A Demonstration of Natural Language Understanding in Embodied Planning Agents

Sachin Grover and Shiwali Mohan

SRI International, CA, USA
{sachin.grover, shiwali.mohan}@sri.com

Autonomous agents operating in human worlds must understand and respond to natural language used by humans to communicate their task needs. In a home environment, an agent must solve the *language-to-action* problem - it should comprehend *put that apple in the refrigerator* by connecting references to objects (*that apple*, *the refrigerator*), instantiating a task that achieves the goal (i.e., the desired spatial relationship between the apple and the refrigerator), and achieve it by applying an action sequence.

Recent work (Sarch et al. 2023) has begun to explore this research challenge, greatly benefiting from the advances in deep learning and large language models (LLMs). These approaches frame the language-to-action problem as a *sequence-to-sequence* mapping problem. In datasets such as ALFRED (Shridhar et al. 2020), a natural language task request is paired with a sequence of natural language 'sub'actions, executing which will achieve the task request. The agent determines (through machine learning-based training) how to map the sequence of tokens in the task request to sequence of 'sub'actions in natural language. An executor that can translate the 'sub'actions into execution in the environment is assumed. Such approaches underplay causal, goal-oriented reasoning that is critical for robust and flexible task performance.

We introduce a distinct way of approaching the language-to-action problem. Central to our approach is a planning-based agent that maintains the *perceive-decide-action* loop with its environment and can achieve a space of plausible goals. Natural Language Understanding (NLU) is, then, framed as identification of relevant environmental elements and construction of a goal that the human intends the agent to perform. Once the goal is constructed, it is performed using the causal reasoning machinery in the planning-based agent. We demonstrate our approach in an embodied planning agent operating in AI2Thor (Kolve et al. 2017).

Our work fits within ongoing research studying the role LLMs in planning systems (Liu et al. 2023b). We use a LLM specifically as an interaction mechanism between a human and an agent while sequential decision making is performed by planning methods. This configuration alleviates the problem of incorrect/incomplete plan inference in LLMs (Valmeekam et al. 2024). Further, extending prior work (Liu et al. 2023a), we bring LLM+planning techniques to taskable, interactive, embodied agents.

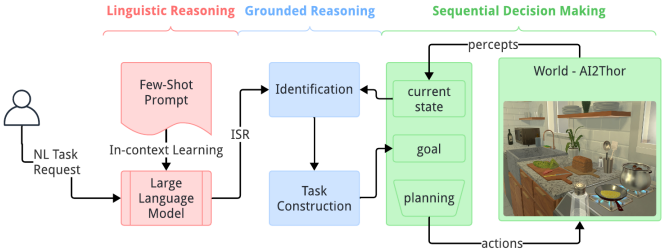

Figure 1: A notional diagram of our approach showing various types of reasoning and decision making.

From Language to Planning

AI2Thor (Kolve et al. 2017) is a live, 3 dimensional, simulated home environment in which embodied agents can interact with and manipulate a variety of objects, tools, and locations. It is equipped with many agents supporting a range of embodiments including robots such as the LoCoBot (Wögerer et al. 2012). It is one of the standard domains for research on conversational embodied agents and has been used in prominent datasets including ALFRED (Shridhar et al. 2020) and Teach (Padmakumar et al. 2022).

Our demonstration is built using AI2Thor in which the agent can be tasked to find certain objects and move them to new locations. Similarly to prior work, we assume a discretized, factored state-space, a fully-observable environment, and a set of primitive, atomic actions. We break down the language-to-action problem as a series of subproblems, each of which is solved using problem-specific mechanism. A notional system diagram is shown in Figure 1. Towards the right (in green) is a standard planning-based agent that receives percepts from the environment and maintains the current state as a set of grounded predicates. Given a goal (described using a conjunction of predicates), it plans a sequence of actions that are executed in the environment. NLU is split into two: one, linguistic reasoning (in red), that translates human natural language into a machine-readable, meaning representation and two, grounded reasoning, that connects information conveyed in the machine representation to elements of the state and the task.

Linguistic Reasoning To translate a task request into a meaning representation, we build upon analogical, case-based reasoning and translation capabilities of LLMs (shown in red in Figure 1). In-context learning in LLMs

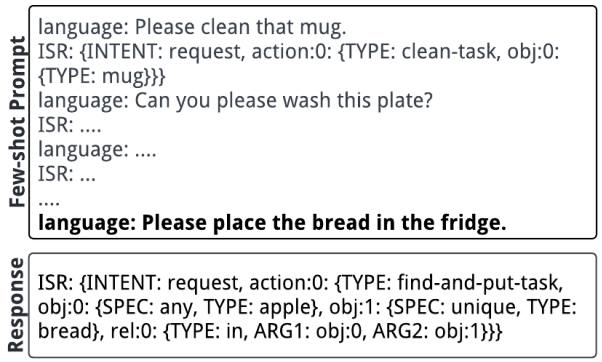

Figure 2: Few-shot prompting for translating a natural language request into a meaning representation

(Dong et al. 2022) operates by introducing a few examples of task performance written as text in the prompt (termed few-shot prompt). The examples are followed by the actual task query. An example is shown in Figure 2. The LLM is shown several example translations (in the Few-Shot Prompt box) of language to intentional semantics representations (ISR). An ISR is a JSON structure explicitly encoding the intent being expressed (a request in this case) as well as the content (`action:0` of TYPE `clean-task` applied to `obj:0` of type `mug`). After presenting examples, the LLM is asked to translate a natural language sentence (shown in bold). State of art LLMs can use examples in the prompt to do the translation task reliably (in the response box). A key observation we make here is that an LLM can be taught to distill task-relevant information from variation in expression of task requests. For example, *can you please clean that mug*, *please clean that mug*, *clean that mug* express the same action. Our demonstration is built with OpenAI GPT3.5 (`gpt-3.5-turbo-0613`).

Grounded Reasoning The ISR in Figure 2 contains information about the task to perform (`find-and-put-task`) and which objects should be selected (e.g. `TYPE: bread`, `TYPE: fridge`). However, these are *ungrounded* descriptions; i.e, they are not connected to specific objects, actions, configurations, or goals instantiated in the world. In the grounded reasoning step, the system fuses its beliefs about the world (and the state of task performance) with information in the ISR to generate a grounded representation of the goal. An example is shown in Figure 3. Beliefs about the current state (Figure 3 top) are represented as predicates and asserted based on the input from the environment (Figure 1).

The first step in grounded reasoning is identification of entities described in the ISR from the current state beliefs. This is done through a filtering process in which the system maps object `TYPE` in ISR to the `isa` predicate. For example, `obj:0` in the ISR is mapped to `o4` in the current state beliefs and `obj:1` to `o3` via the `isa` predicate.

The second step in grounded reasoning is constructing a goal. Each action `TYPE` is mapped to a specific goal construction mechanism. For example, the goal for the action TYPE `find-and-put-task`, the goal is constructed from the spatial relationship described in the ISR instantiated with grounded objects (shown in Figure 3). We envision the agent to encode set of tasks each with its own goal template that various verbs in natural language map to.

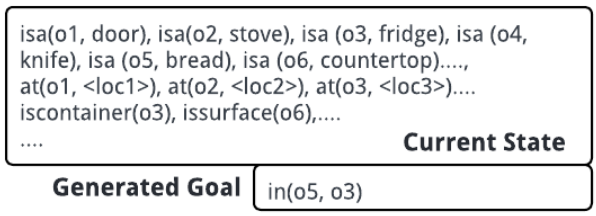

Figure 3: Grounded reasoning for connecting meaning representations to beliefs about the state and the task

Sequential Decision Making Upon constructing the goal, the agent invokes the planning process to determine a sequence of actions (shown in green in Figure 1). The input from the world as well as the goal predicate generated by the grounded reasoning process are written as PDDL (Ghallab et al. 1998) predicates to a problem file. We make some simplifying assumptions to generate the current state, such as the location of each object is encoded as a cell (instead of a numeric 3D location). The agent in AI2Thor has several actions available to it; such as teleport, pickup, put, open, and close. Availability of each action depends of properties of objects provided to the agent as a part of the current state. E.g., `fridge` and `drawer` are `receptacle` entities that can be opened. The agent generates a domain file with action definitions that capture these constraints as pre-conditions and corresponding consequent effects. Actions such as teleport are defined as macro-actions where action application include pose computations. We use state-of-the-art planners such as FF (Hoffmann and Nebel 2001) and Nyx (Piotrowski and Perez 2024) for planning a sequence of actions which is executed in the world using AI2Thor interface.

Demonstrations The demo presents the web-based interface for providing instructions and the simulator where the robot executes the commands. Currently, it shows preliminary results for end-to-end translation of the instruction and the execution of the plan in AI2Thor. The instructions provided to the robot are – `put bread in the fridge`, `put spatula in the drawer`, and `put apple in the fridge`. Demonstration can be checked at – https://bit.ly/icaps24_demo.

At Outlook for the Future

In future, we will extend this demonstration to include a wide class of tasks in the AI2Thor domain, covering those in ALFRED and Teach datasets. We want to demonstrate that integration of ML (LLM-based NLU) and reasoning (planning) methods enables robust embodied interactive behavior that surpasses ML-only approaches. Our eventual goal is to develop *teachable* agents that learn new knowledge from human teaching facilitated by language. For embodied planning agents, this challenge comprises learning new action models, pre-conditions, effects, and new goals by combining information from natural language teaching from humans and sensory information available in the world.

References

Dong, Q.; Li, L.; Dai, D.; Zheng, C.; Wu, Z.; Chang, B.; Sun, X.; Xu, J.; and Sui, Z. 2022. A Survey on In-Context Learning. *arXiv preprint arXiv:2301.00234*.

Ghallab, M.; Howe, A.; Knoblock, C.; McDermott, D.; Ram, A.; Veloso, M.; Weld, D.; Wilkins, D.; Barrett, A.; and Christianson, D. 1998. Pddl— the planning domain definition language.

Hoffmann, J.; and Nebel, B. 2001. The FF planning system: Fast plan generation through heuristic search. *Journal of Artificial Intelligence Research*, 14: 253–302.

Kolve, E.; Mottaghi, R.; Han, W.; VanderBilt, E.; Weihs, L.; Herrasti, A.; Deitke, M.; Ehsani, K.; Gordon, D.; Zhu, Y.; et al. 2017. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*.

Liu, B.; Jiang, Y.; Zhang, X.; Liu, Q.; Zhang, S.; Biswas, J.; and Stone, P. 2023a. Llm+ p: Empowering large language models with optimal planning proficiency. *arXiv preprint arXiv:2304.11477*.

Liu, X.; Yu, H.; Zhang, H.; Xu, Y.; Lei, X.; Lai, H.; Gu, Y.; Ding, H.; Men, K.; Yang, K.; et al. 2023b. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*.

Padmakumar, A.; Thomason, J.; Shrivastava, A.; Lange, P.; Narayan-Chen, A.; Gella, S.; Piramuthu, R.; Tur, G.; and Hakkani-Tur, D. 2022. Teach: Task-driven Embodied Agents that Chat. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2017–2025.

Piotrowski, W.; and Perez, A. 2024. Real-World Planning with PDDL+ and Beyond. *arXiv preprint arXiv:2402.11901*.

Sarch, G.; Wu, Y.; Tarr, M. J.; and Fragkiadaki, K. 2023. Open-ended instructable embodied agents with memory-augmented large language models. *arXiv preprint arXiv:2310.15127*.

Shridhar, M.; Thomason, J.; Gordon, D.; Bisk, Y.; Han, W.; Mottaghi, R.; Zettlemoyer, L.; and Fox, D. 2020. Alfred: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10740–10749.

Valmeekam, K.; Marquez, M.; Olmo, A.; Sreedharan, S.; and Kambhampati, S. 2024. Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change. *Advances in Neural Information Processing Systems*, 36.

Wögerer, C.; Bauer, H.; Rooker, M.; Ebenhofer, G.; Rovetta, A.; Robertson, N.; and Pichler, A. 2012. LOCOBOT-low cost toolkit for building robot co-workers in assembly lines. In *Intelligent Robotics and Applications: 5th International Conference, ICIRA 2012, Montreal, Canada, October 3-5, 2012, Proceedings, Part II 5*, 449–459. Springer.