Asking for Help Using Inverse Semantics

Stefanie Tellex[1,2], Ross A. Knepper[1,3], Adrian Li[4], Daniela Rus[3], and Nicholas Roy[3]

***Abstract*—Robots inevitably fail, often without the ability to recover autonomously. We demonstrate an approach for enabling a robot to recover from failures by communicating its need for specific help to a human partner using natural language. Our approach automatically detects failures, then generates targeted spoken-language requests for help such as "Please give me the white table leg that is on the black table." Once the human partner has repaired the failure condition, the system resumes full autonomy. We present a novel *inverse semantics* algorithm for generating effective help requests. In contrast to forward semantic models that interpret natural language in terms of robot actions and perception, our inverse semantics algorithm generates requests by emulating the human's ability to interpret a request using the Generalized Grounding Graph (G^3) framework. To assess the effectiveness of our approach, we present a corpus-based online evaluation, as well as an end-to-end user study, demonstrating that our approach increases the effectiveness of human interventions compared to static requests for help.**

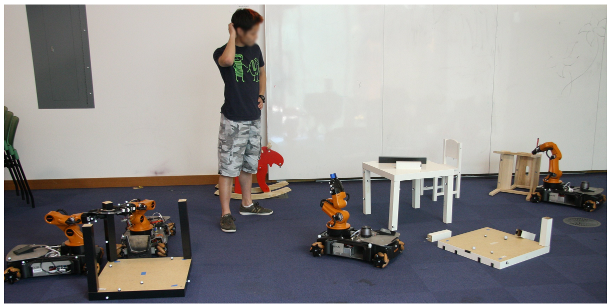


Fig. 1. A robot engaged in assembling an IKEA LACK table requests help using natural language. A vague request such as "Help me" is challenging for a person to understand. Instead, this paper presents an approach for generating targeted requests such as "Please hand me the black table leg that is on the white table."

I. Introduction

Robotic capabilities such as robust manipulation, accurate perception, and fast planning algorithms have led to recent successes such as robots that can fold laundry [15], cook dinner [1], and assemble furniture [11]. However, when robots execute these tasks autonomously, failures often occur, for example failing to pick up an object due to perceptual ambiguity or an inaccurate grasp. A key aim of current research is reducing the incidence of these types of failures, but eliminating them completely remains an elusive goal.

When failures occur, a human can often intervene to help a robot recover. If the human is familiar with the robot, its task, and its common failure modes, then they can provide this help without an explicit request from the robot. However, if a person is unfamiliar with the robotic system, they might not know how to help the robot recover from a failure. This situation will occur frequently when robots interact with untrained users in the home. Moreover, even trained users who are deeply familiar with the robot's capabilities may experience problems during times of high cognitive load, such as a human supervising a large team of robots on a factory floor.

We propose an alternative approach to recovering from the inevitable failures which occur when robots execute complex tasks in real-world environments: when the robot encounters failure, it verbally requests help from a human partner. After receiving help, it resumes autonomous task execution. The contribution of our paper is a family of algorithms for formulating a pithy natural language request so that a human without situational awareness can render appropriate aid.

Generating natural language requests for a robot is challenging because the robot must map from perceptual aspects of the environment to words that a person will understand, in novel environments that are not available at training time. Template-based methods for generating requests do not take into account the ability of a person to understand the request, while existing work in referring expression generation assumes access to a symbolic representation of the environment, including ambiguous spatial relations such as "near" or "under" which may not be directly computed from the robot's perceptual system [13]. We propose an algorithm that addresses this problem by searching for an utterance that maximizes the probability of a correspondence between the words in the language and the action the robot desires the human to perform, making use of a probabilistic model of a person's language understanding faculty [22]. When understanding language, the robot maps from linguistic symbols to low-level motor actions and perceptual features that the robot encounters in the environment. In this paper, we invert that model, mapping from a desired low-level motor action that the robot would like the human to execute to a linguistic description. Inverting the model requires developing a novel algorithm for generating help requests and adapting the request to the specific environment. By modeling the probability of a human misinterpreting the request, the robot is able to generate targeted requests that humans follow more quickly and accurately compared to baselines involving either generic requests (e.g., "Help me") or template-based non-context-specific requests (e.g., "Hand me the <part> ").

As a test domain, we focus on a human-robot team assembling IKEA furniture, shown in Figure 1. We evaluate our approach using a corpus-based experiment with Amazon Me-

[1]The first two authors contributed equally to this paper.
[2]Brown University, stefie10@cs.brown.edu.
[3]Massachusetts Institute of Technology, {rak, rus, nickroy}@mit.edu.
[4]Cambridge University, adrlhli@gmail.com.

chanical Turk as well as a real-world user study. The corpus-based approach allows us to efficiently test the performance of different algorithms. The user study assesses whether we have met our engineering goals in the context of an end-to-end system. Our evaluation demonstrates that the inverse semantics language generation system improves the speed and accuracy of a human's intervention when a human-robot team is engaged in a furniture assembly task and also improves the human's subjective perception of their robotic teammates.

II. Related Work

Traditional methods for generating language rely on a dedicated language-generation system that is not integrated with a language-understanding framework [10, 17]. These approaches typically consist of a sentence planner combined with a surface realizer to guide decision making of what to say, but contain no principled model of how an instruction-follower would comprehend the instruction [21, 7, 2, 19]. These models assume access to a formal symbolic representation of the object to be referenced and its properties and relations to other options [13]. Our approach differs in that it generates language by inverting a module for language understanding. This inversion is required to generate physically grounded language about a particular environment that maps to the robot's perceptual data and communicates it to a human partner. Furthermore, our approach generates references to objects, as well as actions, unlike much previous work which focuses on objects and sets of objects.

Some previous work has approached the generation problem by inverting a semantics model. Golland et al. [8] use a game-theoretic approach combined with a semantics model to generate referring expressions. Our approach, in contrast, uses probabilistic grounded semantics, biasing the algorithm toward shorter sentences unless a longer, more descriptive utterance is unambiguous. Goodman and Stuhlmüller [9] describe a rational speech-act theory of language understanding, where the speaker chooses actions that maximize expected global utility. Similarly, recent work has used Dec-POMDPs to model implicatures and pragmatics in language-using agents [23, 24] but without focusing on grounded, situated language as in this paper, and without implementing an end-to-end robotic system. There is a deep connection between our models and the notion of legibility and predictability for grasping and pointing, as defined by Dragan and Srinivasa [4], pointing toward a unified framework for grounded communication using language and gesture.

Our approach views the language generation problem as inverse language understanding; a large body of work focuses on language understanding for robots [14, 5, 12, 16]. Of these previous approach, we chose to invert the G^3 framework because it is a probabilistic framework which explicitly models the mapping between words in language and aspects of the external world, so metrics based on entropy may be used to assess the quality of generated utterances.

Cooperative human-robot activities, including assembly, have been broadly studied [25, 20, 3, 6]. These architectures

```
function conditions_satisfied(ℓ – list of conditions)
 1: $q \leftarrow$ World state
 2: for all $c \in \ell$ do
 3:     if $c$ not satisfied in q then
 4:         $a \leftarrow$ generate_remedy_action($c$)          ▷ See Section III-B
 5:         generate_help_request($a$)                          ▷ See Section IV
 6:         while $c$ not satisfied do
 7:             if time $\geq$ 60 then                          ▷ wait up to 60 seconds
 8:                 return false
 9: return true
function executive($g$ – goal state)
 1: repeat
 2:     $p \leftarrow$ symbolic_plan($g$)                       ▷ $p$ – list of actions
 3:     $f \leftarrow$ true                                     ▷ $f$ – are we finished?
 4:     while $p \neq \emptyset$ do
 5:         $s \leftarrow p[0]$                                 ▷ first plan step
 6:         if conditions_satisfied($s$.preconditions) then
 7:             $s$.execute()
 8:             if not conditions_satisfied($s$.postconditions) then
 9:                 $f \leftarrow$ false
10:         else
11:             $f \leftarrow$ false
12:         $p$.retire($s$)                                     ▷ $s$ succeeded; remove it from $p$
13: until $f$                                                   ▷ no actions failed
```

Fig. 2. An executive algorithm generates robot actions and help requests.

permit various granularities of human intervention through a sliding autonomy framework. A failure triggers the replay of video of the events preceding failure, from which the human must obtain situational awareness. By contrast, in our approach the robot diagnoses the failure and leverages natural language to convey to the user exactly how the problem should be resolved.

III. Assembling Furniture

Our assembly system consists of a team of KUKA youBots, which collaborate to assemble IKEA furniture, originally described in Knepper et al. [11]. The robots receive assembly instructions encoded in a STRIPS-style planning language. A centralized executive takes as input the symbolic plan and executes each plan step in sequence. Each symbolic action corresponds to a manipulation or perception action to be performed by one or two robots. Assembling an IKEA LACK table requires constructing and executing a 48-step plan. All of the steps are autonomously executable under the right conditions, and the team can assemble the table in approximately ten minutes when no failures occur. Since perception is not a focus of this paper, we employ a VICON motion capture system to track the location of each participating robot, human and furniture part during the assembly process. In our experiments, failures occurred at a rate of roughly one every two minutes, mostly due to mechanical problems, such as grasping failures, perceptual issues, such as a part not being visible on VICON, and planning failures, such as a robot failing to find a path to reach its goal due to obstacles. When the robots detect a failure, one of the robots requests help using one of the approaches described in Section IV. Figure 2 shows the algorithm used to control the robots and request help.

TABLE I
SUMMARY OF COMMON FAILURES AND THE SYMBOLIC GROUNDINGS USED TO FORM A HELP REQUEST.

Failed symbolic condition	Symbolic request
Part is not visible to the robot.	locate_part(*robot, part*)
Robot is not holding the part.	give_part(*robot, part*)
Leg is not aligned with the hole.	align_with_hole(*leg, top, hole*)
Leg is not attached to the hole.	screw_in_leg(*leg, top, hole*)
Table top is not upside down.	flip(*top*)
Legacy software is in infinite loop.	<not detectable>
Risk of hardware damage.	<not detectable>

$$
\begin{aligned}
S &\rightarrow VB \quad NP \\
S &\rightarrow VB \quad NP \quad PP \\
PP &\rightarrow TO \quad NP \\
VB &\rightarrow \text{flip}|\text{give}|\text{pick up}|\text{place} \\
NP &\rightarrow \begin{array}{l}\text{the white leg}|\text{the black leg}|\text{me} \\ \text{the white table}|\text{the black table}\end{array} \\
TO &\rightarrow \text{above}|\text{by}|\text{near}|\text{under}|\text{with}
\end{aligned}
$$

Fig. 3. Part of the context-free grammar defining the linguistic search space.

A. Detecting Failures

To detect failures, the system compares the expected state of the world to the actual state, as sensed by the perceptual system (line 6 of the `executive` function). We represent the state, q, as a vector of values for logical predicates. Elements of the state for the IKEA LACK table include whether the robot is holding each table leg, whether the table is face-up or face-down, and whether each leg is attached to the table. In the furniture assembly domain, we compute the state using the tracked pose of every rigid body known to the VICON system, including each furniture part, each robot chassis and hand, and each human. The system recomputes q frequently, since it may change independently of any deliberate robot action, such as by human intervention or from an unintended side-effect.

Prior to executing each action, the assembly executive verifies the action's preconditions against q. Likewise, following each action, the postconditions are verified. Any unsatisfied condition indicates a failure and triggers the assembly executive to pause the assembly process and initiate error recovery. For example, the robot must be grasping a table leg before screwing it into the hole. If it tries and fails to pick up a leg, then the post-condition for the "pick up" action will not be satisfied in q, which indicates a failure.

B. Recovery Strategy

When a failure occurs, its description takes the form of an unsatisfied condition. The system then asks the human for help to address the problem. The robot first computes actions that, if performed by the human, would resolve the failure and enable the robotic team to continue assembling the piece autonomously. The system computes these actions using a pre-specified model of physical actions a person could take to rectify failed preconditions. Remedy requests are expressed in a simple symbolic language, shown in Table I. This symbolic request, a, specifies the action that the robot would like the person to take to help it recover from failures. However these symbolic forms are not appropriate for speaking to an untrained user. In the following section, we explore a series of approaches that take as input the symbolic request for help and generate a language expression asking a human for assistance.

IV. Asking for Help From a Human Partner

Once the system computes a symbolic representation of the desired action, a, it searches for words, Λ, which effectively communicate this action to a person in the particular environmental context, M, on line 5 of the `conditions_satisfied` function. This section describes various approaches to the `generate_help_request` function which carries out this inference. Formally, we define a function h to score candidate sentences:

$$\underset{\Lambda}{\operatorname{argmax}}\ h(\Lambda, a, M) \qquad (1)$$

The specific function h used in Equation 1 will greatly affect the results. We define three increasingly complex approaches for h, which lead to more targeted natural language requests for help by modeling the ability of the listener to understand it. The contribution of this paper is a definition for h using *inverse semantics*. Forward semantics is the problem of mapping from words in language to aspects of the external world; the canonical problem is enabling a robot to follow a person's natural language commands [14, 12, 22, 16]. Inverse semantics is the reverse: mapping from specific aspects of the external world (in this case, an action that the robot would like the human to take) to words in language. To apply this approach we use the G^3 model of natural language semantics. We build on the work of Tellex et al. [22], who used the G^3 framework to endow the robot with the ability to follow natural language commands given by people. In this paper, instead, we invert the model, to endow the robot with the ability to create natural language requests, which will be understood by people.

The inference process in Equation 1 is a search over possible sentences Λ. We define a space of sentences using a context-free grammar (CFG), shown in Figure 3. The inference procedure creates a grounding graph for each candidate sentence using the parse structure derived from the CFG and then scores it according to the function h. This search space is quite large, and we use greedy search to expand promising nodes first.

A. Speaking by Reflex

The simplest approach from the assembly executive's perspective is to delegate diagnosis and solution of the problem to the human with the simple fixed request, Λ = "Help me." This algorithm takes into account neither the environment or the listener when choosing what to say. We refer to this algorithm as S_0.

B. *Speaking by Template*

As a second baseline, we implemented a template-based algorithm, following traditional approaches to generating language [6, 17]. This approach uses a lookup table to map symbolic help conditions to natural language requests. These generic requests take the following form:

- "Place part 2 where I can see it."
- "Hand me part 2."
- "Attach part 2 at location 1 on part 5." (i.e. screw in a table leg)

Note that the use of first person in these expressions refers to the robot. Since VICON does not possess any semantic qualities of the parts, they are referred to generically by part identifier numbers. Such templates can be effective in simple situations, where the human can infer the part from the context, but do not model how words map to the environment, and thus do not reflect the mapping between words and perceptual data. In constrained interaction scenarios, the programmer could hard-code object names for each part, but this approach becomes impractical as the scope of interaction increases, especially for referring expressions such as "the part on the table."

C. *Modeling Word Meanings*

This section briefly describes how the G^3 framework models word meanings, which has been previously used to understand language [22]. When understanding language, the G^3 framework imposes a distribution over *groundings* in the external world, $\gamma_1 \dots \gamma_N$, given a natural language sentence Λ. Groundings are the specific physical concepts that are referred to by the language and can be objects (e.g., a table leg or a robot), places (e.g., a particular location in the world), paths (e.g., a trajectory through the environment), or events (e.g., a sequence of actions taken by the robot). Each grounding corresponds to a particular constituent $\lambda_i \in \Lambda$, defined by the CFG parse tree for the sentence. For example, for a sentence such as "Pick up the table leg," the grounding for the phrase "the table leg" corresponds to an actual table leg in the external world, and the grounding for the entire sentence corresponds to the actions of a person as they follow the request. Understanding a sentence in the G^3 framework amounts to the following inference problem:

$$\underset{\gamma_1 \dots \gamma_N}{\operatorname{argmax}}\ p(\gamma_1 \dots \gamma_N | \Lambda, M) \tag{2}$$

The environment model M consists of the robot's location along with the locations and geometries of objects in the external world. The computed environment model defines a space of possible values for the grounding variables, $\gamma_1 \dots \gamma_N$. A robot computes the environment model using sensor input; in the domain of furniture assembly, the system creates the environment model using input from VICON.

To factor the model, we introduce a correspondence vector, Φ, as do Tellex et al. [22]. Each entry $\phi_i \in \Phi$ corresponds to whether linguistic constituent $\lambda_i \in \Lambda$ corresponds to the groundings associated with that constituent. For example, the correspondence variable would be $True$ for the phrase "the white table leg" and a grounding of a white leg, and $False$ if the grounding was a different object, such as a black table top. We assume that $\gamma_1 \dots \gamma_N$ are independent of Λ unless Φ is known. Introducing Φ enables factorization according to the structure of language with local normalization at each factor over a space of just the two possible values for ϕ_i.

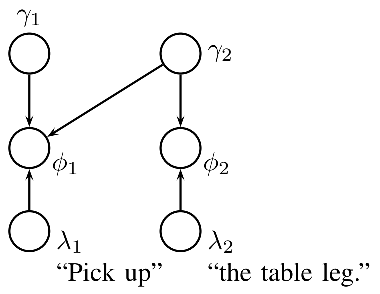


Fig. 4. Grounding graph for the request, "Pick up the table leg." Random variables and edges are created in the graphical model for each constituent in the parse tree. The λ variables correspond to language; the γ variables correspond to groundings in the external world. Edges in the graph are created according to the parse structure of the sentence.

The optimization then becomes:

$$\underset{\gamma_1 \dots \gamma_N}{\operatorname{argmax}}\ p(\gamma_1 \dots \gamma_N | \Lambda, \Phi, M) \tag{3}$$

After factoring using Bayes' rule and ignoring constant terms we have:

$$\underset{\gamma_1 \dots \gamma_N}{\operatorname{argmax}}\ p(\Phi | \Lambda, \gamma_1 \dots \gamma_N, M) \tag{4}$$

We factor the expression according to the compositional syntactic structure of the language Λ, defined by the parse tree.

$$\underset{\gamma_1 \dots \gamma_N}{\operatorname{argmax}}\ \prod_i p(\phi_i | \lambda_i, \gamma_{i_1} \dots \gamma_{i_k}, M) \tag{5}$$

This factorization can be represented as a directed graphical model where random variables and edges in the model are created according to the structure of the language. We refer to one of these graphical models as a *grounding graph*. Figure 4 shows an example graphical model; the details of the factorization are described by Tellex et al. [22]. The system factorizes the distribution according to the well-known hierarchical parse structure of language. When evaluating the model for specific sentences generated by our CFG, we use the parse structure defined by the CFG to factor the distribution. Each factor corresponds to an individual log-linear model for predicting whether the particular node of the CFG corresponds to a particular grounding in the external world. Training model parameters requires an aligned parallel corpus of language paired with groundings; we describe the training procedure used for our furniture assembly domain in Section IV-F.

D. *Speaking by Modeling the Environment*

Next, we describe a more complex model for speaking, that takes into account a model of the environment, but not a model of the listener. We compute this model using the G^3 framework. The system converts the symbolic action request a to a value for the action grounding variable, $\gamma_a \in \Gamma$. This variable, γ_a, corresponds to the entire sentence; we refer to the desired value of γ_a as γ_a^*. It then searches for the most likely sentence Λ according to the semantics model. Equation 1 becomes:

$$\operatorname*{argmax}_{\Lambda} h(\Lambda, \gamma_a^*, M) \tag{6}$$

To speak using a model of the environment, the robot searches for language that best matches the action that the robot would like the human to take. It does not consider other actions or groundings in any way when making this determination. Formally:

$$h(\Lambda, \gamma_a^*, M) = \max_{\Gamma|\gamma_a=\gamma_a^*} p(\Lambda|\Gamma, M) \tag{7}$$

With the correspondence variable, this function is equivalent to:

$$h(\Lambda, \gamma_a^*, M) = \max_{\Gamma|\gamma_a=\gamma_a^*} p(\Phi|\Lambda, \Gamma, M) \tag{8}$$

We refer to this metric as S_1. Note that while the optimization in Equation 8 does assess the quality of the generated sentences, it does not actually model the listener's understanding of those sentences, because it only considers those interpretations where the overall action γ_a matches the desired action γ_a^*. The speaker considers possible sentences and evaluates them using the G^3 framework, selecting the language that best matches the desired action γ_a^* in the environment.

This optimization considers sentences such as "Pick up the black table leg on the white table" even though the grounding for "the white table" is not specified in the symbolic description of the request, because it is added as a disambiguating expression. We handle this by searching over both sentences and paired groundings using beam search.

E. *Speaking by Modeling the Listener and the Environment*

The previous S_1 metric scores shorter, ambiguous phrases more highly than longer, more descriptive phrases. For example, "the white leg" will always have a higher score than "the white leg on the black table" because the corresponding grounding graph for the longer sentence is identical to the shorter one except for an additional factor, which causes the overall probability for the more complex graph to be lower (or at most equal).

However, suppose the robot team needs a specific leg; for example, in Figure 5, the robots might need specifically the leg that is on the black table. In this case, a phrase produced under the S_1 metric such as "Hand me the white leg" will be ambiguous: the person will not know which leg to give to the robot because there are several legs in the environment. If the robot instead said, "Please hand me the white leg that is on the black table," then the person will know exactly which leg to give to the robot.

To address this problem, we augment our robot with a model of the listener's ability to understand a request in the particular environment. Rather than optimizing how well the language in the request fits the desired action, we minimize the uncertainty a listener would experience when using the G^3 model to interpret the request. This metric, which we refer to as S_2, explicitly measures the probability that the listener will correctly understand the requested action γ_a^*:

$$h(\Lambda, \gamma_a^*, M) = p(\gamma_a = \gamma_a^*|\Phi, \Lambda, M) \tag{9}$$

To compute this metric, we marginalize over values of Γ, where $\gamma_a = \gamma_a^*$:

$$h(\Lambda, \gamma_a^*, M) = \sum_{\Gamma|\gamma_a=\gamma_a^*} p(\Gamma|\Phi, \Lambda, M) \tag{10}$$

We factor the model with Bayes' rule:

$$h(\Lambda, \gamma_a^*, M) = \sum_{\Gamma|\gamma_a=\gamma_a^*} \frac{p(\Phi|\Gamma, \Lambda, M)p(\Gamma|\Lambda, M)}{p(\Phi|\Lambda, M)} \tag{11}$$

We rewrite the denominator as a marginalization and conditional distribution on Γ':

$$h(\Lambda, \gamma_a^*, M) = \sum_{\Gamma|\gamma_a=\gamma_a^*} \frac{p(\Phi|\Gamma, \Lambda, M)p(\Gamma|\Lambda, M)}{\sum_{\Gamma'} p(\Phi|\Gamma', \Lambda, M)p(\Gamma'|\Lambda, M)} \tag{12}$$

The denominator is constant so we can move the summation to the numerator:

$$h(\Lambda, \gamma_a^*, M) = \frac{\sum_{\Gamma|\gamma_a=\gamma_a^*} p(\Phi|\Gamma, \Lambda, M)p(\Gamma|\Lambda, M)}{\sum_{\Gamma'} p(\Phi|\Gamma', \Lambda, M)p(\Gamma'|\Lambda, M)} \tag{13}$$

Next we assume that $p(\Gamma|\Lambda, M)$ is a constant, K, for all Γ, so it can move outside the summation. This term is constant because Γ and Λ are independent when we do not know Φ:

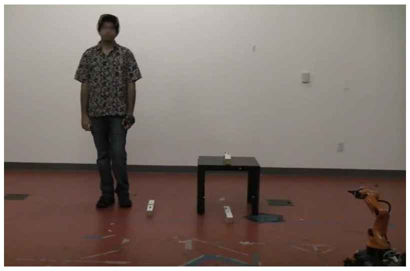
$$h(\Lambda, \gamma_a^*, M) = \frac{K \times \sum_{\Gamma|\gamma_a=\gamma_a^*} p(\Phi|\Gamma, \Lambda, M)}{K \times \sum_{\Gamma'} p(\Phi|\Gamma', \Lambda, M)} \tag{14}$$

The constant K cancels, yielding:

$$h(\Lambda, \gamma_a^*, M) = \frac{\sum_{\Gamma|\gamma_a=\gamma_a^*} p(\Phi|\Gamma, \Lambda, M)}{\sum_{\Gamma'} p(\Phi|\Gamma', \Lambda, M)} \tag{15}$$

This equation expresses the S_2 metric. Note that unlike the S_1 metric, the S_2 metric actually uses the G^3 framework to understand the sentences that it generates: computing the denominator in Equation 15 is equivalent to the problem of understanding the language in the particular environment because the system must assess the mapping between the language Λ and the groundings Γ' for all possible values for the groundings (compare with Equation 4).

In practice, evaluating the S_2 metric is expensive because of the language understanding computation in the denominator. We therefore consider only the best k sentences produced by the S_1 metric, and then re-evaluate them using the S_2 metric. This optimization may remove candidate sentences that the S_1 sentence scores with a low score, but also removes many obviously incorrect sentences and significantly increases overall inference speed.



"Help me" (S_0)	"Help me."
Templates	"Please hand me part 2."
G^3 S_1	"Give me the white leg."
G^3 S_2	"Give me the white leg that is on the black table."
Hand-written	"Take the table leg that is on the table and place it in the robot's hand."

Fig. 5. Scene from our dataset and the requests generated by each approach.

F. *Training*

We trained the model for understanding language following the same procedure as Tellex et al. [22]. We collected a new dataset of natural language requests given by a human to another human in the furniture assembly domain. We created twenty-one videos of a person executing a task involved in assembling a piece of furniture. For example, one video shows a person screwing a table leg into a table, and another shows a person handing a table leg to a second person. Each video has an associated context consisting of the locations, geometries, and trajectories of the people and objects, produced with VICON. We asked annotators on Amazon Mechanical Turk to view the videos and write a natural language request they would give to ask one of the people to carry out the action depicted in the video. Then we annotated requests in the video with associated groundings in the VICON data. The corpus contains 326 requests with a total of 3279 words. In addition we generated additional positive and negative examples for the specific words in our context-free grammar.

V. Evaluation

The goal of our evaluation was to assess whether our algorithms increase the effectiveness of a person's help, or in other words, to enable them to more quickly and accurately provide help to the robot. To evaluate whether our algorithms enable a human to accurately provide help compared to baselines, we use an online corpus-based evaluation. We conducted a real-world user study to assess whether our leading algorithm improves the speed and accuracy of a person's help to a team of autonomous robots engaged in a real-world assembly task.

A. *Corpus-Based Evaluation*

Our online evaluation used Amazon Mechanical Turk (AMT) to measure whether people could use generated help requests to infer the action that the robot was asking them to perform. We presented a worker on AMT with a picture of a scene, showing a robot, a person, and various pieces of furniture, together with the text of the robot's request for help. Figure 5 shows an example initial scene, with several different requests for help generated by different algorithms, all asking the human to carry out the same action. Next, we showed the worker five videos of a human taking various actions in the scene in response to the requests. We asked them to choose the video that best matched the request for help. We chose actions to film based on actions that would recover from typical failures that the robots might encounter. A trial consists of a worker viewing an initial scene paired with a request for help and then choosing a corresponding video.

TABLE II
Fraction of Correctly Followed Requests

Metric	% Success	95% Confidence
Chance	20.0	
"Help me" Baseline (S_0)	21.0	± 8.0
Template Baseline	47.0	± 5.7
G^3 Inverse Semantics with S_1	52.3	± 5.7
G^3 Inverse Semantics with S_2	64.3	± 5.4
Hand-Written Requests	94.0	± 4.7

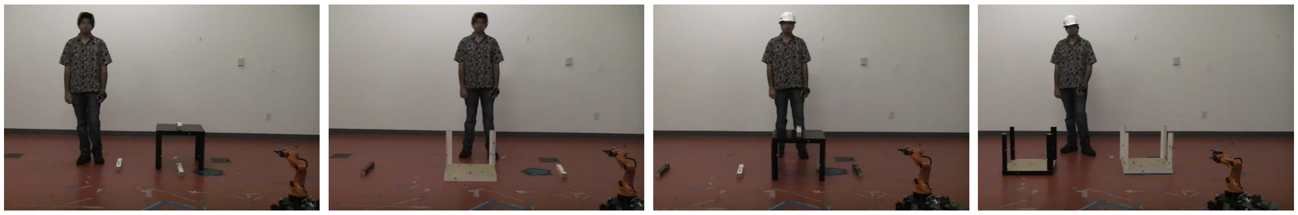
We created a dataset consisting of twenty trials by constructing four different initial scenes and filming an actor taking five different actions in each scene. We present results for the four automatic methods described in Section IV, as well as a baseline consisting of hand-written requests which we created to be clear and unambiguous. Figure 6 shows the four initial scenes paired with handwritten help requests. For the "help me" and hand-written baselines, we issued each of the twenty generated requests to five subjects, for a total of 100 trials. We issued each request in the template and G^3 approaches to fifteen users for a total of 300 trials. We assumed the robot had accurate perceptual access to the objects in the environment and their locations using the VICON system. Results appear in Table II.

Our results show that the "Help me" baseline performs at chance, whereas the template baseline and the G^3 inverse semantics model both improved performance significantly. The S_1 model may have improved performance over the template baseline, but these results do not rise to the level of statistical significance. The S_2 model, however, realizes a significant improvement, $p = 0.002$ by Student's t-test, due to its more specific requests, which model the uncertainty of the listener. These results demonstrate that our model successfully generates help requests for many conditions.

Most failures occurred due to ambiguity in the language, even in sentences generated by the S_2 model. For example, many people confused "the white leg that is near the black table" with "the white leg that is under the black table." Adding more prepositions, such as "next to" would address this issue by enabling the algorithm to generate more specific referring expressions that more accurately match people's expectations.

B. *User Study*

In our experiment, humans and robots collaborated to assemble IKEA furniture. The study split participants into two conditions using a between-subjects design, with 8 subjects in each condition. In the baseline condition, robots requested



Take the table leg that is on the table and place it in the robot's hand.
Take the table leg that is under the table and place it in the robot's hand.
Take the table leg that is next to the table and place it in the robot's hand.
Pick up the table leg that is on the table and hold it.
Take the table leg that is on the table and place it on the floor in front of the robot.

Screw the white table leg into the hole in the table top.
Screw the black table leg into the hole in the table top.
Take the white table leg and insert it in the hole, but do not screw it in.
Move the white table leg over near the table top.
Take the table top and place it near the white table leg on the floor.

Take the white table leg that is next to the table and put it in front of the robot.
Take the black table leg that is next to the table and put it in front of the robot.
Take the black table leg that is far away from the table and put it in front of the robot.
Take the white table leg that is on top of the table and place it in the robot's hand.
Pick up the white table leg next to the table and hold it.

Take the white table, flip it over, and set it down in place.
Take the black table, flip it over, and set it down in place.
Take the white table and move it near the robot, keeping it upside-down.
Pick up the white table and hold it.
Take the white table, flip it over, and put it in the robot's hand.

Fig. 6. The four initial scenes from the evaluation dataset, together with the hand-written help requests used in our evaluation.

help with the S_0 approach, using only the words "Please help me." In the test condition, robots requested help using the S_2 inverse semantics metric. The robots autonomously planned and executed the assembly on two real robots, and all detected failures were real. Our goal was to assess the effect of the choice of help request, made to a user with limited situational awareness, within an end-to-end system. We chose approach S_0 as a baseline to evaluate the magnitude of this effect. The accompanying video is online at http://youtu.be/2Ts0W4SiOfs.

We measure effectiveness by a combination of objective and subjective measures. We report two objective measures: *efficiency* – the elapsed time per help request, and *accuracy* – the number of error-free user interventions. Taken together, these measures show how effectively the human's time is being used by the robots. We also report three subjective measures derived from a post-trial survey, as well as their own written feedback about the system, to gain an understanding of their view of the strengths and weaknesses of our approach.

1) Procedure: Subjects in each condition were gender-balanced and had no significant difference in experience with robots or furniture assembly. To familiarize users with the robot's capabilities, we gave them a list of actions that might help the robots. During preliminary trials, subjects had problems when handing parts to the robot (called a hand-off), so we demonstrated this task and gave each user the opportunity to practice. The entire instruction period lasted less than five minutes, including the demonstration. During the experiment, we instructed users to focus on a different assembly task and only help the robots when requested.

For each subject, the robot team started from the same initial conditions, shown in Figure 7. Some failures were inevitable given the initial conditions (e.g., a table top turned upside down; a part on a table out of the robots' reach.) Other failures happened naturally (e.g., a table leg that slipped out of a robot's gripper.) When a failure occurred during assembly, the failing robot addressed the person by saying, "Excuse me," and generated and spoke a request for help through an on-board speaker, distinguishing itself by color

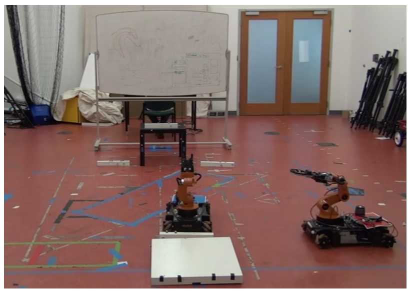


Fig. 7. Initial configuration for the user study. The user is seated behind the whiteboard in the background.

if necessary. We projected all dialogue on a large screen to remove dependence on understanding synthesized speech. The human then intervened in the way they felt was appropriate.

After communicating a help request, the robots waited up to 60 seconds for the user to provide help. If the the environment changed in a way that satisfied the request, the robot said "Thank you, I'll take it from here," and we counted the person's intervention as successful. If the allotted time elapsed, the robot instead said "Never mind, I'll take it from here," and moved on to a different part of the assembly process. These instances were recorded as failed interventions. For each intervention, we recorded the time elapsed and number of actions the human took in attempting to solve the problem.

Each trial ran for fifteen minutes. Although we tried to limit experimenter intervention, there were several problems with the robotic assembly system that required expert assistance. Experimenters intervened when either of two situations arose: potential damage to the hardware (19 times), or an infinite loop in legacy software (15 times). In addition, software running on the robots crashed and needed to be restarted 5 times. In the future, we plan to address these issues using methods for directing requests to the person most likely to satisfy them, rather than only targeting requests at untrained users.

2) Results and Discussion: Over the course of the study, the robots made 102 help requests, of which 76 were satisfied

TABLE III
END-TO-END USER STUDY RESULTS

Objective Metrics	Interventions	S_0	S_2	p (t-test)
Intervention time (sec)	Non-hand-offs	33.3	25.1	0.092
Error-free interventions (%)	All	57.1	77.6	0.039
Successful interventions (%)	All	70.3	80	0.174

Subjective Metrics	p (Kruskal-Wallis)
Robot is effective at communicating its needs	0.001
Prefer working in parallel with the robot	0.056
Easy to switch between and robots' task and user's	0.388

successfully within the 60-second time limit. The most common request type was the hand-off, comprising 50 requests.

Table III gives results from all metrics. Compared to the baseline, S_2 found a decrease in response time for the human to complete non-hand-off interventions and an increased rate of accuracy at performing those interventions. The change in overall success rate was not statistically significant, probably because users were permitted to try multiple actions within the 60-second window when the request was ambiguous. Such retries counted against the accuracy and intervention time metrics.

Qualitatively, subjects preferred the language generation system; Figure 8 shows comments from participants in the study in each condition: even when users successfully helped the robots in the baseline condition, they frequently complained that they did not know what the robot was asking for. Two subjective metrics showed statistically significant results. Users in the S_2 condition reported that the robot was more effective at communicating its needs. They were also more likely to record a preference for assembling two kits in parallel as opposed to assembling one kit at a time together with the robots. These questions were scored on a five-point Likert scale.

Despite these promising successes, important limitations remain. First, hand-offs remained difficult for users even after training. Second, the system required frequent intervention by the experimenters to deal with unexpected failures. Both of these conditions might be modified by a more nuanced model of what help a human teammate could provide. For example, if the robots could predict that handoffs are challenging for people to successfully complete, they might ask for a different action, such as to place the part on the ground near the robot. Similarly, if the robots were able to model the ability of different people to provide targeted help, they might direct some requests to untrained users, and other requests to "level 2" tech support. The different types of interventions provided by the experimenters compared to the subjects points to a need for the robots to model specific types of help that different people can provide, as in Rosenthal et al. [18].

C. Conclusion

The goal of our evaluation was to assess the effectiveness of various approaches for generating requests for help. The corpus-based evaluation compares the inverse semantics method to several baselines in an online evaluation, demonstrating that the inverse semantics algorithm significantly improves the accuracy of a human's response to a natural language request for help compared to baselines. Our end-to-end evaluation demonstrates that this improvement can be realized in the context of a real-world robotic team interacting with minimally trained human users. This work represents a step toward the goal of mixed-initiative human-robot cooperative assembly.

Our end-to-end evaluation highlights the strength of the system, but also its weakness. Robots used a single model for a person's ability to act in the environment; in reality, different people have different abilities and willingness to help the robot. Additionally we assumed that the robot and person both had equal perceptual access to the objects in the environment; in practice many failures may occur due to the perceptual system not detecting an object, leading to the robot being unable to generate an accurate help request, or generating an ambiguous request because it is not aware of all the distractor objects. Developing a dialog system capable of answering questions from people in real time could provide disambiguation when people fail to understand the robot's request. As we move from robot-initiated to mixed-initiative communication, the reliance on common ground and context increases significantly. Since our models can be expected to remain imperfect, the demand for unambiguous sentences becomes less satisfiable. In the long term, we aim to develop robots with increased task-robustness in a variety of domains by leveraging the ability and willingness of human partners to assist robots in recovering from a wide variety of failures.

"Help me" Baseline

"I think if the robot was clearer or I saw it assemble the desk before, I would know more about what it was asking me."

"Did not really feel like 'working together as a team' – For more complex furniture it would be more efficient for robot to say what action the human should do?"

"The difficulty is not so much working together but the robots not being able to communicate the actual problem they have. Also it is unclear which ones of the robots has the problem."

G^3 Inverse Semantics with S_2

"More fun than working alone."

"I was focused on my own task but could hear the robot when it needed help.... However, I like the idea of having a robot help you multitask."

"There was a sense of being much more productive than I would have been on my own."

Fig. 8. Comments from participants in our study.

ACKNOWLEDGMENTS

This work was supported in part by the Boeing Company, and in part by the U.S Army Research Laboratory under the Robotics Collaborative Technology Alliance. The authors thank Dishaan Ahuja and Andrew Spielberg for their assistance in conducting the experiments.

References

[1] M. Bollini, S. Tellex, T. Thompson, N. Roy, and D. Rus. Interpreting and executing recipes with a cooking robot. In *13th International Symposium on Experimental Robotics*, 2012.

[2] David L. Chen and Raymond J. Mooney. Learning to interpret natural language navigation instructions from observations. In *Proc. AAAI*, 2011.

[3] G. Dorais, R. Banasso, D. Kortenkamp, P. Pell, and D. Schreckenghost. Adjustable autonomy for human-centered autonomous systems on mars, 1998.

[4] Anca Dragan and Siddhartha Srinivasa. Generating legible motion. In *Robotics: Science and Systems*, June 2013.

[5] J. Dzifcak, M. Scheutz, C. Baral, and P. Schermerhorn. What to do and how to do it: Translating natural language directives into temporal and dynamic logic representation for goal management and action execution. In *Proc. IEEE Int'l Conf. on Robotics and Automation*, pages 4163–4168, 2009.

[6] T. Fong, C. Thorpe, and C. Baur. Robot, asker of questions. *Journal of Robotics and Autonomous Systems*, 42:235–243, 2003.

[7] K. Garoufi and A. Koller. Combining symbolic and corpus-based approaches for the generation of successful referring expressions. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 121–131. Association for Computational Linguistics, 2011.

[8] D. Golland, P. Liang, and D. Klein. A game-theoretic approach to generating spatial descriptions. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 410–419. Association for Computational Linguistics, 2010.

[9] Noah D Goodman and Andreas Stuhlmüller. Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science*, 5(1):173–184, 2013.

[10] Daniel Jurafsky and James H. Martin. *Speech and Language Processing*. Pearson Prentice Hall, 2 edition, May 2008. ISBN 0131873210.

[11] R. A. Knepper, T. Layton, J. Romanishin, and D. Rus. IkeaBot: An autonomous multi-robot coordinated furniture assembly system. In *Proc. IEEE Int'l Conf. on Robotics and Automation*, Karlsruhe, Germany, May 2013.

[12] Thomas Kollar, Stefanie Tellex, Deb Roy, and Nicholas Roy. Toward understanding natural language directions. In *Proc. ACM/IEEE Int'l Conf. on Human-Robot Interaction*, pages 259–266, 2010.

[13] Emiel Krahmer and Kees Van Deemter. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218, 2012.

[14] Matt MacMahon, Brian Stankiewicz, and Benjamin Kuipers. Walk the talk: Connecting language, knowledge, and action in route instructions. In *Proc. Nat'l Conf. on Artificial Intelligence (AAAI)*, pages 1475–1482, 2006.

[15] J. Maitin-Shepard, J. Lei, M. Cusumano-Towner, and P. Abbeel. Cloth grasp point detection based on multiple-view geometric cues with application to robotic towel folding. In *Proc. IEEE Int'l Conf. on Robotics and Automation*, Anchorage, Alaska, USA, May 2010.

[16] 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.

[17] Ehud Reiter and Robert Dale. *Building Natural Language Generation Systems*. Cambridge University Press, January 2000. ISBN 9780521620369.

[18] Stephanie Rosenthal, Manuela Veloso, and Anind K. Dey. Learning accuracy and availability of humans who help mobile robots. In *Proc. AAAI*, 2011.

[19] D. Roy. A trainable visually-grounded spoken language generation system. In *Proceedings of the International Conference of Spoken Language Processing*, 2002.

[20] R. Simmons, S. Singh, F. Heger, L. M. Hiatt, S. C. Koterba, N. Melchior, and B. P. Sellner. Human-robot teams for large-scale assembly. In *Proceedings of the NASA Science Technology Conference*, May 2007.

[21] K. Striegnitz, A. Denis, A. Gargett, K. Garoufi, A. Koller, and M. Theune. Report on the second second challenge on generating instructions in virtual environments (give-2.5). In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 270–279. Association for Computational Linguistics, 2011.

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

[23] Adam Vogel, Max Bodoia, Christopher Potts, and Dan Jurafsky. Emergence of gricean maxims from multi-agent decision theory. In *Proceedings of NAACL 2013*, 2013.

[24] Adam Vogel, Christopher Potts, and Dan Jurafsky. Implicatures and nested beliefs in approximate Decentralized-POMDPs. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[25] R.H. Wilson. Minimizing user queries in interactive assembly planning. *IEEE Transactions on Robotics and Automation*, 11(2), April 1995.