A Framework for Proactive and Adaptive Natural Language Interaction in Dynamic and Unstructured Environments

Siddharth Patki   Jacob Arkin   Ethan Fahnestock   Thomas M. Howard

Abstract— **Humans should be able to collaborate with robots without sacrificing their own operational tempo, and consequently the communication should be efficient and uninhibitive. Physically-grounded language interfaces provide a generally-accessible solution that allows the human to communicate with their robot teammates. However, both robot perception and grounded language understanding are computationally expensive in unstructured, dynamic environments and thus impose an efficiency bottleneck for collaboration. One recent approach attempts to address this problem by reactively building compact, task relevant world models by exploiting the information in the received utterance. Another recent approach exploits the idle time before receiving an utterance to proactively generate and ground the likely phrases. In this work, we propose an integrated framework that leverages both of these seemingly contradictory models in order to maximize the runtime performance of language understanding.**

I. Introduction

Collaborative robots that seamlessly interact as part of human-robot teams in dynamic and unstructured environments will transform our daily lives. A significant barrier to this is situated natural language understanding, wherein the robot must be capable of accurately and swiftly comprehending complex relationships in uncertain or incomplete world models so as to not hinder team performance. Contemporary models [1–9] pose the language understanding problem as one of associating linguistic constituents of a parsed instruction with perceived entities or actions the robot should take. Such physically-grounded language understanding systems have two main computationally expensive components: producing a sufficient representation of the world via perception and mapping language to a representation of meaning (symbols) that can be interpreted by the robot. The efficiency of grounded language understanding is constrained by the runtime of perception which provides environmental context to these models. Generating highly detailed world representations in dynamic and unstructured environments and reasoning in their context is slow and inhibits realtime interaction with these robots.

Towards addressing this problem, a recent work has proposed conditionally adapting the robot's perception pipeline by exploiting the information in the language utterance to more quickly provide a minimal, approximate, task-relevant world model that is sufficiently expressive for accurate language grounding [10]. Thus, the system provides a more temporally and contextually relevant world model. Experiments conducted using a synthetic corpus in the manipulation domain have demonstrated that adapting perception resulted into an approximate four times reduction in the runtime for perception and nine times reduction in runtime of symbol grounding per instruction without a loss in the accuracy of the latter.

Reactively mapping language to symbols similarly imposes a computational bottleneck. Another recent work has proposed the proactive generation and grounding of language in anticipation of what a human teammate might say in order to reuse solutions at the time of interaction [11]; thus, the size of the reactive language grounding problem is reduced. This facilitates faster inference, which in the best case is a solution look-up and in the worst is identical to the baseline but with a look-up overhead. Proactive inference exploits the pre-utterance idle time for computation but assumes the presence of a world model.

On the surface, the advantages of these two recent works seem mutually exclusive as the adaptive perception is done after the human has given a command whereas proactive language grounding is done during the idle time before the human says anything. In this paper, we propose a concept for an integrated framework that combines the Adaptive Perception (AP) model with the Proactive Symbol Grounding (PSG) model to exploit the advantages of both.

II. Technical Approach

We frame the problem of natural language understanding as inference over a learned distribution that associates linguistic elements with their corresponding symbolic representation Γ_s that represents objects, places, spatial relations, actions, and other concepts. The distribution over symbols is conditioned by the parse of the utterance Λ and a world model Υ expressing environment knowledge extracted from sensor measurements $z_{1:t}$ using classifiers in the robot's perception pipeline P. We use Distributed Correspondence Graphs (DCGs) [4] in the formulation of our two learned models.

$$\Phi_s^* = \underset{\phi_{ij} \in \Phi_s}{\arg\max} \prod_{i=1}^{|\Lambda|} \prod_{j=1}^{|\Gamma_s|} p(\phi_{ij} | \gamma_{ij}, \lambda_i, \Phi_{ci}, \Upsilon). \quad (1)$$

Formally, DCG inference involves searching for the most likely correspondence variables Φ_s^* in the context of the phrases $\lambda_i \in \Lambda$, groundings $\gamma_{ij} \in \Gamma_s$, child phrase correspondences Φ_{ci}, and the world model Υ by maximizing the factorization in Equation 1.

Siddharth Patki, Jacob Arkin, Ethan Fahnestock and Thomas M. Howard are with the Electrical and Computer Engineering Department at the University of Rochester, Rochester, NY USA, spatki@ur.rochester.edu, jarkin@ur.rochester.edu, efahnest@u.rochester.edu, thomas.howard@rochester.edu

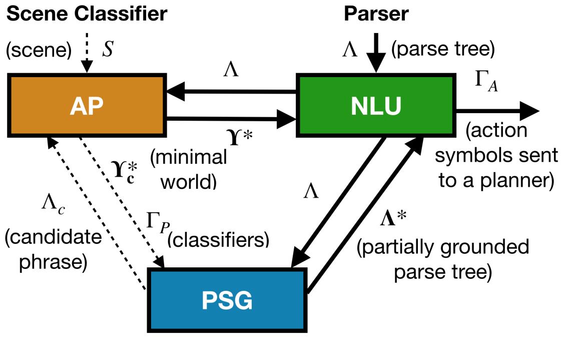

Fig. 1. The system architecture for combined proactive and adaptive model for language understanding. Learned models are highlighted in color. Solid arrows denote the information flow after parser receives an utterance, while the dotted arrows denote the idle time transactions.

Adaptive Perception [10] improves the runtime efficiency of Equation 1 by building a compact environment representation Υ^* that is sufficient for interpreting the utterance. Central to this approach is the ability to infer a subset of perceptual classifiers $P^* \in P$ conditioned on the utterance. The method exploits language to guide the generation of a instruction-specific pipeline $P^* = f(P, \Lambda)$ resulting in a minimal world model $\Upsilon^* = f(z_{1:t}, P^*)$. Inference in the context of the minimal world representation takes the form of equation:

$$\Phi_s^* = \underset{\phi_{ij} \in \Phi_s}{\arg\max} \prod_{i=1}^{|\Lambda|} \prod_{j=1}^{|\Gamma_s|} p(\phi_{ij} | \gamma_{ij}, \lambda_i, \Phi_{ci}, \Upsilon^*). \tag{2}$$

Proactive Symbol Grounding [11] improves the runtime of Equation 1 by generating and grounding candidate language phrases during idle time in anticipation of what a human teammate might say in order to reuse solutions at the time of interaction. Formally, PSG reduces the number of phrases that need evaluation during reactive language grounding. This reduction is a function of the utterance's parse tree and the proactively grounded phrases $\Lambda^* = f(\Lambda, \Lambda_{PSG})$.

Figure 1 illustrates the system architecture of the proposed framework. It has two operational phases: one for the proactive computations done during idle time and another for the reactive computations done after an instruction is received. Since it is necessary that the world model used by the PSG process is consistent with the world model used by NLU during reactive inference, AP must be leveraged both proactively and reactively. To achieve this, AP runs a sequenced subset of detectors as informed by the scene semantics to incrementally build a world representation during the idle time. Concurrently, PSG generates candidate phrases and queries the AP node to receive the corresponding minimal world models for those phrases, skipping the phrase if no minimal world model exists yet. On receiving an utterance, a parser converts it to a parse tree and sends it to the NLU node which coordinates with the AP and PSG nodes to perform the symbol grounding inference. NLU first sends the parse tree to the AP node which returns the associated minimal world model and an indication of whether the world has changed as compared to the world used by PSG. If the world is sufficiently similar, NLU sends the parse tree to PSG which fills in the solutions for known proactively grounded candidate phrases. A partially grounded parse tree is returned to the NLU node which performs inference on any remaining ungrounded phrases as conditioned on the compact world model. Therefore, the inference in our full model combining AP and PSG is formulated as:

$$\Phi_s^* = \underset{\phi_{ij} \in \Phi_s}{\arg\max} \prod_{i=1}^{|\Lambda^*|} \prod_{j=1}^{|\Gamma_s|} p(\phi_{ij} | \gamma_{ij}, \lambda_i, \Phi_{ci}, \Upsilon^*). \tag{3}$$

We hypothesize that the runtime for solving Equation 3 will be less than that for Equations 1 and 2 in the cases where proactively computed solutions are valid at the time of inference. In the worst case, it would be same as that for Equation 2 with a small overhead cost of PSG look-up.

III. Discussion

Presented here is a theoretical framework for proactive and adaptive natural language interaction. As part of future work, we intend to investigate and quantify the expected performance gains in a series of ablation experiments targeting different scenarios in static and dynamic worlds.

We are also interested in contributing further improvements to our framework. The utility of PSG depends on how much of the reactive language grounding problem it can provide solutions for; but, since many phrases are world-dependent, the precomputed solutions may become invalid due to world dynamics. Notably, our symbols fall along a spectrum of environmental-sensitivity. Some are world-invariant and thus will never become invalid while others may be invariant within a definable set of world configurations. We are interested in fast, robust ways of determining where along that spectrum particular symbols land and for which sets of worlds their expression is valid in order to maximize the resuability of each proactive computation.

In addition to maximizing solution reusability, PSG and proactive AP can be made more effective by increasing the likelihood that precomputed solutions will be relevant for the utterance. The current implementation uses a two-stage search whereby AP iteratively perceives the world using a sequence of detectors biased according to scene semantics, and PSG searches in a strict bottom-up, breadth-first process conditioned on which detectors have been used. For future work, we are interested in further biasing both stages according to information expressed in prior utterances, thereby improving the chance that the robot has predicted what the human will say.

IV. Acknowledgements

This work was supported in part by the National Science Foundation under the grant IIS-1637813 and by the Robotics Consortium of the U.S. Army Research Laboratory under the Collaborative Technology Alliance Program Cooperative Agreement W911NF-10-2-0016.

References

[1] S. Tellex, T. Kollar, S. Dickerson, M. R. Walter, A. G. Banerjee, S. Teller, and N. Roy, "Understanding natural language commands for robotic navigation and mobile manipulation," in *Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.

[2] C. Matuszek, E. Herbst, L. Zettlemoyer, and D. Fox, "Learning to parse natural language commands to a robot control system," in *Experimental Robotics*. Springer, 2013, pp. 403–415.

[3] C. Matuszek, N. FitzGerald, L. Zettlemoyer, L. Bo, and D. Fox, "A joint model of language and perception for grounded attribute learning," *arXiv preprint arXiv:1206.6423*, 2012.

[4] R. Paul, J. Arkin, D. Aksaray, N. Roy, and T. M. Howard, "Efficient grounding of abstract spatial concepts for natural language interaction with robot platforms," *Int'l J. of Robotics Research*, 2018.

[5] A. Boularias, F. Duvallet, J. Oh, and A. Stentz, "Grounding spatial relations for outdoor robot navigation," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2015, pp. 1976–1982.

[6] J. H. Oh, A. Suppé, F. Duvallet, A. Boularias, L. Navarro-Serment, M. Hebert, A. Stentz, J. Vinokurov, O. Romero, C. Lebiere *et al.*, "Toward mobile robots reasoning like humans," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[7] A. Boteanu, T. Howard, J. Arkin, and H. Kress-Gazit, "A model for verifiable grounding and execution of complex natural language instructions," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 2649–2654.

[8] D. Yi, T. M. Howard, K. Seppi, and M. Goodrich, "Expressing homotopic requirements for mobile robot navigation through natural language instructions," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, Oct. 2016, pp. 1462–1468.

[9] J. Arkin, M. Walter, A. Boteanu, M. Napoli, H. Biggie, H. Kress-Gazit, and T. M. Howard, "Contextual awareness: Understanding monologic natural language instructions for autonomous robots," in *IEEE International Symposium on Robot and Human Interactive Communication*, Aug. 2017.

[10] S. Patki and T. M. Howard, "Language-guided adaptive perception for efficient grounded communication with robotic manipulators in cluttered environments," in *Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 2018.

[11] J. Arkin, R. Paul, S. Roy, D. Park, N. Roy, and T. M. Howard, "Real-time human-robot communication for manipulation tasks in partially observed environments," in *International Symposium on Experimental Robotics*, Nov. 2018.