Computer Science and Artificial Intelligence Laboratory

Technical Report

MIT-CSAIL-TR-2011-007 February 1, 2011

Towards Understanding Hierarchical Natural Language Commands for Robotic Navigation and Manipulation

Thomas Kollar, Steven Dickerson, Stefanie Tellex, Ashis Gopal Banerjee, Matthew R. Walter, Seth Teller, Nicholas Roy

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

Towards Understanding Hierarchical Natural Language Commands for Robotic Navigation and Manipulation

Thomas Kollar[1], Steven Dickerson[1], Stefanie Tellex[1],
Ashis Gopal Banerjee, Matthew R. Walter, Seth Teller, and Nicholas Roy
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139
{tkollar, stevend, stefie10, ashis, mwalter, teller, nickroy} @csail.mit.edu

ABSTRACT

We describe a new model for understanding hierarchical natural language commands for robot navigation and manipulation. The model has three components: a semantic structure that captures the hierarchical structure of language; a cost function that maps the command's semantic structure to the robot's sensorimotor capabilities; and an efficient search method for finding the lowest-cost plan. We present a proof-of-concept system that carries out navigation commands in a simulated setting.

1. INTRODUCTION

To be useful teammates to human partners, robots must be able to follow spoken instructions expressed in natural language. An essential class of instructions involves commanding a robot to navigate and manipulate objects within an environment. For example, a human supervisor may issue the following instruction to an autonomous forklift: "Pick up the pallet of boxes from the truck in receiving, and go to the porter in issue" (Fig. 1). A key challenge in understanding these types of commands is capturing the hierarchical structure of the language and mapping each component in this hierarchy onto the robot's representation of the environment. The robot must map phrases such as "the pallet of boxes on the truck" and "the porter in issue" to specific world objects, then plan actions corresponding to phrases like "pick up" based on the grounded objects.

We address this challenge through a new probabilistic model for understanding navigation and manipulation commands. We introduce a semantic structure, called *Extended Spatial Description Clauses* (ESDCs) that captures the hierarchical semantics of natural language commands. The new results presented here build on previous approaches [Kollar et al., 2010] which convert a natural language instruction into a flat sequence of components. Our new semantic structure enables us to understand relationships between components and the hierarchical information inherent in them. We provide an algorithm for grounding ESDCs to entities in the environment, enabling the robot to plan a sequence of actions that satisfies the natural language command.

2. EXTENDED SPATIAL DESCRIPTION CLAUSES (ESDCS)

ESDCs abstract away from the detailed linguistic structure of the language, enabling factorization of the cost function and efficient search for a plan. A single ESDC corresponds to a clause of the linguistic input and consists of three fields: a *figure*, a *relation*, and a list of *landmarks*. Any field can be unspecified. For example, for the command "Pick up the tire pallet," the figure is an implied "you," the relation is "Pick up," and the landmark is "the tire pallet." ESDCs are hierarchical: any ESDC field can be another ESDC. Each ESDC has one of the following types:

- **EVENT** Something that takes place (or should take place) in the world (e.g., "Forklift, stop!", "Pick up the tire pallet").
- **OBJECT** A thing in the world. This category includes people and the forklift as well as physical objects (e.g., "Forklift," "the tire pallet," "the truck," "the person").
- **PLACE** Places in the world (e.g., "on the truck," "next to the tire pallet").
- **PATH** Paths through the world (e.g., "past the truck," "toward receiving").

ESDCs are automatically constructed from the Stanford dependencies, which are extracted using the Stanford Parser [de Marneffe et al., 2006]. ESDCs for the command "Go to the pallet next to the truck" appear in Figure 2.

(a) Robotic Forklift (b) Robotic Porter

Figure 1: Real-world robotic platforms that we are targeting for natural language command and control interfaces.

3. FROM LANGUAGE TO ACTION

In order to find a plan corresponding to the natural language command, we define a cost function over state sequences. The cost function is defined as the negative log probability of the states S, the language (represented as ESDCs), and correspondences between the language and parts of the world, Γ. Our goal is to find the sequence of states

[1]The first three authors contributed equally to this paper.

$EVENT_1(r = \texttt{Go}$
$\quad l = PATH_2(r = \texttt{to},$
$\quad\quad l = OBJECT_3(f = \texttt{the pallet},$
$\quad\quad\quad\quad r = \texttt{next to},$
$\quad\quad\quad\quad l = OBJECT_4(f = \texttt{the trailer}))))$

Figure 2: ESDC tree for "Go to the pallet next to the trailer."

and correspondences that minimize the cost function:

$$\underset{S,\Gamma}{\text{argmin}}\ C(\text{command}, S, \Gamma) \triangleq -log(p(\text{ESDCs}, S, \Gamma)) \quad (1)$$

The ESDC hierarchy for the command defines a graphical model corresponding to a factorization of this cost function. The model contains random variables corresponding to the natural language command and the groundings of those variables in the world as follows:

- λ_i^f The text of the figure field of the i^{th} ESDC.
- λ_i^r The text of the relation field of the i^{th} ESDC.
- λ_i^l The contents of the landmark field of the i^{th} ESDC; if non-empty, always a child ESDC.
- $\gamma_i \in \Gamma$ The grounding associated with the i^{th} ESDC. For EVENTs and PATHs, γ_i is a sequence of robot states. For OBJECTs, γ_i is a specific object in the environment. For PLACEs, γ_i is a specific place in the environment.

Given these variables, the model has the following links:

- $\forall i \quad link(\gamma_i, \lambda_i^f)$
- $\forall i, j \quad \lambda_i^f = \lambda_j^l \implies link(\gamma_i, \gamma_j, \lambda_i^r)$

The first rule connects a physical object such as a pallet to the text describing it, such as "the pallet." The second rule connects a relation, such as "next to," to its arguments. Because the model is built dynamically according to the structure of the language, the framework supports multi-argument verbs such as "Put the tire pallet on the truck" and nested noun phrases such as "the pallet in receiving next to the truck."

Figure 3 gives the model created for the phrase 'to the pallet near the trailer." The ESDCs for this phrase are shown in Figure 2. Here γ_3 and γ_4 correspond to specific objects in the environment that match the corresponding phrase in the ESDC; γ_2 is a path through the environment. This model leads to the following factorization of the cost function:

$$p(\text{ESDCs}, S, \Gamma) = p(ESDCs|S, \Gamma) \times p(S, \Gamma) \quad (2)$$

$$= p(\lambda_1^r|\gamma_2, \gamma_3) \times p(\lambda_3^f|\gamma_3) \times p(\lambda_3^r|\gamma_3, \gamma_4) \times$$
$$p(\lambda_4^f|\gamma_4) \times p(\gamma_2) \times p(\gamma_3) \times p(\gamma_4) \quad (3)$$

We estimate each of these distributions separately from training data, as described in Kollar et al. [2010], and assume a uniform prior.

Once the structure of the model is defined, the system needs to find the assignment of objects and paths to the text that minimizes the total cost. The system searches over the space of possible state sequences, S, as well as bindings for each γ_i in Γ. We use breadth-first beam search with a fixed beam width to bound the number of candidate nodes considered at any particular tree level.

The search combined with the cost function enables the system to infer action sequences that correspond well to hierarchical natural language commands. Figure 4 shows the action identified in response to the command "Go to the pallet next to the trailer."

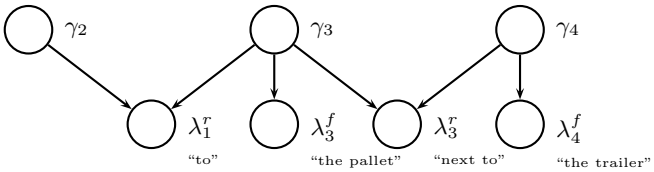

Figure 3: Graphical model for "to the pallet next to the trailer," part of the ESDC tree shown in Figure 2.

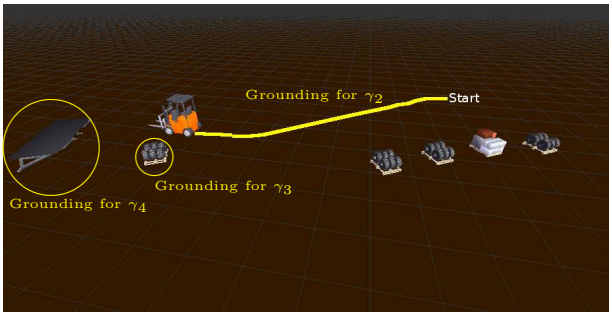

Figure 4: Example execution of the command "Go to the pallet next to the trailer." The system finds the objects corresponding to the phrases "the trailer," "the pallet next to the trailer," and "to the pallet next to the trailer."

4. CONCLUSION

For the generative version of the model described here, models for each factor are trained from separate corpora of examples. Training these models is problematic for verbs of motion such as "pick up" and "move," because suitable corpora are hard to obtain, and because verbs are open-class parts of speech. To address this problem, we are developing a conditional random field (CRF) version of the model which is capable of learning from labeled corpora of commands, rather than models for each word. To train the CRF, we are collecting a corpus of natural language commands paired with robot actions. The corpus currently consists of more than three hundred commands and nearly six thousand words. Commands are quite diverse, varying from short directives such as "Lift the tire pallet," to detailed instructions such as "Back up forklift and turn it around. Move forklift slowly to the pallet with tires on it. With the mechanical fork is underneath the pallet (*sic*), lift the pallet up." We will use this corpus with a CRF to train the system to carry out navigation and manipulation commands without requiring separate models to be learned for each factor. This approach should result in a system that can robustly handle a wide variety of commands.

This paper presents a novel semantic structure and decomposition of the cost function in order to ground natural language commands for navigation and manipulation tasks. We demonstrate a proof-of-concept version of our framework which uses hierarchical structures in the language to resolve objects and infer a plan corresponding to the command.

References

M. de Marneffe, B. MacCartney, and C. Manning. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC-06*, pages 449–454, Genoa, Italy, 2006.

T. Kollar, S. Tellex, D. Roy, and N. Roy. Toward understanding natural language directions. In *Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International Conference on*, pages 259–266, Osaka, Japan, 2010.