5-2013

Learning Environmental Knowledge from Task-Based Human-Robot Dialog

Thomas Kollar
Carnegie Mellon University

Vittorio Perera
University of Roma "La Sapienza"

Daniele Nardi
University of Roma "La Sapienza"

Manuela M. Veloso
Carnegie Mellon University

Published In
Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), 2013, 4304-4309.

Learning Environmental Knowledge from Task-Based Human-Robot Dialog

Thomas Kollar[1] Vittorio Perera[2] Daniele Nardi[2] Manuela Veloso[1]

***Abstract*— This paper presents an approach for learning environmental knowledge from task-based human-robot dialog. Previous approaches to dialog use domain knowledge to constrain the types of language people are likely to use. In contrast, by introducing a joint probabilistic model over speech, the resulting semantic parse and the mapping from each element of the parse to a physical entity in the building (e.g., grounding), our approach is flexible to the ways that untrained people interact with robots, is robust to speech to text errors and is able to learn referring expressions for physical locations in a map (e.g., to create a semantic map). Our approach has been evaluated by having untrained people interact with a service robot. Starting with an empty semantic map, our approach is able ask 50% fewer questions than a baseline approach, thereby enabling more effective and intuitive human robot dialog.**

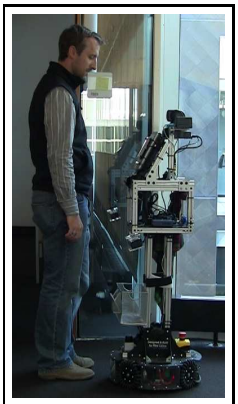
I. INTRODUCTION

As robots move out of the lab and into the real world, it is critical that humans will be able to specify complex task requirements in an intuitive and flexible way. In this paper, we enable untrained people to instruct robots via dialog, which is challenging for several reasons. First text to speech results are inherently noisy and are prone to errors due to out-of-domain speech input (e.g., "Kristina" vs. "Christina"). Second, human speech is highly-variable ("the kitchen" vs "the kitchenette"). Finally, the mapping from the natural language expression of locations and objects onto new environments is unknown. Instead of presupposing that the robot has access to a large repository of environmental knowledge to understand the commands received, our approach interactively learns by executing robot tasks. For example, when the robot is in a new environment and someone says, "Go to the kitchen." our approach is able to learn that "go to" refers to the GoTo task and that "the kitchen" refers to a set of locations in the environment. After learning, the robot is able to execute natural language commands, as in Figure (1).

We address these issues by developing a dialog system that is able to learn interactively from people. Our dialog system is aimed at capturing environmental knowledge from untrained users. There are three main components: a probabilistic model that connects speech to the locations in the physical environment, a dialog system which acquires knowledge that the robot does not know a priori, and a knowledge base which stores the acquired knowledge.

[1] Thomas Kollar (tkollar@cmu.edu) and Manuela Veloso (mmv@cs.cmu.edu) are with the Computer Science Department, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA, USA.

[2] Vittorio Perera (vittorio.perera@gmail.com) and Daniele Nardi (daniele.nardi@dis.uniroma1.it) are with Department of Computer, Control and Management Engineering Sapienza, University of Rome, Rome, Italy



(a)

Commands
- Go to the bridge.
- Go to the lab.
- Bring me to the elevator.
- Go to Christina's office.
- Take me to the meeting room.

(b)

Fig. 1: In (a) is CoBot, a mobile service robot and the test platform used in our experiments. In (b) are examples of commands that our approach is able to successfully follow.

The probabilistic model parses the speech to text candidates to an intermediate meaning representation (parsing) and maps elements in the intermediate representation to aspects of the physical environment (grounding). The parse consists of linguistic constituents, including actions, people, and locations. The grounding consists of locations in the physical environment (e.g., places where people want the robot to go). When the mapping to the physical environment is unknown, the dialog system initiates an interaction with the person to acquire the knowledge necessary to achieve the task. The results of the interaction are stored in the knowledge base that is used for future interactions with people. When the mapping to a physical location is known, then the robot executes the corresponding action.

To evaluate our approach, a dialog system was developed for a mobile service robot [1], [2], enabling it to be commanded to move anywhere in three floors of a large office building. The aim of these experiments was two-fold. First, we show that as people give commands, the robot is able to learn the environmental knowledge necessary to interact with untrained people. As a result, the robot needs to ask fewer questions to understand new commands correctly. Over both actions and locations, our approach reduces the required questions in half, compared with a baseline approach which asks about the action and the destination. Second, we have demonstrated that the robot is able to learn meaningful referring expressions for the locations where it was sent. We have found that, although the system will sometimes infer the wrong location (e.g., there are multiple "Tom's offices"

in the environment), a person is able to correct the behavior of the robot, resulting in a system that is able to consistently understand task-constrained speech as given from untrained users.

II. Related Work

This paper builds off of recent work in semantic mapping and grounded language acquisition. Typically, semantic maps (maps with the location and type of objects in them) are created without interacting with people. Visual and laser features are used to classify the type of location or the scene type using classification algorithms, such as AdaBoost [3]–[6]. Smoothing may be applied (e.g., a Hidden Markov Model / Voronoi Random Fields) to make the final classification more robust [3], [6]. Visual semantic mapping is often cast as either visual place categorization or object recognition [7]. Sometimes the aim is to provide anchoring of high-level semantic elements that are grounded in perception [4]. Unlike these approaches, we focus on learning a semantic map directly from human-robot dialog.

When humans are involved in semantic mapping the aim is to build a *shared* representation between the robot and human via human-augmented mapping (HAM). This is sometimes done using a tangible user interface in conjunction with speech [8] or just with speech alone [9]. Our approach moves beyond this work by taking into account the uncertainty of the speech, parsing and grounding jointly, being robust to errors in speech recognition and words and phrases that are previously unseen.

Beginning with SHRDLU [10], many systems have exploited the compositional structure of language to statically generate a plan corresponding to a natural language command [11]–[16]. Although our work is more narrow in the scope of semantic parses presented in this paper (e.g., a single task), our approach is less restrictive since any untrained person can interact with the robot and teach it about actions and locations, enabling the robot interpret arbitrary language about the task. Our work focuses on task-based dialog and specifically on the GoTo task, complementing other work that has focused on dialog specific to a task [17]–[24].

III. Approach

In this section, we present an approach to human-robot dialog in the presence of speech to text errors and the high-variability that comes from untrained human users. Figure (2a) shows an example of a dialog between the robot and a user that our system is able to understand. Figure (2b) shows the information that the robot has extracted and stored. Even with just a task that has the robot go from place to place in the environment (the *GoTo* task), our approach is able to learn referring expressions for actions and locations.

The main components of our approach are 1) a probabilistic model that is able to interpret natural language commands for service tasks, 2) a knowledge base that represents the mapping from referring expressions to locations and actions 3) a dialog system that is able to add facts to the knowledge base after interacting with people. Each of the following sections describes these components of our approach.

A. Probabilistic Model

We model the problem of understanding natural language as inference in joint probabilistic model over the groundings Γ, a parse P and speech S, given access to a (potentially empty) knowledge base KB:

$$\arg\max_{\Gamma,P,S} p(\Gamma, P, S|\text{KB}) \tag{1}$$

The probabilistic model is composed of three components: 1) a speech to text model that provides multiple speech to text candidates 2) a parsing model that that is able to extract the high-level structure of the command and 3) a grounding model that maps the referring expressions from the components of a parse (e.g., action, location, person) to actions that the robot can execute and places where the robot can go. Formally, this becomes:

$$p(\Gamma, P, S|\text{KB}) = p(\Gamma|P, \text{KB}) \times p(P|S) \times p(S) \tag{2}$$

We describe the parsing model, the knowledge base and the grounding model the following sections. The speech model is obtained from a recognizer able capable of understanding free-form speech[1].

1) Parsing Model: Given natural language text from the speech recognizer, the system parses it into a semantic representation (frames), which consist of an action and a variable number of arguments. The arguments can be among three different types, including actions, people and locations. Actions include the referring expressions for actions that the robot can execute such as "go to," "bring me", and "let's go". Locations include referring expressions like "Rashid Auditorium" or "the lab", people include "Tom" or "Joydeep", times include "now" or "at 2PM" and messages include "hi" or "I'm running a few minutes late" (e.g., from "Tell Tom that I'm running a few minutes late."). An example of a parse can be seen in Figure 3b.

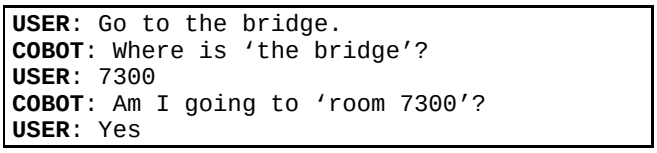
If $l_i \in \{\text{action}, \text{location}, \text{person}\}$ is the label of the ith word in the natural language command and there are N words s_i in the command, then the parsing model is represented as a linear function of weights w and features ϕ:

$$p(P|S) \triangleq p(l_1 \dots l_N | s_1 \dots s_K) \tag{3}$$

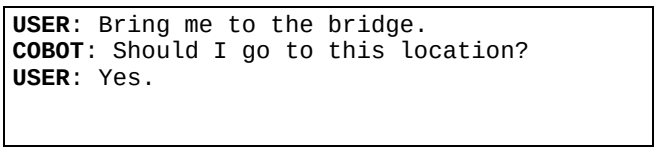
$$= \frac{1}{Z} \exp\left(\sum_i^N w \cdot \phi(l_i, s_{i-1}, s_i, s_{i+1})\right) \tag{4}$$

To extract a frame, such as $f = \{a = \text{"go to"}, e_1 = \text{"the kitchen"}\}$ (where a is the action and e_1 is the first argument to that action), the system greedily groups the labels together for "go" and "to" into an action. The same happens for "the" and "kitchen", which are both labeled as a location and are passed to the frame as an argument. The model is learned as a conditional random field (CRF); we use gradient descent (LBFGS) to optimize

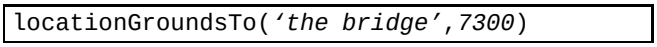
[1]The Google speech recognizer was used for our experiments.



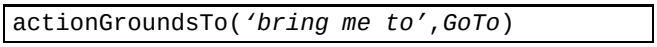
(a) Unknown Location



(b) Unknown action



(c) Location Predicate



(d) Action Predicate

Fig. 2: Sample dialogs when the robot has not seen the action or location referring expressions. In (a/b) are turns in the dialog. In (c) is the fact that is added from the dialog in (a). In (d) is the fact added from the dialog from (b). In both cases the robot would execute the corresponding action after the dialog completes.

the parameters w[2]. The features ϕ are binary features and include the part of speech tags for the current, next and previous word as well as the current, next and previous word in the natural language command.

2) Grounding Model: Given a parse of the natural language command and a knowledge base, the grounding model produces a distribution over the groundings of the referring expressions for location, person or action in the command. This grounding happens by using the knowledge base to provide correct coupling between natural language expressions and task the robot should execute or the location the robot should go to. To learn this model, we rewrite it using Bayes rule:

$$p(\Gamma|P;\text{KB}) = \frac{p(P|\Gamma;\text{KB}) \times p(\Gamma;\text{KB})}{\sum_P p(P|\Gamma;\text{KB}) \times p(\Gamma;\text{KB})} \quad (5)$$

The prior over groundings $p(\Gamma;\text{KB})$ is computed by looking at the counts of each element of Γ in the knowledge base (the category predicates). The other term $p(P|\Gamma;\text{KB})$ is computed by grounding referring expressions for actions, locations, people and objects (the relation predicates). Since our mobile robot can only move to places, we approximated the grounding of people and objects as being at a particular location in the environment. We make the assumption that the natural language command is well-approximated by a sequence of frames. Thus, if f_i is a frame, then the probability of a parse P given the groundings Γ can be written as:

$$p(P|\Gamma;\text{KB}) = \prod_i p(f_i|\Gamma;\text{KB}) \quad (6)$$

Each frame f_i consists of an action a and its arguments e, which are grounded separately, such that:

$$p(f|\Gamma) = p(a|\Gamma,\text{KB}) \times p(e|\Gamma;\text{KB}) \quad (7)$$

To compute the first term, the model assumes access to a knowledge base (Section III-B) that contains predicates and frequency counts. Assuming access to a *groundsTo* predicate (does the first argument ground to the second argument) and a corresponding $count$, then for a grounding $\gamma \in \Gamma$ the first term can be computed as:

$$p(a|\gamma;\text{KB}) = \frac{groundsTo(a,\gamma).count}{\sum_i groundsTo(a_i,\gamma).count} \quad (8)$$

In Equation 8, a is a multinomial random variable that ranges over the possible referring expressions in the knowledge base and γ ranges over the possible groundings in the environment.

To compute the second term in Equation (7), if $e_{i,j}$ is a multinomial random variable over the ith element of the parse and referring expression j and further that $k(i)$ is the realized referring expression for the ith element in the semantic parse, then we can write the probability of the arguments e given a grounding γ as:

$$p(e|\gamma;\text{KB}) = \frac{\sum_i groundsTo(e_{i,k(i)},\gamma).count}{\sum_{i,j} groundsTo(e_{i,j},\gamma).count} \quad (9)$$

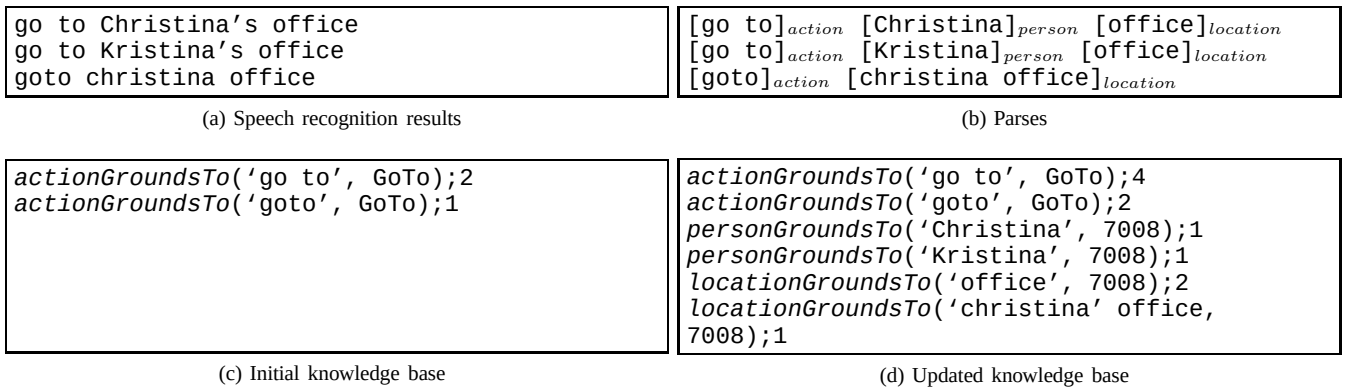
B. Knowledge Base

To maintain and re-use knowledge that the robot acquires as a part of the dialogs that has with humans, we define a knowledge base that consists of categories and relations. Categories are single argument predicates, which include $action(X)$, $location(X)$ and $person(X)$, corresponding to the labels that are extracted by the parser. Argument types have corresponding relations that determine when a referring expression (argument) X corresponds to a grounding Y: $groundsTo(X,Y)$. For person, location and action referring expressions, there are the corresponding $personGroundsTo$, $locationGroundsTo$ and $actionGroundsTo$ predicates.

To each relation instance in the knowledge base, a number is attached to keep track of how many times the specific arguments of the relation have been correctly grounded together; in the rest of the paper we will refer to this number using a dotted notation such as $locationGroundsTo(X,Y).count$ or simply as *count*. Some examples in our knowledge base include $person(\text{`}Tom\text{'})$, $location(\text{`}kitchen\text{'})$ and $locationGroundsTo(\text{`}kitchen\text{'},7602)$.

There are multiple ways for facts to get added to the knowledge base. First, a fact may be added when a user explicitly confirms the name of an action or an argument

[2]We used CRF++ to perform this optimization



(a) Speech recognition results

(b) Parses

(c) Initial knowledge base

(d) Updated knowledge base

Fig. 3: (a) Top three results of the speech recognizer. (b) Parses for each of the top three speech recognition results. For the first element, "go to" is parsed as an action, "Christina" is parsed as a person and "office" is parsed as a location (c) The initial knowledge base for state A, which contains two facts; "go to" and "goto" refer to the GoTo action. (d) The updated knowledge base for state A. "Christina" and "Kristina" are added as candidate people and "christina office" is a candidate location for room 7008.

(e.g., location, object). Second, the knowledge base may be updated when a user confirms that a task should be executed in response to a natural language command.

In either case, since the action is confirmed, the knowledge base is updated by adding new category and relation predicates or updating the counts of ones already presents. For each of the parsed actions and arguments, a corresponding category predicate is added to the knowledge base (e.g., if "go to" is parsed as an action, then *action('go to')* is added). For each of the parsed actions or predicates, the corresponding *groundsTo(X, Y)* relations are either added to the knowledge base and initialized to a count of 1 or, if already present, their counts are incremented by one. An example of this can be seen in Figure (3).

C. Human-robot dialog for task execution

In order to execute tasks, the robot performs dialog with people to fill in unknown components of the plan, as in Figure (2). Given a natural language command, which is parsed into a sequence of frames, the dialog with humans will proceed by filling in gaps in the knowledge of the robot. If any part of the frame is missing the robot will ask a question. If there is no action, then it will ask the person to say the action and ground it to an action that the robot can execute. If there is no argument, then the robot will ask for an argument according to the action template defining the command (e.g., for the "go to" action, it will ask for the location argument). In the case where the frame template is filled, but there is no grounding for either the action or the arguments, then the system will ask for the grounding of these components. For the action field, the robot will ask for the grounding to one of the actions (e.g., "go to" or "bring object"). For the location field, the robot will ask for the grounding to a room number in the building. At the end of the dialog, for safety reasons, the robot always asks for confirmation before executing an action. At this stage the robot will execute the action corresponding to the command.

In order to give a better insight on how the algorithm modifies the knowledge base after each interaction we describe a simple but meaningful example. We assume that only one parse is available for each speech interpretation and we will focus on the grounding relations leaving the categories aside. In this example the user gives the following command: *"go to Christina's office"*; Figure (3a) shows the results of the speech recognizer while Figure (3b) shows the parse of each of them.

The initial knowledge of the robot, collected from previous interactions, is shown in Figure (3c). The algorithm queries the knowledge base for possible groundings of actions and parameters of the three transcriptions returned by the speech recognizer. The query returns the same results for the action, but nothing for the parameters; therefore, the robot asks the user to spell the room number of its destination. The user spells "7008" and, after asking for confirmation, the algorithm updates the knowledge base to:

- *actionGroundsTo*('go to', GoTo); 4
- *actionGroundsTo*('goto', GoTo); 2

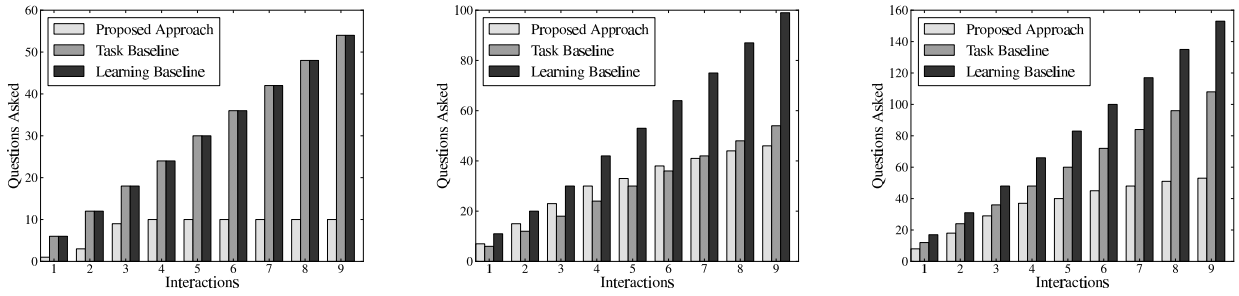
Second, the following relations are added to the knowledge base:

- *locationGroundsTo*('office', 7008); 2
- *locationGroundsTo*('christina office', 7008); 1
- *personGroundsTo*('Cristina', 7008); 1
- *personGroundsTo*('Kristina', 7008); 1

This example illustrates an important aspect of the algorithm. Once a grounding is retrieved, all high probability speech interpretations are added to the knowledge base. Doing this allows us to generalize over different plausible speech results. In this way we also allows other reasonable groundings into the knowledge base, such as *locationGroundsTo*('christina office', 7008). This is done in order to cope with the uncertainty in the speech recognizer, which might provide multiple reasonable interpretations.

IV. Results

Our approach is evaluated in two ways. First, we show that the robot can learn the meaning of natural language



(a) Action-related question (b) Location-related question (c) Total number of question

Fig. 4: Comparison between our approach and the two baseline proposed (using the cumulative number of question grouped by user). The graphs are cumulative (e.g., the knowledge base resulting from subject 1 is used to interact with subject 2) (a) shows the number of questions required to understand the action parameter across subjects and over time. (b) shows the number of questions required to understand the location parameter across subjects over time. (c) shows the number of questions required to understand all parameters and execute the task.

commands from dialog. Second, we show that using our algorithm the robot learns a reasonable referring expression across multiple groundings, by aggregating results in the knowledge base.

A. Learning from dialog

To evaluate our approach, we asked 9 different people to give a mobile service robot a command to go to destinations in a real-world environment. The robot had the capability of going anywhere across three floors of an office building [1], [2]. Although the task was fixed (e.g., going to a destination), people could use whatever language was natural to them. The subjects ranged between an age of 21 and 54 and were both native and non-native English speakers, which made the task more challenging. We provided each person with the same map of the seventh floor of our building. Six locations were marked on the map and we asked the people to give the robot commands to go to the marked destinations. Since the people had different degrees of familiarity with the building, the map was also annotated with room numbers. The aim was to test the ability of our algorithm to learn the referring expressions for different groundings through dialog, therefore the initial knowledge base was empty. After each person interacted with the robot, the knowledge was aggregated and used as starting point for the following participants.

We compared our algorithm with two different baseline. The first baseline, called the *Task Baseline*, enables the robot to execute the task without learning any semantic information about the environment. Although less natural than the proposed approach since the person must explicitly define the room number and action, only two questions are required before the robot can execute the task. The second baseline proposed, called *Learning Baseline*, tries to execute the assigned task while learning semantic knowledge about the environment. However, this baseline does not use this knowledge about the environment for the dialog. In this case, people can use whatever language they like for the locations, but the robot will always ask three questions.

Figure (4) shows the results of this experiment. On the horizontal axis there are the nine people who interacted with the robot and on the vertical axis are the cumulative number of questions asked over all sessions. For example, session 9 includes the knowledge base acquired from sessions 1-8, and the vertical axis corresponds to all of the questions asked during those previous sessions as well as the current session. Figure (4a) shows results for the action parameter. Specifically, we have shown that the number of questions asked for actions stops increasing after the first few interactions. This happens because there is limited diversity in the ways that a person can command the robot to perform a task. Out of 54 instructions, only three different verbs were used to command the robot to go to a place (*go to*, *bring me*, *take me*). Figure (4a) additionally shows how our approach performs better then both baselines since, on average, after a few examples, the robot will stop asking the person about whether it should execute the GoTo task.

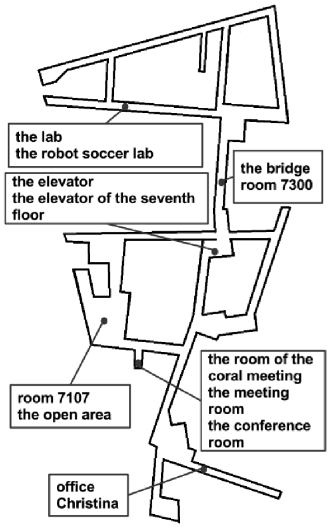


Fig. 5: The semantic map after interacting with all nine subjects. Plotted on the map are the most frequently occurring referring expressions for each location.

Figure (4b) shows how frequently the robot had to

ask about the location or person argument of the parse. Specifically, the vertical axis corresponds to the number of questions required to retrieve the correct grounding for referring expressions of locations and persons. The number of questions asked about this argument is greater because people refer to the same location in many different ways and therefore the algorithm needs more examples to learn the correct grounding. In this case, in the worst case, our approach must ask two questions of a person, whereas the *Task Baseline* must ask only one (note, however, that the task baseline is less intuitive than our approach). Nevertheless, as the number of the interactions increases, the algorithm learns how people address different places and the number of questions needed decreases. When the seventh person has interacted with the robot our approach started to outperform both of the baselines. Figure (4c) shows the aggregation of the grounding of all action and argument parameters, which shows that the overall system always performs better than both baselines.

B. Learning referring expressions

We also wanted to evaluate how well our system learned referring expressions for people and locations across multiple people who were not primed to speak in particular way. In order to perform this experiment, we evaluated the referring expressions (and their corresponding grounding) from the previous experiment. Looking at the most common referring expressions, we found that for five out of the six locations, the robot had learned a suitable expression such as *"the soccer lab"* or *"conference room"*, while for the last one the two most common referring expression are *"Christina"* and *"Office"*. These two labels come from the expression *"Christina's Office"* and were correctly understood by the parser in order to represent the fact that the room is an office and that we are likely to find Christina in it. We have also plotted the resulting most frequent referring expressions on a semantic map in Figure (5). Using these referring expressions, Figure 1 shows commands that our CoBot robot is able to successfully follow.

V. Conclusions

In this paper, we have presented a dialog system which is able to learn environmental knowledge from task-based human-robot dialog. We have defined a joint probabilistic model that consists of a speech model, a parsing model and a grounding model. Further, we have shown how this model can be used as a part of a dialog system to learn the correct interpretations of referring expressions involving actions, locations and people by adding new facts to its knowledge-base. The experiments show that our approach is a more effective interface and is able to reduce the number of questions asked by the robot by 50% compared to a baseline approach.

Acknowledgments

This research was supported by the National Science Foundation award number NSF IIS-1012733 and NSF IIS-1218932. The views and conclusions contained in this document are those of the authors only. We would also like to acknowledge Robin Soetens for his contributions to later versions of this system.

References

[1] S. Rosenthal, J. Biswas, and M. Veloso, "An Effective Mobile Robot Through Symbiotic Human-Robot Interaction," in *AAMAS*, 2010.

[2] J. Biswas, B. Coltin, and M. Veloso, "Corrective Gradient Refinement for Mobile Robot Localization," in *IROS*, 2011.

[3] A. Rottmann, O. Martínez Mozos, C. Stachniss, and W. Burgard, "Place classification of indoor environments with mobile robots using boosting," in *AAAI*, 2005.

[4] C. Galindo, A. Saffiotti, S. Coradeschi, and P. Buschka, "Multi-hierarchical semantic maps for mobile robotics," in *IROS*, 2005.

[5] E. Brunskill, T. Kollar, and N. Roy, "Topological mapping using spectral clustering and classification," in *IROS*, 2007.

[6] S. Friedman, H. Pasula, and D. Fox, "Voronoi random fields: extracting the topological structure of indoor environments via place labeling," in *IJCAI*, 2007.

[7] J. Wu, H. I. Christensen, and J. M. Rehg, "Visual place categorization: Problem, dataset, and algorithm." in *IROS*, 2009.

[8] G. Randelli, T. M. Bonanni, L. Iocchi, and D. Nardi, "Knowledge acquisition through human-robot multimodal interaction," *Intelligent Service Robotics*, pp. 1–13, 2012.

[9] G. Kruijff and H. Zender, "Clarification dialogues in human-augmented mapping," in *Proceedings of HRI*, 2006, pp. 282–289.

[10] T. Winograd, "Procedures as a representation for data in a program for understanding natural language." Ph.D. dissertation, MIT, 1970.

[11] K. Hsiao, S. Tellex, S. Vosoughi, R. Kubat, and D. Roy, "Object schemas for grounding language in a responsive robot," *Connection Science*, vol. 20, no. 4, pp. 253–276, 2008.

[12] M. MacMahon, B. Stankiewicz, and B. Kuipers, "Walk the talk: language, knowledge, and action in route instructions," in *AAAI*, 2006.

[13] 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, May 2004.

[14] 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 *ICRA*, 2009.

[15] C. Matuszek, D. Fox, and K. Koscher, "Following directions using statistical machine translation," in *HRI*, 2010.

[16] X. Chen, J. Ji, J. Jiang, G. Jin, F. Wang, and J. Xie, "Developing high-level cognitive functions for service robots," in *AAMAS*, 2010.

[17] R. Stiefelhagen, H. K. Ekenel, C. Fugen, P. Gieselmann, H. Holzapfel, F. Kraft, K. Nickel, M. Voit, and A. Waibel, "Enabling multimodal human-robot interaction for the karlsruhe humanoid robot," *IEEE Transactions on Robotics*, vol. 23, no. 5, pp. 840–851, 2007.

[18] H. Holzapfel, D. Neubig, and A. Waibel, "A dialogue approach to learning object descriptions and semantic categories," *Robotics and Autonomous Systems*, vol. 56, no. 11, pp. 1004–1013, 2008.

[19] S. Lemaignan, R. Ros, E. A. Sisbot, R. Alami, and M. Beetz, "Grounding the interaction: Anchoring situated discourse in everyday human-robot interaction." *International Journal of Social Robotics*, vol. 4, no. 2, pp. 181–199, 2012.

[20] M. Scheutz, R. Cantrell, and P. Schemerhorn, "Toward humanlike task-based dialogue processing for human robot interaction," *AI Magazine*, vol. 34, no. 4, pp. 64–76, 2011.

[21] G.-J. M. Kruijff, P. Lison, T. Benjamin, H. Jacobsson, H. Zender, I. Kruijff Korbayová, and N. Hawes, *Situated Dialogue Processing for Human-Robot Interaction*, ser. Cognitive Systems Monographs. Springer Berlin Heidelberg, April 2010, vol. 8, pp. 311–364.

[22] D. Bohus and E. Horvitz, "Facilitating multiparty dialog with gaze, gesture, and speech," in *ICMI*, 2010.

[23] T. Kollar, S. Tellex, D. Roy, and N. Roy, "Toward understanding natural language directions," in *HRI*, 2010.

[24] S. Tellex, T. Kollar, S. Dickerson, M. Walter, A. Banerjee, S. Teller, and N. Roy, "Understanding natural language commands for robotic navigation and mobile manipulation," in *AAAI*, 2011.