Designing for Long Horizon Social Interactions

Ifrah Idrees
Department of Computer Science
Brown University United States
ifrah_idrees@brown.edu

Abstract: Conversational assistive robots have the potential to guide humans to accomplish various daily tasks such as cooking meals, performing exercises, or operating machines. However, to interact effectively over long-horizon tasks, the robot must be able to infer the human's goal from their interactions with the surrounding environment. A few key challenges in inferring the world and user state are that the sensors are noisy, and the robot has partial observability of the environment and the human's intention. These problems compound as the robot collects more noisy observations about the user and environment over long periods of time. Our proposed work aims to bridge the gap in enabling the robot to perform accurate online inference of the world and user state in a partially observable environment, especially over extended periods.

Keywords: CoRL, Long-Horizon Planning, Long-Horizon State Estimation, Human-Robot Interaction, Assistive Robotics

1 Introduction

A conversational social robot can help humans perform daily activities, such as cooking dinner more independently and safely. Imagine a scenario where the human mistakenly performs sub-optimal actions while cooking dinner, e.g., forgetting to turn off the stove. In such a situation, an assistive robot must be able to suggest corrective next steps based on its understanding of the world and the user [1, 2]. An assistive robot guide to turning off the stove will be effective in the given case. To interact effectively, the robot must infer the human's goal and current step in the activity given the observations of the state of the appliances, e.g. (the microwave is off, the stove is on) and the state of the attributes of objects (e.g., dishes are dirty). Such an assistive robot will benefit people with dementia or cognitive impairment. It can also be helpful to an operator trying to build a machine, a child with autism trying to do their homework, or a child learning to do a chore.

The problem of inferring the world and user state is challenging because the sensors are noisy, and the robot has partial observability of the environment and the human's intention. The uncertainty in state estimation compounds as the robot collects more noisy observations about the user and environment over long periods of time [3]. Existing approaches do model the user and the world state but do not handle uncertainty from both the world sensors' observation and human language for state estimation over extended periods. Our proposed work aims to introduce methods to manage these sources of uncertainty. Our work aims to enable the robot to perform accurate and time-efficient online inference of the world and user state in a partially observable environment, especially over extended periods.

2 Handling Different Sources of Uncertainty

Long-term social interactions require an estimation of the world and user state given a sequence of observations of human's interaction with the environment. In this section, I will describe the different sources of uncertainty for estimating the world state, the user state and the task specification

6th Conference on Robot Learning (CoRL 2022), Auckland, New Zealand.

for long-horizon assistive robotics. I will also describe briefly our proposed methods for handling uncertainty in long-horizon tasks.

2.1 Uncertainty in the World State

Home-service robots have a great potential to assist human users by retrieving spatial-temporal [1] information about the objects in the environment from their long-term observations. Imagine a home robot monitoring the environment over long periods of time. Such a robot will have a massive amount of observations of the objects in the environment. The human user can then ask the home robot to assist them in finding objects in the environment by asking a simple query such as *"What are the favorite places in the house where I can place my keys?"* To answer such a simple query, the service robots must have a situational understanding of the environment over extended periods, not just for days but weeks and months. This requires the robot to estimate the state of objects in the environment over an extended time in a partially observable environment. Service robots with such an ability will be well-suited to help the elderly, especially those with dementia.

The previous approaches for long-term object state estimation and retrieval either assume **1.** static cameras [4, 5, 6], **2.** full observability [6], or **3.** short-time horizon for object search [7, 8]. For long-term object retrieval, the above assumptions leave a partially-observable robot searching over countless detections of objects in visual sensor data from many different time slices. These detections will contain partial views [2] of different object instances. Further, storing all these partial-view detections will take up a lot of memory space and time. Existing approaches will search all of these partial views of objects even when they are not irrelevant to the query. What we propose is a **D**etection-based **3**-level hierarchical **A**ssociation approach, **D3A** that allows for a compact and query-able spatial-temporal state estimation representation [9, 10]. Our algorithm performs online incremental and hierarchical learning to identify keyframes that best represent the unique objects in the environment. This spatial-temporal representation of objects in the environment also enables the answering of queries for object retrieval from long-term observations. D3A demonstrates high accuracy and time efficiency when queried. For a set of 150 queries, D3A returns a small set of candidate keyframes (which occupy only 0.17% of the total sensory data) with 81.98% mean accuracy for answering object retrieval queries in 11.7 ms. Our algorithm is powerful enough that it does not need to know the object detection algorithm and the natural language query processing algorithm during data collection and can still build a queryable spatial-temporal representation of objects over long periods of time.

2.2 Uncertainty in the User State

Next, for effective long-term social interactions, we must handle uncertainty in both the user and the world state over long-horizon tasks. We aim to establish a method to enable assistive robots to infer the goal the human tries to achieve based on their interaction with the environment. Even in an environment where the robot has maximal sensor information, like a smart home, the robot still needs to figure out what it is human doing. To be an effective assistant, the robot must interpret the goal the human is trying to achieve and the action the human should take to complete their plan. Previously, human progress during hierarchical tasks has been modeled using hierarchical task networks (HTNs) [2, 11]. However, these plan/goal recognition techniques do not allow the agent to leverage its ability to use language to reduce uncertainty by asking questions. Further, interpreting language input from the human is challenging because of the vast space of observations — language utterances spoken by the human. The existing solution to this problem is heuristics [12, 13, 14], which are prone to fail as the tasks get complex and the environment sensors become complex or noisy.

[1] Spatial-temporal information of object refers to the whereabouts of the object such as where the object has been identified in the physical environment of the robot and at what times

[2] Partial views of objects refers to the partially occluded objects seen from different viewpoints by the robot

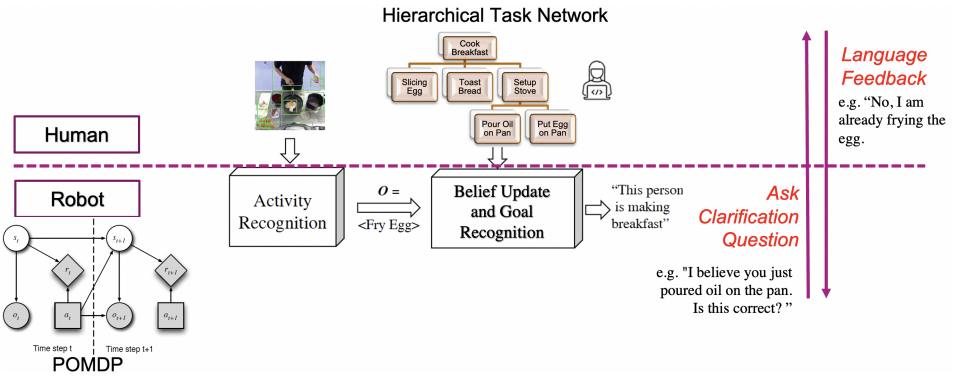

Figure 1: Architecture of **D**ialogue **G**oal **R**ecognition POMDP (**DGR**-POMDP)

We propose to solve this problem by combining the Hierarchical Task Networks (HTNs) with Partially Observable Markov Decision Process (POMDP) in our **D**ialogue **G**oal **R**ecognition POMDP (**DGR**-POMDP). This is particularly challenging because the POMDP does not assume a hierarchical human mental model or task specification structure [15]. Further, the state and observation spaces are large for modeling users and the world. For our work, we assume the person is a planner with hierarchically-described goals and subgoals. The agent is represented by a POMDP model updating its belief in human progress by asking clarification questions about noisy sensor data. An illustration of our approach's architecture and a sample example of a clarification question is shown in Fig-1.

We evaluate the performance of DGR-POMDP over various cooking tasks in a simulated environment and perform a robot demonstration. We show that by incorporating language feedback from the human along with the world state information in a hierarchical task model, our decision-theoretic dialogue framework achieves 86% accuracy in goal recognition and step recognition by only asking 44% more questions than a baseline that also models human as HTN but does not incorporate language feedback. Our oracle that always asks the right question only outperforms our policy by 1%.

2.3 Uncertainty in the Task Specifications

Next, we deal with the uncertainty in the task specifications. For human goal and plan recognition, our previous work assumes that a handcrafted hierarchical task specification exists for the tasks humans aim to achieve. However, hand-coding an HTN for complex tasks is difficult, as it requires significant domain knowledge and engineering effort [16]. Hence, lastly, I propose a learning algorithm that learns task specifications from natural language and demonstrations to enable generalized goal inference. Large Language Models have previously been shown to successfully generate stories (Rashkin et al., 2020), summaries (Lewis et al., 2020), and commonsense facts (Bosselut et al., 2019; Hwang et al., 2020). Here we investigate their application to Hierarchical Task Networks. We plan to use common sense knowledge in large language models and demonstrations of human completing their tasks to extract hierarchical temporal task representations of various goals. We propose to generate the action-graph structure of the available actions and objects in an environment by prompting LLM for subtasks of new goals.

3 Conclusion

Our proposed line of work aims to test the hypothesis that using hierarchical spatial-temporal structures to model the world and user state in long-term social interactions can enable the robot to have a more accurate & time-efficient online state inference. Our proposed algorithms aim to improve the robot's user experience over extended periods by performing better state estimation for long-term social interactions. In our work, we describe three sources of uncertainty - world state, user state,

and task specifications. Then we make the following contributions. First, we bring forward a novel algorithm that performs efficient world state estimation over long periods while handling uncertainty due to noisy sensors and partial observability. Next, we present a novel formulation for accurate inference of the latent variable of the user's goals and plans from noisy world sensors and language feedback. Finally, we provide an outline of learning hierarchical task specifications from demonstrations and natural language using common sense knowledge in large language models. Our proposed work aims to enable generalized goal inference of human progress during long-horizon task completion. Our introduced methods help manage uncertainty in world, user, and task specifications. A home robot with enhanced capabilities of state estimation can assist humans during task completion, improving social interactions over long periods.

References

[1] K. Erol, J. Hendler, and D. S. Nau. Htn planning: Complexity and expressivity. In *AAAI*, volume 94, pages 1123–1128, 1994.

[2] D. Wang and J. Hoey. Hierarchical task recognition and planning in smart homes with partial observability. In *International Conference on Ubiquitous Computing and Ambient Intelligence*, pages 439–452. Springer, 2017.

[3] A. Adu-Bredu, N. Devraj, P.-H. Lin, Z. Zeng, and O. C. Jenkins. Probabilistic inference in planning for partially observable long horizon problems. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3154–3161. IEEE, 2021.

[4] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *2007 IEEE CVPR*, pages 1–8, June 2007. doi: 10.1109/CVPR.2007.383172.

[5] P. Yadav and E. Curry. Vidcep: Complex event processing framework to detect spatiotemporal patterns in video streams. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 2513–2522. IEEE, 2019.

[6] D. Kang, P. Bailis, and M. Zaharia. Challenges and opportunities in dnn-based video analytics: A demonstration of the blazeit video query engine. In *CIDR*, 2019.

[7] R. Ambruş, N. Bore, J. Folkesson, and P. Jensfelt. Meta-rooms: Building and maintaining long term spatial models in a dynamic world. In *2014 IEEE/RSJ IROS*, pages 1854–1861. IEEE, 2014.

[8] N. Bore, P. Jensfelt, and J. Folkesson. Retrieval of arbitrary 3d objects from robot observations. In *2015 European Conference on Mobile Robots (ECMR)*, pages 1–8. IEEE, 2015.

[9] I. Idrees, Z. Hasan, S. P. Reiss, and S. Tellex. Where were my keys? - aggregating spatial-temporal instances of objects for efficient retrieval over long periods of time. *CoRR*, abs/2110.13061, 2021. URL `https://arxiv.org/abs/2110.13061`.

[10] I. Idrees, S. P. Reiss, and S. Tellex. Robomem: Giving long term memory to robots. *CoRR*, abs/2003.10553, 2020. URL `https://arxiv.org/abs/2003.10553`.

[11] D. Höller, G. Behnke, P. Bercher, and S. Biundo. Plan and goal recognition as htn planning. In *2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 466–473. IEEE, 2018.

[12] M. A. Razzaq, W. A. Khan, and S. Lee. Intent-context fusioning in healthcare dialogue-based systems using jdl model. In *International Conference on Smart Homes and Health Telematics*, pages 61–72. Springer, 2017. doi:10.1007/978-3-319-66188-9_6.

[13] J. Fasola and M. J. Matarić. A socially assistive robot exercise coach for the elderly. *Journal of Human-Robot Interaction*, 2(2):3–32, 2013.

[14] C. D. Kidd and C. Breazeal. Robots at home: Understanding long-term human-robot interaction. In *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3230–3235. IEEE, 2008.

[15] R. P. Goldman. Solving pomdps online through htn planning and monte carlo tree search. *HPlan 2021*, page 57, 2021.

[16] K. Chen, N. S. Srikanth, D. Kent, H. Ravichandar, and S. Chernova. Learning hierarchical task networks with preferences from unannotated demonstrations. In *Conference on Robot Learning*, pages 1572–1581. PMLR, 2021.