Computer Science and Artificial Intelligence Laboratory

Technical Report

MIT-CSAIL-TR-2012-002 January 23, 2012

Toward a Probabilistic Approach to Acquiring Information from Human Partners Using Language

Stefanie Tellex, Pratiksha Thaker, Robin Deits , Dimitar Simeonov, Thomas Kollar, and Nicholas Roy

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

Toward a Probabilistic Approach to Acquiring Information from Human Partners Using Language

Stefanie Tellex[a] Pratiksha Thaker[a] Robin Deits[b] Dimitar Simeonov[a] Thomas Kollar[c] Nicholas Roy[a]
[a]Massachusetts Institute of Technology [b]Battelle [c]Carnegie Mellon University

Categories and Subject Descriptors

I.2.9 [**Computing Methodologies**]: Artificial Intelligence

Keywords

dialog, robotics, question-asking

1. INTRODUCTION

Our aim is to make robots that can naturally and flexibly interact with a human partner via natural language. Understanding natural language commands is a challenging problem because of the highly variable nature of unstructured linguistic input, and the use of ambiguous referring expressions that do not have a mapping to a unique object in the external world. When humans encounter this ambiguity in dialog with each other, a key strategy is to ask questions in order to better understand the situation. These questions are a crucial mechanism to clarify misunderstandings, explain problems, and gather more information. For example, if a person was asked to "Put the pallet on the truck," they might ask "Which one?" or "Do you mean the pallet on my right?" A response might be "Yes, the tire pallet" or "the red one near the truck."

In order for a robot to acquire information through dialog, it must have two capabilities: first, it must create a question that elicits useful information from the human partner, and second, it must exploit that information in order to infer better actions. These two problems are related, because what question to choose depends both on what the robot is uncertain about, and also on what types of answers the robot is likely to understand. This paper focuses on the second problem: how to understand answers to questions, leaving question generation to future work.

Understanding a human teammate's answers is difficult because of the large variety of responses the human could produce in response to the robot's questions. Previous approaches to robotic dialog understanding use a POMDP [Doshi and Roy, 2008, Rosenthal et al., 2011] with a fixed state/action space and a limited space of actions in response to the human's input. However, natural language is a powerful, compositional modality for expressing complex ideas that the robot may never have encountered during training. In our previous work [Tellex et al., 2011], we addressed this problem by introducing the G^3 framework, which converts a natural language command into a probabilistic graphical model or *grounding graph* that factors according to the compositional and hierarchical structure of language. Random variables in the model correspond to linguistic constituents in the command; inference in the model corresponds to finding *groundings*, or objects, places, paths, or events in the external world that map to linguistic constituents in the command.

HRI'12, March 5–8, 2012, Boston, Massachusetts, USA.
ACM 978-1-4503-1063-5/12/03.

```
Lift the tire pallet in the air, then proceed to deposit it to the right of the tire
pallet already on the table right in front of you.

Place a second pallet of tires on the trailer.
```

Figure 2: Sample commands from the corpus.

In this paper we describe an approach to asking questions that leverages the G^3 framework. Given a natural language command, a question, and an answer, the robot extracts grounding graphs for each linguistic construct. Then it finds linguistic constituents in the natural language command that refer to the same grounding in the external world, or that corefer. This problem, called *coreference resolution*, is a well-studied problem in computational linguistics [Jurafsky and Martin, 2008, section 18.1]. Next it merges variables in the graphs based on inferred coreference. Finally, it performs inference in the merged model, enabling it to infer the best set of groundings corresponding to the command, the question, and the answer.

Our approach is applicable not only to questions and answers, but also to understanding a sequence of commands that use coreference. For example, one command from our corpus is "Go to the second crate on the right. Pick it up and place it beside lonely crate." In order to pick up the correct pallet, the robot can use linguistic coreference to infer that "it" refers to "the second crate on the right," and then perform joint inference to find an action that corresponds to picking up that crate. An earlier version of this work appeared in Simeonov et al. [2011]; this work is an extension with evaluation results.

2. TECHNICAL APPROACH

When faced with a command, a question, and an answer, the system extracts grounding graphs from the spatial language input, finds coreferences, merges variables in the graphs according to the coreference information, and finally performs inference in the merged graph. To train and evaluate our system, we used the aligned parallel corpus of language paired with robot action described in Tellex et al. [2011]. This corpus consists of natural language commands paired with robot actions. Sample commands from the corpus appear in Figure 2.

Resolving linguistic coreferences involves identifying linguistic constituents that refer to the same entity in the external world. Although there are several existing software packages to address this problem, most are developed for large corpora of newspaper articles and generalize poorly to language in our corpus. Instead, we created a coreference system which is trained on language from our corpus. Following typical approaches to coreference [Stoyanov et al.,

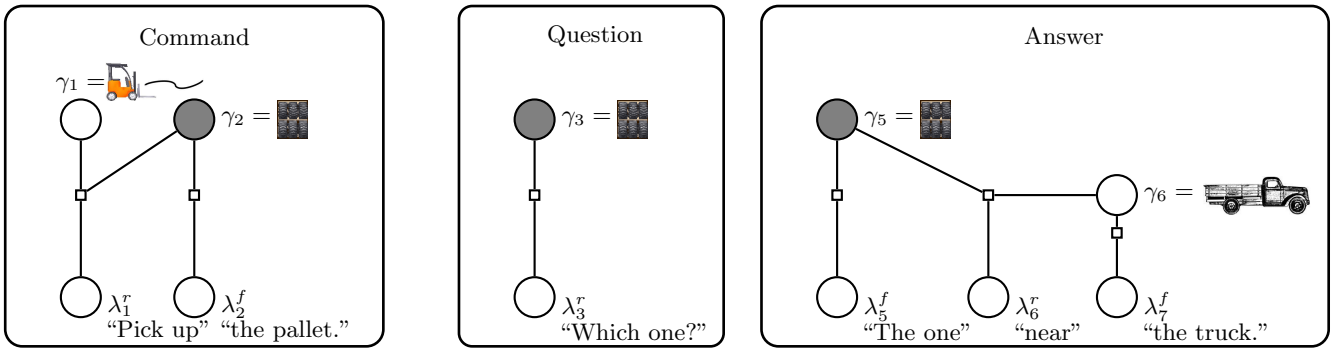

Figure 1: Grounding graphs for a three-turn dialog. The robot merges the three shaded variables.

2010], our system consists of a classifier to predict pairwise coreference, combined with a clustering algorithm that enforces transitivity. For the pair-wise classifier we used a log-linear model which is trained on the alignment annotations in our corpus: two constituents are coreferent if they are annotated with the same grounding in the external world. The model uses bag-of-words features. Since the pair-wise classifier does not guarantee that predicted coreferences are transitive, we post-process the classification results by taking the transitive closure of all links. On a corpus of commands, using ground-truth parses, our algorithm achieves an F-score of 0.889 on a held-out test set, and it achieves an F-score of 0.793 on a dialog corpus of commands paired with questions and answers.

Once coreferring variables have been identified, a merging algorithm creates a single unified grounding graph. Figure 1 shows a merged graph created from a command, a question, and an answer by our system. The λ variables correspond to language; the γ variables correspond to groundings in the external world, and the ϕ variables are $True$ if the groundings correspond to the language, and $False$ otherwise. For more information, see Tellex et al. [2011].

3. RESULTS

To evaluate the system, we used a subset of the corpus of natural language commands paired with robot actions described by Tellex et al. [2011], consisting of commands in which coreference is relevant. To focus the evaluation on the impact of the coreference algorithm, we used ground-truth parses. Within the test set, we see an increase in object grounding accuracy with merging. Inference using merging resulted in a total of 31 grounding variables having different inferred values, compared to not using merging. Of these, 13 were correct and resulted in changes from an incorrect grounding to a correct grounding. Many of these commands involved anaphoric expressions such as "Place it on the truck," in which coreference resolution enabled the system to find the correct grounding for the pronoun "it." Of the other nodes, many involved changes from one incorrect grounding to another and ambiguous noun phrases such as "the tire pallet" which could refer to more than one object, even while incorporating information from the merging algorithm. The robot could disambiguate these by asking questions. Others involved genuine errors, such as confusing "left" and "right," due to limitations in the system's semantic models.

We have also demonstrated the system end-to-end on an example dialog involving a question and an answer:

- **Person** "Pick up the pallet."
- **Robot** "Which one?"
- **Person** "The one that has boxes."

The system is able to parse the language, automatically extract coreference, and create a merged graph. By performing inference in the merged graph, it infers that the person was referring to a pallet of boxes, and not a tire pallet, which happened to be a slightly more likely grounding for the phrase "the pallet."

4. CONCLUSION

In this paper we presented preliminary results for a robot dialog understanding system that is able to ask the human user targeted questions and incorporate the inference into a probabilistic graphical model that factors according to the structure of language. We demonstrated that our framework is able to use information from commands, questions, and answers in order to infer more accurate actions from a corpus of realistic dialog collected on Amazon Mechanical Turk.

Our immediate next steps are to implement an algorithm for asking a question based on the entropy of variables in the grounding graph. By generating questions dynamically based on the robot's uncertainty about the mapping between the command and the external world, it can collect exactly the information it needs from the human user, maximizing the value of each interaction.

Our larger vision is to extend the framework to support question answering, enabling the robot answer questions about why it is acting in a certain way. Furthermore, this paper focuses on relatively short clarification dialogs. Supporting multi-turn dialogs in which complex activities are designed remains a challenging problem which requires more complicated discourse-level semantic structures, that represent the conversation at a more abstract level.

5. ACKNOWLEDGEMENTS

This work was sponsored under the Robotics Collaborative Technology Alliance by the U.S Army Research Laboratory, the Office of Naval Research under MURI N00014-07-1-0749, and Battelle. Their support is gratefully acknowledged.

References

F. Doshi and N. Roy. Spoken language interaction with model uncertainty: An adaptive human-robot interaction system. *Connection Science*, 20(4):299–319, 2008.

D. Jurafsky and J. H. Martin. *Speech and Language Processing*. Pearson Prentice Hall, 2 edition, May 2008. ISBN 0131873210.

S. Rosenthal, M. Veloso, and A. K. Dey. Learning accuracy and availability of humans who help mobile robots. In *Proc. AAAI*, 2011.

D. Simeonov, S. Tellex, T. Kollar, and N. Roy. Toward interpreting spatial language discourse with grounding graphs. In *2011 RSS Workshop on Grounding Human-Robot Dialog for Spatial Tasks*, 2011.

V. Stoyanov, C. Cardie, N. Gilbert, E. Riloff, D. Buttler, and D. Hysom. Reconcile: A coreference resolution research platform. Technical report, Cornell University, 2010.

S. Tellex, T. Kollar, S. Dickerson, M. Walter, A. Banerjee, S. Teller, and N. Roy. Understanding natural language commands for robotic navigation and mobile manipulation. In *Proc. AAAI*, 2011.