Enabling Human-like Language-Capable Robots Through Working Memory Modeling

Rafael Sousa Silva
rsousasilva@mines.edu
Colorado School of Mines
Golden, Colorado, USA

Tom Williams
twilliams@mines.edu
Colorado School of Mines
Golden, Colorado, USA

ABSTRACT

Working Memory (WM) is a central component of cognition. It has direct impact not only on core cognitive processes, such as learning, comprehension, and reasoning, but also language-related processes, such as natural language understanding and referring expression generation. Thus, for robots to achieve human-like natural language capabilities, we argue that their cognitive models should include an accurate WM representation that plays a similarly central role. Our research investigates how different WM models from cognitive psychology affect robots' natural language capabilities. Specifically, we explore the limited capacity nature of WM and how different information forgetting strategies, namely decay and interference, impact the human-likeness of utterances formulated by robots.

CCS CONCEPTS

• **Computing methodologies → Cognitive science**; **Cognitive robotics**; **Natural language generation**.

KEYWORDS

working memory, forgetting, referring expression generation

ACM Reference Format:
Rafael Sousa Silva and Tom Williams. 2023. Enabling Human-like Language-Capable Robots Through Working Memory Modeling. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI '23 Companion), March 13–16, 2023, Stockholm, Sweden.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3568294.3579967

1 MOTIVATION AND BACKGROUND

Working memory (WM) is the component of human cognition responsible for the temporary storage and maintenance of information necessary for core cognitive tasks [1]. These tasks include processes such as learning [2], reasoning [13, 28], and comprehension [11]. Moreover, research has shown that language-related processes are also greatly affected by WM (e.g., natural language understanding [25], acquisition [7], and generation [10]). For instance, WM may directly affect the process of Referring Expression Generation (REG) [3, 8, 9], in which an agent selects a set of properties to describe an object to other agents [29].

While there exist many theories of WM within cognitive psychology, three characteristics seem to be widely accepted by researchers. First, that WM has limited capacity, with early research presenting estimations in the range from four to nine items [6, 17]. Second, that the information within it is volatile and may be forgotten at faster rates than information contained in long-term memory [30]. Finally, WM contents are readily accessible to other cognitive processes and may immediately influence deliberative processes [19], including those related to language. Therefore, to arrive at better language-capable robots, our research is focused on the limited capacity of WM and how different information forgetting strategies impact robotic natural language generation.

Two models of forgetting have become popular within this discussion. On one hand, the theory of decay [5] proposes that information in WM fades away with time if not rehearsed. On the other hand, the theory of interference [30] defends that older items within WM are replaced by newer items that enter WM buffers. Both decay and interference have empirical evidence to support their claims (see [12, 18, 21, 23], for example) and have been implemented into computational models, such as TBRS [22] and SOB-CS [4, 20].

However, these computational models of WM forgetting are often implemented as individual systems that are not integrated with other components and processes of cognition. For robots capable of natural language, this connection is important because WM should not only serve as a temporary storage for information, but also as a mechanism of cognition that directly affects language processes, such as REG. In addition, complex cognitive architectures like SOAR [14] and ACT-R [24] are often concerned with the storage of entities rather than the storage of the properties that apply to that entity [16]. Recent HRI research has started to remedy these issues through the maintenance of properties that belong to activated entities within architectural WM buffers [34]. Yet, there is little exploration on how different forgetting strategies might be implemented through this perspective and how they might affect natural language processes.

The central aim of our work is to address this knowledge gap by exploring how models of decay and interference can be optimally integrated into robot cognitive architectures. Our overarching hypothesis is that mechanisms of decay and interference can lead to a better robotic language generation by allowing these processes to leverage WM buffers containing the salient object properties most likely to be natural and effective to use in natural language descriptions. More specifically, our research aims to answer three key research questions:

RQ1 – What level of decay and interference will result in optimal performance with respect to (1) accuracy, (2) naturalness, (3)

HRI '23 Companion, March 13–16, 2023, Stockholm, Sweden

ACM ISBN 978-1-4503-9970-8/23/03.
https://doi.org/10.1145/3568294.3579967

computational efficiency, (4) ease of cognitive processing, and (5) human-likeness of referring expression generation?

RQ2 – In order to optimize each key metric from RQ1, should resources be distributed according to a limit imposed at an entity, consultant, or architectural level?

RQ3 – In order to optimize each key metric from RQ1, how should the architecture decide which entities for which to maintain representations in Working Memory?

2 RESEARCH APPROACH

To answer the research questions described above, we are developing computational models of WM dynamics that account for the interference and decay forgetting strategies. These models are implemented using the Agent Development Environment (ADE) middleware [26] in which the Distributed, Integrated, Affect, Reflection, and Cognition (DIARC) architecture [27] was implemented. DIARC is a cognitive architecture consisting of components that implement key theories and concepts from linguistics and cognitive psychology in order to enable language-capable robots. The ADE middleware allows DIARC components to operate in parallelism and communicate with other components asynchronously.

DIARC's long-term memory is organized by a set of components classified as *consultants* [31]. Each consultant serves as a Distributed Heterogeneous Knowledge Base (DHKB) [33] for a certain type of entity (e.g., people, animals, places). For this work, each DHKB is equipped with an additional WM representation that maintains a set of activated properties for each entity within the consultant. These WM buffers are handled by the WM Manager component [32], which detects DHKBs that are in operation within the architecture and creates connections with them in order to (1) add activated properties to the buffers of appropriate objects and (2) remove properties from these buffers according to the forgetting strategy that is in use. If decay is in use, properties are removed from buffers after a specified amount of seconds, denoted by the parameter δ. Otherwise, when interference is in effect, the buffer for each entity is limited to a maximum capacity of α items, and the least-recently-added properties give space to newer entries.

3 PRELIMINARY WORK

To validate the efficacy of our forgetting models within the cognitive architecture, we conducted a human-subjects study (N = 90) in which participants interacted with our robotic architecture in the context of a "Guess Who" reference game involving sixteen known people, knowledge of each of which was stored in a "face consultant" in the form of a set of logically specified properties.

To start addressing RQ2, the resources within WM buffers were organized at an *entity level*, meaning that each entity within the consultant had a dedicated WM buffer whose design and dynamics differed based on the WM model employed. Under the interference model, each buffer was capable of holding a total of $\alpha \in \{2, 3, 4\}$ properties at any given time[1]. Under the decay model, buffer size was unlimited, but the δ parameter according to which items were forgotten was set to either 10, 15, or 20 seconds.

During the game, participants alternated between rounds in which they had to (1) select properties to describe a face to the robot or (2) guess which face a given robot description was referring to. Each property used in face descriptions had its salience updated in the appropriate WM buffers. The robot descriptions were generated through DIARC's WM-enabled REG algorithm, SD-PIA [34], which attempts to describe an entity with the properties that are present within that entity's WM buffer. If those properties are not sufficient to create a description that can rule out all distractors, then SD-PIA adds properties from the given entity's long-term memory buffer until there are no distractors left. The list of properties returned by SD-PIA was then processed by a template-based sentence realization system, which outputted a string with the final description formatted in standardized American English.

To validate the architectural implementation of these forgetting models, we collected data for a subset of the measures listed for RQ1 above, including naturalness (assessed every five rounds), and accuracy, response time, and human-likeness (assessed through transcripts). Our results showed that mean human accuracy readings of at least 95% were obtained for all conditions, mean perceived naturalness readings were uniformly above 3.5 out of 5, and Jensen-Shannon Divergence [15] values demonstrated high similarity between robot and human descriptions. However, ease of listener cognitive processing was uniformly poor, with an average response time above 14 seconds, suggesting participants were relatively slow to respond regardless of model. These preliminary results provide insights into how forgetting models can be implemented into robot cognitive architectures and allow future research to explore new model parameterizations and resource management strategies.

4 FUTURE WORK

In future work we will pursue three key directions. First, to address RQ1, due to the overall success of our model implementations, we will explore a wider array of model parameterizations to better determine which are most effective for each of our key measures of interest. Since these preliminary results were not yet matched to a control group, this comparison will be done in relation to an experimental condition that uses no WM forgetting in order to emphasize the benefits of using the proposed method. Second, to address RQ2, we will explore the implementation of two other resource management strategies into our cognitive architecture. On one hand, a *consultant-based* resource management strategy will limit the amount of properties that can be stored by each consultant, as opposed to limiting the amount of resources available to each entity. On the other hand, a *global* resource management strategy will limit the total amount of properties that can be stored in WM at any given time, independently of how many consultants are in operation. Finally, to address RQ3, we will combine the best parameterizations with the best resource management strategies to identify the overall best architectural approach. The answers to these research questions will lead to a better understanding of how WM representations need to be maintained and will help towards the development of better language-capable robots.

ACKNOWLEDGMENTS

This work was funded in part by NSF CAREER grant IIS-2044865.

[1] The chosen values of α were small because, for an *entity-based* strategy, even the lowest value of two properties per buffer imposed an upper bound of thirty-two properties within WM at any given time, which is well above the speculated range for human WM capacity, a value within the range of four to nine items [6, 17].

REFERENCES

[1] Alan Baddeley. 1992. Working memory. *Science* 255, 5044 (1992), 556–559.

[2] Alan Baddeley. 2010. Working memory. *Current biology* 20, 4 (2010), R136–R140.

[3] Julie Bannon. 2019. Working memory and referential communication: An investigation of the cognitive factors affecting the production of overspecified referring expressions.

[4] Pierre Barrouillet, Sophie Bernardin, and Valérie Camos. 2004. Time constraints and resource sharing in adults' working memory spans. *Journal of Experimental Psychology: General* 133, 1 (2004), 83.

[5] John Brown. 1958. Some tests of the decay theory of immediate memory. *Quarterly journal of experimental psychology* 10, 1 (1958), 12–21.

[6] Nelson Cowan. 2001. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and brain sciences* 24, 1 (2001), 87–114.

[7] Nadiia Denhovska, Ludovica Serratrice, and John Payne. 2016. Acquisition of second language grammar under incidental learning conditions: The role of frequency and working memory. *Language Learning* 66, 1 (2016), 159–190.

[8] Albert Gatt, Martjin Goudbeek, and Emiel Krahmer. 2011. Attribute preference and priming in reference production: Experimental evidence and computational modeling. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 33.

[9] Martijn Goudbeek and Emiel Krahmer. 2011. Referring under load: Disentangling preference-based and alignment-based content selection processes in referring expression generation. *Proceedings of PRE-Cogsci: Bridging the gap between computational, empirical and theoretical approaches to reference* (2011).

[10] Jeanette K Gundel, Nancy Hedberg, and Ron Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language* (1993), 274–307.

[11] Graeme S Halford, William H Wilson, and Steven Phillips. 1998. Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral and brain sciences* 21, 6 (1998), 803–831.

[12] John Jonides, Richard L Lewis, Derek Evan Nee, Cindy A Lustig, Marc G Berman, and Katherine Sledge Moore. 2008. The mind and brain of short-term memory. *Annual review of psychology* 59 (2008), 193.

[13] Patrick C Kyllonen and Raymond E Christal. 1990. Reasoning ability is (little more than) working-memory capacity?! *Intelligence* 14, 4 (1990), 389–433.

[14] John E Laird. 2019. *The Soar cognitive architecture.* MIT press.

[15] Jianhua Lin. 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory* 37, 1 (1991), 145–151.

[16] Wei Ji Ma, Masud Husain, and Paul M Bays. 2014. Changing concepts of working memory. *Nature neuroscience* 17, 3 (2014), 347–356.

[17] George A Miller. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review* 63, 2 (1956), 81.

[18] Paul Muter. 1980. Very rapid forgetting. *Memory & Cognition* 8, 2 (1980), 174–179.

[19] Donald A Norman and Tim Shallice. 1986. Attention to action. In *Consciousness and self-regulation.* Springer, 1–18.

[20] Klaus Oberauer and Stephan Lewandowsky. 2011. Modeling working memory: A computational implementation of the Time-Based Resource-Sharing theory. *Psychonomic bulletin & review* 18, 1 (2011), 10–45.

[21] Klaus Oberauer and Stephan Lewandowsky. 2014. Further evidence against decay in working memory. *Journal of Memory and Language* 73 (2014), 15–30.

[22] Klaus Oberauer, Stephan Lewandowsky, Simon Farrell, Christopher Jarrold, and Martin Greaves. 2012. Modeling working memory: An interference model of complex span. *Psychonomic bulletin & review* 19, 5 (2012), 779–819.

[23] Ehud Reiter and Robert Dale. 1997. Building applied natural language generation systems. *Natural Language Engineering* 3, 1 (1997), 57–87.

[24] Frank E Ritter, Farnaz Tehranchi, and Jacob D Oury. 2019. ACT-R: A cognitive architecture for modeling cognition. *Wiley Interdisciplinary Reviews: Cognitive Science* 10, 3 (2019), e1488.

[25] Jerker Rönnberg, Mary Rudner, Thomas Lunner, Adriana A Zekveld, et al. 2010. When cognition kicks in: Working memory and speech understanding in noise. *Noise and Health* 12, 49 (2010), 263.

[26] Matthias Scheutz. 2006. ADE: Steps toward a distributed development and runtime environment for complex robotic agent architectures. *Applied Artificial Intelligence* 20, 2-4 (2006), 275–304.

[27] Matthias Scheutz, Thomas Williams, Evan Krause, Bradley Oosterveld, Vasanth Sarathy, and Tyler Frasca. 2019. An overview of the distributed integrated cognition affect and reflection diarc architecture. *Cognitive architectures* (2019), 165–193.

[28] Heinz-Martin Süß, Klaus Oberauer, Werner W Wittmann, Oliver Wilhelm, and Ralf Schulze. 2002. Working-memory capacity explains reasoning ability—and a little bit more. *Intelligence* 30, 3 (2002), 261–288.

[29] Kees Van Deemter. 2016. *Computational models of referring: a study in cognitive science.* MIT Press.

[30] Nancy C Waugh and Donald A Norman. 1965. Primary memory. *Psychological review* 72, 2 (1965), 89.

[31] Tom Williams. 2017. A Consultant Framework for Natural Language Processing in Integrated Robot Architectures. *IEEE Intell. Informatics Bull.* 18, 1 (2017), 10–14.

[32] Tom Williams, Torin Johnson, Will Culpepper, and Kellyn Larson. 2020. Toward forgetting-sensitive referring expression generationfor integrated robot architectures. *arXiv preprint arXiv:2007.08672* (2020).

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

[34] Tom Williams, Ravenna Thielstrom, Evan Krause, Bradley Oosterveld, and Matthias Scheutz. 2018. Augmenting robot knowledge consultants with distributed short term memory. In *International Conference on Social Robotics*. Springer, 170–180.