The Importance of Memory for Language-Capable Robots

Robots need to be able to communicate with people through natural language. But how should their memory systems be designed to facilitate this communication?

By Rafael Sousa Silva, Zhao Han, and Tom Williams

DOI: 10.1145/3611687

OPEN ACCESS

As robots become more widely available to the public and play more prominent roles in people's day-to-day routines, those robots will increasingly need to communicate through natural language. As specified by Tellex, language-capable robots will not only need to effectively understand human language but will also need to generate natural and human-like utterances to communicate with humans [1]. While large language models have recently gained popularity, on their own these models are insufficient for human-robot communication, due to the need for robots to ground their language in knowledge of the specific environment in which it is situated.

Figure 1, for example, displays the interaction between a human and a robot playing a "guess who" game. In this context, the robot needs to be aware of all characters in the game and their properties in order to be able to understand and create sentences about those characters.

Robots collect world information through different perception systems. For instance, object detection systems collect information about objects, mapping systems collect information about different locations, and face detection systems collect information about encounters with people. To maintain and use all of the collected information appropriately, robots need carefully constructed memory systems. Thus, in language-capable human-robot interaction (HRI), memory becomes the key component in a robot's cognitive system to incorporate situational knowledge about the world into its communication.

To create robotic systems that are capable of communicating in a natural and human-like way, we can draw inspiration from what is known about human cognition. Specifically, we can implement cognitively inspired models of long-term memory and working memory to keep track of what the robot knows and what knowledge the robot is currently attending to. In addition, just like humans, robots may need to mentally model their interlocutors' cognitive states. In this article, we will discuss the ways that our laboratory has drawn inspiration from across cognitive science research in order to achieve these capabilities in robot cognitive architectures. We will then use this analysis to understand the open questions and future directions that

Image by vs148 / Shutterstock.com

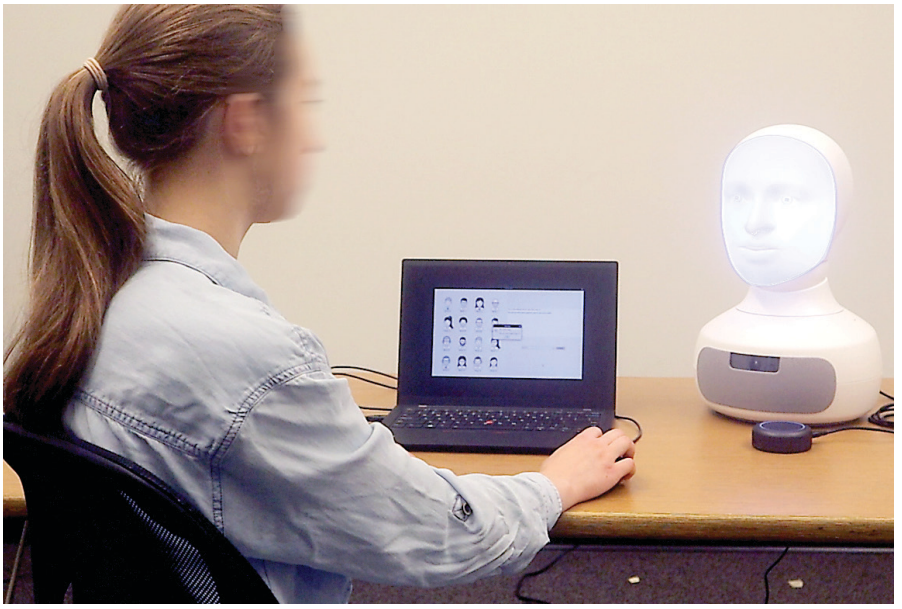
are available for new researchers to enter and contribute to this field.

LONG-TERM MEMORY: HOW IS KNOWLEDGE ORGANIZED?

If the central problems of language-capable robots revolve around how robots can communicate with people about their shared environment, the key questions we first need to answer are: (1) What are the domains of knowledge that robots will need to talk about? (2) What parts of a robot architecture are responsible for representing information in those domains of knowledge; that is, for encoding the long-term memories associated with that domain? (3) How do those architectural components represent information from those domains as long-term memories? and (4) How can the general-purpose mechanisms for language understanding and generation make use of disparate sources of knowledge (i.e., distributed long-term memory stores) despite their distribution throughout a robot architecture and their heterogeneous modes of representation of the knowledge?

To answer these questions, we looked to the child development community, where Spelke's theory of core knowledge postulates that humans are born with separable systems of core knowledge for representing space, objects, actions, numbers, and possibly social partners [2]. As roboticists, we can leverage this theory to answer our second question and understand the distinct knowledge-bearing components that our robots' language systems may need to connect with. Even though we may not want our robots to reason about num-

Figure 1. A human and a robot play a "guess who" game together.



bers as humans do (since, for example, they have built-in "super-human" capacities for numerical manipulation), by this theory, we will likely need distinct architectural components for reasoning about space, objects, actions, and people. For our third question, the precise knowledge representations used likely depend on other advances in robotics and AI such as the representations used by the common systems for metric-topological mapping, object recognition, planning and goal reasoning, and agent modeling.

Our laboratory has spent substantial time answering the fourth question. The consultant framework we use in our robot architecture requires that each domain of core knowledge provide the following capabilities to language understanding and generation: (1) list known entities (e.g., the object consultant should be able to provide IDs for particular objects it knows of); (2) list known properties and relations (e.g., the object consultant should say that it knows about blueness, largeness, etc.); (3) provide answers about the properties of specific entities and (4) create new, abstract representations for entities mentioned in dialogue that have not previously been observed, and may be merely hypothetical.

Inspired by the core knowledge theory, in our framework a general knowledge clearinghouse interfaces with distinct systems of core knowledge while allowing those systems to represent information in whatever way is preferable or feasible. Using this framework, we have been able to develop generalized, principled algorithms for reference resolution [3] (i.e., grounding a referring expression to knowledge representations stored in memory to determine what is being referenced) and referring expression generation [4] (i.e., selecting the properties that will be used in a generated expression such as choosing to highlight the redness, or the boxiness, of a red box, among other possible properties). Moreover, the algorithms for reference resolution that we have developed are inherently open world: They allow robots to understand descriptions of entities they've never heard of before and use such descriptions as opportunities for learning about the world, rather than as failure modes.

Memory becomes the key component in a robot's cognitive system to incorporate situational knowledge about the world into its communication.

Finally, in this approach, each core knowledge component provides identifiers for known entities and a consistent property-based interface for questioning and asserting knowledge of those entities. This approach provides a natural framework for building systems that exhibit human-like dynamics inspired by how human knowledge is organized and stored to form long-term memory.

WORKING MEMORY: WHAT DOES THE ROBOT KEEP IN CACHE?

Now that we have established how theories from cognitive science can be used to model robots' overarching long-term memory systems, how can we take similar inspiration to model robots' working memory systems? Determining which objects are being referred to can be computationally taxing when a robot knows of thousands of distinct objects. In our work, we have argued that theories of working memory from cognitive psychology may be used to enable real-time HRI while minimizing the computational costs of long-term memory processes such as information retrieval. Working memory can provide a "cache" of task-relevant objects in order to avoid the need to assess all known objects. To understand the nature of a working memory inspired cache, we can consider the psychological origins of working memory. Working memory is an important component of human cognition that has evolved from the concept of short-term memory. While short-term memory refers to a unitary system that is responsible for the management of information that is readily available, working memory is divided into subcomponents that handle different types of knowledge (e.g., visual, verbal, or episodic information).

Many of the different models of human working memory from cognitive psychology share the following three assumptions. First, the capacity of working memory is limited. Second, the information within working mem-

ory is volatile and may be forgotten at faster rates than information within long-term memory. And finally, working memory contents are readily accessible to other cognitive processes and may immediately influence deliberative processes, such as those related to natural language.

These assumptions reflect the fundamentally resource-limited nature of working memory and the ways that the process of forgetting is used to account for those resource limits. The dynamics of our own robot architecture's working memory system are inspired by two of the most widely accepted theories of forgetting from cognitive psychology: the theory of decay and the theory of interference. The theory of decay claims information leaves working memory with time, if not rehearsed or reinforced. The theory of interference claims older, unrehearsed information is replaced by newer information. In addition to these models of forgetting, we explore different ways in which information is organized within working memory. While older research on working memory emphasized the number of entities that could be maintained at any given time, recent work has instead focused on the number of properties of those entities that can be maintained, a perspective that emphasizes the quality of working memory representations.

WHAT ARE OTHERS THINKING ABOUT?

Thus far, we have focused on the way that human memory systems can inform the design of robot memory systems. Careful attention to the cognitive science literature can also bolster robots' communicative capabilities by providing an understanding of how people change the way they communicate based on their assumptions about other people's memory systems (first-order theory of mind modeling), as well as their assumptions about what others assume to be in their own memory systems (second-order theory of mind modeling).

One theory that can assist us in this way is Gundel, Hedberg, and Zacharski's givenness hierarchy theory [5]. The givenness hierarchy (GH) seeks to explain why people use different sorts of referring forms, like "it," "this," "that," "this <description>," "that <description>," "the <description>," and "a <description>." According to GH, these decisions are based on cognitive status: Speakers can permissibly use "it" if they think their referent is already "in focus" for their interlocutor; they can use "this," "that," and "this <description>" if they think their referent is "activated" for their interlocutor; they can use "that <description>" (a form that can be used to describe things that are not physically present, and that in fact may not have been seen in some time) if they think their referent is "familiar" to their interlocutor, and so on. Figure 2 shows the six cognitive statuses of GH.

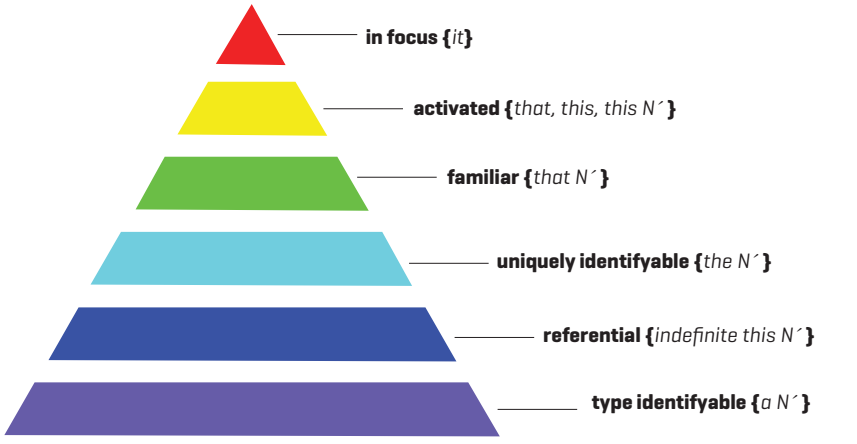
In most cases a robot can quickly find the representation corresponding to the language it is hearing by merely searching through the concise GH-informed data.

Critically, these different levels of cognitive status seem to make implicit assumptions about where things are presumed to be in memory; "in focus" roughly corresponds with the focus of attention, "activated" roughly corresponds with working memory, and so forth. Based on this alignment, we have leveraged GH for a variety of purposes.

Part 1: Generating human-like language through first order theory of mind modeling. The first way we have used GH is to enable robots to better generate concise, yet natural, referring forms. To do so, we began by developing a cognitive status engine, which tracks the cognitive status of each entity within a robot's task environment [6]. The engine is comprised of a set of cognitive status Bayes filters. Based on how entities are mentioned in dialogue, each Bayes filter recursively estimates the probability distribution over cognitive statuses for a given entity.

Leveraging this cognitive status engine, we then proposed an explainable decision tree-based machine learning model for referring form selection [7]. When the robot needs to describe an object, this model leverages both the most likely cognitive status for that object, as well as a number of other situated features—like the number of distractor objects in the scene and the object's physical distance—in order to predict the referring form the robot should use. We have evaluated this model both in previous benchmark environments used in the HRI community [8], as well as in novel task environments developed in our own laboratory that force users to engage in open-world tasks, where not all objects

Figure 2.The six cognitive statuses of the givenness hierarchy framework. Each status entails all lower statuses.



that need to be referred to can be seen or are even known about [9, 10].

Part 2: Understanding human-like language through second order theory of mind modeling. Next, we have leveraged GH for natural language understanding through second-order theory of mind modeling. While generating language requires a robot to estimate the likely cognitive status of an entity for its interlocutor, understanding language requires a robot to estimate which objects its interlocutor would likely consider, to have a specific cognitive status in its own mind. That is, when a robot's interlocutor uses "it," GH suggests they expect their referent to be in focus for the robot. The robot's first step, then, is to determine which objects would conversational partners expect me to be focusing on?

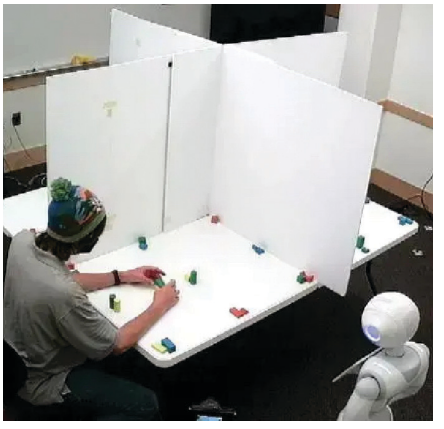
Our approach to enabling this type of reasoning relies on several GH-informed data structures [11]. When entities are mentioned in dialogue (by the robot or its conversational partner), they are stored in a set of focused or activated entities according to certain linguistic rules. If these entities are no longer mentioned as dialogue carries on, they "decay" into a set of merely familiar entities. When a dialogue ends, the focused, activated, and familiar data structures are flushed.

These data structures are then used as follows: First, when an utterance is heard, it is parsed into a tree structure, with each leaf corresponding to a different entity mentioned by the speaker. Second, based on how each entity was referred to, a presumed cognitive status is associated with each entity. Third, a search plan is constructed for each referenced entity to locate it in memory based on its presumed cognitive status. For example, if the entity is presumed to be activated, the plan would be to check the set of activated entities for an entity that matches its ascribed properties, and if that fails, to check the set of focused entities. If multiple entities are mentioned in the utterance, these search plans are combined by taking their cross-product to create a plan for searching. Fourth, in the worst case, such as when "the <description>" is used, this search plan may require searching for a representation in long-term memory. When this is the case, the reference resolution component performs a search across the architecture's distributed set of core knowledge components. Finally, if the robot determines that a described referent is something it did not previously know about, it requests the relevant core knowledge component to create a new representation for that entity based on the way it was described.

Our use of these data structures allows for enhanced efficiency as it essentially represents a caching strategy. Rather than always searching long-term memory, in most cases a robot can quickly find the representation corresponding to the language it is hearing by merely searching through the concise GH-informed data structures. Experimental analysis also showed this approach was able to appropriately handle about 55% of correctly parsed references. Of the remaining 45%, 10% represented genuine room for improvement, while 35% represented cases we did not intend to handle such as plural noun phrases, references to vague spatial regions, cases requiring gesture to disambiguate, and idioms.

Finally, we have used GH for high-level dialogue planning. In many domains, robots will need to plan sequences of utterances they want to say such as when teaching interactants how to perform novel tasks. For example, a robot may need to instruct a human to perform a complex task over several steps (see Figure 3). Using GH can as-

Figure 3. A robot provides instructions to a human on how to assemble a block structure.



Understanding language requires a robot to estimate which objects its interlocutor would likely consider, to have a specific cognitive status in its own mind.

sist robots in generating more natural task instructions. Typically, classical planning approaches have been used to generate the most efficient series of utterances to teach novel tasks. In contrast, we use GH to encourage planners to find solutions that involve references to objects that are already in focus or activated [12]. We have shown that this approach yields a number of key benefits:

1. Object reuse. The instructions from the GH-informed planner avoided switching physical tools (e.g., a screwdriver) during a task while the classical planner hops back and forth between separate tools.

2. Separating subtasks. The GH-informed planner completes work on one subtask before starting another one to keep objects in focus, whereas the classical planner generated instructions that switch back and forth between different subtasks.

3. More concise referring forms. Finally, the GH-informed planner, as designed, enables the robot to generate shorter utterances, as it can use "it," "this," and "that" more frequently, rather than constantly introducing or re-introducing objects into the dialogue.

This work showed that modeling cognitive status is useful for referring form selection and for the higher-level task of dialog planning.

FUTURE DIRECTIONS AND CONCLUSIONS

There are a number of key directions for future work that build on the foundations described in this article. Future work should consider:

- How can nonverbal cues like gaze and gesture be used both in the understanding and generation of referring forms? The use of gaze and gestures is very important to help robots convey messages through embodied language and can also impact the way in which a robot formulates referring expressions. For example, a robot that refers to an object while pointing to it will be able to use "this" to refer to it instead of the object's name.
- How can a robot's memory system design impact other aspects of language understanding and generation beyond reference? For example, how does a robot's memory system impact the way the robot uses specific gaze patterns and gestures?
- How can episodic memory systems be integrated into robot memory architectures to enable understanding and generation of references to specific past situations? Information in episodic memory can help the robot with making decisions about how to refer to entities. For example, if a robot is having a conversation with an interlocutor about an object that both of them have previously encountered together, the robot's referring expressions about the object might be more specific (i.e., that N'), because that object is familiar to both parties (and both parties know it to be familiar to each other).
- How can actions and goals be represented as core knowledge components to enable understanding and generation of references to them? A robot's procedural memory is important for performing tasks consistently and improving upon previous experiences. Thus, the representation of actions and goals within memory must be carefully designed.

In conclusion, memory plays a critical role in enabling language-capable robots to communicate effectively with humans. By drawing inspiration from human cognition and leveraging theories from cognitive psychology, researchers can develop models of long-term and working memory that facilitate grounding of language in the robot's knowledge of its situated environment. Finally, a number of key open questions and challenges present opportunities for future work.

ACKNOWLEDGMENTS

This work was funded in part by NSF CAREER grant IIS-2044865 and Office of Naval Research grant N00014-21-1-2418.

References

[1] Tellex, S., Gopalan, N., Kress-Gazit, H., and Matuszek, C. Robots that use language. *Annual Review of Control, Robotics, and Autonomous Systems* 3, 1 (2020), 25–55.

[2] Spelke, E., and Kinzler, K., Core knowledge. *Developmental Science*, 2007, 89–96.

[3] Williams, T., and Scheutz, M. A framework for resolving open-world referential expressions in distributed heterogeneous knowledge bases. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*. 2016.

[4] Williams, T., and Scheutz, M. Referring expression generation under uncertainty: Algorithm and evaluation framework. In *Proceedings of the 10th International Conference on Natural Language Generation*. 2017.

[5] Gundel, J., Hedberg, N., and Zacharski, R., Cognitive status and the form of referring expressions in discourse. *Language* 69, 2 (1993), 274–307.

[6] Pal, P., Zhu, L., Golden-Lasher, A., Swaminathan, A., and Williams, T. Givenness hierarchy theoretic cognitive status filtering. In *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*. 2020.

[7] Pal, P., Clark, G., and Williams, T. Givenness hierarchy theoretic referential choice in situated contexts. In *Proceedings of the Annual Meeting of the Cognitive Science Society (Cogsci)*. 2021.

[8] Bennett, M., Williams, T., Thames, D., and Scheutz, M. Differences in interaction patterns and perception for teleoperated and autonomous humanoid robots. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, 6589–6594.

[9] Han, Z., Rygina, P., and Williams, T. Evaluating referring form selection models in partially-known environments. In *Proceedings of the 15th International Conference on Natural Language Generation (INLG)*. 2022.

[10] Han, Z., and Williams, T. Evaluating cognitive status-informed referring form selection for human-robot interactions. In *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*. 2023.

[11] Williams, T., and Scheutz, M. Reference in robotics: A givenness hierarchy theoretic approach. In J. Gundel and B. Abbott (Eds.), *The Oxford Handbook of Reference*. Oxford University Press, 2019, 456–474.

[12] Spevak, K., Han, Z., Williams, T., and Dantam, N. T. Givenness hierarchy informed optimal document planning for situated human-robot interaction. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, 6109–6115.

Biographies

Rafael Sousa Silva is a Ph.D. student at the MIRRORLab at Colorado School of Mines. Sousa Silva is originally from Brazil and is starting his third year in the robotics program at Mines.

Zhao Han is formerly a post-doctoral fellow at Colorado School of Mines. he is currently an assistant professor of computer science at the University of South Florida.

Tom Williams is an associate professor of computer science at Colorado School of Mines, where he directs the Mines Interactive Robotics Research Lab.

1528-4972/23/09 $15.00