Information-Theoretic Dialog to Improve Spatial-Semantic Representations

Sachithra Hemachandra Matthew R. Walter

***Abstract*— We propose an algorithm that enables robots to improve their spatial-semantic representation of an environment by engaging users in dialog during a guided tour. The algorithm selects the best information gathering actions in the form of targeted questions that reduce the ambiguity over the grounding of user-provided natural language descriptions (e.g., "The kitchen is down the hallway"). These questions include those that query the robot's local surround (e.g., "Are we in front of the kitchen?") as well as areas distant from the robot (e.g., "Is the lounge near the conference room?"). Our algorithm treats dialog as an optimization problem that seeks to balance the information-theoretic value of candidate questions with a measure of cost associated with dialog. In this manner, the algorithm determines the best questions to ask based upon the expected entropy reduction, while accounting for the burden on the user. We evaluate entropy reduction for a joint distribution over a hybrid metric, topological, and semantic representation of the environment learned from user-provided descriptions and the robot's sensor data during the guided tour. We demonstrate that, by asking deliberate questions of the user, the method significantly improves the accuracy of the learned map.**

I. Introduction

Robots are increasingly being deployed in human-occupied environments. In order to be effective partners, robots need to reason over representations of these environments that model the spatial, topological, and semantic properties (e.g., room types and names) that people associate with their environment. An efficient means of learning these representations is through a guided tour in which a human provides natural language descriptions of the environment [1, 2, 3, 4, 5]. With these approaches, the robot takes a passive role, whereby it infers information from the descriptions that it fuses with its onboard sensor stream.

The challenge to learning is largely one of resolving the high-level knowledge that language conveys with the low-level observations from the robot's sensors. Human descriptions tend to be ambiguous, with several possible interpretations (*groundings*) for a particular environment. For example, the user may describe the location of the kitchen as being "down the hallway," yet there may be several hallways nearby, each leading to a number of different rooms. Furthermore, grounding language typically requires a complete map, however, the robot may not yet have visited the regions that the user is referring to. The user may be describing a location known to the robot or a new location outside the field-of-view of its sensors.

S. Hemachandra is with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA USA sachih@csail.mit.edu

M.R. Walter is with the Toyota Technological Institute at Chicago, Chicago, IL USA mwalter@ttic.edu

Fig. 1. A user gives a tour to a robotic wheelchair designed to assist residents in a long-term care facility. (Left) The guide provides an ambiguous description of the kitchen's location. (Right) When the robot is near one of the likely locations, it asks the guide a question to resolve the ambiguity.

In this paper, we propose an active approach whereby the robot asks targeted questions of the user as a means of gathering information (Fig. 1). Engaging the user in dialog requires choosing which question to ask and when to ask it, while balancing the benefits of asking questions with the cost that comes from interrupting the tour and burdening the guide. There are three primary challenges to using dialog as an information gathering action. First, the robot needs to ask questions that provide enough context to the guide to be understood. Second, the questions should be structured such that the answers are as informative as possible. Third, the decision of if and when to ask questions should account for the social cost incurred by engaging the user in dialog.

We address these challenges by modeling human-robot dialog during the tour as a decision process. During the tour, the robot maintains a distribution over a *semantic graph* [3, 4], a metric, topological, and semantic representation of the environment, using a Rao-Blackwellized particle filter. Taking an information-theoretic approach, the algorithm decides whether to follow the guide or ask questions at each timestep in order to update the distribution. The algorithm reasons over the natural language descriptions and the current learned map to identify potential questions that best reduce ambiguity in the map. The algorithm considers egocentric (situated) and allocentric (non-situated) binary (yes/no) questions that express spatial relations between pairs of regions. These regions may be local to the robot in the case of egocentric dialog (e.g., "Is the lab on my right?")

or distant in the case of allocentric dialog (e.g., "Is the lounge next to the conference room?"). We associate a cost with each question that reflects the burden on the user and a reward based on the answer's expected information gain. We formulate the decision process as a QMDP [6], where we evaluate actions as a Markov Decision Process (MDP) for each possible configuration of the world (particle), and select the best action using the QMDP heuristic. This allows us to balance the value of the information gained by asking questions with their associated cost and only ask useful questions. We outline experiments that demonstrate the ability of our approach to reduce the ambiguity in natural language descriptions and, in turn, learn more accurate semantic maps of the environment than the current state-of-the-art.

II. Related Work

Several approaches exist that construct semantic environment models using traditional robot sensors [7, 1, 2, 8], while others have looked at additionally integrating natural language descriptions to improve the semantic representations [3, 9, 4]. Duvallet et al. [10] build upon this approach in order to follow instructions without any prior knowledge of the environment by exploiting environment knowledge implicit in the command and treating this knowledge as observations in a language-based mapping algorithm. With most of these techniques, however, the robot only passively receives observations, whether they are from traditional sensors or user-provided descriptions.

Related work endows robots with the ability to ask questions of the user in the context of following guided tours [11], executing route instructions [12], and interpreting a user's commands [13]. Kruijff et al. [11] outline a question-asking procedure mainly to determine robust room segmentation by asking about the presence of doorways. However, they do not consider allocentric descriptions, maintain multiple hypotheses, reason about entropy when choosing questions, or consider uncertainty over groundings. More recently, Deits et al. [13] looked at question-asking from an information-theoretic perspective in the scenario of following natural language manipulation commands. They determine the best questions to ask based upon their ability to reduce the entropy over the grounding for a given command. However, they do not reason over when to ask the questions, which instead immediately follow the corresponding command. While we use a similar information gain metric to drive our approach, we formulate the problem as a decision problem, where the robot has to decide between continuing the tour or interrupting the user to ask a question. In our case, a question can simultaneously refer to areas that the user described at distant points in time. This necessitates that we consider when it is most meaningful to ask the question and how it should be structured so as to provide sufficient context. Tellex et al. [14] propose an improved dialog method that employs a learned probabilistic language understanding model to ask questions that are easier for the user to understand and whose answers are more informative. With regards to improving the map, we use a similar reward metric to Stachniss et al. [15], who decide the best exploration-based motion actions that improve the entropy over the map.

III. Semantic Mapping Algorithm

During the guided tour, the robot constructs a spatial-semantic representation based on the algorithm we outlined in Hemachandra et al. [4]. Next, we reiterate our representation of the environment and the mechanism that we use to integrate natural language descriptions, and highlight how ambiguity in the descriptions can affect the utility of the information inferred from natural language.

A. Spatial-Semantic Representation

We define the semantic graph $S_t = \{G_t, X_t, L_t\}$ as a tuple containing topological, metric, and semantic representations of the environment. The topology G_t is composed of nodes n_i that denote the robot's trajectory through the environment (with a fixed 1 m spacing) and edges that denote metric constraints (e.g., via odometry). We associate a set of observations with each node that include laser scans z_i^l and distributions over the semantic scene class a_i inferred from laser scans a_i^l and camera images a_i^c. The algorithm assigns nodes to regions $R_\alpha = \{n_1, .., n_m\}$ that represent spatially coherent areas in the environment intended to be compatible with human concepts (e.g., rooms and hallways). We do so via spectral clustering using the overlap between laser scans to express the similarity between pairs of nodes [4].

The vector X_t consisting of the pose x_i of each node n_i constitutes the metric map, which takes the form of a pose graph [16] according to the structure of the topology. The semantic map L_t is modeled as a factor graph with variables that represent the type (e.g., office or lounge) and colloquial name (e.g., "Carrie's office") of each region in the environment. The algorithm infers this information using the aforementioned scene classifiers as well as the user's natural language descriptions Λ_t [4].

We maintain a distribution over the semantic graph using a Rao-Blackwellized particle filter, where each particle $S_t^{(i)}$ consists of a sampled topology, analytical distributions over the metric and semantic maps, and an associated weight $w_t^{(i)}$.

B. Grounding Natural Language Descriptions

We consider two types of natural language descriptions that the guide provides the robot, those that refer to the robot's current region (egocentric) and those that describe non-local, possibly distant regions in the environment (allocentric). We parse each utterance into its corresponding Spatial Description Clause (SDC), a structured language representation that includes a figure $\gamma_\mathcal{F}$, spatial relation r, and a (possibly null) landmark $\gamma_\mathcal{L}$ [17]. For example, the allocentric description "the lounge is down the hallway" results in an SDC in which the figure is the "lounge," the spatial relation is "down," and the landmark is the "hallway." With egocentric descriptions, the figure is implicitly the robot's current position and there is a null landmark.

In order to ground each expression Λ_k, the algorithm identifies the likelihood that each region is the landmark based on

its semantic distribution. We normalize these likelihoods to compute the landmark grounding probability for each region

$$p(\gamma_{\mathcal{L}} = R_j|\Lambda_k) = \frac{p(\phi^{\mathcal{L}}_{R_j} = \mathtt{T}|\Lambda_k)}{\sum_{R_k} p(\phi^{\mathcal{L}}_{R_k} = \mathtt{T}|\Lambda_k)}, \tag{1}$$

where $\gamma_{\mathcal{L}}$ is a grounding variable identifying the landmark region and $\phi^{\mathcal{L}}_{R_j}$ is a binary correspondence variable that is TRUE if region R_j is the landmark. For each potential landmark region, the algorithm then calculates the likelihood of each region in the map being the figure (i.e., $\gamma_{\mathcal{F}} = R_i$ and $\phi^{\mathcal{F}}_{R_i} = \mathtt{T}$) based on a model for the spatial relation r. We arrive at the overall figure grounding likelihood by marginalizing over the landmarks.

$$p(\phi^{\mathcal{F}}_{R_i} = \mathtt{T}|\Lambda_k) = \sum_{R_j} p(\phi^{\mathcal{F}}_{R_i} = \mathtt{T}|\gamma_{\mathcal{L}} = R_j, \Lambda_k) p(\gamma_{\mathcal{L}} = R_j|\Lambda_k) \tag{2}$$

We normalize this likelihood for each potential figure region

$$p(\gamma_{\mathcal{F}} = R_i|\Lambda_k) = \frac{p(\phi^{\mathcal{F}}_{R_i} = \mathtt{T}|\Lambda_k)}{\sum_{R_k} p(\phi^{\mathcal{F}}_{R_k} = \mathtt{T}|\Lambda_k)}. \tag{3}$$

We use this likelihood to update the semantic knowledge associated with figure region R_i in the map. However, the information gained from the description is diluted when the figure or landmark groundings are uncertain.

Our previous approach [3, 4] commits to a description once the likelihood of its grounding exceeds a threshold. Here, we improve upon this by continuously re-grounding the language when relevant regions of the map change. These changes can be in the form of updates to the metric position of the figure or landmark regions (e.g., due to a loop closure), or the addition of new potential landmark or figure regions to the map as they are visited.

IV. Action Selection Algorithm

Algorithm 1 outlines the process by which the robot updates its representation and chooses the optimal action $\mathcal{A}^*_t$. At each time step, the algorithm first updates the distribution over the semantic graph based on new odometry u_t, sensor observations z^l_t and a_t, natural language descriptions Λ_t, and the answer $z^{\mathcal{A}}_t$ to the last question that was asked $\mathcal{A}_{t-1}$. This includes reevaluating previous language descriptions and question-answer pairs based on newly visited regions.

Next, the algorithm formulates the guided tour as a decision process and selects the best action to take, which can either be to follow the user or to ask a particular question. We define the set of candidate questions according to the allocentric descriptions since their grounding is more ambiguous than egocentric descriptions. We model the value of the robot's next state using an information gain-based metric such that the robot values asking useful questions. We define the information gain as the reduction in entropy in the groundings for the natural language description based on the question $\mathcal{A}_t$ asked and the answer $z^{\mathcal{A}}_{t+1}$ received at the next time step. We associate a cost with these questions to model the social burden of asking a question.

Algorithm 1: Semantic Mapping and Action Selection

Input: $S_{t-1} = \left\{S^{(i)}_{t-1}\right\}$, and $\left(u_t, z^l_t, a_t, \mathcal{A}_{t-1}, z^{\mathcal{A}}_t, \Lambda_t\right)$, where $S^{(i)}_{t-1} = \left\{G^{(i)}_{t-1}, X^{(i)}_{t-1}, L^{(i)}_{t-1}, w^{(i)}_{t-1}\right\}$

Output: $\left\{\mathcal{A}^*_t, S_t = \left\{S^{(i)}_t\right\}\right\}$

1) Update the distribution over the semantic graph.

for $i = 1$ *to* n **do**
- a) Employ proposal distribution to propagate the graph sample $G^{(i)}_t$ based on u_t, Λ_t and a_t.
- b) Update the Gaussian distribution over the node poses $X^{(i)}_t$ conditioned on the topology.
- c) If A_{t-1} was a question, add $\left\{A_{t-1}, z^{\mathcal{A}}_t\right\}$ to the corresponding description.
- d) Reevaluate language descriptions and update the semantic layer $L^{(i)}_t$.
- e) Update the particle weights.

end

2) Normalize the weights and resample if needed.

3) Select the best action $\mathcal{A}^*_t$ (Eqn. 5).

We define the state as the robot's representation of the environment, i.e., the semantic graph S_t. If the robot asks a question and receives an answer, the algorithm updates the semantic graph S_t based on the question-answer tuple $\left\{\mathcal{A}_t, z^{\mathcal{A}}_{t+1}\right\}$. When the robot chooses to ask a question, it will stop and wait for the answer before continuing to follow the user. As such, the next state S_{t+1} depends only on S_t, the robot's question, and the guide's response. Since each question queries the figure region referenced by a description, the update to the semantic graph modifies only the distribution over this grounding. If the robot chooses to follow the guide, the algorithm updates the semantic graph according to subsequent observations and descriptions. In this work, we do not predict the change in state that results from following the guide since it would require reasoning over the unobserved part of the world.

The algorithm utilizes a Rao-Blackwellized particle filter to maintain the semantic graph distribution, which we represent as a collection of weighted particles $S_t = \{S^{(i)}_t\}$. Thus, we model the action selection as a MDP for each particle $S^{(i)}_t$ and infer the optimal action given the distribution over the semantic graph using the QMDP heuristic [6].

The Q value for each particle $S^{(i)}_t$ is defined as

$$Q(S^{(i)}_t, \mathcal{A}_t) = \sum_{S^{(i)}_{t+1}} \gamma V(S^{(i)}_{t+1})\, p(S^{(i)}_{t+1}|S^{(i)}_t, \mathcal{A}_t) - \mathcal{C}(\mathcal{A}_t), \tag{4}$$

where $\gamma = 1$ is a discount factor, $V(S^{(i)}_{t+1})$ is the value of the state at the next timestep, and $\mathcal{C}(\mathcal{A}_t)$ is the cost associated with action $\mathcal{A}_t$, each of which we define shortly. For question-asking actions, we represent the value of the next state in terms of the information gain associated with getting an answer to a question that seeks to resolve the figure

grounding for a particular description. This information-theoretic metric biases the decision process to value actions that in expectation reduce the uncertainty over the language groundings in the semantic graph.

At each time step, the robot chooses the best action $\mathcal{A}_t^*$ according to the QMDP heuristic

$$\mathcal{A}_t^* = \arg\max_{\mathcal{A}_t} \sum_{S_t^{(i)}} p(S_t^{(i)}) Q(S_t^{(i)}, \mathcal{A}_t), \tag{5}$$

where $p(S_t^{(i)})$ is the particle weight $w_t^{(i)}$. The following paragraphs explain this process in detail.

A. Action Set

The available actions include the FOLLOWPERSON action A_F and the set of valid question-asking actions. We focus on questions with a limited set of answers ("yes" and "no"), which allows us to more easily model the information gain for each question and answer pair, and reason over the likelihood of receiving each answer given the question. We use a template to compose questions for each grounding figure entity in a natural language description. These templates can be categorized into two basic types:

I *Egocentric Questions*: This template chooses from a set of spatial relations ("at," "near," "away," "in front," "behind," "left," or "right") and a grounding variable to create a question (e.g., "Is the kitchen in front of me?").

II *Allocentric Questions*: This template defines questions in terms of spatial relations between non-local figure and landmark regions in the environment (e.g., "Is the lounge next to the conference room?"). We select landmark regions that have a high likelihood of a particular semantic label and generate questions that reference other regions in relation to these landmarks.

The robot can only use egocentric questions to ask about spatial regions in its immediate vicinity. As such, the ability to receive useful information is limited to instances when the robot is near a potential hypothesized location. Allocentric questions allow the robot to reduce its uncertainty even when a hypothesized location is not within view. However, the need for the user to situate this question in their understanding of the environment may impose a higher mental burden.

B. Value Function

We define the value of the next state in Equation 4 as a function of the information gain resulting from an action

$$V(S_{t+1}^{(i)}) = f(I(\mathcal{A}_t, z_{t+1}^{\mathcal{A}})). \tag{6}$$

In the case of question-asking actions, the next state S_{t+1} depends only on S_t, the asked question $\mathcal{A}_t$, and the user's answer $z^{\mathcal{A}_t}$. Thus, the space of possible next states $S_{t+1}^{(i)}$ for semantic graph particle $S_t^{(i)}$ is limited by the set of question and answer pairs.

We assume that there is no information gain for the FOLLOWPERSON action A_F, since the motion is assumed not to affect the distribution over the language grounding. Thus, the associated Q value (Eqn. 4) only expresses the cost of the action.

C. Information Gain

We define the information gain $I(\mathcal{A}_t, z_t^{\mathcal{A}})$ for a question-answer pair as the reduction in entropy H over the figure grounding variable $\gamma_{\mathcal{F}}$ for a description Λ_k provided by the guide.[1]

$$I(\mathcal{A}_t, z_t^{\mathcal{A}}) = H(\gamma_{\mathcal{F}}|\Lambda_k) - H(\gamma_{\mathcal{F}}|\Lambda_k, \mathcal{A}_t, z_t^{\mathcal{A}}) \tag{7}$$

We calculate the new entropy $H(\gamma_{\mathcal{F}}|\Lambda_k, \mathcal{A}_t, z_t^{\mathcal{A}})$ for the updated distribution over the figure grounding based on the question and answer pair

$$p(\gamma_{\mathcal{F}}{=}R_i|\Lambda_k, \mathcal{A}_t, z_t^{\mathcal{A}}){=}\frac{p(z_t^{\mathcal{A}}|\mathcal{A}_t, R_i)\, p(\gamma_{\mathcal{F}}{=}R_i|\Lambda_k)}{\sum_{R_j} p(z_t^{\mathcal{A}}|\mathcal{A}_t, R_j) p(\gamma_{\mathcal{F}}{=}R_j|\Lambda_k)} \tag{8}$$

where $p(z_t^{\mathcal{A}}|\mathcal{A}_t, R_i)$ is the likelihood of the answer $z_t^{\mathcal{A}}$ to question $\mathcal{A}_t$ referring to region R_i. We compute the probability of a "yes" answer as the likelihood that region R_i is consistent with the landmark and spatial relation for the question (Eqn. 9).

D. Answer Likelihood

We define the answer probability by introducing a latent random variable v that specifies whether the question is valid given the referenced region ($p(z_t^{\mathcal{A}}|v, \mathcal{A}_t, R_i)$)

$$p(z_t^{\mathcal{A}}|\mathcal{A}_t, R_i) = \sum_{v\in 0,1} p(z_t^{\mathcal{A}}|v, \mathcal{A}_t, R_i)\, p(v|\mathcal{A}_t, R_i). \tag{9}$$

The prior $p(v|\mathcal{A}_t, R_i)$ expresses the likelihood that the question is contextually relevant to figure region R_i given the user's location. In our experiments, we only considered regions sufficiently close to the user[2] to be contextually valid. We use this to limit the entropy calculation (Eqn. 8) to regions for which the question is expected to be meaningful.

When question $\mathcal{A}_t$ (e.g., "Is the kitchen in front of me?") using spatial relation r is valid ($v = 1$) for a potential grounding region R_i, we define the likelihood of receiving a "yes" answer as

$$p(z_t^{\mathcal{A}}{=}\text{"yes"}|v{=}1, \mathcal{A}_t, R_i) = p(\phi_{R_i}^{\mathcal{F}}{=}\mathtt{T}|\gamma_{\mathcal{L}}{=}R_{\mathcal{L}}, r), \tag{10}$$

where $R_{\mathcal{L}}$ is the landmark region specified in the question (null for egocentric questions). We represent this likelihood using probabilistic spatial language models learned from natural language corpora [17], in the same manner as with the user's descriptions (Sec. III-B).

E. Transition Likelihood

For a question $\mathcal{A}_t$ about a figure region γ_F that was originally described in the description Λ_k, the transition likelihood for a single particle $p(S_{t+1}^{(i)}|S_t^{(i)}, \mathcal{A}_t)$ is equivalent to the likelihood of receiving a particular answer given the state and the question-asking action. In order to arrive at the overall likelihood, we calculate the probability of receiving the answer for each potential region (Eqn. 9) and marginalize

[1]The information gain calculation is limited to grounding regions in the environment that have currently been explored.

[2]In the experiments presented in the paper, we use 15 m as the threshold.

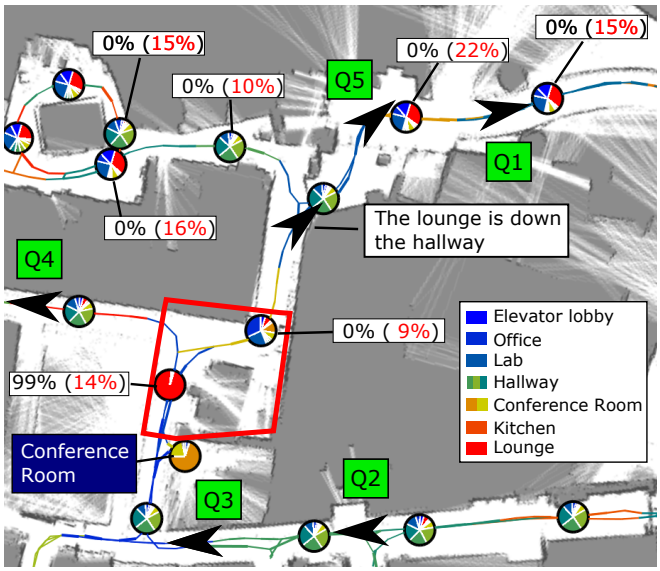

Fig. 2. Experiment I: Language groundings for the expression "The lounge is down the hallway." Values in red denote the grounding likelihood without questions, while those in black express the likelihood with questions. Question (answer): Q1: "Is the lounge near me?" ("No"); Q2: "Is the lounge near me?" ("No"); Q3: "Is the lounge near me?" ("Yes"); Q4: "Is the lounge near me?" ("Yes"); Q5: "Is the lounge near the conference room?" ("Yes"). The ground truth region boundary is in red. The robot's locations at the time of descriptions and questions are denoted with black arrows. Path color denotes region segmentation while pie charts denote a region's type.

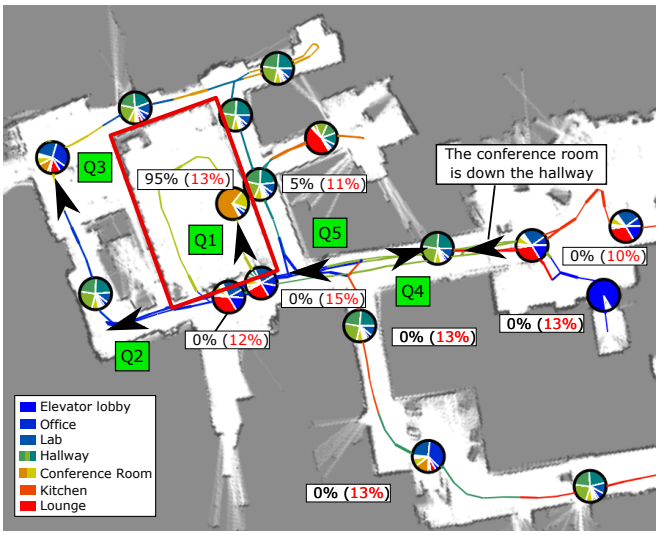

Fig. 3. Experiment I: Language groundings for the expression "The conference room is down the hallway." Values in red denote the grounding likelihood without questions, while those in black express the likelihood with questions. Question (answer): Q1: "Is the conference room near me?" ("Yes"); Q2: "Is the conference room behind me?" ("No"); Q3: "Is the conference room near me?" ("Yes"); Q4: "Is the conference room near me?" ("No"); Q5: "Is the conference room near me?" ("Yes"). The ground truth region boundary is in red. The robot's locations at the time of descriptions and questions are denoted with black arrows. Path color denotes region segmentation while pie charts denote a region's type.

over the regions based on their prior likelihood of being the figure grounding (Eqn. 3).

$$p(S_{t+1}^{(i)}|\mathcal{A}_t, S_t^{(i)}) = p(z_t^{\mathcal{A}}|\mathcal{A}_t, S_t^{(i)}) \tag{11a}$$

$$= \sum_{R_j} p(z_t^{\mathcal{A}}|\mathcal{A}_t, S_t^{(i)}, \gamma_{\mathcal{F}} = R_j)\, p(\gamma_{\mathcal{F}} = R_j|\Lambda_k) \tag{11b}$$

We include the description Λ_k to make the dependence on language explicit. This results in higher transition likelihoods for new states with figure regions that both have a higher grounding likelihood for the original description as well as an increased likelihood for the answer.

F. Cost Function

While questions are useful in resolving ambiguity in the map, they also require the user's time and effort. We account for the latter with a hand-crafted cost function $\mathcal{C}(\mathcal{A}_t)$ that encodes the burden of asking a given question at each timestep. The cost is a function of the following features:

i. Time since the last question was asked
ii. Time since last question was asked about the grounding

We use a linear combination of these features to arrive at the cost function. We set the weights such that they balance the burden of asking questions with reducing the ambiguity in the description groundings. We additionally associate a fixed (negative) cost with the person-following action A_F such that only a reasonably high expected information gain will result in a question being asked.

G. Integrating Answers to the Representation

Each question and answer modifies the distribution over the grounding variable $\gamma_{\mathcal{F}}$. In turn, an informative answer will improve the semantic graph distribution. We use the question-answer pair and the corresponding language distribution Λ_k to calculate the grounding likelihood (Eqn. 8) and subsequently, the information gain (Eqn. 7). Unlike the case of the entropy calculation, which only considers contextually valid regions, we consider all potential groundings, including regions outside the meaningful area for the question.

When new valid grounding regions are added, we reevaluate both the original description as well as the likelihood of generating the received answer for each new region, and update the language grounding. Figure 2 shows the grounding likelihoods before and after asking three questions.

V. Results

We evaluated our algorithm on two indoor datasets in which a human gives a robotic wheelchair (Fig. 1) [2] narrated tours of different floors of MIT's Stata Center. For these experiments, we purposefully injected natural language descriptions at locations where their groundings are ambiguous. Dataset I consists of seven egocentric descriptions that the algorithm grounds to the robot's current region, and three allocentric expressions that describe regions with relation to either landmarks (e.g., "The elevator lobby is down the hall") or the robot's location (e.g., "The lounge is behind you"). Dataset II consists of two egocentric descriptions that the algorithm correctly grounds to the robot's current region, and three allocentric expressions that describe regions relative to

TABLE I

EXPERIMENT I: ENTROPY OVER FIGURE GROUNDINGS WITH AND WITHOUT QUESTIONS

		Without Questions		With Questions				
				# of	Random		Our Method	
Dataset	Utterance	Entropy	Accuracy	Questions	Entropy	Accuracy	Entropy	Accuracy
I	(A) "The lounge is down the hallway" (Fig. 2)	2.100	15.3%	5	1.301	54.6%	0.050	92.5%
	(B) "The elevator lobby is down the hallway"	1.396	24.1%	1	1.148	35.7%	0.000	83.8%
	(C) "The lounge is behind you"	0.414	83.1%	1	0.422	80.5%	0.126	87.3%
II	(D) "The lab is down the hall"	2.047	20.3%	4	1.612	37.9%	0.384	90.3%
	(E) "The conference room is down the hallway" (Fig. 3)	2.154	11.3%	5	1.252	44.3%	0.226	89.3%
	(F) "The lounge is in front of us"	1.199	23.3%	2	1.086	24.2%	0.254	43.8%

either landmarks or the robot's pose. We ran the algorithm on the datasets offline in real-time with a human providing answers to the questions. The algorithm operated with four particles, which were sufficient to accurately capture valid topologies in these relatively small environments.

We quantify the results using two metrics, the reduction of entropy and improvement in accuracy for the language grounding of each utterance. We express the accuracy of each grounding k in terms of the spatial overlap $\mathbb{O}_{R_i}$ of each inferred figure region with its ground truth location. We weight the overlap by the grounding likelihood and sum to arrive at the accuracy score

$$\mathbb{S}_k = \sum_{R_i} \mathbb{O}_{R_i}\, p(\gamma_f = R_i | \Lambda_k, \{\mathcal{A}, z^{\mathcal{A}}\}). \qquad (12)$$

The score penalizes situations in which the groundings are assigned to regions outside the ground truth region or the grounded regions contain only a part of the ground truth region (i.e., due to improper segmentation).

A. Experiment I: Egocentric and Allocentric Questions

The first experiment that we consider enables the algorithm to ask both egocentric and allocentric questions. For comparison, we consider two baselines. The first considers the case in which the robot can not ask questions, which is equivalent to our previous semantic mapping algorithm [4]. For the second baseline, the robot asks the same number of questions as our method, but the questions are constructed from figure regions, landmark regions, and spatial relations chosen at random. For the first dataset, the robot asked a total of seven questions, including one allocentric question. On the second dataset, the robot asked eleven questions of the guide, including one allocentric question.

Table I compares the performance of our algorithm with the two baselines. By asking targeted questions, our method significantly reduces the entropy over the figure groundings while increasing their accuracy. On average, there is a 86% reduction in the entropy and a 309% improvement in the accuracy of the figure groundings compared to the baseline that does not ask questions. When compared against asking random questions, we see a 67% reduction in entropy and a 138% improvement in accuracy on average. This supports the claim that deliberately choosing when and which question to ask is integral to our algorithm's success.

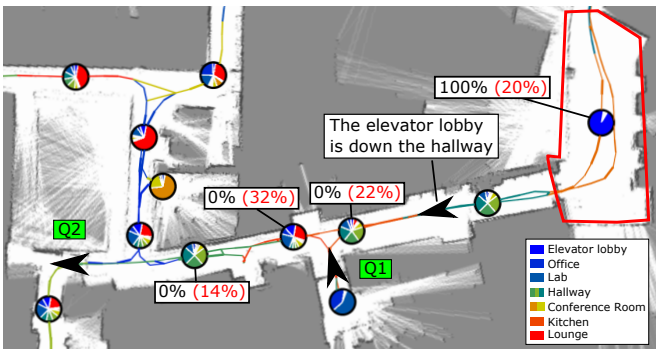

Fig. 4. Experiment II: Language groundings for the utterance "The elevator lobby is down the hallway." Values in red denote the grounding likelihood without questions, while those in black express the likelihood with questions. Question (answer): Q1: "Is the elevator lobby near the lab?" ("No"); Q2: "Is the elevator lobby near the conference room?" ("No"). The ground truth region boundary is in red. The robot's locations at the time of descriptions and questions are denoted with black arrows. Path color denotes region segmentation while pie charts denote a region's type.

B. Experiment II: Allocentric Questions Only

We also applied the algorithm to the first dataset with the restriction that the system only ask allocentric questions. By only allowing the robot to ask allocentric questions, we better demonstrate the potential to reduce the ambiguity even when the robot doesn't revisit the locations it is uncertain about. However, the reduction in entropy depends upon the presence of salient landmarks (which came from language annotations in this experiment). Table II compares the performance of only asking allocentric questions with asking both egocentric and allocentric questions. The robot is still able to reduce the ambiguity in the map, but not to the level that it achieves by asking both types of questions. The reasons are two-fold: firstly, there are only a few salient landmarks that can be used to generate valid questions; secondly, allocentric questions were only generated using the "near" spatial relation limiting the available set of useful questions. Figure 4 shows the resulting grounding likelihoods for the utterance "The elevator lobby is down the hallway."

TABLE II

FIGURE GROUNDING ENTROPY WITH EGOCENTRIC AND ALLOCENTRIC QUESTIONS (NUMBER OF QUESTIONS IN PARENTHESES)

Utterance	Entropy		Accuracy	
	Question Type		Question Type	
	Both	Allocentric	Both	Allocentric
(A)	0.050 (5)	1.632 (1)	92.5%	70.0%
(B)	0.000 (1)	0.000 (2)	83.8%	96.4%
(C)	0.126 (1)	0.441 (0)	87.3%	78.5%

VI. CONCLUSION

We described a framework that enables a robot to engage a human in dialog to improve its learned semantic map during a guided tour. The method takes an information-theoretic approach to deciding when and which questions to ask in order to best reduce ambiguity in its model of the environment. We evaluated the algorithm on a pair of experiments that demonstrate that by asking targeted questions, the algorithm is able to reduce the entropy and improve the accuracy of the semantic map compared to asking no questions as well as to a baseline in which random questions were asked.

While the framework results in more accurate environment models, there are limitations to the proposed method. First, the framework currently assumes that the figure region that the user refers to when answering a question is in the robot's current model of the environment, which is not necessarily the case. Instead, the framework should model the likelihood that the figure grounds to unvisited regions in the environment (e.g., via a learned spatial prior). Second, the cost function is integral to the decision of whether or not to ask a question. We currently define this cost in terms of a linear combination of hand-selected features and weights. A more effective approach would be to employ a richer function that captures the complex set of factors that influence the effort in answering questions. This cost function should be developed and learned based on user studies. Third, the space of actions can be expanded beyond questions to allow the robot to physically explore the environment as another means of resolving ambiguity. These, together with human-subject studies to better understand the generalizability and effectiveness of the algorithm are directions for future work.

REFERENCES

[1] 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.

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

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

[4] S. Hemachandra, M. R. Walter, S. Tellex, and S. Teller, "Learning spatial-semantic representations from natural language descriptions and scene classifications," in *Proc. IEEE Int'l Conf. on Robotics and Automation (ICRA)*, May 2014.

[5] M. R. Walter, S. Hemachandra, B. Homberg, S. Tellex, and S. Teller, "A framework for learning semantic maps from grounded natural language descriptions," *International Journal of Robotics Research*, vol. 33, no. 9, pp. 1167–1190, August 2014.

[6] M. L. Littman, A. R. Cassandra, and L. P. Kaelbling, "Learning policies for partially observable environments: Scaling up," in *Proc. Int'l Conf. on Machine Learning (ICML)*, 1995.

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

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

[9] T. Williams, R. Cantrell, G. Briggs, P. Schermerhorn, and M. Scheutz, "Grounding natural language references to unvisited and hypothetical locations," in *Proc. Nat'l Conf. on Artificial Intelligence (AAAI)*, 2013.

[10] F. Duvallet, M. R. Walter, T. Howard, S. Hemachandra, J. Oh, S. Teller, N. Roy, and A. Stentz, "Inferring maps and behaviors from natural language instructions," in *Proc. Int'l. Symp. on Experimental Robotics (ISER)*, Marrakech/Essaouira, Morocco, June 2014.

[11] G.-J. M. Kruijff, H. Zender, P. Jensfelt, and H. I. Christensen, "Clarification dialogues in human-augmented mapping," in *Proc. ACM/IEEE Int'l. Conf. on Human-Robot Interaction (HRI)*, Salt Lake City, UT, 2006.

[12] H. Shi and T. Tenbrink, "Telling Rolland where to go: HRI dialogues on route navigation," *Spatial Language and Dialogue*, 2009.

[13] R. Deits, S. Tellex, P. Thaker, D. Simeonov, T. Kollar, and N. Roy, "Clarifying commands with information-theoretic human-robot dialog," *J. Human-Robot Interaction*, vol. 2, no. 2, pp. 58–79, 2013.

[14] S. Tellex, R. Knepper, A. Li, D. Rus, and N. Roy, "Asking for help using inverse semantics," in *Proc. Robotics: Science and Systems (RSS)*, Berkeley, CA, July 2014.

[15] C. Stachniss, G. Grisetti, and W. Burgard, "Information gain-based exploration using rao-blackwellized particle filters," in *Proc. Robotics: Science and Systems (RSS)*, Cambridge, MA, USA, 2005.

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

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