribution

Given the posterior distribution over the semantic graph at time $t-1$, we first compute the prior distribution over the graph G_t. We do so by sampling from a proposal distribution that is the predictive prior of the current graph given the previous graph and sensor data, and the recent odometry and language:

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

We formulate the proposal distribution by first augmenting the graph to reflect the robot's motion. Specifically, we add a node v_t to the graph that corresponds to the robot's current pose with an edge to the previous node v_{t-1} that represents the temporal constraint between the two poses. We denote this intermediate graph as G_t^-. Similarly, we add the new pose as predicted by the robot's motion model to the vector of poses X_t^- and the node's label to the label vector L_t^- according to the process described in Subsection III-E.[2]

We formulate the proposal distribution (5) in terms of the likelihood of adding edges between nodes in this modified graph G_t^-. The system considers two forms of additional edges: first, those suggested by the spatial distribution of nodes and second, by the semantic distribution for each node.

1) Spatial Distribution-based Constraints: We first propose connections between the robot's current node v_t and others in the graph based upon their metric location. We do so by sampling from a distance-based proposal distribution biased towards nodes that are spatially close. Doing so requires marginalization over the distances d_t between node pairs, as shown in equation (6) (we have omitted the history of language

[2]The label update explains the presence of the latest language λ_t.

observations λ^t, metric measurements z^{t-1}, and odometry u^t for brevity). Equation (6a) reflects the assumption that additional edges expressing constraints involving the current node $e_{tj} \notin E^-$ are conditionally independent. Equation (6c) approximates the marginal in terms of the distance between the two nodes associated with the additional edge.

$$p_a(G_t|G_t^-, z^{t-1}, u^t, \lambda^t) = \prod_{j:e_{tj}\notin E^-} p(G_t^{tj}|G_t^-) \tag{6a}$$

$$= \prod_{j:e_{tj}\notin E^-} \int_{X_t^-} p(G_t^{tj}|X_t^-, G_t^-, u_t) p(X_t^-|G_t^-) \tag{6b}$$

$$\approx \prod_{j:e_{tj}\notin E^-} \int_{d_{tj}} p(G_t^{tj}|d_{tj}, G_t^-) p(d_{tj}|G_t^-), \tag{6c}$$

The conditional distribution $p(G_t^{tj}|d_{tj}, G_{t-1}, z^{t-1}, u^t)$ expresses the likelihood of adding an edge between nodes v_t and v_j based upon their spatial location. We represent the distribution for a particular edge between vertices v_i and v_j a distance $d_{ij} = |x_i - x_j|_2$ apart as

$$p(G_t^{ij}|d_{ij}, G_t^-, z^{t-1}, u^t) \propto \frac{1}{1+\gamma d_{ij}^2}, \tag{7}$$

where γ specifies distance bias. For the evaluations in this paper, we use $\gamma = 0.2$. We approximate the distance prior $p(d_{tj}|G_t^-, z^{t-1}, u^t)$ with a folded Gaussian distribution. The algorithm samples from the proposal distribution (6) and adds the resulting edges to the graph. In practice, we use laser scan measurements to estimate the corresponding transformation.

2) Semantic Map-based Constraints: A fundamental contribution of our method is the ability for the semantic map to influence the metric and topological maps. This capability results from the use of the label distributions to perform place recognition. The algorithm identifies loop closures by sampling from a proposal distribution that expresses the semantic similarity between nodes. In similar fashion to the spatial distance-based proposal, computing the proposal requires marginalizing over the space of labels:

$$p_a(G_t|G_t^-, z^{t-1}, u^t, \lambda^t) = \prod_{j:e_{tj}\notin E^-} p(G_t^{tj}|G_t^-, \lambda_t) \tag{8a}$$

$$= \prod_{j:e_{tj}\notin E^-} \sum_{L_t^-} p(G_t^{tj}|L_t^-, G_t^-, \lambda_t) p(L_t^-|G_t^-) \tag{8b}$$

$$\approx \prod_{j:e_{tj}\notin E^-} \sum_{l_t^-, l_j^-} p(G_t^{tj}|l_t^-, l_j^-, G_t^-) p(l_t^-, l_j^-|G_t^-), \tag{8c}$$

where we have omitted the metric, odometry, and language inputs for clarity. The first line follows from the assumption that additional edges that express constraints to the current node $e_{tj} \notin E^-$ are conditionally independent. The second line represents the marginalization over the space of labels, while the last line results from the assumption that the semantic edge likelihoods depend only on the labels for the vertex pair. We model the likelihood of edges between two nodes as non-zero for the same label:

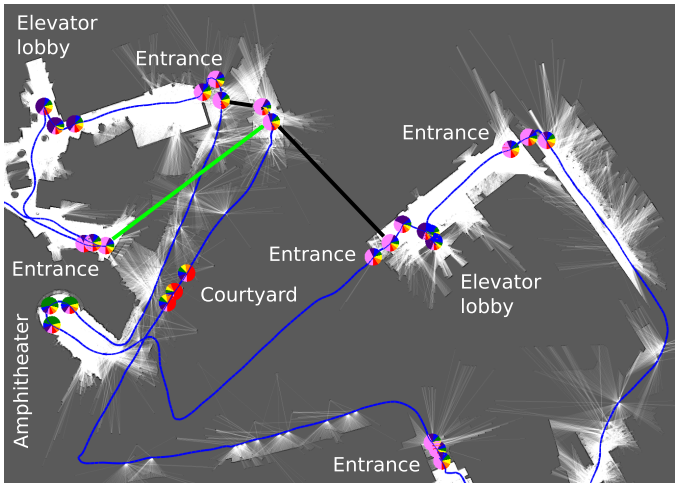


Fig. 3. Depicted as pie charts, the nodes' label distributions are used to propose new graph edges. The algorithm rejects invalid edges that result from ambiguous labels (black) and adds the valid edge (green) to the graph.

$$p(G_t^{tj}|l_t, l_j) = \begin{cases} \theta_{l_t} & \text{if } l_t = l_j \\ 0 & \text{if } l_t \neq l_j \end{cases} \tag{9}$$

where θ_{l_t} denotes the label-dependent likelihood that edges exist between nodes with the same label. In practice, we assume a uniform saliency prior for each label. Equation (8c) then measures the cosine similarity between the label distributions.

We sample from the proposal distribution (8a) to hypothesize new semantic map-based edges. As with distance-based edges, we estimate the transformation associated with each edge based upon local metric observations. Figure 3 shows several different edges sampled from the proposal distribution at one stage of a tour. Here, the algorithm identifies candidate loop closures between different "entrances" in the environment and accepts those (shown in green) whose local laser scans are consistent. Note that some particles may add invalid edges (e.g., due to perceptual aliasing), but their weights will decrease as subsequent measurements become inconsistent with the hypothesis.

D. Updating the Metric Map Based on New Edges

The proposal step results in the addition, to each particle, of a new node at the current robot pose, along with an edge representing its temporal relationship to the previous node. The proposal step also hypothesizes additional loop-closure edges. Next, the algorithm incorporates these relative pose constraints into the Gaussian representation for the marginal distribution over the map

$$p(X_t|G_t, z^t, u^t, \lambda^t) = \mathcal{N}^{-1}(X_t; \Sigma_t^{-1}, \eta_t), \tag{10}$$

where Σ_t^{-1} and η_t are the information (inverse covariance) matrix and information vector that parametrize the canonical form of the Gaussian. We utilize the iSAM algorithm [22] to update the canonical form by iteratively solving for the QR factorization of the information matrix.

E. Updating the Semantic Map Based on Natural Language

Next, the algorithm updates each particle's analytic distribution over the current set of labels $L_t = \{l_{t,1}, l_{t,2}, \ldots, l_{t,t}\}$. This update reflects label information conveyed by spoken descriptions as well as that suggested by the addition of edges to the graph. In maintaining the distribution, we assume that the node labels are conditionally independent:

$$p(L_t|X_t, G_t, z^t, u^t, \lambda^t) = \prod_{i=1}^{t} p(l_{t,i}|X_t, G_t, z^t, u^t, \lambda^t). \quad (11)$$

This assumption ignores dependencies between labels associated with nearby nodes, but simplifies the form for the distribution over labels associated with a single node. We model each node's label distribution as a Dirichlet distribution of the form

$$\begin{aligned} p(l_{t,i}|\lambda_1 \ldots \lambda_t) &= Dir(l_{t,i}; \alpha_1 \ldots \alpha_K) \\ &= \frac{\Gamma(\sum_1^K \alpha_i)}{\Gamma(\alpha_1) \times \ldots \times \Gamma(\alpha_K)} \prod_{k=1}^{K} l_{t,i,k}^{\alpha_k - 1}. \end{aligned} \quad (12)$$

We initialize parameters $\alpha_1 \ldots \alpha_K$ to 0.2, corresponding to a uniform prior over the labels. Given subsequent language, this favors distributions that are peaked around a single label.

We consider two forms of natural language inputs. The first are *simple* utterances that refer only to the robot's current position, such as "This is the gym." The second are expressions that convey semantic information and spatial relations associated with possibly distant regions in the environment, such as "The kitchen is down the hall," which include a figure ("the kitchen") and landmark ("the hall"). We have implemented our complex language system with the words "through," "down," "away from," and "near."

To understand the expression "The kitchen is down the hall," the system must first ground the landmark phrase "the hall" to a specific object in the environment. It must then infer an object in the environment that corresponds to the word "the kitchen." One can no longer assume that the user is referring to the current location as "the kitchen" (referent) or that the hall's (landmark) location is known. We use the label distribution to reason over the possible nodes that denote the landmark. We account for the uncertainty in the figure by formulating a distribution over the nodes in the topology that expresses their likelihood of being the referent. We arrive at this distribution using the G^3 framework [2] to infer groundings for the different parts of the natural language description. In the case of this example, the framework uses the multinomial distributions over labels to find a node corresponding to "the hall" and induces a probability distribution over kitchens based on the nodes that are "down the hall" from the identified landmark nodes.

For both types of expressions, the algorithm updates the semantic distribution according to the rule

$$\begin{aligned} &p(l_{t,i}|\lambda_t = (k,i), l_{t-1,i}) = \\ &\frac{\Gamma(\sum_1^K \alpha_i^{t-1} + \Delta\alpha)}{\Gamma(\alpha_1^{t-1}) \times \ldots \times \Gamma(\alpha_k^{t-1} + \Delta\alpha) \times \ldots \times \Gamma(\alpha_K)} \prod_{k=1}^{K} l_{t,i,k}^{\alpha_k - 1}, \end{aligned} \quad (13)$$

where $\Delta\alpha$ is set to the likelihood of the grounding. In the case of simple language, the grounding is trivial, and we use $\Delta\alpha = 1$ for the current node in the graph. For complex expressions, we use the likelihood from G^3 for $\Delta\alpha$.

G^3 creates a vector of grounding variables Γ for each linguistic constituent in the natural language input λ. The top-level constituent γ_a corresponds to the graph node to which the natural language input refers. Our aim is to find:

$$\Delta\alpha = p(\gamma_a = x_i|\lambda) \quad (14)$$

We compute this probability by marginalizing over groundings for other variables in the language:

$$\Delta\alpha = \sum_{\Gamma/\gamma_a} p(\Gamma|\lambda). \quad (15)$$

G^3 computes this distribution by factoring according to the linguistic structure of the natural language command:

$$\Delta\alpha = \sum_{\Gamma/\gamma_a} \frac{1}{Z} \prod_m f(\gamma^m|\lambda^m) \quad (16)$$

Tellex et al. [2] describe the factorization process in detail.

In addition to input language, we also update the label distribution for a node when the proposal step adds an edge to another node in the graph. These edges may correspond to temporal constraints that exist between consecutive nodes, or they may denote loop closures based upon the spatial distance between nodes that we infer from the metric map. Upon adding an edge to a node for which we have previously incorporated a direct language observation, we propagate the observed label to the newly connected node using a value of $\Delta\alpha = 0.5$.

F. Updating the Particle Weights

Having proposed a new set of graphs $\{G_t^{(i)}\}$ and updated the analytic distributions over the metric and semantic maps for each particle, we update their weights. The update follows from the ratio between the target distribution over the graph and the proposal distribution, and can be shown to be

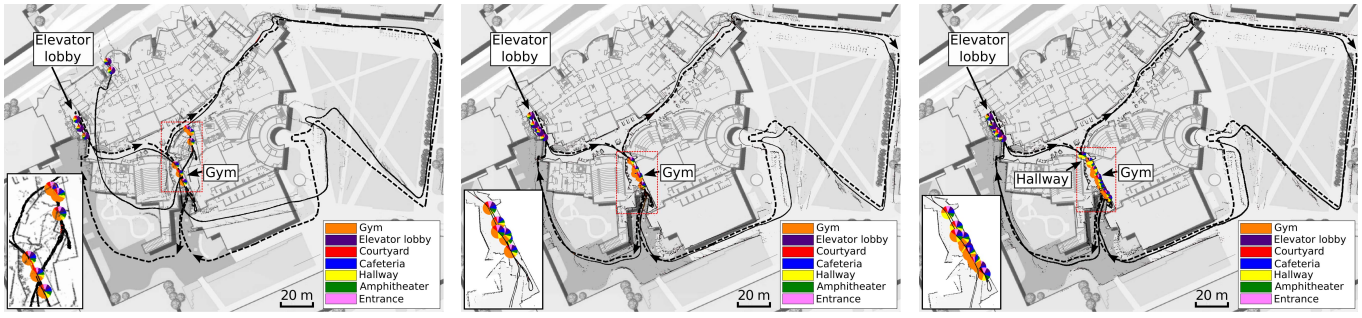
$$\tilde{w}_t^{(i)} = p(z_t|G_t^{(i)}, z^{t-1}, u^t, \lambda^t) \cdot w_{t-1}^{(i)}, \quad (17)$$

where $w_{t-1}^{(i)}$ is the weight of particle i at time $t-1$ and $\tilde{w}_t^{(i)}$ denotes the weight at time t. We evaluate the measurement likelihood (e.g., of LIDAR) by marginalizing over the node poses

$$\begin{aligned} p(z_t|G_t^{(i)}, z^{t-1}, u^t, \lambda^t) = \int_{X_t} &p(z_t|X_t^{(i)}, G_t^{(i)}, z^{t-1}, u^t, \lambda^t) \\ &\times p(X_t^{(i)}|G_t^{(i)}, z^{t-1}, u^t, \lambda^t) dX_t, \end{aligned} \quad (18)$$

which allows us to utilize the conditional measurement model. In the experiments presented next, we compute the conditional likelihood by matching the scans between poses.

After calculating the new importance weights, we periodically perform resampling in which we replace poorly-weighted particles with those with higher weights according to the algorithm of Doucet et al. [6].



(a) No language constraints (b) Simple language (c) Complex language

Fig. 4. Maximum likelihood semantic graphs for the small tour. In contrast to (a) the baseline algorithm, our method incorporates key loop closures based upon (b) simple and (c) complex descriptions that result in metric, topological, and semantic maps that are noticeably more accurate. The dashed line denotes the approximate ground truth trajectory. The inset presents a view of the semantic and topological maps near the gym region.

IV. Results

We evaluate our algorithm through three experiments in which a human gives a robotic wheelchair (Fig. 1) [11] a narrated tour of buildings on the MIT campus. The robot was equipped with a forward-facing LIDAR, wheel encoders, and a microphone. In the first two experiments, the robot was manually driven while the user interjected textual descriptions of the environment. In the third experiment, the robot autonomously followed the human who provided spoken descriptions. Speech recognition was performed manually.

A. Small Tour

In the first experiment (Fig. 4), the user started at the elevator lobby, visited the gym, exited the building, and later returned to the gym and elevator lobby. The user provided textual descriptions of the environment, twice each for the elevator lobby and gym regions. We compare our method with different types of language input against a baseline algorithm.

1) No Language: We consider a baseline approach that directly labels nodes based upon simple language, but does not propose edges based upon label distributions. The baseline emulates typical solutions by augmenting a state-of-the-art iSAM metric map with a semantic layer without allowing semantic information to influence lower layers.

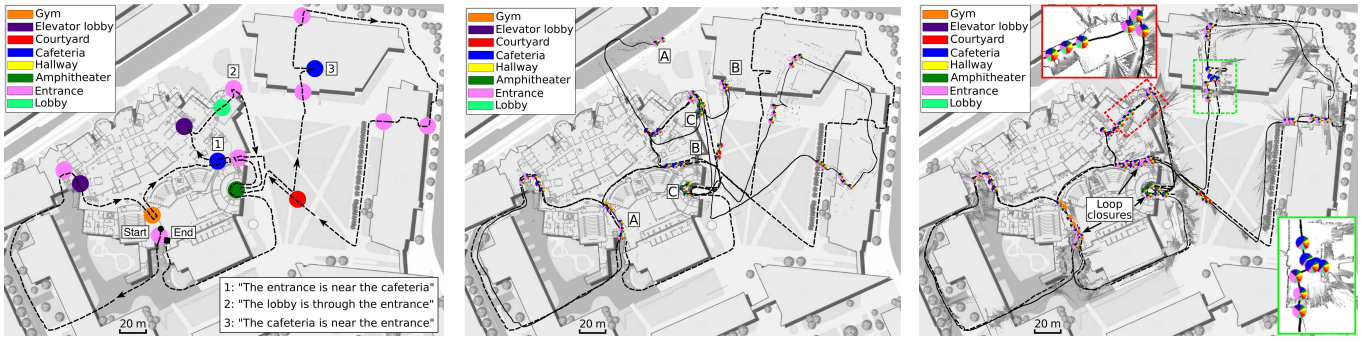
Figure 4(a) presents the resulting metric, topological, and semantic maps that constitute the semantic graph for the highest-weighted particle. The accumulation of odometry drift results in significant errors in the estimate for the robot's pose when revisiting the gym and elevator lobby. Without reasoning over the semantic map, the algorithm is unable to detect loop closures. This results in significant errors in the metric map as well as the semantic map, which hallucinates two separate elevator lobbies (purple) and gyms (orange).

2) Simple Language: We evaluate our algorithm in the case of simple language with which the human references the robot's current position when describing the environment.

Figure 4(b) presents the semantic graph corresponding to the highest-weighted particle estimated by our algorithm. By considering the semantic map when proposing loop closures, the algorithm recognizes that the second region that the user labeled as the gym is the same place that was labeled earlier in the tour. At the time of receiving the second label, drift in the odometry led to significant error in the gym's location much like the baseline result (Fig. 4(a)). The algorithm immediately corrects this error in the semantic graph by using the label distribution to propose loop closures at the gym and elevator lobby, which would otherwise require searching a combinatorially large space. The resulting maximum likelihood map is topologically and semantically consistent throughout and metrically consistent for most of the environment. The exception is the courtyard, where only odometry measurements were available, causing drift in the pose estimate. Attesting to the model's validity, the ground truth topology receives 92.7% of the probability mass and, furthermore, the top four particles are each consistent with the ground truth.

3) Complex Language: Next, we consider algorithm's performance when natural language descriptions reference locations that can no longer be assumed to be the robot's current position. Specifically, we replaced the initial labeling of the gym with an indirect reference of the form "the gym is down the hallway," with the hallway labeled through simple language. The language inputs are otherwise identical to those employed for the simple language scenario and the baseline evaluation.

The algorithm incorporates complex language into the semantic map using the G^3 framework to infer the nodes in the graph that constitute the referent (i.e., the "gym") and the landmark (i.e., the "hallway"). This grounding attributes a non-zero likelihood to all nodes that exhibit the relation of being "down" from the nodes identified as being the "hallway." The inset view in Fig. 4(c) depicts the label distributions that result from this grounding. The algorithm attributes the "gym" label to multiple nodes in the semantic graph as a result of the ambiguity in the referent as well as the G^3 model for the "near" relation. When the user later labels the region after returning from the courtyard, the algorithm proposes a loop closure despite significant drift in the estimate for the robot's pose. As with the simple language scenario, this results in a semantic graph for the environment that is accurate topologically, semantically, and metrically (Fig. 4(c)).



(a) Ground Truth (b) No language constraints (c) Complex language

Fig. 5. Maximum likelihood semantic graphs for (a) the large tour experiment. (b) The result of the baseline algorithm with letter pairs that indicate map components that correspond to the same environment region. (c) Our method with inset views that indicate the inclusion of two complex language descriptions.

B. Large Tour

The second experiment (Fig. 5) considers an extended tour of MIT's Stata Center, two neighboring buildings, and their shared courtyard. The robot visited several places with the same semantic attributes (e.g., elevator lobbies, entrances, and cafeterias) and visited some places more than once (e.g., one cafeteria and the amphitheater). We accompanied the tour with 20 descriptions of the environment that included both simple and complex language.

As with the shorter tour, we compare our method against the baseline semantic mapping algorithm. Figure 5(b) presents the baseline estimate for the environment's semantic graph. Without incorporating complex language or allowing semantic information to influence the topological and metric layers, the resulting semantic graph exhibits significant errors in the metric map, an incorrect topology, and aliasing of the labeled places that the robot revisited. In contrast, Fig. 5(c) demonstrates that, by using semantic information to propose constraints in the topology, our algorithm yields correct topological and semantic maps, and metric maps with notably less error. The resulting model assigns 93.5% of the probability mass to the ground truth topology, with each of the top five particles being consistent with ground truth.

The results highlight the ability of our method to tolerate ambiguities in the labels assigned to different regions of the environment. This is a direct consequence of the use of semantic information, which allows the algorithm to significantly reduce the number of candidate loop closures that is otherwise combinatorial in the size of the map. This enables the particle filter to efficiently model the distribution over graphs. While some particles may propose invalid loop closures due to ambiguity in the labels, the algorithm is able to recover with a manageable number of particles.

For utterances with complex language, the G^3 framework was able to generate reasonable groundings for the referent locations. However, due to the simplistic way in which we define regions, groundings for "the lobby" were not entirely accurate (Fig. 5(c), inset) as grounding valid paths that go "through the entrance" is sensitive to the local metric structure of the landmark (entrance).

C. Autonomous Tour

In the third experiment, the robot autonomously followed a user during a narrated tour along a route similar to that of the first experiment [24]. Using a headset microphone, the user provided spoken descriptions of the environment that included ambiguous references to regions with the same label (e.g., elevator lobbies, entrances). The descriptions included both simple and complex utterances that were manually annotated. Figure 6 presents the maximum likelihood semantic graph that our algorithm estimates. By incorporating information that the natural language descriptions convey, the algorithm recognizes key loop closures that result in accurate semantic maps. The resulting model assigns 82.9% of the probability mass to the ground truth topology, with each of the top nine particles being consistent with ground truth.

V. Conclusion

We described a semantic mapping algorithm enabling robots to efficiently learn metrically accurate semantic maps from natural language descriptions. The algorithm infers rich models

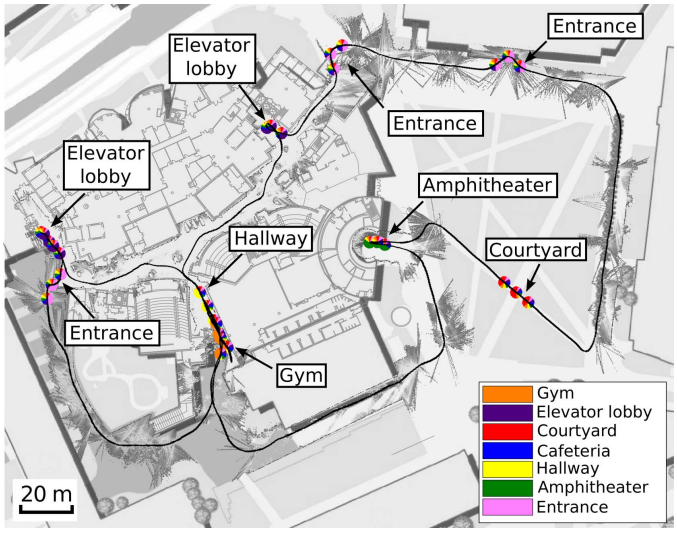


Fig. 6. Maximum likelihood map for the autonomous tour.

of an environment from complex expressions uttered during a narrated tour. Currently, we assume that the robot has previously visited both the landmark and the referent locations, and that the user has already labeled the landmark. As such, the algorithm can incorrectly attribute labels in situations where the user refers to regions that, while they may be visible, the robot has not yet visited. This problem results from the algorithm needing to integrate the spoken information *in situ*. We are currently working on modifying our approach to allow the user to provide a stream of spoken descriptions, and for the robot to later ground the description with sensor observations as needed during environment exploration. This description need not be situated; such an approach offers the benefit that the robot can learn semantic properties of the environment without requiring that the user provide a guided tour.

At present, our method uses traditional sensors to observe only geometric properties of the environment. We are building upon techniques in scene classification, appearance modeling, and object detection to learn more complete maps by inferring higher-level semantic information from LIDAR and camera data. We are also working toward automatic region segmentation in order to create more meaningful topological entities.

Spoken descriptions can convey information about space that includes the types of places, their colloquial names, their locations within the environment, and the types of objects they contain. Our current approach supports assigning labels and spatial relationships to the environment. A direction for future work is to extend the scope of admissible descriptions to include those that convey general properties of the environment. For example, the robot should be able to infer knowledge from statements such as "you can find computers in offices," or "nurses' stations tend to be located near elevator lobbies." Such an extension may build upon existing data-driven efforts toward learning ontologies that describe properties of space.

In summary, we proposed an approach to learning human-centric maps of an environment from user-provided natural language descriptions. The novelty lies in fusing high-level information conveyed by a user's speech with low-level observations from traditional sensors. By jointly estimating the environment's metric, topological, and semantic structure, we demonstrated that the algorithm yields accurate representations of its environment.

VI. Acknowledgments

We thank Nick Roy, Josh Joseph, and Javier Velez for their helpful feedback. We gratefully acknowledge Quanta Computer, which supported this work.

References

[1] 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.

[2] 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.

[3] 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.

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

[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] A. Doucet, N. de Freitas, K. Murphy, and S. Russell, "Rao-Blackwellised particle filtering for dynamic bayesian networks," in *Proc. Conf. on Uncertainty in Artificial Intelligence (UAI)*, 2000, pp. 176–183.

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

[8] 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.

[9] 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.

[10] T. Kollar and N. Roy, "Utilizing object-object and object-scene context when planning to find things," in *Proc. IEEE Int'l Conf. on Robotics and Automation (ICRA)*, 2009, pp. 4116–4121.

[11] 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.

[12] S. Harnad, "The symbol grounding problem," *Physica D*, vol. 42, pp. 335–346, 1990.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] R. Cantrell, K. Talamadupula, P. Schermerhorn, J. Benton, S. Kambhampati, and M. Scheutz, "Tell me when and why to do it!: Run-time planner model updates via natural language instruction," in *Proc. ACM/IEEE Int'l. Conf. on Human-Robot Interaction (HRI)*, 2012, pp. 471–478.

[19] S. Se, D. G. Lowe, and J. J. Little, "Vision-based global localization and mapping for mobile robots," *Trans. on Robotics*, vol. 21, no. 3, pp. 364–375, 2005.

[20] M. Cummins and P. Newman, "FAB-MAP: Probabilistic localization and mapping in the space of appearance," *Int'l J. of Robotics Research*, vol. 27, no. 6, pp. 647–665, 2008.

[21] J.-S. Gutmann and K. Konolige, "Incremental mapping of large cyclic environments," in *Proc. IEEE Int'l. Symp. on Computational Intelligence in Robotics and Automation*, 1999.

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

[23] A. Ranganathan and F. Dellaert, "Online probabilistic topological mapping," *Int'l J. of Robotics Research*, vol. 30, no. 6, pp. 755–771, 2011.

[24] S. Hemachandra, M. R. Walter, S. Tellex, and S. Teller, "Learning semantic maps from natural language descriptions," 2013. [Online]. Available: http://vimeo.com/67438012