Learning Semantic Maps Through Dialog for a Voice-Commandable Wheelchair

Sachithra Hemachandra and Matthew R. Walter
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139 USA
{sachih, mwalter}@csail.mit.edu

***Abstract*—In this paper, we propose an algorithm that enables a voice-commandable wheelchair to learn a semantic model of a user's environment by engaging them in dialog. The algorithm reduces the entropy in maps formulated based upon user-provided natural language descriptions (e.g., "The kitchen is down the hallway"). The robot's available information-gathering actions take the form of targeted questions intended to reduce the entropy over the grounding of the user's descriptions. These questions include those that query the robot's local surround (e.g., "Are we in the kitchen?") as well as areas distant from the robot (e.g., "Is the lab near the kitchen?"). Our algorithm treats dialog as an optimization problem that seeks to balance information-theoretic value of candidate questions with a measure of cost associated with dialog. In this manner, the method determines the best questions to ask based upon expected entropy reduction while accounting for the burden on the user. We evaluate the entropy reduction based upon 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. We demonstrate that, by asking deliberate questions of the user, the method results in significant improvements in the accuracy of the resulting map.**

I. Introduction

The Boston Home (TBH) is a long-term assisted living residence in Boston, Massachusetts USA for adults with multiple sclerosis (MS) and other progressive neurological disorders. TBH has approximately 100 residents who, like others living with MS, are inhibited in their ability to move about and interact with their environment. MIT has worked together with TBH to develop and deploy a variety of assistive technologies in an effort to improve the quality of life of its residents. These include a localization system that uses the facility's existing WiFi infrastructure to monitor the safety and location of the residents [1, 2], and dialog interfaces [3, 4] that allow residents to use speech to request information (e.g., regarding daily events or weather), make phone calls, and send e-mail [5].

The majority of TBH residents require power wheelchairs to move within and on the grounds of the facility. The way in which users drive their wheelchairs depends upon their physical capabilities. Residents often use hand-operated joysticks initially, but, as their muscular control deteriorates, so does their ability to accurately steer their chair. Head-actuated switches and sip and puff arrays offer safer alternatives, but reduce the user's level of control, significantly impact operating speeds, and can be physically taxing. Taking advantage of advancements in robot navigation, semi-autonomous wheelchairs [6, 7, 8] seek to overcome these limitations by augmenting a user's ability with automatic wall following and obstacle avoidance through a shared control interface.

Fig. 1. A staff member gives a robotic wheelchair a tour of The Boston Home, a long-term care residence for adults with neurological diseases.

While a resident's motor skills may deteriorate, many afflicted with MS and similar neurological disorders retain the ability to speak, albeit with potentially significant speech pathologies. Building on our lab's earlier work in automatic speech recognition (ASR) [9] and language understanding [10], we developed a voice-commandable autonomous wheelchair (Fig. 1) that enables users to navigate simply by instructing their wheelchair using natural language speech (e.g., "take me to the room across from the nurse's station").

In order to understand these natural language instructions, the wheelchair needs to reason over environment representations that model the spatial, topological, and semantic properties (e.g., room types and names) that users associate with their environment. An effective means of sharing this knowledge with the wheelchair is for a staff member to lead the wheelchair on a guided tour (Fig. 1), using natural language speech to describe the environment [11, 12, 13, 14]. With these approaches, the robot takes a passive role, whereby it infers information from descriptions and 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. The guide's descriptions tend to be ambiguous, with several possible interpretations (*groundings*) for a particular environment. For example, the guide may describe the location of the kitchen as being "down

the hall," yet there may be several hallways nearby, each leading to a number of different rooms. Furthermore, language grounding typically requires a complete map, however the robot may not yet have visited the regions that the guide is referring to. It may be that the guide is describing a location known to the robot or a new location outside the field-of-view of its sensors.

Rather than try to passively resolve the ambiguity in the inferred map, the robot can take active information-gathering actions, either physically exploring the environment or, as we consider in this paper, asking questions of the guide. There are several challenges to using dialog in order to improve the accuracy of the inferred map in an effective manner. The first involves context. It would be beneficial if the algorithm was not restricted to questions that query the robot's current location. However, asking the guide about temporally and spatially distant locations necessitates that the questions provide the guide with sufficient context. Second, the questions should be structured in such a way that the answers are as informative as possible. Third, it is important that the method accounts for the social cost incurred by engaging the guide in dialog, for example, by not asking too many questions.

This paper considers the scenario in which the wheelchair acquires a model of the world through a guided tour [13, 14], where a human shows the robot around the facility while providing natural language descriptions. During the tour, the robot maintains a distribution over the *semantic graph*, which is a metric, topological and semantic representation of the environment, using a Rao-Blackwellized particle filter [13]. The robot also decides between an action that either follows the guide or asks a question to improve its representation. We formulate the decision process as a QMDP [15], where the actions are evaluated as a Markov Decision Process (MDP) for each possible configuration of the world (particle), and the best action is selected using the QMDP heuristic. This allows us to balance the information gained by asking questions of the guide with the cost of each action. The algorithm reasons over the natural language descriptions and the current learned map to identify the (possibly null) question that best reduces the ambiguity in the map. The algorithm considers egocentric and allocentric binary (yes/no) questions that consist of spatial relations between pairs of regions. These regions may be local to the robot in the case of situated dialog (e.g., "Are we in the kitchen?", "Is the nurse's station on my right?") or distant in the case of non-situated dialog (e.g., "Is the lounge next to the kitchen?"). We assign a cost to each action based on the interaction and a reward based on the information gain for each possible answer. The algorithm then selects the best action using the expected Q value of each action using the QMDP formulation.

We demonstrate that this question asking policy reduces the ambiguity in natural language descriptions and, in turn, results in semantic maps of the environment that are more accurate than the current state-of-the-art.

II. Related Work

Several approaches exist that construct semantic environment models using traditional robot sensors [11, 12, 16], while others have looked at also integrating natural language descriptions to improve the semantic representations [13, 17, 14]. With most of these techniques, however, the robot only passively receives observations, whether they are from traditional sensors or user-provided descriptions.

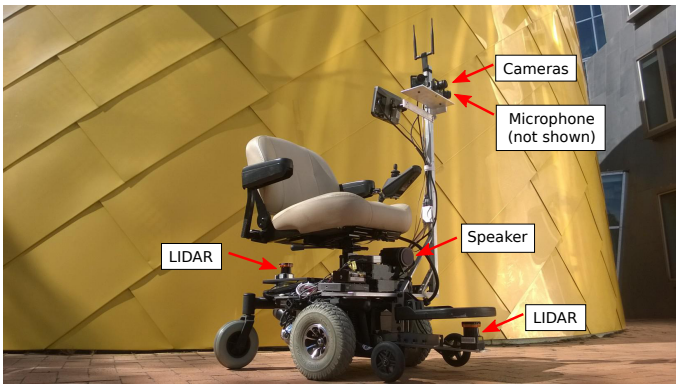

Fig. 2. The prototype robotic wheelchair.

Related work exists that endows robots with the ability to ask questions of a human in the context of following guided tours [18] and understanding the user's commands [19]. Kruijff et al. [18] outline a procedure that asks about the presence of doorways in order to more robustly segment the environment. However, their work is limited to egocentric utterances and does not account for ambiguity in the descriptions. More recently, Deits et al. [19] have looked at asking questions in the context of following natural language manipulation commands. They use an information gain-based metric to evaluate the best questions to ask in order to reduce the entropy over the grounding for a given command. However, the questions they ask are more straightforward and do not explicitly provide context to the human. 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 tour to ask a question. Furthermore, Deits et al. [19] do not reason over when to ask the questions, since they immediately follow the corresponding command. In our case, a question can refer to areas that the guide described at distant points in time. This necessitates that we consider when it is most meaningful to ask the question and that it be phrased in a manner that provides sufficient context.

III. The Wheelchair Platform

Based upon our interactions with clinicians and residents at TBH, we built a prototype robotic wheelchair (Fig. 2), by taking an off-the-shelf power wheelchair and adding circuitry to control its drive motors and sensors to perceive its surround. The platform is equipped with two Hokuyo UTM-30LX planar LIDARs, both with horizontal scanning planes. The forward-facing LIDAR is positioned a few inches off the ground to observe building structure (e.g., for mapping), obstacles, and people in front of the robot. The robot employs this sensor to detect and track the location of the staff member who is conducting the tour. The rearward-facing LIDAR is positioned slightly higher and is used to detect walls and obstacles. Additional exterocpetive sensing includes three cameras mounted approximately 1 m above the ground, which provide a nearly 180 degree field-of-view in front of the wheelchair. These cameras as well as the LIDARs allow the robot to identify a

region's type (e.g., hallway, lounge, etc.) based upon its image-space appearance and local structure. A directional microphone located near the cameras enables the wheelchair to receive spoken commands from users as well as descriptions that the tour guide provides[1]. The wheelchair uses a speaker to engage in dialog, notably to ask questions of the tour guide.

IV. Semantic Graph Representation

A. Spatial-Semantic Representation

We define the semantic graph [13] 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 (sampled at a fixed 1 m spacing) and edges that represent connectivity. We associate with each node a set of observations that include laser scans z_i, semantic appearance observations a_i based on laser l_i and camera i_i models, and available language observations λ_i. We assign nodes to regions $R_\alpha = \{n_1, .., n_m\}$ that represent spatially coherent areas in the environment intended to be compatible with human concepts (e.g., rooms and hallways).

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 [20] according to the structure of the topology. The semantic map L_t is modeled as a factor graph with variables that represent the type C_r (e.g., resident's room, lounge) and label Λ_r (e.g., "John's room") for each region r in the environment. This information is inferred from observations made from scene classifiers (image and laser) as well as by grounding the guide's natural language descriptions [13]. In this paper, we consistently segment groups of nodes into regions using spectral clustering (compared to sampling segments in Hemachandra et al. [14]). We also use a template-based door detector to segment regions.

B. Grounding Natural Language Descriptions

We consider two broad types of natural language descriptions provided by the guide. Egocentric descriptions that refer to the robot's immediate surround are directly grounded to the region in which the description was provided. Allocentric descriptions that provide information about distant regions require more careful handling.

We parse each natural language command into its corresponding Spatial Description Clauses (SDCs), a structured language representation that includes a figure, a spatial relation and possibly a landmark [10]. 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 from," and the landmark is the "hallway." With egocentric descriptions, the landmark or figure are implicitly the robot's current position.[2]

The algorithm grounds the expression by inducing a distribution over the figure's location. It does so by treating the location of the landmark as a latent variable, calculating the normalized likelihood that a region R_j is the landmark based upon that region's label distribution according to the semantic map

$$p(\gamma_l = R_j) = \frac{p(\phi^l_{R_j} = \mathtt{T})}{\sum_{R_j} p(\phi^l_{R_j} = \mathtt{T})}, \quad (1)$$

where γ_l is the region that the description's landmark reference grounds to, and $\phi^l_{R_j}$ denotes the binary correspondence variable that specifies whether region R_j is the landmark. For every potential landmark, the algorithm then calculates the likelihood of each region in the map as being the corresponding figure based on a model for the spatial relation SR. We arrive at the overall likelihood that this region is the figure grounding by marginalizing over the landmarks

$$p(\phi^f_{R_i} = \mathtt{T}) = \sum_{R_j} p(\phi^f_{R_i} = \mathtt{T} | \gamma_l = R_j, SR)\, p(\gamma_l = R_j), \quad (2)$$

where $\phi^f_{R_i}$ is the correspondence variable for the figure. We normalize these likelihoods for each potential figure region

$$p(\gamma_f = R_i) = \frac{p(\phi^f_{R_i} = \mathtt{T})}{\sum_{R_i} p(\phi^f_{R_i} = \mathtt{T})}. \quad (3)$$

This expresses the likelihood of the corresponding variable being true for each figure region R_j in the factor graph in the semantic layer. However, when there is uncertainty over the landmark or figure grounding, the likelihood of the label associated with the figure region can become diluted.

In our previous approaches [13, 14], we commit to a description once the likelihood of its grounding exceeds a pre-specified threshold. In this paper, we improve upon this by continuously re-grounding the language when relevant regions of the map change. These changes could be in the form of

Algorithm 1: Semantic Mapping Algorithm

Input: $P_{t-1} = \left\{ P^{(i)}_{t-1} \right\}$, and $(u_t, z_t, a_t, \lambda_t)$, where $P^{(i)}_{t-1} = \left\{ G^{(i)}_{t-1}, X^{(i)}_{t-1}, L^{(i)}_{t-1}, w^{(i)}_{t-1} \right\}$

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

1) Update Particles with odometry and sensor data.

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

1) Employ proposal distribution to propagate the graph sample based on u_t, λ_t and a_t.
   a) Segment regions
   b) Sample region edges
   c) Merge newly connected regions
2) Update the Gaussian distribution over the node poses $X^{(i)}_t$ conditioned on topology.
3) Reevaluate language groundings and answered question and update the semantic layer L_t.
4) Update particle weights.

end

2.) Normalize weights and resample if needed.

3.) Evaluate action costs and carry out minimum cost action.

[1]The guide can also speak to the wheelchair using a wireless, head-worn microphone

[2]We make the assumption that the descriptions are provided with respect to the robot's reference frame and not that of the guide.

updates to the metric position of the figure or landmark regions (e.g., due to a loop closure), or new potential landmark or figure regions being visited and added to the map.

V. Learning from Dialog

Algorithm 1 outlines the process by which robot updates its representation and decides on the optimal action. At each time step, the system integrates the odometry and sensor information to update the distribution over semantic graphs. This includes reevaluating the language descriptions and answers received for questions asked from the guide. Then, the algorithm evaluates the cost of each valid dialog action, and executes the one with the highest expected Q value. The following section elaborates on our action selection procedure.

A. Action Selection

In this section, we outline the action selection procedure employed by the algorithm. We treat the guided tour as an MDP, with associated costs for taking each action. These actions include following the person, staying in place, and asking a particular question. We define an additional set of question asking actions dependant on the current number of allocentric descriptions provided by the guide. We introduce a cost function for these question asking actions based upon the expected information gain for each question as well as a measure of social burden.

We define the state S_{t+1} as a tuple of $\{P_t^{(i)}, a_t, z_t^a\}$, where $P_t^{(i)}$ is particle i at time t, a_t is the action taken, and z_t^a is the resulting observation. For a single particle, we define the Q value as

$$\begin{aligned} Q(S_t, a_t) &= \sum_{S_{t+1}} \gamma V(S_{t+1}) \times p(S_{t+1}|S_t, a_t) - \mathcal{C}(a_t) \\ &= \sum_{S_{t+1}} \gamma \mathbb{E}(V(S_{t+1})) - \mathcal{C}(a_t), \end{aligned} \tag{4}$$

where the value of S_{t+1},

$$V(S_{t+1}) = \mathcal{F}(I(a_t)), \tag{5}$$

is a function of the information gain, and the cost of question asking action a_t

$$\mathcal{C}(a_t) = \mathcal{F}(f(a_t)) \tag{6}$$

is a function of the feature set of each action. We use a discounting factor $\gamma = 1$.

At each time step, the robot takes the best action a_t^B from the available set of actions using the QMDP heuristic.

$$a_t^B = \arg\max_{a_t} \sum_{S_t} p(S_t) Q(S_t, a_t), \tag{7}$$

where $p(S_t)$ is the particle weight $w_t^{(i)}$.

1) Action Set: The action set consists of the "Follow Person" action $A_{\mathcal{F}}$, "Stay-In-Place" action $A_{\mathcal{S}}$, and the valid set of question asking actions. The "Follow Person" action $A_{\mathcal{F}}$ is available at all times except when the robot is waiting for an answer to a question, when only $A_{\mathcal{S}}$ is available for selection. We derive our questions from a templated set for each grounding entity in a natural language description. These templates can be categorized into two basic types:

i Situated questions employ a spatial relation (near, away, front, behind, left, right) relative to the robot's pose to query a region by its label or type (e.g., "Is the kitchen in front of me?"). The answer can be "yes," "no," or "invalid" (for questions that do not make sense).

ii Non-situated questions consider spatial relations between two non-local regions, referred to by their label or type (e.g., "Is the lounge in front of the conference room?"). The answer can be "yes," "no," or "invalid."

The robot can only use questions of the first type to ask about 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. Questions of the second type allow the robot to reduce its uncertainty even when a hypothesized location is not within its immediate vicinity. In general, these questions are asked when the robot is confident about the location of one region but uncertain about the other. We note that these questions may place a higher mental burden on the guide, who must then reason about spatial entities outside their immediate perception range.

2) Value Function: We define the value of the next state as a linear function of the information gain for each action. We define the next state S_{t+1} as the question and answer pair. Each next state is assigned a value based on the information gain for the related language grounding. Since the answer for a given question is unknown, we evaluate the expected likelihood of transitioning to a particular state given a question. The likelihood of transitioning to each state is the likelihood of receiving a particular answer given the question.

We define the information gain $I(a, z^a)$ for action a (8) the reduction in entropy by taking action a and receiving observation z^a. In our framework, the entropy is over a grounding variable γ_f created for a natural language description provided by the guide. Calculating the exact entropy is infeasible since the map might not yet be complete, and also because it is inefficient to calculate the likelihood of some spatial regions that are too far outside the local area. Therefore, we approximate the distribution based on the spatial regions considered during the language grounding step for the language description.

$$I(a, z^a) = H(\gamma_f|\Lambda) - H(\gamma_f|\Lambda, a, z^a) \tag{8}$$

In this paper, we concentrate on questions that can result in a discrete set of answers. This allows us to better model the expected change in entropy given the answer to the question (unlike an open ended answer which could be drawn from a large space of possible answers).

Given the answer, we evaluate the change it has on the distribution over the particular grounding variable. For most spatial relations, we define a range over which a particular question can be applied in a meaningful manner. For example, we only consider regions within a 20 m distance when evaluating a question. As such, we limit the entropy calculation to the regions for which the question is expected to be meaningful.

$$p(\gamma_f = R_i|\Lambda, a, z^a) = \frac{p(z^a|a, R_i) \times p(\gamma_f = R_i|\Lambda)}{\sum_{R_i} p(z^a|a) \times p(\gamma_f = R_i|\Lambda)} \tag{9}$$

The expected value of the next state is based on the transition function from the current state to the next state

$$\mathbb{E}(V(S_{t+1})) = \sum_{z_j^a} \mathcal{F}(I(a|z_j^a)) \times p(z_j^a|S_t, a). \quad (10)$$

For the action $A_{\mathcal{F}}$, we assume that there is no change in the entropy as we are not modeling the expected change in the language groundings based on spatial exploration. Thus, the Q value for $A_{\mathcal{F}}$ is only the cost of the action.

3) Transition Likelihood: The transition function captures the likelihood of receiving each answer, given the state and the question asking action. We arrive at this value by marginalizing out the grounding variable. This results in a higher expected likelihood of receiving a particular answer if there were spatial regions that had a high likelihood of being the grounding and also fit the spatial relation in the question.

$$p(z_j^a|S_t, a) = \sum_{R_i} p(z_j^a|S_t, R_i, a) \times p(R_i|\Lambda) \quad (11)$$

4) Cost Function Definition: We define a hand-crafted cost function that encodes the desirability of asking a given question at each timestep. The cost of each question asking action is a function of several relevant features. For this implementation, we have used the following:

i Time since last question asked
ii Time since last question asked about grounding
iii Number of questions asked about entity

In our current implementation, we use a linear combination of these features to arrive at the cost function. The weights have been set empirically such that they result in negligible burden on the guide and do not impeded the conducting of the tour. Ideally, these weights would be learned from user preferences based upon human trials.

For the person following action $A_{\mathcal{F}}$, we assign a fixed cost such that only a reasonably high expected information gain will result in a question being asked. The value was set empirically to achieve a reasonable level of questions.

5) Integrating Answers to the Representation: We couple each of the answers with the original question to arrive at an equivalent natural language description. However, since the question was tied to a particular spatial entity, we treat the question and answer pair together with the original description, according to Equation 9. As such, each new answer modifies the distribution over that grounding variable, and any informative answer improves the map distribution.

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 3 shows the grounding likelihoods before and after asking three questions.

VI. Results

We evaluated our algorithm on an indoor dataset in which a human gives our wheelchair a narrated tour of MIT's Stata Center building. We injected three natural language descriptions at locations where the descriptions contained a level of ambiguity. We ran the algorithm on the dataset and a human provided answers to the robot's questions. We outline the resulting semantic map and compare it with a semantic map that did not integrate language, and one that integrated language but did not ask questions of the guide.

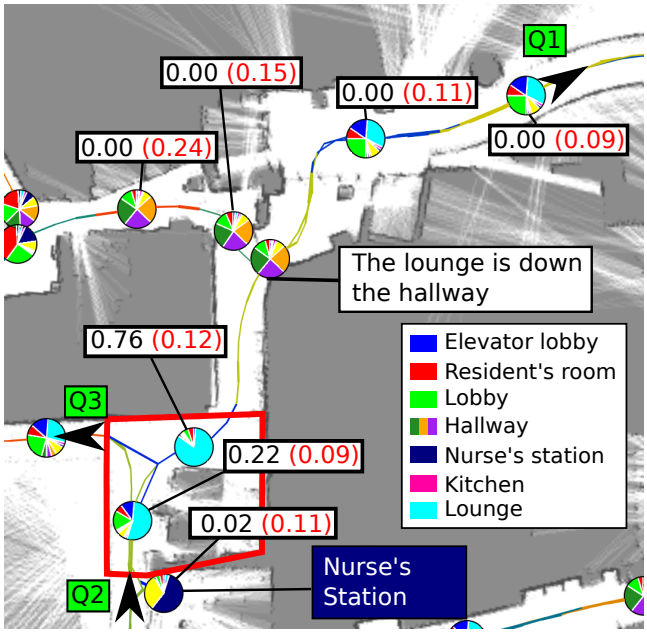

Fig. 3. Language groundings for the expression "The lounge is down the hall." Grounding likelihood with and without questions is shown in black and red, respectively. Questions asked (answers), Q1: "Is the lounge near the conference room?" ("Yes"); Q2: "Is the lounge on my right?" ("No"); Q3: "Is the lounge behind me?" ("Yes"). The ground truth region boundary is in red. Pie charts centered in each region denote its type, while path color denotes different regions.

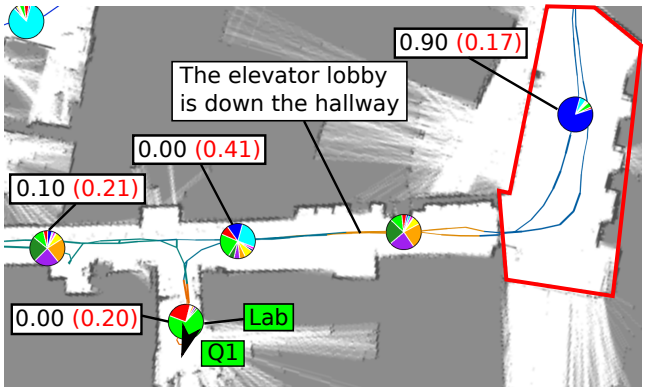

Fig. 4. Language groundings for the expression "The elevator lobby is down the hall." Grounding likelihood with questions is shown in black and without questions in red. Question asked (and answer), Q1: "Is the elevator lobby near me?" ("No"). The ground truth region is outlined in red.

Overall, the dataset contained six descriptions of the robot's location that the algorithm grounded to the current region, and three allocentric expressions that describe regions with relation to either landmarks in the environment (e.g., "the elevator lobby is down the hall"), or to the robot (e.g., "the lounge is behind you"). The robot asked a total of five questions of the guide, four of which were in relation to itself, and one in relation to a landmark in the environment.

TABLE I. ENTROPY OVER FIGURE GROUNDINGS WITH AND WITHOUT QUESTIONS

	Entropy		
Original Description	Without Questions	With Questions	No. of Questions
"The lounge is down the hallway" (Fig. 3)	2.015	0.620	3
"The elevator lobby is down the hallway" (Fig. 4)	1.320	0.325	1
"The lounge is behind you" (Fig. 5)	0.705	0.056	1

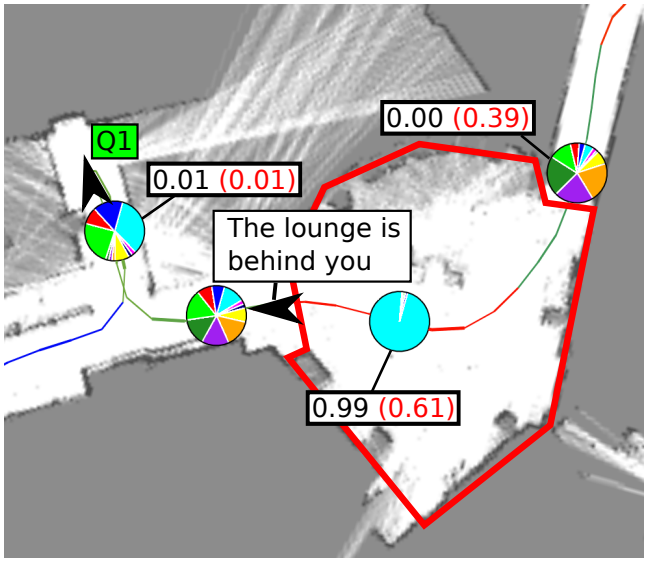

Fig. 5. Language groundings for the expression "The lounge is behind you." Grounding likelihood with questions is shown in black and without questions in red. Question asked (and answer), Q1: "Is the lounge near me?" ("Yes"). The ground truth region is outlined in red.

As can be seen in Table I, the semantic map that results from integrating the answers received from the guide exhibits much less uncertainty (and lower entropy) over the figure groundings than those for which no questions were asked. For all three descriptions, the robot was able to significantly reduce the entropy over the figure groundings by asking one to three questions each.

VII. CONCLUSION

We are working with The Boston Home to develop a voice-commandable wheelchair that allows the mobility impaired to independently move about their environment, without relying on the assistance of others. In order for the wheelchair to correctly follow spoken directions, it must reason over the spatial and semantic properties that users associate with their environment. This paper proposed a framework that allows users and caregivers to share this information with the robot in an efficient, intuitive manner, by leading the wheelchair on a tour and engaging it in natural language dialog, much like they would with a new colleague. With this semantic understanding, users can then command their wheelchair to navigate simply by speaking to it.

Our approach treats automatic speech recognition (ASR) and language understanding as separate processes, whereby we use a continuously-running probabilistic recognizer [9] to convert audio to the highest likelihood text and then infer the user's intention (i.e., desired destination) based on this text. Such a decoupled approach generally works well for people with unimpaired speech in acoustically clean environments. However, people with MS and other neurological conditions often exhibit speech pathologies, such as rapid fatigue, prolonged speaking style, or dysarthria, that are not captured by the acoustic or language models or in the data used for training. Hence, standard ASR systems often fail to recognize portions of the speech, resulting in erroneous text that the language understanding component will (incorrectly) ground. A better alternative would be to consider the top N (for some N) most likely outputs of the recognizer and the distribution over their parses when inferring their meaning. This would provide some robustness to failures in the ASR. The resulting distribution over groundings would provide a measure of confidence in the system's ability to infer the user's intent that can be used to decide whether to proceed or to ask clarifying questions to resolve ambiguity. Our collaborators have recently taken a similar approach by using dialog for speech-based interfaces at TBH [4, 5]. Further, the ASR should employ acoustic and language models that better represent speech pathologies and should be trained on users with similar speech patterns.

Our system allows users to command their wheelchair to a desired location by referring to it by its colloquial name, type, and/or relation to other regions in the environment. It would be useful if the user were able to convey other intentions, such as a desire to perform a certain activity. For example, the user may say "I want to watch television" or "I want something to eat." In the case of the former, the wheelchair would bring the user to the lounge, position them in front of the television and possibly turn it on. Following commands like these requires a richer model of the environment that includes objects, their type, their location and relationship to regions (i.e., that televisions are commonly found in lounges), and their utility. Natural language understanding also requires models that capture the relationship between a particular activity, the state of the world, and the actions of the robot. For example, this could be a pre- and post-condition model in which an activity is defined by a certain allocation of states (post-condition) and the robot's actions are a means of satisfying these post-conditions.

In the context of the current system, we hope to conduct extensive experiments with a number of different guides, and TBH users who will then attempt to use the learned maps to navigate around the facility. This is critical to understanding the extent to which our framework can model the different types of information that people typically associate with their environment. It is also necessary to assess the extent to which this information allows users to independently navigate within their residence. We anticipate that the latter will involve close, on-site collaboration with clinicians and the residents in order to understand the aforementioned limitations of the ASR and understanding pipeline. This would be tremendously valuable for developing a more effective speech-based interface.

REFERENCES

[1] J.-G. Park, B. Charrow, D. Curtis, J. Battat, E. Minkov, J. Hicks, S. Teller, and J. Ledlie, "Growing an organic indoor location system," in *Proc. Int'l Conf. on Mobile Systems, Applications, and Services (MobiSys)*, San Francisco, CA, June 2010, pp. 271–284.

[2] F. Doshi-Velez, W. Li, Y. Battat, B. Charow, D. Curtis, J. Park, S. Hemachandra, J. Velez, C. Walsh, D. Fredette, B. Reimer, N. Roy, and S. Teller, "Improving safety and operational efficiency in residential care settings with WiFi-based localization," *J. American Medical Directors Association*, vol. 13, no. 6, pp. 558–563, July 2012.

[3] F. Doshi and N. Roy, "Spoken language interaction with model uncertainty: An adaptive human-robot interaction system," *Connection Science*, vol. 20, no. 4, pp. 299–318, November 2008.

[4] W. Li, J. Glass, N. Roy, and S. Teller, "Probabilistic dialogue modeling for speech-enabled assistive technology," in *Proc. Work. on Speech and Language Processing for Assistive Technologies (SPLAT)*, Grenoble, France, August 2013.

[5] W. Li, D. Fredette, A. Burnham, B. Lamoureux, M. Serotkin, and S. Teller, "Making speech-based assistive technology work for a real user," in *Proc. Work. on Speech and Language Processing for Assistive Technologies (SPLAT)*, Grenoble, France, August 2013.

[6] H. Yanco, "Wheelesley: A robotic wheelchair system: Indoor navigation and user interface," *Assistive Technology and Artificial Intelligence*, vol. 1458, pp. 256–268, 1998.

[7] S. Parikh, V. Grassi, V. Kumar, and J. Okamoto, "Integrating human inputs with autonomous behaviors on an intelligent wheelchair platform," *IEEE Intelligent Systems*, vol. 22, no. 2, pp. 33–41, March–April 2007.

[8] D. Sinyukov, R. Desmond, M. Dickerman, J. Fleming, J. Schaufeld, and T. Padir, "Multi-modal control framework for a semi-autonomous wheelchair using modular sensor designs," *Intelligent Service Robotics*, vol. 7, no. 3, pp. 145–155, July 2014.

[9] J. R. Glass, "A probabilistic framework for segment-based speech recognition," *Computer Speech & Language*, vol. 17, no. 2–3, pp. 137–152, April–July 2003.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] M. Kaess, H. Johannsson, R. Roberts, V. Ila, J. J. Leonard, and F. Dellaert, "iSAM2: Incremental smoothing and mapping with fluid relinearization and incremental variable reordering," in *Proc. IEEE Int'l Conf. on Robotics and Automation (ICRA)*, 2011, pp. 3281–3288.