A Framework for Learning Semantic Maps from Grounded Natural Language Descriptions

Matthew R. Walter*,[1] Sachithra Hemachandra*,[1] Bianca Homberg* Stefanie Tellex† Seth Teller*

*Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139 USA
{mwalter, sachih, bhomberg, teller}@csail.mit.edu

†Department of Computer Science
Brown University
Providence, RI 02912 USA
stefie10@cs.brown.edu

Abstract

This paper describes a framework that enables robots to efficiently learn human-centric models of their environment from natural language descriptions. Typical semantic mapping approaches are limited to augmenting metric maps with higher-level properties of the robot's surroundings (e.g., place type, object locations) that can be inferred from the robot's sensor data, but do not use this information to improve the metric map. The novelty of our algorithm lies in fusing high-level knowledge that people can uniquely provide through 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 information that the user communicates (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 detail the means by which the framework incorporates knowledge conveyed by the user's descriptions, including the ability to reason over expressions that reference yet unknown regions in the environment. We evaluate the algorithm's ability to learn human-centric maps of several different environments and analyze the knowledge inferred from language and the utility of the learned maps. The results demonstrate that the incorporation of information from free-form descriptions increases the metric, topological and semantic accuracy of the recovered environment model.

1 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 [31, 43, 8, 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.

Figure 1: A user gives a tour to a robotic wheelchair designed to assist residents in a long-term care facility.

Semantic mapping [23, 51, 36] addresses this need by providing robots with human-centric models of their environment. Approaches often take as input low-level sensor (e.g., LIDAR, images) and odometry streams and infer metric, topological, and semantic properties of the environment. Existing algorithms populate the semantic map with scene attributes (e.g., room type) that can be inferred from image- and LIDAR-based classifiers. However, general purpose classifiers are unable to identify many useful properties of an environment, such as the colloquial names associated with each region. For example, it would be difficult to recognize and infer the meaning of the question mark (Fig. 2) that indicates the location of the information desk at MIT's Stata Center, which people frequently use

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

Figure 2: General-purpose sensor-based classifiers would find it difficult to recognize the question mark that indicates the location of an information desk.

as a reference point. Furthermore, the dependence on onboard sensor streams prevents the algorithms from reasoning about parts of the world that are outside the field-of-view of these sensors. This has implications on the efficiency with which robot's can learn human-centric representations of their environment.

We describe an approach first presented by the authors [50] that enables robots to efficiently learn human-centric models of the 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 advantage is that the robot can learn concepts that people are arguably better-able to convey from its opportunistic interaction with humans. The challenge lies in effectively combining these noisy, disparate sources of information. A user's 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 a representation referred to as the *semantic graph* that jointly combines metric, topological, and semantic models of the environment. The metric layer takes the form of a vector of poses for each region in the environment together with the resulting occupancy-grid map that captures the 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 a user's descriptions with the robot's metric observations of its surroundings.

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 [7] 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 descriptions and sensor measurements as they arrive. We model the likelihood of natural language utterances with the Generalized Grounding Graph (G^3) framework [43]. Given a description, the G^3 model induces a learned 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.

This paper builds on our earlier work [50], which presents the initial semantic graph framework. We better place the contributions of our method in the context of the current state-of-the-art in semantic mapping and provide a more detailed description of our estimation framework, including the means by which we interpret natural language descriptions. Additionally, we describe a new capability whereby the method reasons over and learns from *anticipatory* descriptions that refer to regions in the environment not currently in the map. The user can then describe locations that the robot may or may not have previously visited, enabling the robot to more efficiently learn semantic maps of the environment.

We evaluate our algorithm through six "guided tour" experiments that take place within mixed indoor-outdoor environments. We show that, by maintaining a joint distribution over the metric, topological, and semantic maps, the algorithm learns models of the environment that are richer and more accurate than can be achieved with existing language-based semantic mapping algorithms. We analyze the effectiveness with which the algorithm integrates semantic knowledge from natural language descriptions and demonstrate the utility of the learned maps for navigation.

2 Related Work

The field is at the point where robots need to reason over human-centric models of space due in large part to the extensive progress that has been made in solving the Simultaneous Localization and Mapping (SLAM) problem. Not only have contributions to SLAM allowed robots to operate robustly in unstructured environments like our homes, but many existing

approaches to semantic mapping are built upon SLAM algorithms. Beginning with the seminal work of Smith and Cheeseman [42], SLAM is predominantly concerned with building either globally metric or, to a lesser extent, topological maps for the purpose of navigation. Our algorithm differs in that we represent the map as a hierarchy that jointly models the metric, topological, and semantic properties of the environment. The latter two layers are particularly useful for human-centric mapping as the semantic map models properties useful in grounding natural language commands [43], while the topology is consistent with the representation that humans use to model space [27]. We use the topological layer in our semantic graph to induce a pose graph in the same fashion as many of the state-of-the-art SLAM algorithms [21, 9, 34]. Given this topology, we represent the distribution over the metric map as a Gaussian and, like information filter-based SLAM algorithms [46, 9, 49, 18], parametrize the distribution in the canonical form for computational efficiency.

Unlike the SLAM problem, semantic mapping [23, 51, 19, 15, 36] is primarily interested in learning higher-level properties of the robot's environment. These properties include spatial attributes (like metric mapping) as well as concepts such each room's type (e.g., "hallway," or "kitchen"), their colloquial names (e.g., "Carrie's office"), or the objects that they contain. This information is useful for navigation, but also facilitates human-robot interaction, including more efficient command and control mechanisms.

Early work in semantic mapping includes the Spatial Semantic Hierarchy (SSH) proposed by Kuipers [23] that represents a robot's spatial knowledge as a coupled hierarchy. At the lowest level, the local environment is modeled as a collection of control laws, each expressing the relationship between sensory input and motor output, that facilitate localization and generating local geometric maps. Above the control level is the causal level, which provides a discrete model of the actions that transition between each of the control laws. The topological level represents the environment as a collection of regions, places, and paths that abstract states and actions from the causal level. While the topology serves as the primary global map of the environment, the local geometric maps from the control level can be merged via the topology to formulate a global metric map.

Kuipers et al. [24] describe an extension to the SSH that employs a hybrid metric and topological representation to better represent environments at both small and large scales. The Hybrid SSH treats the environment as a collection of interconnected locations, each being small in scale. The method employs metric maps to model the local geometry of distinct regions from which they use local paths to induce a symbolic global topology that describes the large-scale environment. By decoupling the map in this manner, this approach more efficiently models ambiguities in large-scale loop closure with multiple compact topologies, without requiring that the set of local metric maps be registered consistently in a single global reference frame. This is a distinct benefit over submap approaches to SLAM [25, 2] that similarly employ local metric maps but also seek to ensure that these submaps are consistent in a global reference frame. The authors have shown [33, 1] that the representation allows uncertainties to be handled more effectively by factoring them into individual components that capture local metrical, global topological, and globally metrical uncertainties. Our algorithm also consists of a hybrid metric and topological representation and factors the joint distribution into separate metrical and topological terms, employing different hypotheses over the topological map to represent the distribution over the space of loop closures. However, we maintain a globally metric map of the environment with respect to a single frame of reference, which can make our algorithm sensitive to global inconsistencies within large environments. Another difference lies in our definition of regions, which we segment based upon distance traveled. The Hybrid SSH, in contrast, defines regions based upon their local geometric structure (e.g., separated by doorways). Unlike our approach, however, the Hybrid SSH does not model the semantic labels or colloquial names associated with different regions of the environment.

More recent efforts similarly take a hierarchical approach to representing semantic and spatial properties of a robot's environment. Many existing solutions [11, 29, 32, 48, 22, 51, 15, 36] build on the effectiveness of SLAM by augmenting a low-level metric map with layers that encode the topological and semantic properties of the environment. Typically, an off-the-shelf SLAM implementation is used to build the metric layer. One level up in the hierarchy is the topological map, taking the form of a graph, where vertices denote different *places* in the environment and edges model their connectivity. Layered above the topology is the semantic map that represents abstract properties associated with each place, such as their type or the objects that they contain.

One distinction among existing approaches to semantic mapping, and topological mapping in general, is the means by which the environment is segmented into different places. Thrun et al. [45], for example, rely on a user to push a button each time the robot transitions to a new region. A straightforward automatic strategy is to segment regions based upon distance, placing vertices at a fixed spacing as the robot travels in the environment [51]. This is the approach that we take in this paper. An alternative is to use heuristics such as door detections to separate regions [36], which yields segmentations that can be more semantically meaningful within indoor environments. Meanwhile, others have found success partitioning the environment by clustering observations of local metric [3] and semantic [29] properties. Of particular relevance to this work, Ranganathan and Dellaert [38] employ multiple methods to define regions, including manual segmentation at the location of gateways (e.g., doorways, junctions) and automatic segmentation based upon changes in visual appearance [37].

The properties contained within the semantic map are most often inferred from the robot's sensor data (e.g., LIDAR scans and camera images), using scene classifiers [35] and object detectors [47, 19]. For example, Martínez Mozos et al. [29] use a combination of boosted laser range features and image-based object detections to classify the robot's surround as it navigates and show how this can be used to induce a topology for the environment. Similarly, Meger et al. [32] layer a visual attention system and image-based object recognition on top of a

SLAM occupancy grid map to build semantic maps that encode the locations of objects of interest within the environment. Vasudevan and Siegwart [48] describe a probabilistic framework that uses clustered object detections to learn conceptual models of space that express their hierarchical structure (e.g., that an "office" may include a "workspace" and "meeting area") and the objects that they contain. They argue that this model is amenable to a hierarchical metric-topological-semantic SLAM framework, though they leave that for future work.

These solutions rely upon scene classifiers and object detectors to infer the properties that make up the semantic map. The effectiveness of these approaches is a function of the richness of the training data. As such, they perform best when the environments have similar appearance and regular geometry, and when the objects are drawn from a common set. Even in structured settings, it is not uncommon for the regions to be irregular and for the objects to be difficult to recognize, either because they are out of context or are singletons (Fig. 2). Furthermore, scene classification doesn't provide a means to infer the specific labels that humans use for a location, such as "Carrie's office" or the "Kiva conference room."

Our algorithm, on the other hand, is capable of learning the class and colloquial name for different spaces in the environment from a human's description. Recognizing the efficiency of human supervision during learning, several researchers have proposed methods that incorporate user-provided spoken cues into the semantic map [51, 15, 36]. Of particular relevance, Zender et al. [51] use labels assigned by people to identify objects in the robot's surround. They combine these labels with a LIDAR-based place classifier to learn semantic maps for office environments that encode the relationship between room categories and the objects that they contain. Similarly, Pronobis and Jensfelt [36] describe a multi-modal probabilistic framework capable of incorporating semantic information from a wide variety of modalities. These include a user's speech, which is modeled just like any other sensor. The method seeks to learn richer, more descriptive environment models by fusing semantic cues from object detections, place appearance and geometry, as well as human input into a single model.

Our algorithm differs from the existing state-of-the-art in semantic mapping in three fundamental ways. First, our framework employs a learned model of free-form utterances to reason over expressions that are less constrained than those handled by other methods. To be precise, we currently assume that these descriptions involve labels for and spatial relations between one or two locations, though the structure of these expressions is only limited by rules of grammar and the amount of training data. Second, by using scene classification, existing methods can only infer semantic properties of areas that are within the field-of-view of the robot's sensors. Similarly, previous efforts to incorporate user-provided labels assume that the object is within view or that the user is referencing the robot's current location. In contrast, our method reasons over more expressive descriptions that enable robots to learn concepts like labels and spatial relations for distant areas as well as regions of the environment that the robot has not yet visited. Third, existing methods allow updates to the metric layer to influence the topological and semantic maps, but don't use information at the semantic layer improve the rest of the hierarchy. By maintaining a joint distribution over the metric, topological, and semantic properties of the environment, our framework uses updates to any one layer to improve the other layers in the hierarchy. For example, we show how the semantic map can be used to recognize loop closures, a fundamental problem in SLAM, and thereby add edges to the topology that, in turn, correct errors in the metric map.

We note that semantic observations are not the only source of information that is useful for place recognition. Many mapping algorithms build local laser scan patches for each region and correlate these patches to identify loop closures [12]. However, these techniques are prone to perceptual aliasing when the local geometry is not distinctive, such as in the case of hallways. More recent methods consider a region's visual appearance as a more discriminative means of performing place recognition [40, 6, 38]. Of particular note, Cummins and Newman [6] learn a generative model of region appearance using a bag-of-words representation that expresses the commonality of certain features. By effectively modeling this perceptual ambiguity, the authors are able to reject invalid loop closures despite significant aliasing, while correctly recognizing valid loop closures. This and related approaches in effect choose the maximum likelihood loop closure, relying on the assumption that the place model is sufficiently descriptive that the resulting distribution over the space of loop closures is peaked around the true correspondence. Our approach differs in that it uses semantic information to maintain a distribution over the space of loop closures rather than only that which is most likely.

This work incorporates user-provided labels and spatial relations by interpreting the free-form descriptions in the context of the semantic graph. As such, it is important to note existing efforts to solve what Harnad [13] refers to as the symbol grounding problem, the problem of mapping linguistic elements to their corresponding manifestation in the external world. In the robotics domain, the grounding problem has primarily been addressed in the context of following route directions and other natural language commands. One class of solutions [41, 28, 8, 5, 31, 30] considers the problem as one of parsing free-form commands into their formal language equivalent, which a planner takes as input. Other approaches [20, 43, 44] function by mapping free-form utterances into their corresponding object and action referents in the robot's world model. With the exception of MacMahon et al. [28] and Matuszek et al. [30], existing methods assume the map is known a priori. We take the latter approach to grounding spatial language by inferring the locations that the user is referring to in the map that is learned online. However, we allow these locations to be unknown to the robot when the user provides the descriptions.

3 The Semantic Graph

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

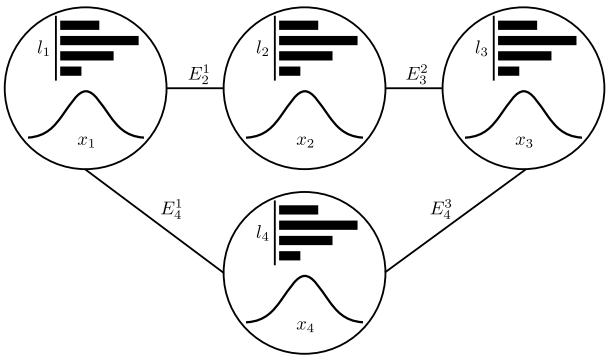


Figure 3: An example of a semantic graph.

The metric map models information contained in the robot's low-level sensor readings. The topological map models the connectivity between regions that can be inferred from navigation as well as natural language descriptions. The semantic map represents categories that the user conveys.

3.1 Semantic Graphs

We model the environment as a set of *places*, regions in the environment a fixed distance apart[2] 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 $\{G_t, X_t, L_t\}$ that constitutes the semantic graph. The graph $G_t = (V_t, E_t)$ denotes the environment topology with a vertex $V_t = \{v_1, v_2, \ldots, v_t\}$ for each place that the robot has visited, and undirected edges E_t that signify observed relations between vertices, based on metric or semantic information. The vector $X_t = [x_1, x_2, \ldots, x_t]$ encodes the pose associated with each vertex. The set $L_t = \{l_1, l_2, \ldots, l_t\}$ includes the semantic label l_i associated with each vertex. The semantic graph (Fig. 3) 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 [18].

Our goal is to induce a distribution over the semantic graph, including the locations, topology, and semantic labels given information about an environment obtained from a robot's range sensors, odometry readings, and the user's descriptions of the environment.

3.2 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 $\lambda^t = \{\lambda_1, \lambda_2, \ldots, \lambda_t\}$:

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

[2]We use 5 m spacing for the results presented in this paper.

Table 1: Semantic Graph Notation

Symbol	Description
$G_t = (V_t, E_t)$	Graph representation of the topology at time t that consists of a set of vertices $V_t = \{v_1, v_2, \ldots, v_t\}$ connected by undirected edges E_t.
L_t	Set of labels $l_{t,i}$ associated with each place.
$l_{t,j}^{(i)}$	Label distribution for node j in particle i.
λ_t	Parsed natural language description of the environment at time t.
X_t	Vector of landmark poses $[x_1, ..., x_t]$
z^t	Set of sensor readings made up to time t by sensors onboard the robot.
u^t	Set of odometry readings up to time t.

Each language variable λ_i denotes a (possibly null) utterance, such as "This is the kitchen," or "The gym is down the hall." Table 1 outlines our notation. We factor the joint posterior into a distribution over the graphs and a conditional distribution over the node poses and labels,

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

The left-most expression in this factorization explicitly models the dependence of the labels on the topology and the location of each region. The middle term encodes the conditional distribution over the metric map given the topology and, in this way, mimics pose graph formulations to SLAM, given the loop closure (i.e., the topology). The right-most expression denotes the distribution over the graph conditioned upon the sensor history and language.

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 [38] 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 [38]. In addition, conditioning the graph on the descriptions further increases the peakedness of the distribution, thereby increasing the validity of this assumption, 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 [38] 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

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

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

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

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

4 Building Semantic Maps with Language

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. In the first step, we propagate each sample $G_{t-1}^{(i)}$, which represents the posterior $p(G_{t-1}|z^{t-1}, u^{t-1}, \lambda^{t-1})$ at time $t-1$, by adding a node for the robot's new pose (connected by an edge to the previous node) and proposing additional loop-closure edges according to the current metric and label distributions. This results in a sample-based estimate for the prior at time t, $p(G_t|z^{t-1}, u^t, \lambda^t)$. Next, we update the Gaussian distribution over the node poses by incorporating the constraints induced by the new loop-closure edges. We then proceed to update the Dirichlet distributions based upon the structure of the graph and parsed language λ_t, if available. Finally, we update the weight $w_t^{(i)}$ according to the likelihood of new metric measurements z_t and resample if needed. We repeat these steps for each particle, yielding the particle set representation $\mathcal{P}_t$ of the new posterior distribution at time t, $p(G_t, X_t, L_t|z^t, u^t, \lambda^t)$. The following sections explain each step in detail.

4.1 Graph Augmentation 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:

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

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 4.3.[3]

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.

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. Propagate the graph sample $G_{t-1}^{(i)}$ using the proposal distribution $p(G_t|G_{t-1}^{(i)}, z^{t-1}, u^t, \lambda^t)$, using 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.

4.1.1 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 marginalizing over the distances d_t between node pairs, as shown in equation (6), where we omit the history of language 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

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

two nodes associated with the additional edge.

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

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

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

The conditional distribution $p(G_t^{tj}|d_{tj}, G_t^-, z^{t-1}, u^t, \lambda^t)$ 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

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

where γ specifies distance bias. For the evaluations in this paper, we use $\gamma = 0.2$. We approximate the distance prior $p(d_{tj}|G_t^-, z^{t-1}, u^t, \lambda^t)$ with a folded Gaussian distribution,

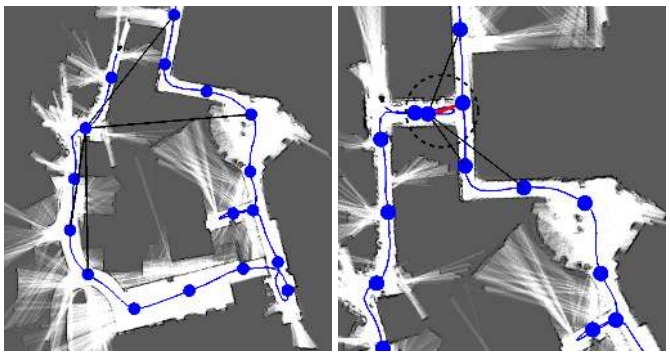
$$p(d_{ij};\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(-d_{ij}-\mu)^2}{2\sigma^2}\right) + \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(d_{ij}-\mu)^2}{2\sigma^2}\right) \tag{8}$$

where μ is the the mean and σ is the standard deviation, approximated based upon a linearized model for the distance between the normally distributed positions x_i and x_j. The probability is 0 for $d_{ij} < 0$.

The algorithm samples from the proposal distribution (6) to identify candidate edges. Before adding these to the graph, we use laser scans to build local maps around each node and compare the maps associated with the two nodes using scan-matching (Fig. 4). This matching allows the method to reject most invalid edges, however it may still yield false positives for areas with ambiguous local geometry. In order to reduce the effects of this perceptual aliasing, we evaluate the likelihood of the scan-matched estimates of the inter-region transformations under our distribution over the metric map. The algorithm retains edges according to their Mahalanobis distance and adds edges deemed to be valid along with their estimated transformations.

4.1.2 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:



(a) Rejected Edges (b) Rejected and Accepted Edges

Figure 4: In the proposal step, the algorithm hypothesizes the addition of new edges in the graph based upon the estimated distance between nodes. Candidate edges are (a) rejected (black) or (b) accepted (red) based upon scan-matching.

$$p_s(G_t|G_t^-, z^{t-1}, u^t, \lambda^t) = \prod_{j:e_{tj}\notin E^-} p(G_t^{tj}|G_t^-, \lambda_t) \tag{9a}$$

$$= \prod_{j:e_{tj}\notin E^-} \sum_{L_t^-} p(G_t^{tj}|L_t^-, G_t^-, \lambda_t)p(L_t^-|G_t^-) \tag{9b}$$

$$\approx \prod_{j:e_{tj}\notin E^-} \sum_{l_t^-, l_j^-} p(G_t^{tj}|l_t^-, l_j^-, G_t^-)p(l_t^-, l_j^-|G_t^-), \tag{9c}$$

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

$$p(G_t^{tj}|l_t, l_j) = \begin{cases} \theta_{l_t} & \text{if } l_t = l_j \\ 0 & \text{if } l_t \neq l_j \end{cases} \tag{10}$$

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 (9c) then measures the cosine similarity between the label distributions.

We sample from the proposal distribution (9) to hypothesize new semantic map-based edges. As with distance-based edges, we validate proposed edges by building local maps for each region and performing scan-matching between these maps. In practice, we additionally introduce a bias that penalizes matches between frequently occurring regions like hallways. Figure 5 shows several different edges sampled from the proposal distribution at one stage of a tour.[4] Here, the algorithm identifies candidate loop closures between different "entrances" in the environment and accepts those (shown in green) whose

[4]Throughout the paper, we only visualize the semantic distribution for nodes whose distribution is not uniform.

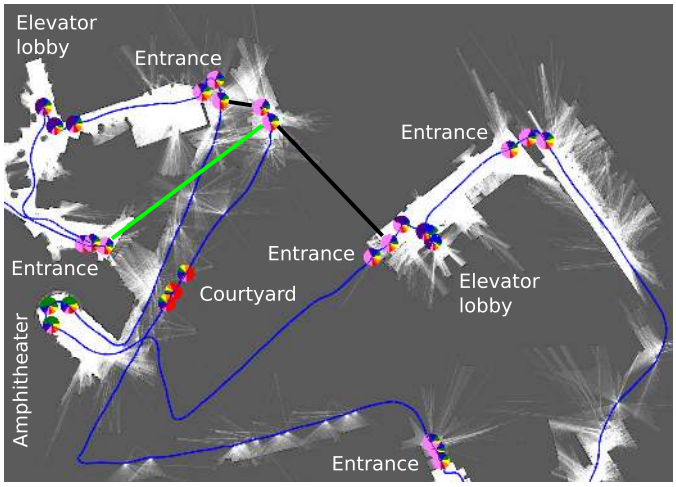


Figure 5: The algorithm proposes new graph edges between node pairs based on their label distributions, which are depicted as pie charts for nodes whose distribution is not uniform. It rejects invalid edges that result from ambiguous labels (black) and adds the edge (green) that denotes a valid loop closure.

local laser scans are consistent. Note that some particles may add invalid edges (e.g., due to perceptual or semantic aliasing), but their weights will decrease as subsequent measurements become inconsistent with the hypothesis.

4.2 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

$$p(X_t|G_t, z^t, u^t, \lambda^t) = \mathcal{N}^{-1}(X_t; \Sigma_t^{-1}, \eta_t), \qquad (11)$$

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 [18] to update the canonical form by iteratively solving for the QR factorization of the information matrix. We omit the details of the algorithm for lack of space and refer the reader to Kaess et al. [18] for more information. Figure 6 shows the resulting metric poses and their uncertainties.

4.3 Updating the Semantic Map Based on Natural Language

Next, the algorithm updates the distribution over the current set of labels $L_t = \{l_{t,1}, l_{t,2}, \ldots, l_{t,t}\}$ associated with each particle. This update reflects information regarding labels and spatial relations that spoken descriptions convey, as well as semantic concepts that are suggested by the addition of edges to the graph. In maintaining the label distribution, we make the assumption that the node labels are conditionally independent

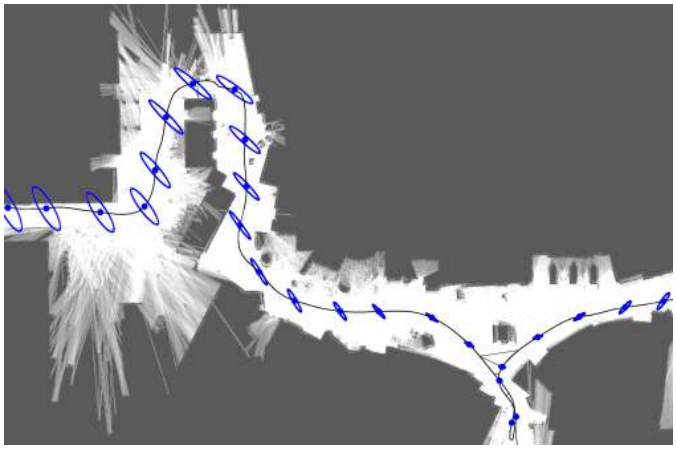


Figure 6: The mean position and 1σ uncertainty ellipse for each node, along with the resulting occupancy grid map.

given the topology and node poses

$$p(L_t|X_t, G_t, z^t, u^t, \lambda^t) = \prod_{i=1}^{t} p(l_{t,i}|X_t, G_t, z^t, u^t, \lambda^t). \quad (12)$$

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

$$\begin{aligned} p(l_{t,i}|\lambda_1 \ldots \lambda_t) &= Dir(l_{t,i}; \alpha_1 \ldots \alpha_K) \\ &= \frac{\Gamma(\sum_1^K \alpha_i)}{\Gamma(\alpha_1) \times \ldots \times \Gamma(\alpha_K)} \prod_{k=1}^{K} l_{t,i,k}^{\alpha_k - 1}, \end{aligned} \quad (13)$$

where $l_{t,i,k}$ for $k \in \{1, \ldots, K\}$ is the k^{th} label associated with node i at time t. We initialize the parameters $\alpha_1 \ldots \alpha_K$ to 0.2, which results in a prior that is uniform over the different labels. Given subsequent language input, this favors distributions that are peaked around a single label.

We consider user-provided expressions that use spatial relations to describe one or two locations in the environment. The first form are *egocentric* utterances (e.g., "This is the gym") that assign labels to the robot's current location. A contribution of our work is the ability to incorporate information from *allocentric* spatial language that express spatial relations and labels that are associated with non-local, potentially distant regions in the environment. By interpreting these expressions, such as "The kitchen is through the cafeteria," our framework enables robots to learn rich semantic maps of their environment more efficiently.

Learning from allocentric expressions is challenging because their groundings are ambiguous—the places to which the user refers are often not obvious. Consider the scenario outlined in Figure 7. The semantic map includes an area that has a high likelihood of being a "lobby" and a second believed to be a "hallway." As the robot (triangle) continues to explore the environment, the user utters the description "The gym is down the hall." Descriptions like these are often ambiguous. For example, there may be multiple "hall" regions in the map or it may

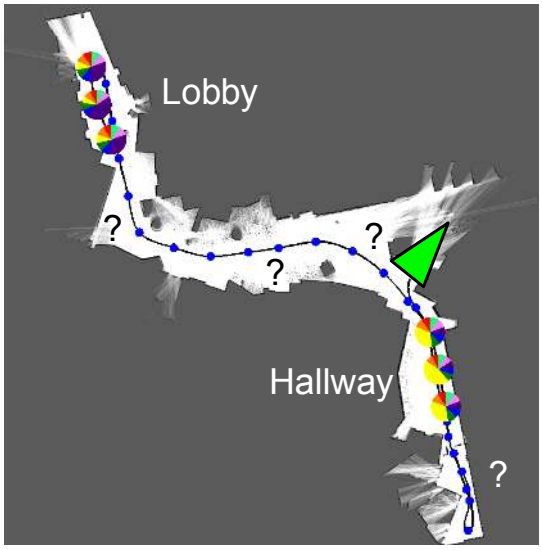


Figure 7: The user utters the description "The gym is down the hall" when the robot is at the location indicated by the triangle.

be that the robot has yet to visit the region that the user is referring to, or if it has, it is not aware of its label. Similarly, several regions in the map are candidates for being the "gym," but the user may also be identifying a region that is not yet in the map.

In order to understand an expression like "The gym is down the hall," the system must first ground the landmark phrase "the hall" to a specific entity in the environment. It must then infer an entity in the environment that corresponds to the word "the gym." One can no longer assume that the user is referring to the current location as "the gym" (the *figure*[5]) or that the location of the "hall" (the *landmark*) is known (e.g., there are likely many "halls" in the environment). We use the label distribution to reason over the possible nodes that denote the landmark. In doing so, we make the additional assumption that the landmark exists in the graph and normalize the likelihoods for candidate "hall" nodes. We later relax this assumption as we describe shortly. 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 figure. Formally, we model the likelihood that each node v_i is the figure by marginalizing over the space of candidate landmarks

$$p(\phi^f_{v_i} = \texttt{T}) = \sum_{v_j} p(\phi^f_{v_i} = \texttt{T} | \phi^l_{v_j} = \texttt{T})\, p(\phi^l_{v_j} = \texttt{T}), \quad (14)$$

where $\phi^l_{v_i}$ and $\phi^f_{v_i}$ are binary-valued random variables that indicate that node v_i is the landmark and figure, respectively. The landmark likelihood $p(\phi^l_{v_j} = \texttt{T})$ follows from the normalized label distributions, as described above. We arrive at the conditional distribution $p(\phi^f_{v_i} = \texttt{T} | \phi^l_{v_j} = \texttt{T})$ using the G^3 framework to infer groundings for the different parts of the description. In the case of this example, the framework induces a probability distribution over nodes whose location is consistent with being "down the hall" from each of the conditioned landmark nodes, based upon the robot's pose at the time the user offers the communication. In this manner, we make the assumption that the person is describing the environment in the robot's frame of reference. The grounding likelihood (14) simplifies with egocentric language as the figure is implicitly the robot's current location.

[5]In spatial linguistic theory, this is often referred to as the *trajector*.

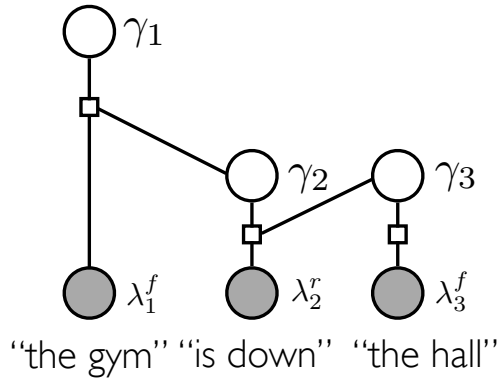


Figure 8: The factor graph model for the utterance "The gym is down the hall" that is used by the G^3 algorithm.

4.3.1 Grounding Natural Language Descriptions with G^3

Before proceeding, we briefly describe the G^3 algorithm, which was initially proposed by Tellex et al. [43]. Given natural language text Λ, G^3 provides a distribution over the space of possible mappings between each word in the parsed description and the corresponding groundings in the external model. This distribution takes the general form

$$p(\Phi | \Gamma, \Lambda, M), \quad (15)$$

where $\Gamma = \{\gamma_1, \ldots, \gamma_n\}$ denotes the set of possible groundings and M represents the robot's world model, which includes the robot's pose and a map of the environment. The correspondence variable Φ contains boolean-valued variables ϕ for each linguistic element $\lambda \in \Lambda$ and grounding $\gamma \in \Gamma$, such that $\phi = \texttt{True}$ iff γ corresponds to λ. In our application, the groundings are the locations of the nodes in the semantic graph the paths between nodes according to the metric map.

Taking advantage of the compositional, hierarchical structure of natural language [17], G^3 parses the utterance into a set of Spatial Description Clauses (SDCs). Each SDC is assigned a type (event, object, place, or path) and consists of landmark λ^l_i, figure λ^f_i, and relation λ^r_i phrases. For the purposes of this work, we parse descriptions into place and path SDCs using a learned grammar that includes possible labels and spatial relations. G^3 then factors the distribution (15) into individual terms, one for each linguistic element

$$p(\Phi | \Gamma, \Lambda, M) = \prod_i p(\phi_i | \lambda_i, \Gamma, M). \quad (16)$$

This factored distribution is represented as a graphical model using a factor graph, such as the one shown in Figure 8 for the "the gym is down the hall" utterance. The G^3 algorithm uses a

log-linear model for each of the factors

$$p(\phi_i|\lambda_i, \Gamma, M) \propto \exp\left(\sum_j \mu_j s_j(\phi_i, \lambda_i, \Gamma, M)\right), \quad (17)$$

where μ_j are weights and s_j are features that encode the relationship between the linguistic element λ_i and the groundings Γ. For example, we use a feature that relates the length of the path through the map from the landmark grounding γ_i^l and figure grounding γ_i^f when the relation λ_i^r is "down from"

$$s(\gamma_i^l, \gamma_i^f, \lambda_i^r) \triangleq |x_{\gamma_i^l} - x_{\gamma_i^f}| \wedge (\text{"down from"} \in \lambda_i^r). \quad (18)$$

Similarly, features for other relations express the consistency of the path between pairs of nodes with the uttered relation. The set of relations for which we have trained feature weights include "through," "down from," "near," and "away from." Additional features include the likelihood of the landmark label λ_i^l under the multinomial associated with the node's γ_i^l label distribution.

The G^3 model learns the weights μ_k associated with each feature by training on a corpus of SDCs from natural language descriptions and the known groundings Γ and correspondences Φ. In particular, we train our G^3 model using a route directions corpus [20] that includes a set spoken directions through an office building and positive and negative examples of paths through the environment.

Given a particular spoken description, we use G^3 to infer groundings for the different parts of the utterance. In the case of the current example, the framework uses the multinomial distributions over labels to find a node corresponding to the "hall" landmark and induces a probability distribution over "gyms" based on the nodes that are "down the hall" from the identified landmark nodes. We ground relational utterances λ_i^r by considering the shortest path that travels from the robot's pose at the time of the description through the pair of landmark γ_i^l and figure γ_i^f node groundings. We use the A* algorithm [39] to solve for the shortest path through the semantic graph topology. We then use features over these paths (18) to evaluate their consistency with the uttered relation (e.g., "down from," "near," and "through"). The likelihood of this path is calculated for each possible figure and landmark pair. We marginalize out the landmarks to arrive at the likelihood of the figure region having the described label

$$p(f_j) = \sum_{l_i} p(\phi_j|f_j, l_i, p_j)p(l_i), \quad (19)$$

where f_j is the figure being evaluated, p_j is the path from the robot's location at the time of the description to the figure, ϕ_j is the corresponding likelihood of the grounding, and l_i is a corresponding landmark.

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

$$p(l_{t,i}|\lambda_t = (k,i), l_{t-1,i}) = \frac{\Gamma(\sum_1^K \alpha_i^{t-1} + \Delta\alpha)}{\Gamma(\alpha_1^{t-1}) \times \ldots \times \Gamma(\alpha_k^{t-1} + \Delta\alpha) \times \ldots \times \Gamma(\alpha_K)} \prod_{k=1}^{K} l_{t,i,k}^{\alpha_k - 1}, \quad (20)$$

where $\Delta\alpha$ is the likelihood of the figure grounding. In the case of egocentric language, when the robot's position is implicitly the figure, we set this likelihood to $\Delta\alpha = 1$ for the current node in the graph. When the descriptions are ambiguous, we set $\Delta\alpha$ to the landmark likelihood computed via Equation 14.

An advantage of having a probabilistic model over the space of groundings is that it provides a means of recognizing when there is not enough information contained in the semantic graph to ground the language. This allows us to recognize many of the situations in which the user describes areas that either the robot hasn't yet visited or they reference landmarks whose labels were never added to the map. For example, it's not uncommon for the user to mention regions that are within sight but they have yet to reach (e.g., the user may say "The lab is across the lobby," but the robot has never been to the region being referred to as "the lab."). We refer to descriptions of this form as *anticipatory*.

We identify instances of anticipatory descriptions by using our distributions over the landmark and figure locations to evaluate the likelihood that the landmark matches a labeled region in the graph and that there are one or mode candidate figure regions consistent with the language. When the method is sufficiently confident in the ability to ground the language (we use a threshold of 0.2), we update the label distributions as described above. However, when the grounding likelihoods suggest an anticipatory description, the algorithm adds the expression along with its timestamp to a per-particle queue of anticipatory descriptions. As the robot proceeds through the environment and new nodes and semantic information are added to the map, the algorithm continues to evaluate the grounding likelihood (14) for the queued descriptions. Specifically, we consider candidate pairs of landmark and figure nodes and determine the landmark's likelihood according to its label distribution. We express the figure likelihood as the probability of the landmark-to-figure path under the learned language model (16), where we consider the shortest path that runs from the robot's pose at the time of the description, through the landmark region, and on to the figure node. The logic is that the description is most useful when the robot has visited the regions to which the user refers and, thereby, the map has regions whose labels and inter-region paths that are consistent with the expression. The algorithm performs this process separately for each particle, which may result in some particles incorporating the description sooner than others.

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 $\Delta\alpha = 0.5$.

4.4 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

$$w_t^{(i)} = \frac{\text{Target distribution}}{\text{Proposal distribution}} \tag{21a}$$

$$= \frac{p(G_t^{(i)}|z^t, u^t, \lambda^t)}{p(G_t^{(i)}|G_{t-1}^{(i)}, z^{t-1}, u^t, \lambda^t)} w_{t-1}^{(i)} \tag{21b}$$

$$= \frac{p(z_t|G_t^{(i)}, z^{t-1}, u^t, \lambda^t)}{p(z_t|z^{t-1})} \cdot p(G_{t-1}^{(i)}|z^{t-1}, u^t, \lambda^t) \tag{21c}$$

$$\propto p(z_t|G_t^{(i)}, z^{t-1}, u^t, \lambda^t) \cdot p(G_{t-1}^{(i)}|z^{t-1}, u^t, \lambda^t) \tag{21d}$$

$$\tilde{w}_t^{(i)} = p(z_t|G_t^{(i)}, z^{t-1}, u^t, \lambda^t) \cdot w_{t-1}^{(i)}, \tag{21e}$$

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

$$p(z_t|G_t^{(i)}, z^{t-1}, u^t, \lambda^t) = \int_{X_t} p(z_t|X_t^{(i)}, G_t^{(i)}, z^{t-1}, u^t, \lambda^t) \times p(X_t^{(i)}|G_t^{(i)}, z^{t-1}, u^t, \lambda^t) dX_t, \tag{22}$$

which allows us to utilize the conditional measurement model. In the experiments presented next, we model the measurement as an observed transformation between poses, which we compute via scan-matching. We model this distribution (first term in the integral) as Gaussian, which we have empirically found to be accurate.

After calculating and normalizing the new importance weights, we periodically perform resampling based upon the effective number of particles, as proposed by Liu [26],

$$N_{eff} = \frac{1}{\sum_{i=0}^{n} w_i^2}. \tag{23}$$

When the effective number of particles N_{eff} falls below the threshold $N/2$, where N is the number of particles, we resample using the algorithm described by Doucet et al. [7].

5 Results

We evaluate our algorithm through six experiments that involve a human giving a robotic wheelchair (Fig. 1) [15] a narrated tour of several buildings and courtyards on the MIT campus. The robot was equipped with forward- and rearward-facing LIDARs, wheel encoders, and an IMU. Speech was recorded using a wireless microphone worn by the user. 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.

5.1 Indoor/Outdoor: Small Tour

The first experiment (Fig. 9) took place on the first floor of the Stata Center at MIT, which includes lecture halls, elevator lobbies, a gym, and a cafeteria, as well as the adjacent courtyard. Starting at one of the elevator lobbies, the user proceeded to visit the gym, exited the building and, after navigating the courtyard, returned to the gym and finished at the elevator lobby. The user provided textual descriptions of the environment, twice each for the elevator lobby and gym regions. We compare the performance of our method based upon different forms of language input against a baseline algorithm that emulates the current state-of-the-art in language-augmented semantic mapping. In all cases, the algorithms were run with 10 particles to approximate the distribution over the space of topologies. The final topology contained 137 nodes.

5.1.1 No Language Constraints

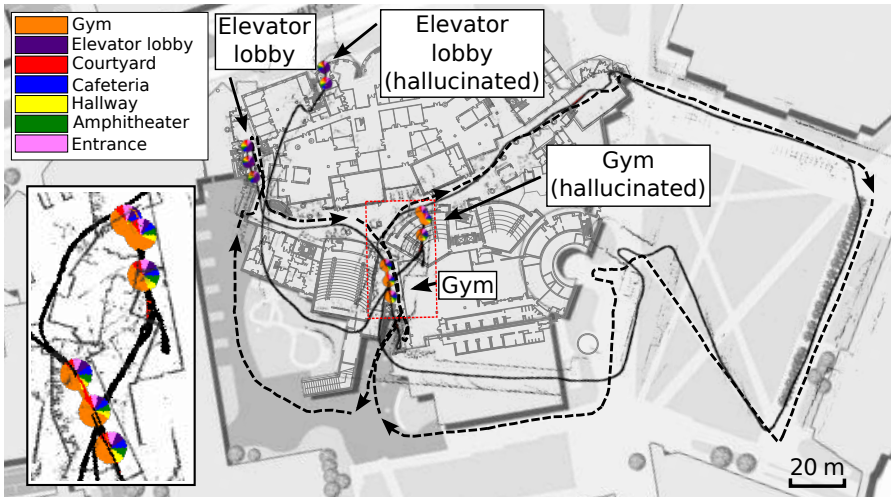
We consider a baseline approach that directly labels nodes based upon egocentric language, but does not propose edges based upon label distributions. It does, however, propose loop closures based upon the distribution over the metric map (Section 4.1.1). The baseline emulates typical solutions by augmenting a state-of-the-art iSAM metric map with a semantic layer without allowing semantic information to influence lower layers.

Figure 9(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).

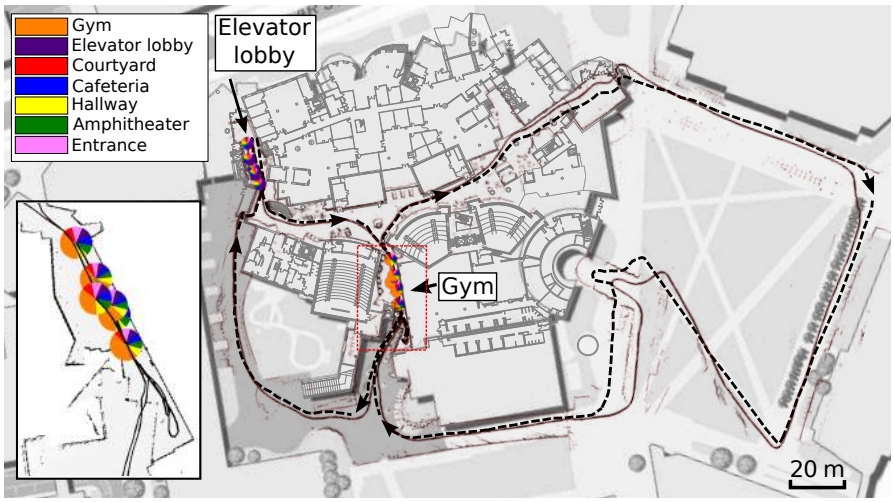
5.1.2 Egocentric Language

We evaluate our algorithm when the user provides descriptions in the form of egocentric language, in which case there is no ambiguity in the landmark and figure that are implicitly the robot's current location.

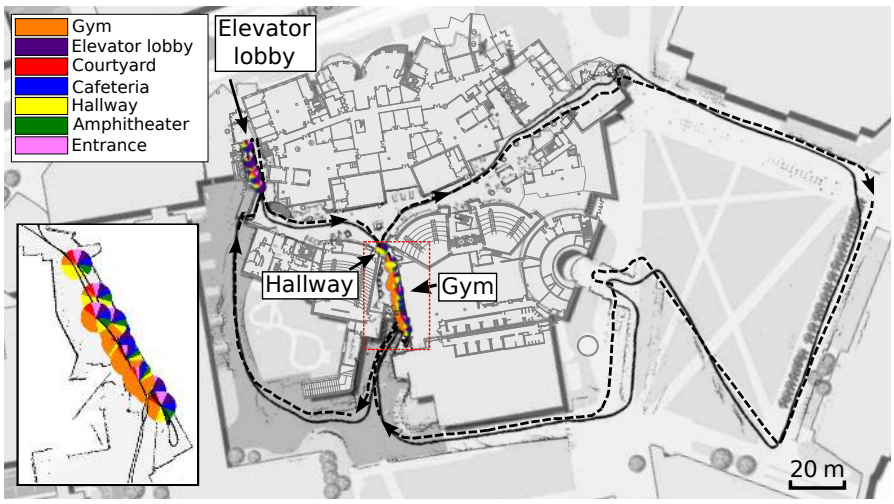
Figure 9(b) presents the semantic graph corresponding to the highest-weighted particle that our algorithm estimates. By considering the semantic map when proposing loop closures, the algorithm recognizes that the second region that the user labeled as the gym is the same place that was labeled earlier 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. 9(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



(a) No language constraints



(b) Egocentric language



(c) Allocentric language

Figure 9: Maximum likelihood semantic graphs for the small tour. In contrast to (a) the baseline algorithm, our method incorporates key loop closures based upon (b) egocentric and (c) allocentric 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.

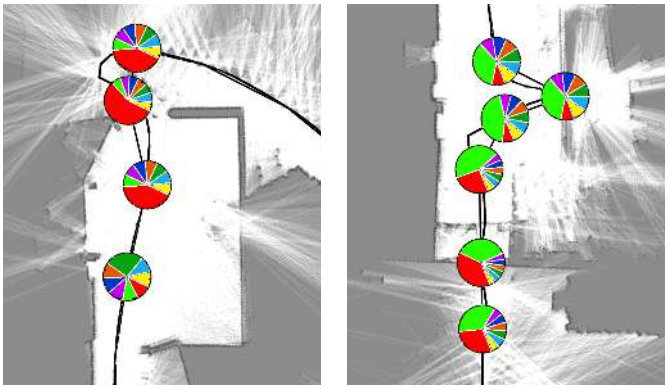
validity, the ground truth topology receives 92.7% of the probability mass and, furthermore, the top four particles are each consistent with the ground truth.

5.1.3 Allocentric Language

Next, we consider the algorithm's performance when the figure and landmark regions that the user's descriptions reference 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 egocentric language. The language inputs are otherwise identical to those employed for the egocentric language scenario and the baseline evaluation.

(a) Egocentric language (b) Allocentric language

Figure 11: Pie charts that compare the semantic map label distributions that result from (a) the egocentric language description "This is the gym" with that of (b) the allocentric language description "The gym is down the hall."



(a) Lobby (b) Cafeteria

Figure 12: Inset views of the (a) lobby and (b) cafeteria portions of the semantic graph for the large tour experiment (Fig. 10(c)).

The algorithm incorporates allocentric language into the semantic map using the G^3 framework as described in Section 4.3

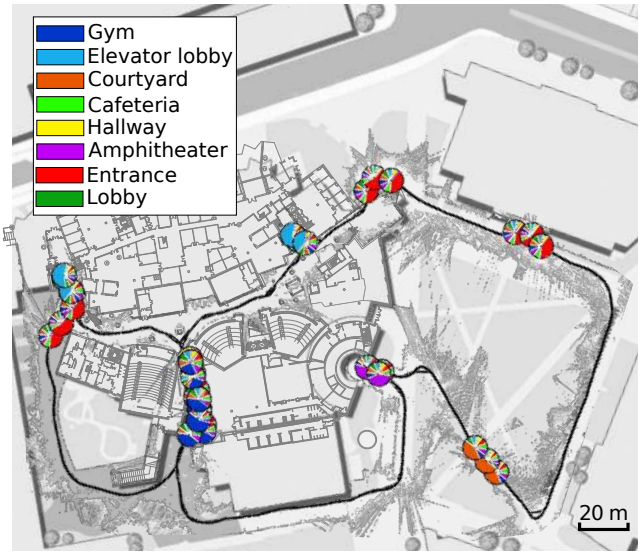


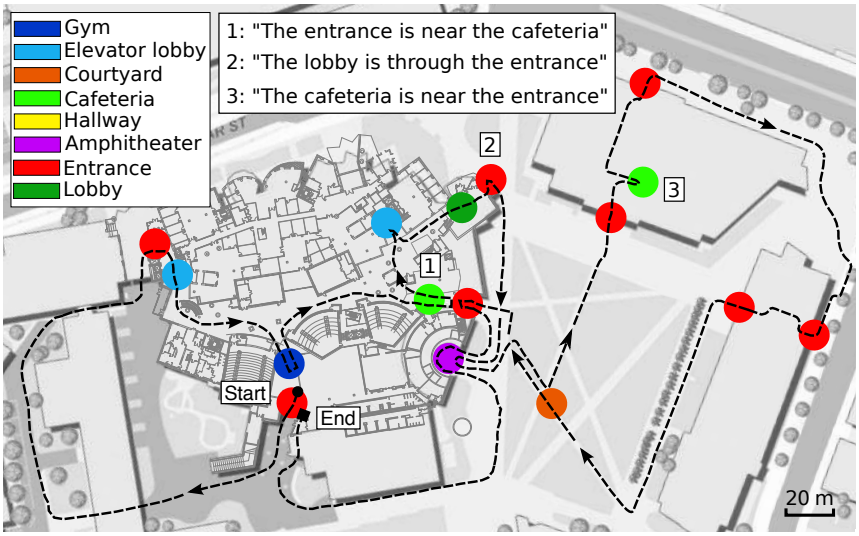
Figure 13: Maximum likelihood map for the autonomous tour.

to infer the nodes in the graph that constitute the figure (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." Figure 11 compares the label distributions that result from this grounding with those from egocentric language. The algorithm attributes the "gym" label to multiple nodes in the semantic graph as a result of the ambiguity in the figure's location as well as the G^3 model, which yields high likelihoods for several paths as being "down from" the landmark nodes. 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 egocentric language scenario, this results in a semantic graph for the environment that is accurate topologically, semantically, and metrically (Fig. 9(c)).

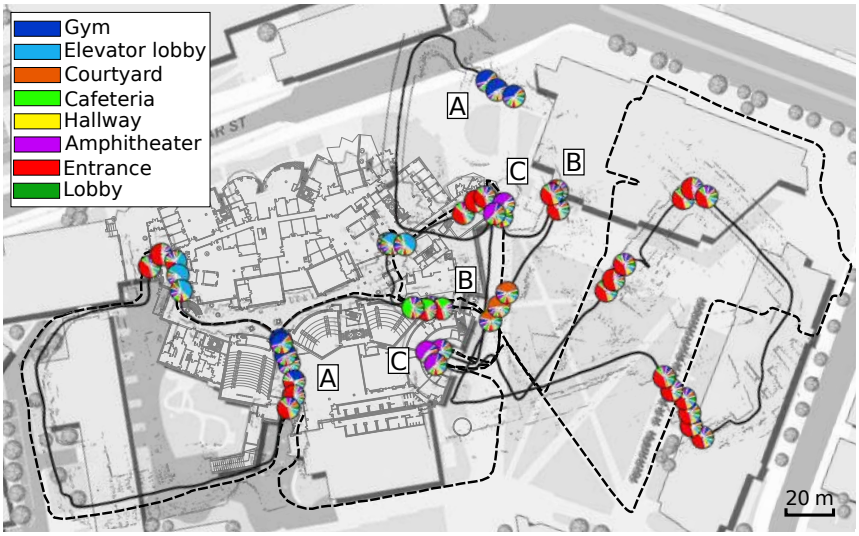
5.2 Indoor/Outdoor: Large Tour

The second experiment (Fig. 10) considers an extended tour of MIT's Stata Center as well as two neighboring buildings and their shared courtyard. In order to evaluate the algorithm's ability to deal with ambiguity in the labels, 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 took the form of both egocentric and allocentric language.

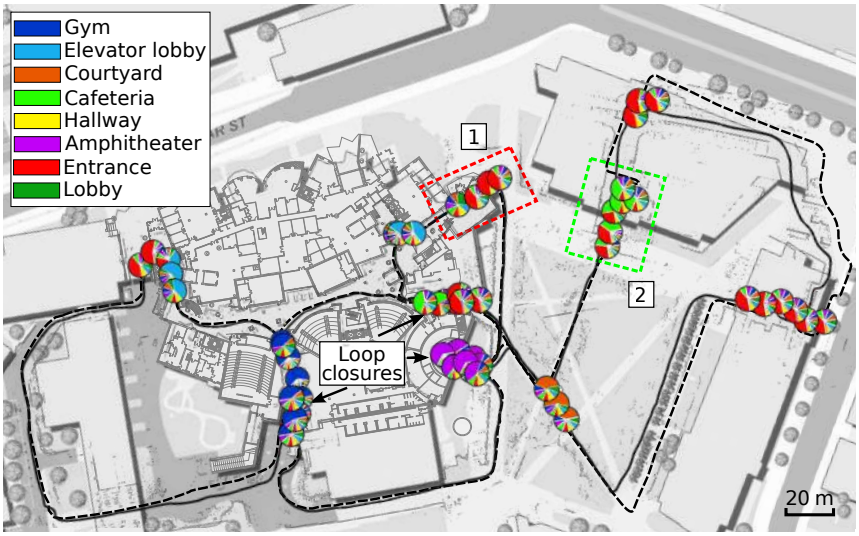
As with the smaller tour, we compare our method against the baseline semantic mapping algorithm. Figure 10(b) presents the baseline estimate for the environment's semantic graph. Without incorporating allocentric language or allowing semantic information to influence the topological and metric layers, 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, Figure 10(c) demonstrates that, by using semantic information to propose con-



(a) Ground Truth



(b) No language constraints



(c) Allocentric language

Figure 10: 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) The result produced by our method based upon allocentric language descriptions, with an indication of the loop closures recognized based upon the semantic map.

straints in the topology, our algorithm yields correct topological and semantic maps, and metric maps with notably less error. Figure 12 presents the inset views for the lobby and second cafeteria portion of the map that were labeled with allocentric descriptions. 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. In this experiment, the algorithm employed 10 particles to approximate the distribution over topologies. The final topology contained 213 nodes.

For utterances with allocentric language, our algorithm was able to generate reasonable groundings for the figure and landmark locations. However, due to the simplistic way in which we define regions, groundings for "the lobby" were not entirely accurate due to the sensitivity to the local metric structure of the environment when grounding paths that go "through the entrance." We discuss this in more detail in Section 6.1.

5.3 Indoor/Outdoor: 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 [16]. 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 utterances included both egocentric and allocentric descriptions of the environment. The speech was recorded as it was uttered in synchronization with the LIDAR and odometry data. The audio was later manually transcribed into text that was inserted alongside the sensor observations according to the time that the audio was initially recorded. In this manner, the algorithm handled the text, LIDAR, and odometry data as they were received, emulating a scenario in which a speech recognizer was used to parse the user's utterances during the tour.

The algorithm operated in this fashion using 10 particles to approximate the distribution over the space of topologies. The final topology contained 135 nodes. Figure 13 presents the maximum likelihood semantic graph that our algorithm estimates. By incorporating information that the 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.

5.4 Stata Center Lab Tour

We consider an additional experiment in which the robot was driven throughout different labs on the third floor of MIT's

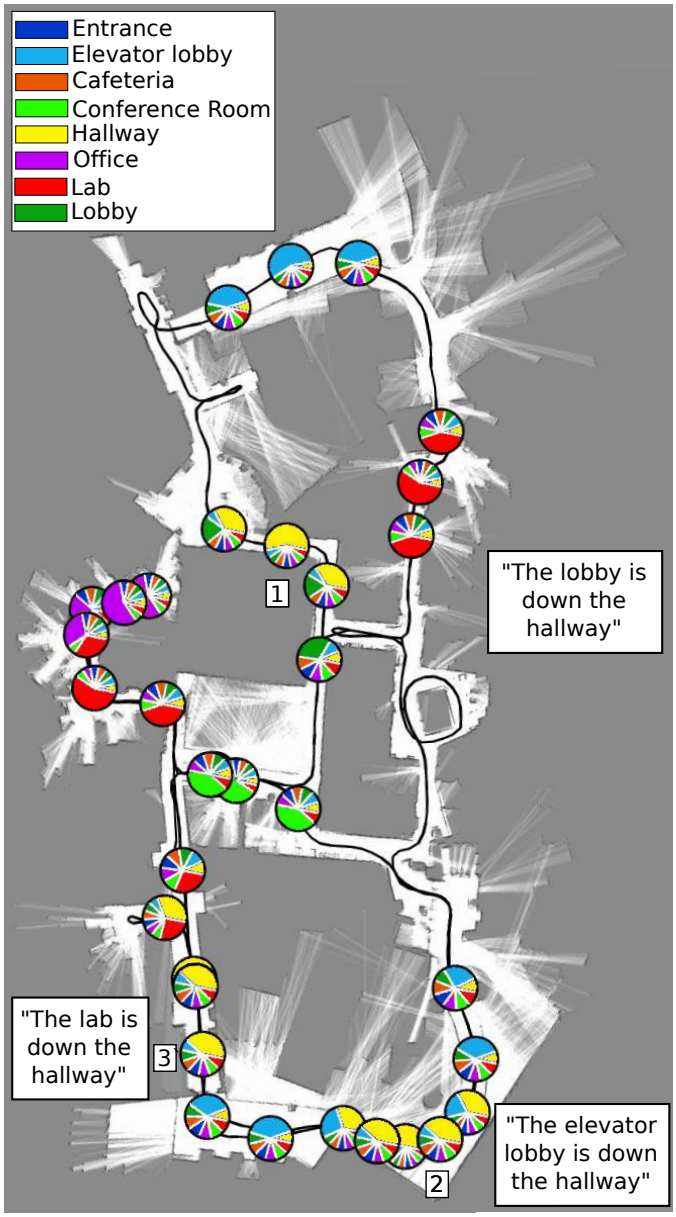


Figure 14: Maximum likelihood semantic graph inferred from the narrated tour of the Stata Center lab.

Stata Center. The narrated tour involved both egocentric and allocentric descriptions of the environment, the latter of which were anticipatory in nature with the user referencing locations in the environment that the robot had not yet visited. Figure 14 presents the maximum likelihood semantic map that our framework learned from the narrated tour using a total of 10 particles. The final topology contained 71 nodes. The system correctly grounds each of the allocentric descriptions despite the ambiguity that exists in the landmark and figure locations, as we discuss in more detail shortly.

5.5 MIT 32-36-38 Tour

In order to verify the validity of the algorithm in different environments, we consider an extended tour of three connected

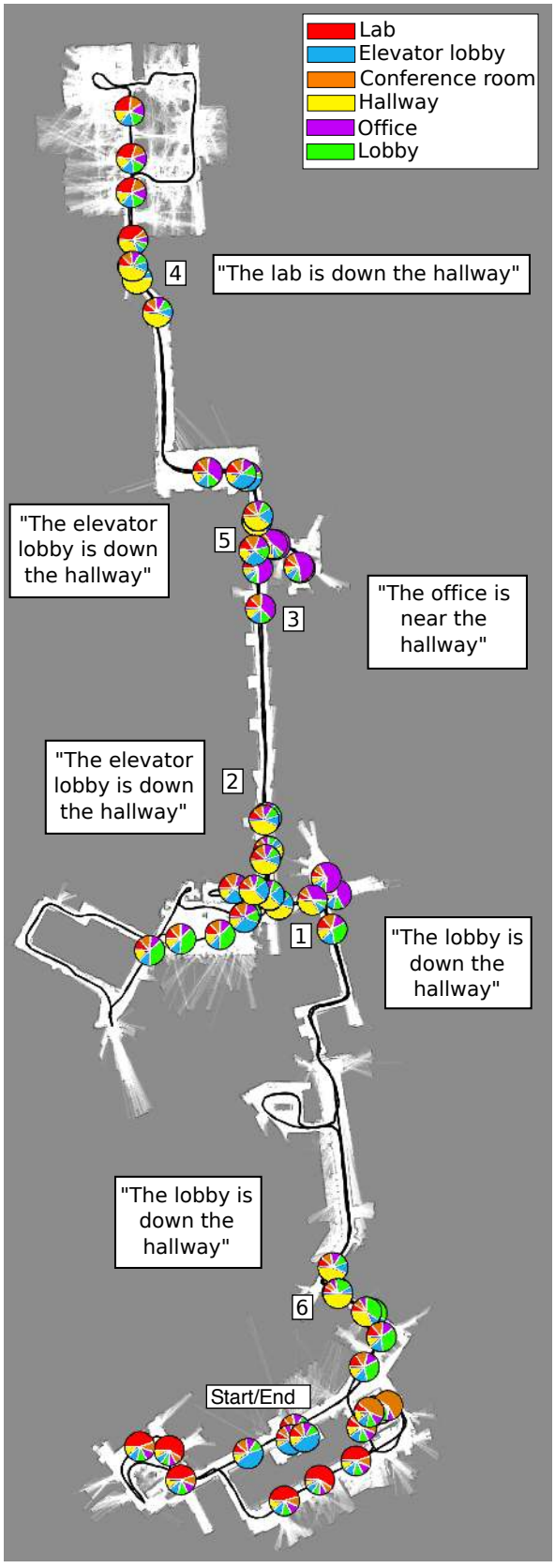


Figure 15: Maximum likelihood semantic graph for the MIT 32-36-38 tour. The allocentric descriptions are shown with numbers indicating their order.

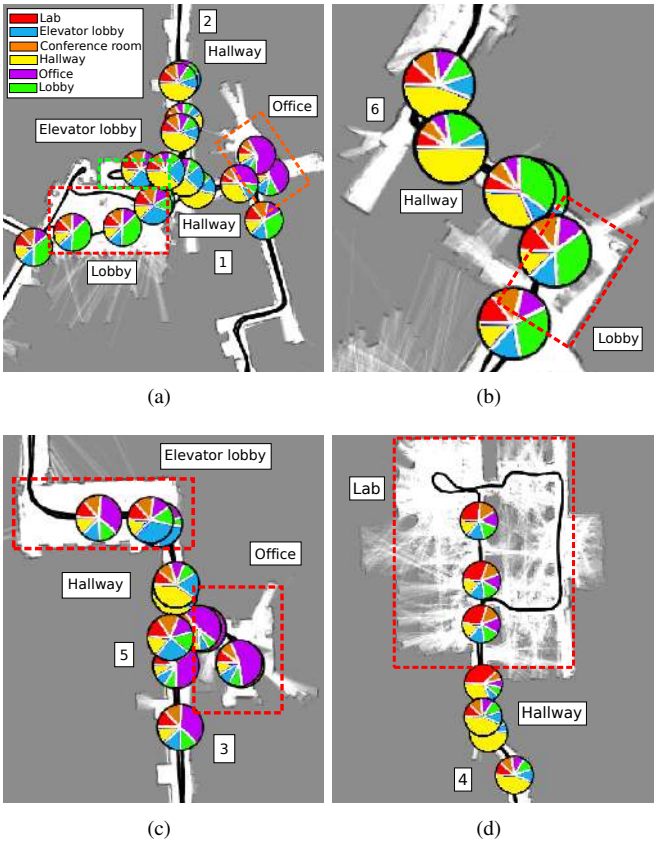


Figure 16: Inset views for the MIT 32-36-38 tour (Fig. 15) that demonstrate the way in which the algorithm learns from allocentric descriptions (a) "The lobby is down the hallway" (anticipatory, location 1) and "The elevator lobby is down the hallway" (location 2), (b) "The lobby is down the hallway" (anticipatory), (c) "The office is near the hallway" (anticipatory, location 3) and "The elevator lobby is down the hallway" (location 5), and (d) "The lab is down the hallway" (anticipatory). The dashed boxes denote the ground-truth boundaries for the regions.

buildings on the MIT campus (buildings 32, 36, and 38). The robotic wheelchair was manually driven throughout the office-like environment, visiting offices, elevator lobbies, conference rooms, and lab spaces whose appearance and structure varied between each building. Text was added at several points throughout the tour to emulate recognized natural language descriptions. We provided both egocentric and allocentric utterances, including several instances of anticipatory descriptions when the robot had not yet visited the referenced portions of the environment (both the figure and the referent). We ran our framework with 10 particles to model the distribution over topologies. The final topology contained 148 nodes. Figure 15 denotes the maximum likelihood semantic graph that resulted from our algorithm. The text indicates the allocentric descriptions that were given to the system in the numbered order.

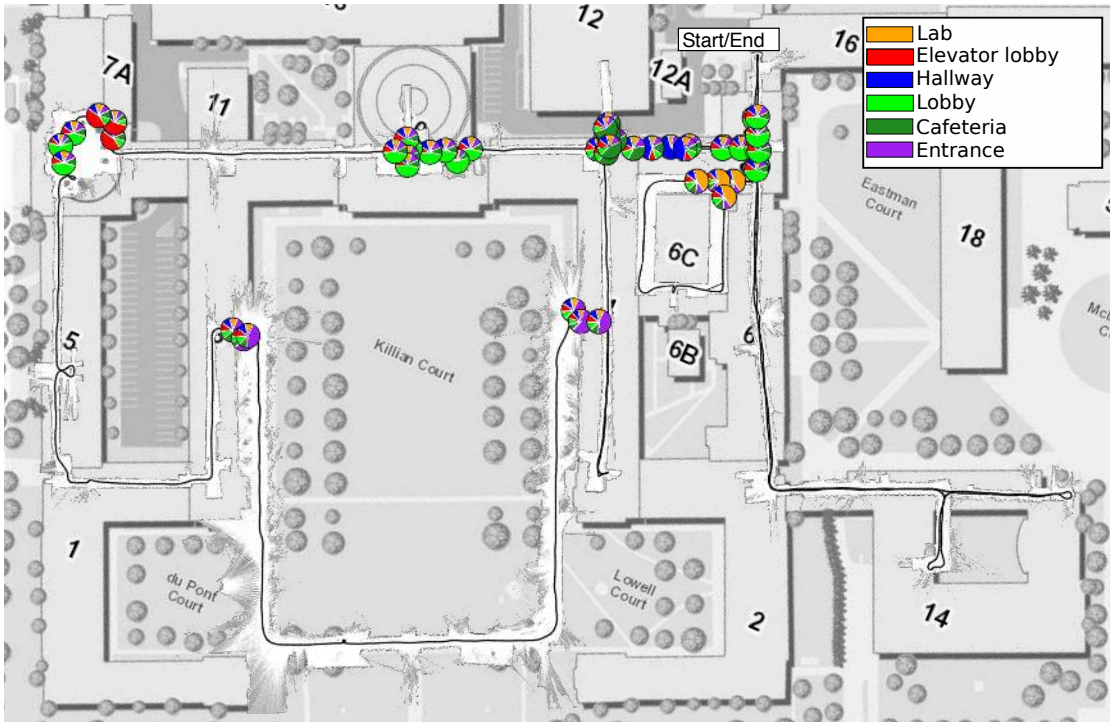


Figure 17: The maximum likelihood semantic map that results from a tour of MIT's Killian Court.

5.6 Killian Court Tour

The final experiment considers a tour of Killian Court, a set of interconnected buildings on MIT's campus, which has served as a benchmark environment for previous mapping algorithms. We consider this environment in an effort to see how the algorithm performs when tasked with mapping larger spaces that involve significant geometric and semantic aliasing. Specifically, this part of the MIT campus consists primarily of several long hallways with nearly identical structure, including the so-called "infinite corridor" that serves as one of the main hallways at MIT.

Starting in the north-east corner (Fig. 17), we gave the robot a tour along the infinite corridor that spans from left to right in the Figure. After entering one of the main lobbies (upper-left), we proceeded through buildings 5 and 3 and then exited into the courtyard. We took a U-shaped path outside, entered building 4, and then traveled through buildings 6, 6C, and 14 before returning to the start. We provided both egocentric and allocentric language descriptions at different points during the tour to assign labels to and spatial relations between different regions. These descriptions took the form of text that was interjected in synchronization with the LIDAR and odometry streams as the data was post-processed.

The algorithm learned a distribution over semantic maps from the stream of descriptions, odometry, and LIDAR data, using 10 particles to hypothesize the different topologies. The final topology contained 276 nodes. Figure 17 shows the resulting maximum likelihood semantic graph overlaid on an approximately aligned map of the MIT campus. Qualitatively, the map is metrically, topologically, and semantically accurate with the exception of the map of building 14 where a glass hallway between buildings 2 and 14 forced the algorithm to use odometry for the inter-pose constraints. As with the previous evaluations, we ran our framework without language-based constraints to emulate the current state-of-the-art in language-augmented semantic mapping. While we omit the figure for space, we note that the resulting map is significantly warped.

Table 2: Average Delay in Adding Node

Dataset	Average Delay (s)	Standard Deviation (s)
MIT 32-36-38	0.532	3.138
Killian Court Tour	0.682	2.726
Indoor/Outdoor Large Tour	2.186	4.670

5.7 Computational Requirements

We analyze the computational cost of the algorithm by considering the delay between when a node is first proposed (i.e., based on distance traveled) and the time at which it is added to the map. This measure reflects the overall time required of the algorithm, since it will not add nodes until it has finished incorporating the most recent description and proposed loop closures. We consider the delay for the three longest datasets, namely the indoor/outdoor large tour, the MIT 32-36-38 tour and the Killian Court tour. Table 2 summarizes the performance for each of these datasets. Note that the implementation has not been optimized to run in real-time, and each particle is currently processed sequentially (i.e., particle updates are not

Table 3: Semantic Map Accuracy

	Indoor/Outdoor Large Tour		Killian Court Tour		MIT 32-36-38 Tour	Stata Center Tour
Type	Baseline	SG	Baseline	SG	SG	SG
Cafeteria	20%	36%	23%	45%	-	-
Entrance	43%	46%	12%	47%	-	-
Elevator Lobby	46%	46%	49%	49%	34%	40%
Hallway	8%	8%	18%	19%	36%	30%
Lobby	8%	13%	34%	47%	29%	21%
Lab	-	-	0%	47%	42%	37%
Amphitheater	25%	53%	-	-	-	-
Courtyard	12%	47%	-	-	-	-
Office	-	-	-	-	53%	56%
Conference Room	-	-	-	-	51%	56%
Gym	33%	48%	-	-	-	-

parallelized). The variance in the delays is due to periods of increased computation that correspond to instances when language annotations are processed. This delay is dominated by two components of the algorithm. The first is the time required to ground allocentric descriptions using the G^3 framework for all particles. The second is the time taken to scan-match the semantic-based loop closures that are subsequently proposed between nodes with updated label distributions. Allocentric language grounding requires computational effort that is linear in the number of unique particles. Similarly, the scan-match verification is linear in the number of nodes that are updated with new label information, which is independent of the size of the map. The computational requirements for verification are dominated by a scan-match procedure that is exhaustive in its search due to the potentially large error in the prior pose-to-pose transform.

5.8 Semantic Accuracy

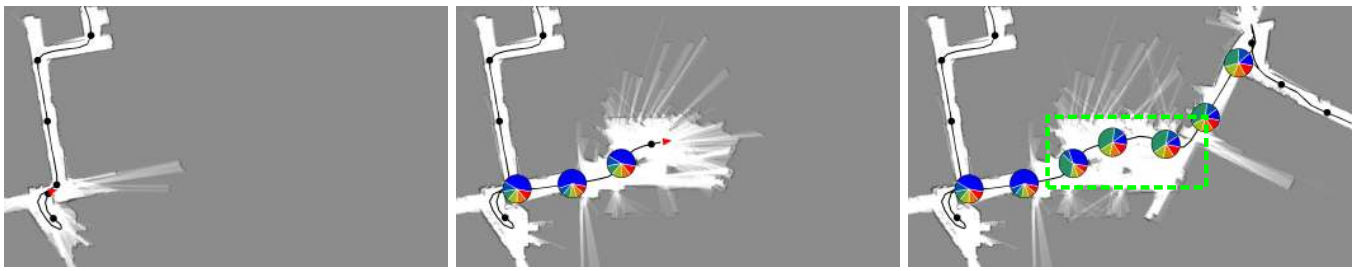
Table 3 outlines the accuracy of the resulting semantic maps for four datasets, where we calculate the accuracy as follows. First, we identify the regions for which language contributed to their label distributions. We compute the ground truth label for each of these regions and compute the cosine similarity between the ground truth multinomial (assumed to have a likelihood of 1.0 for the true label) and that of the label distribution.

For the indoor/outdoor large tour and the Killian Court tour, we also compared the results for the maps that did not propose language edges. Since large segments of these maps were metrically and topologically inaccurate, we assigned a minimum score for regions that were significantly inaccurate. In effect, this corresponds to assigning these regions a uniform multinomial over labels. As can be seen for the first two datasets, the use of our approach improves the semantic accuracy of a number of regions. This improvement stems both from the metric and topological accuracy of the learned maps as well as the algorithm's ability to integrate allocentric language. In the MIT 32-36-38 and the Stata Center tours, we also achieve reasonable accuracy for most categories. We do note that in case of allocentric language, some expressions can be ambiguous, either due to the presence of multiple potential landmarks or due to the ambiguity in the expression. For example, given the description "The lobby is down the hallway," there may be multiple regions whose location is consistent with being "down" the hallway, of which only one is the lobby. In these situations, each of these regions will receive high likelihood of being the figure and the label distributions for each will be updated accordingly. Additionally, we find that the accuracy of the semantic maps is sensitive to our choice for region decomposition. For example, hallways score fairly low under our fixed-size segmentation, which can significantly underestimate their spatial extent. We see these issues as inherent to our definition of regions that would be diminished with a more sophisticated segmentation strategy that takes into account local appearance [3, 36, 37, 14] and semantic [29] properties of the environment.

6 Discussion

6.1 Learning from Allocentric, Anticipatory Language

A contribution of our work is the use of natural language descriptions to produce consistent semantic maps from spatial relations and labels inferred from language. The advantage of this capability is that it allows robots to more efficiently acquire human-centric maps of their environment. The challenge to learning from these expressions is that their groundings are ambiguous—the user may refer to regions that may be distant from the robot and outside the field-of-view of its sensors. Additionally, it may be that the descriptions are anticipatory, when the robot has yet to visit the figure that the user is describing or the landmark that they are referencing. Figure 18 depicts



(a) Anticipatory Description Given (b) Landmark is Added (c) Figure is Grounded

Figure 18: A depiction of the process of learning from an anticipatory description. (a) The user describes the "lobby" as being "down the hallway," yet the hallway has not been labeled and there is no node for the elevator lobby in the topology. (b) The user labels the current region as the "hallway," providing the landmark location. (c) Once nodes are added that are consistent with the description, the algorithm updates the labels. The green box indicates the actual location of the lobby.

the process of learning from an anticipatory description as part of the Stata Center lab tour (Fig. 14). Figure 18(a) shows the robot traversing a hallway when the user states that "The elevator lobby is down the hallway." At this point, the semantic graph includes several nodes with a high likelihood of having the label "hallway." However, the robot has yet to visit the specific hallway that the person is using as the landmark and, as a result, the semantic graph does not include nodes for this region. The graph also lacks nodes for the region that the user refers to as the "elevator lobby." The algorithm attempts to ground the description using the language model as described in Section 4.3, which yields a likelihood for each pair of nodes as being the landmark and the figure.

This algorithm performs this grounding process for each particle, and updates those for which the likelihood of the top pair is sufficiently high (0.2). In this example, the likelihood of the candidate groundings for most of the particles is low and the algorithm postpones language integration. As the tour proceeds (Fig. 18(b)), the guide labels the robot's position as being the "hallway," which updates the label distribution for the adjacent node. The algorithm again attempts to ground the language, this time using the newly added hallway nodes as the landmark. However, paths that start at the pose from which the description was first given and pass through the landmark to other nodes do not resemble the learned model for the "down" relation. After the robot and user continue and more nodes are added to the topology (Fig. 18(c)), the framework again attempts to ground the description, this time returning highly-confident estimates for the locations of the landmark and the figure, per the induced path. However, not all of the inferred locations are correct, which is consistent with what we see with other allocentric expressions. In this case, the system assigns "elevator lobby" labels to nodes that preceded the hallway as well as several nodes beyond the true location of the lobby (green box). We attribute this to the difficulty in dealing with frame-of-reference when grounding language as well as to using features for the "down" relation that attempt to accommodate a wide range of scales (i.e. the length of hallways differs significantly across the environments that we consider).

In an effort to better understand the accuracy with which the algorithm learns from environment descriptions, we consider regions whose semantic properties were inferred from allocentric utterances. Figure 16 presents close-up views of the regions that were labeled as part of the multi-building tour (Fig. 15). The portion of the semantic graph shown in Figure 16(a) results from two descriptions, "The lobby is down the hallway" and "The elevator lobby is down the hallway," which were uttered at the locations indicated by the numbers "1" and "2," respectively. The former utterance was anticipatory as the robot had not yet visited the lobby area when the description was given. Nonetheless, the framework successfully labels that region of the environment when the robot later visits it, without any aliasing effects. However, grounding the second utterance results in high likelihoods associated with some nodes that are not actually in the elevator lobby, causing the label to "bleed" into other areas. We attribute this to the ambiguity that results from not reasoning over frame-of-reference without which the nodes are consistent with being "down" the hallway. The performance improves for the anticipatory utterance in Figure 16(d) where the algorithm waits to infer the location of the lab until it is visited. We see similar effects for the descriptions in Figure 16(c) where the system correctly infers the location of another elevator lobby but attributes the "office" label to nodes that are actually in a hallway. This results from a simple set of features that encode the "near" relation based upon distance. Additionally, our algorithm uses a fixed separation to define regions and does not reason over their geometry (e.g., the shape of hallways is typically distinct from that of offices.) Meanwhile, Figure 12(a) depicts the semantic information inferred for the utterance "The lobby is through the entrance" from the large indoor/outdoor tour where we see that the algorithm correctly grounds the location of the lobby without any aliasing.

6.2 Navigation

A consequence of maintaining a joint distribution over each layer of the semantic graph is that the framework is able to use knowledge of the semantic properties of the environment to update the topology and metric map. This improves the accuracy of the resulting semantic graph and, in turn, facilitates navigation. To better understand the effects on navigation efficiency, we consider the task of finding the optimal path between two

Table 4: Average Length of the Optimal Path

Experiment	Baseline	SG
Small Indoor/Outdoor	41.59 m	23.50 m
Large Indoor/Outdoor	68.14 m	35.52 m
Autonomous	43.49 m	25.70 m
Killian Court	63.08 m	40.76 m

nodes in the topology, as if the robot were asked to use the semantic graph to navigate from its current location to a named region in the environment.

We examine the semantic graphs that we learned with and without language-based constraints for the two indoor/outdoor scenarios, the autonomous tour, and the Killian Court dataset. For each, we randomly picked 1000 pairs of start and goal nodes in the graph and used a graph search algorithm to find the shortest path through the topology, with equal cost for each edge in the graph. The same node pairs were used for each of the semantic graphs for a given environment. Table 4 compares the average optimal path length through the graphs that result from our method and the baseline, which does not infer constraints from the descriptions. The graphs that we estimate when language influences only the semantic layer give rise to optimal paths that are noticeably longer than the paths reflected in the graphs that we learn by jointly estimating the semantic graph. This difference stems from the fact that our representation provides semantic-based edges that allow the planner to identify shortcuts in the topology that are otherwise not suggested by the baseline map, which mimics the current state-of-the-art in language-augmented semantic mapping.

6.3 Possibility of aliasing with language

When proposing edges to the topology based upon the label distributions, we perform exhaustive scan-matching to check the validity of each proposed loop closure. While this helps to filter out the large majority of erroneous edges, the matching may yield false positives in regions that are perceptually aliased (Fig. 19). However, since the hypothesis space of potential language edges is large, the likelihood that all particles sample invalid edges is low, confining such occurrences to a small subset of particles. Empirically, we have found that the weight of these particles is quickly reduced as their metric maps are inconsistent with subsequent sensor measurements. These particles then tend to be removed during resampling.

7 Conclusion

We have described an algorithm that estimates metrically accurate semantic maps from a user's natural language descriptions. The novelty lies in learning the joint distribution over the metric, topological, and semantic properties of the environment, which enables the method to fuse the robot's sensor stream with knowledge inferred from the descriptions. We have presented results from several experimental evaluations that demonstrate the algorithm's ability to infer accurate metric, topological, and semantic maps. However, there are several limitations to our current approach.

A known issue with sample-based methods such as ours is the problem of particle depletion [7] whereby a majority of samples evolve to support regions of the distribution with negligible likelihood. This results in a poor approximation to the target distribution and can cause the filter to diverge. Resampling the particles based upon a measure of the variance in their weights, as we do, reduces the likelihood of particle depletion. In practice, we have not found depletion to occur, as suggested by the results. We partially attribute this to using the distribution over the semantic map as part of the proposal, which reduces the frequency of erroneous samples. Nonetheless, particle depletion may occur and can be mitigated by adding additional particles to hypothesize new topologies in the event that the distribution appears to misrepresent the target distribution, for example, as suggested by the particle weights [10].

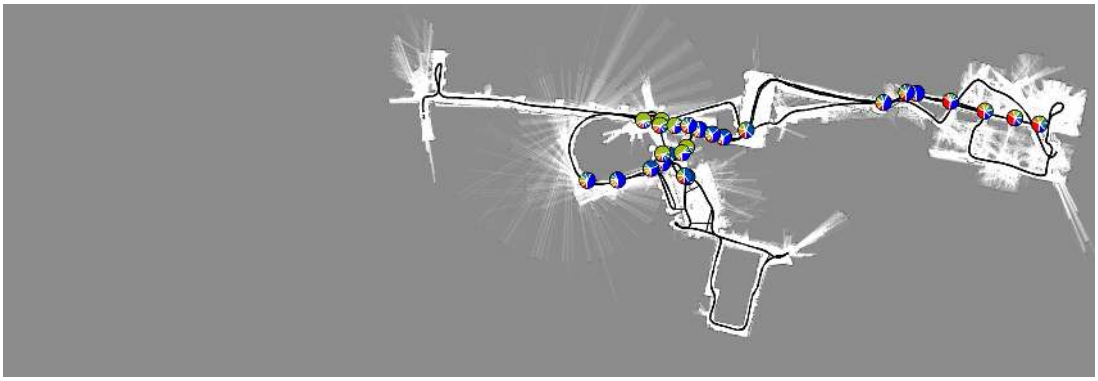
Our method is capable of inferring semantic information only from the user's descriptions. This means that the algorithm can only model a region's label if it was specifically referenced by the user. Further, it precludes the method from incorporating allocentric descriptions for which the user never labels the landmark. For example, the algorithm can not learn from the description "The gym is down the hall" unless the user identifies the location of the hallway. Our recent work [14] alleviates this requirement by also using geometric- and appearance-based scene classifiers to infer semantic information from LIDAR and vision.

An additional limitation of the algorithm is that it treats a region's colloquial name (e.g., "Carrie's office") and its type ("office") jointly as being labels. Thus, any subsequent reference to a labeled region is required to use the label that was originally given. We have recently extended our representation to model the type-name hierarchy for each region, where we infer the type from the aforementioned scene classifiers [14].

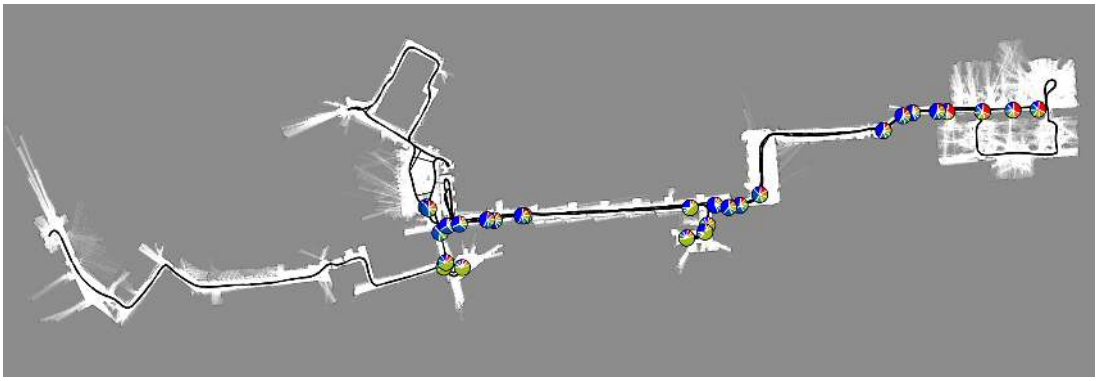
The algorithm partions the environment by instantiating regions at a fixed distance apart as the robot travels. This results in regions that are not semantically meaningful, with multiple regions being used to model the same area. Consequently, the algorithm may ground language to the wrong node in the topology, either by inferring an incorrect landmark or by diffusing labels across multiple nodes. Our latest work [14] employs spectral clustering to segment the environment into regions based upon the consistency of their local LIDAR scans, yielding regions that are more meaningful.

The information that we are currently able to infer from a user's descriptions is limited to a region's colloquial name and its relation to another region in the environment. It does not support a user's ability to convey general properties of the environment, such as "You can find computers in offices," or "nurses' stations tend to be located near elevator lobbies."

Our current framework was designed to learn a semantic map of an environment from an initial tour, with the idea that this map can then be used for localization, navigation, and grounding natural language commands during long term operation. As with most pose graph approaches to SLAM, our algorithm may



(a) Incorrect loop closure added



(b) No incorrect loop closures

Figure 19: A demonstration of the effects of perceptual aliasing for the three building tour (Fig. 15) in which (a) the algorithm accepts an invalid edge between different regions that have similar geometry for one particle. However, the majority of the particles did not propose erroneous edges and the weight of this map soon decreases to $1/10^{th}$ of that of the correct particle and is removed upon resampling.

add regions that duplicate the same part of the environment, building maps that grow with time rather than space. This is particularly undesirable when the robot operates for extended periods of time within the same environment. We have recently updated our representation so as to reuse existing nodes in the graph, updating their metric, topological, and semantic properties, rather than adding new redundant nodes [14].

In summary, we described 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.

8 Acknowledgments

We thank Faruh Paerhati for his help with the experimental evaluation and Nick Roy for his feedback. This work was supported in part by Quanta Computer and 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] P. Beeson, J. Modayil, and B. Kuipers, "Factoring the mapping problem: Moble robot map-building in the Hybrid Spatial Semantic Hierarchy," *International Journal of Robotics Research*, vol. 29, no. 4, pp. 428–459, April 2010.

[2] M. Bosse, P. Newman, J. Leonard, and S. Teller, "Simultaneous localization and map building in large-scale cyclic environments using the atlas framework," *International Journal of Robotics Research*, vol. 23, no. 12, pp. 1113–1139, 2004.

[3] E. Brunskill, T. Kollar, and N. Roy, "Topological mapping using spectral clustering and classification," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2007, pp. 3491–3496.

[4] G. Bugmann, E. Klein, S. Lauria, and T. Kyriacou, "Corpus-based robotics: A route instruction example," *Proceedings of 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 *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, San Francisco, CA, August 2011, pp. 859–865.

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

[7] A. Doucet, N. de Freitas, K. Murphy, and S. Russell, "Rao-Blackwellised particle filtering for dynamic Bayesian networks," in *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, Stanford, CA, July 2000, pp. 176–183.

[8] 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 *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2009, pp. 4163–4168.

[9] R. Eustice, H. Singh, and J. Leonard, "Exactly sparse delayed-state filters," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Barcelona, Spain, April 2005, pp. 2417–2424.

[10] P. Fearnhead and P. Clifford, "Online inference for hidden markov models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 65, no. 4, pp. 887–889, November 2003.

[11] C. Galindo, A. Saffiotti, S. Coradeschi, P. Buschka, J. Fernandez-Madrigal, and J. Gonzalez, "Multi-hierarchical semantic maps for mobile robotics," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2005, pp. 2278–2283.

[12] J.-S. Gutmann and K. Konolige, "Incremental mapping of large cyclic environments," in *Proceedings of the IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA)*, Monterey, CA, November 1999, pp. 318–325.

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

[14] S. Hemachandra, M. R. Walter, S. Tellex, and S. Teller, "Learning spatial-semantic representations from natural language descriptions and scene classifications," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, May 2014.

[15] S. Hemachandra, T. Kollar, N. Roy, and S. Teller, "Following and interpreting narrated guided tours," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2011, pp. 2574–2579.

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

[17] R. Jackendoff, *Semantics and Cognition*. The MIT Press, September 1985.

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

[19] T. Kollar and N. Roy, "Utilizing object-object and object-scene context when planning to find things," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2009, pp. 4116–4121.

[20] T. Kollar, S. Tellex, D. Roy, and N. Roy, "Toward understanding natural language directions," in *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Osaka, Japan, March 2010, pp. 259–266.

[21] K. Konolige, "Large-scale map-making," in *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, San Jose, CA, July 2004, pp. 457–463.

[22] B. Krieg-Brückher, U. Frese, K. Lüttich, C. Mandel, T. Massakowski, and R. J. Ross, "Specification of an ontology for route graphs," *Spatial Cognition IV: Reasoning, Action, Interaction*, vol. 3343, pp. 390–412, 2005.

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

[24] B. Kuipers, J. Modayil, P. Beeson, and M. MacMahon, "Local metrical and global topological maps in the Hybrid Spatial Semantic Hierarchy," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, April 2004, pp. 4845–4851.

[25] J. Leonard and P. Newman, "Consistent, convergent, and constant-time SLAM," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, Acapulco, Mexico, August 2003, pp. 1143–1150.

[26] J. Liu, "Metropolized independent sampling with comparisons to rejection sampling and importance sampling," *Statistics and Computing*, vol. 6, pp. 113–119, 1996.

[27] K. Lynch, *The Image of the City*. MIT Press, 1960.

[28] M. MacMahon, B. Stankiewicz, and B. Kuipers, "Walk the talk: Connecting language, knowledge, and action in route instructions," in *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, Boston, MA, July 2006, pp. 1475–1482.

[29] O. Martínez Mozos, R. Triebel, P. Jensfelt, A. Rottmann, and W. Burgard, "Supervised semantic labeling of places using information extracted from sensor data," *Robotics and Autonomous Systems*, vol. 55, no. 5, pp. 391–402, May 2007.

[30] C. Matuszek, E. Herbst, L. Zettlemoyer, and D. Fox, "Learning to parse natural language commands to a robot control system," in *Proceedings of the International Symposium on Experimental Robotics (ISER)*, Québec City, June 2012, pp. 403–415.

[31] C. Matuszek, D. Fox, and K. Koscher, "Following directions using statistical machine translation," in *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Osaka, Japan, March 2010, pp. 251–258.

[32] D. Meger, P.-E. Forssén, K. Lai, S. Helmer, S. McCann, T. Southey, M. Baumann, J. J. Little, and D. G. Lowe, "Curious George: An attentive semantic robot," *Robotics and Autonomous Systems*, vol. 56, no. 6, pp. 503–511, June 2008.

[33] J. Modayil, P. Beeson, and B. Kuipers, "Using the topological skeleton for scalable global metrical map-building," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, September 2004, pp. 1530–1536.

[34] E. Olson, J. Leonard, and S. Teller, "Spatially-adaptive learning rates for online incremental SLAM," in *Proceedings of Robotics: Science and Systems (RSS)*, Atlanta, GA, June 2007.

[35] A. Pronobis, O. Martínez Mozos, B. Caputo, and P. Jensfelt, "Multi-modal semantic place classification," *International Journal of Robotics Research*, vol. 29, no. 2–3, pp. 298–320, 2010.

[36] A. Pronobis and P. Jensfelt, "Large-scale semantic mapping and reasoning with heterogeneous modalities," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2012, pp. 3515–3522.

[37] A. Ranganathan and F. Dellaert, "Bayesian suprise and landmark detection," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, May 2009, pp. 2017–2023.

[38] ——, "Online probabilistic topological mapping," *International Journal of Robotics Research*, vol. 30, no. 6, pp. 755–771, 2011.

[39] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 2nd ed. Upper Saddle River, NJ: Prentice Hall, 2003, pp. 97–104.

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

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

[42] R. Smith and P. Cheeseman, "On the representation and estimation of spatial uncertainty," *International Journal of Robotics Research*, vol. 5, no. 4, pp. 56–68, 1986.

[43] 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 *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, San Francisco, CA, August 2011, pp. 1507–1514.

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

[45] S. Thrun, J.-S. Gutmann, D. Fox, W. Burgard, and B. J. Kuipers, "Integrating topological and metric maps for mobile robot navigation: A statistical approach," in *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, Madison, WI, July 1998, pp. 989–995.

[46] S. Thrun, Y. Liu, D. Koller, A. Ng, Z. Ghahramani, and H. Durrant-Whyte, "Simultaneous localization and mapping with sparse extended information filters," *International Journal of Robotics Research*, vol. 23, no. 7–8, pp. 693–716, 2004.

[47] A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin, "Context-based vision system for place and object recognition," in *Proceedings of the International Conference on Computer Vision (ICCV)*, Nice, France, October 2003, pp. 273–280.

[48] S. Vasudevan and R. Siegwart, "Bayesian space conceptualization and place classification for semantic maps in mobile robotics," *Robotics and Autonomous Systems*, vol. 56, no. 6, pp. 522–537, June 2008.

[49] M. R. Walter, R. M. Eustice, and L. J. J., "Exactly sparse extended information filters for feature-based SLAM," *International Journal of Robotics Research*, vol. 26, no. 4, pp. 335–359, 2007.

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

[51] 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.