Situated Human–Robot Collaboration: predicting intent from grounded natural language

Jake Brawer, Olivier Mangin, Alessandro Roncone, Sarah Widder, and Brian Scassellati[1]

***Abstract*— Research in human teamwork shows that a key element of fluid and fluent interactions is the interpretation of implicit verbal and non-verbal cues in context. This poses an issue to robotic platforms, however, as they have historically worked best when controlled through explicit commands that have employed structured, unequivocal representations of the external world and their human partners. In this work, we present a framework for effectively grounding situated and naturalistic speech to action selection during human-robot collaborative activities. This is accomplished by maintaining and incrementally updating separate "speech" and "context" models that jointly classify a collaborator's utterance. We evaluate the efficacy of the system on a collaborative construction task with an autonomous robot and human participants. We first demonstrate that our system is capable of acquiring and deploying new task representations from limited and naturalistic data sets, and without any prior domain knowledge of language or the task itself. Finally, we show that our system is capable of significantly improving performance on an unfamiliar task after a one-shot exposure.**

I. Introduction

The field of Human–Robot Collaboration (HRC) is tasked with designing proactive and autonomous robot collaborators able to complement the superior capabilities of human workers to maximize throughput, improve safety of the workplace, and reduce cognitive load on humans. The general application domain for HRC is composed of a robot that collaborates with humans on a joint task such as furniture assembly [1], [2], assembly lines [3], or other factory-related applications [4], [5]. However, state of the art technologies still rely on sterile and rigid interactions that resort to turn-taking behaviors [3], tele-operation, or more generally limited autonomy and decision making capabilities [6].

Conversely, human–human interaction (HHI) during teamwork does not show this friction. Fluent and natural HHIs are multimodal [7], highly contextual and situated [8]. This is particularly true when coordination during teamwork is attended through natural language. Humans resolve the natural ambiguities of speech by integrating verbal with non-verbal cues and, importantly, by grounding speech to the physical domain of the interaction—e.g. through implicature [9] or lexical entrainment [10], [11].

Yet, despite evidence of the importance of situated natural language in HHI, achieving the same level of richness still represents a significant challenge for HRI in general and HRC in particular. Reasons for this are specific to HRC, e.g. the presence of noise in environments such as those commonly found in factories. Noisy environments may result in failure to recognize significant portions of an utterance—if not the totality of it. This not only leads to erroneous naming of specific actions and objects, but also makes the structure of sentences harder to parse by natural language understanding (NLU) algorithms that exploit syntax. Most notably, impediments to deploying effective HRC interactions are also to be found in the very nature of the communication itself. Communication during collaboration often occurs in a time-constrained context, is highly goal-oriented, typically requires a high success rate in order to be effective, is domain-dependent, and often features mutual adaptation between peers. The time constraint during collaboration pressures agents to make shorter utterances that might not be well-formed sentences; the noise and the need for unambiguity favor some classes of words over others, often resulting in a highly domain-specific language. All these factors greatly hamper the deployment of standard NLU techniques to HRC. State-of-the-art technologies resort extensively to hand-coded domain knowledge, or require training on large datasets—most of which are taken from descriptive text and are

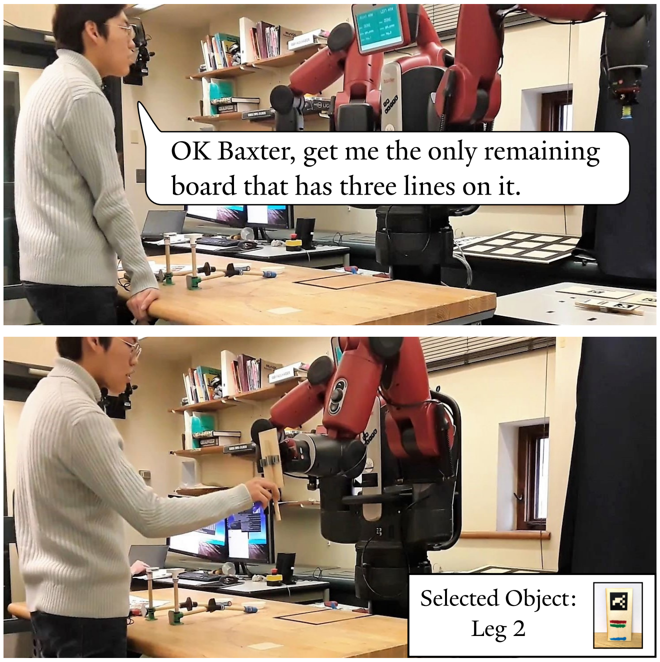

OK Baxter, get me the only remaining board that has three lines on it.
Selected Object:
Leg 2

Fig. 1. Depiction of the human-robot interaction. After the human participant requests an object to the robot via natural language, the proposed systems integrates speech and contextual information to autonomously select the optimal part.

[1] The authors are with the Social Robotics Lab, Computer Science Department, Yale University, New Haven, CT 06511, USA `name.surname@yale.edu`.

borrowed from different contexts that do not necessarily leverage the specific domain knowledge. Still, to achieve the level of fluency seen in HHIs, a core ability of future generations of robots will be for them to collaborate with humans through the situated interactions with which humans are most comfortable [12].

Robots, in order to become proficient collaborators, should be able to exploit *context* in order to ground ambiguous and referential speech. It is worth noting that this is however more complex than straightforwardly deploying NLU algorithms to HRC scenarios. Effective collaboration requires fast adaptation to different tasks and/or user preferences, a feature for which systems trained on corpora composed of billions of sentences do not allow. Further, a different kind of context awareness is needed. Conversely, an HRC scenario constrains the verbal interaction to a very specific physical environment, and this affords the unique opportunity to latch a narrower context to the bigger NLP problem, which becomes then more tractable.

In this work, we implement a situated HRC system that integrates verbal instructions from a human partner with contextual information in the form of a task model. The proposed system learns task representations from demonstration, without requiring hand-coded domain linguistic or a priori task knowledge, on a task that is designed to trigger ambiguous, referential speech due to the use of parts and tools that are challenging to refer to verbally. Our experiments demonstrate that the system dynamically leverages linguistic and contextual information to provide support to the human worker; furthermore, it is capable of accomplishing this given a minimal set of noisy and naturalistic data. Additionally, the system is capable of learning online effective representations of tasks from one shot exposures, over real collaborative interactions.

In the following Sections, we provide an overview of related work and how our approach is positioned with respect to research in NLP and HRI (Section II). We then proceed to describe the experimental setup (Section III-A, see also Fig. 1), the proposed algorithm (Section III-B), and to detail the collection of training data (Section III-C). Results are presented in Section IV, followed by a discussion in Section V, and conclusions and future work in Section VI.

II. Background and Related Work

Although communication is pervasive in collaborative activities, most HRC systems are not yet capable of handling the variability of natural communication. On the other hand, the large research in NLP and NLU mostly focuses on corpora of written language, that, by nature, differ from the short-term, context bounded, goal-oriented typical utterances that arise during collaborative activities. Indeed, collaborative scenarios typically include resource constraints that change the nature of the communication strategies by, for example, favoring explicit or implicit references (e.g. time pressure, [8]). Overall, communication serves various roles during collaborative activities, such as sharing and aligning mental states [13], providing confirmation [14], assigning roles and allocating subtasks [15], [2], or asking for help [16]. However, due to the limitations of current technology, it is often necessary to develop strategies for the robot to handle misunderstanding: a robot may for example provide feedback on instructions that triggers users to adapt their speech and gestures [11], or rely on more elaborate dialog templates [17].

Most research on natural language processing relies on pre-coded domain knowledge, or requires large datasets; in addition, it typically focuses on specific tasks like classification or translation, that are studied in isolation from real-world interactions with humans [18]. In robotics and human–computer interaction, instead, the interaction with users is paramount and data collection is expensive; for this reason, past works have augmented the amount of information available to such systems by *integrating language with context*. One example is seen in *multimodal fusion* approaches, that demonstrate how visual and acoustic information improve the understanding of commands from a human [19]. *Compositional instructions* also constitute a powerful knowledge representation to ground natural language commands: for example, Wang and colleagues demonstrate how an autonomous agent can learn to ground an unknown language in a simulated block assembly tasks from demonstrations [20]. Of great interest to this work is the field of *pragmatic modeling*, which introduces a model of the speaker's intentions to improve interpretation of goal-oriented utterances [5], [16], [20]. In this work, we explore a similar problem to these but in the context of a realistic HRC, where the language component is acquired from the human peer and not from typed text or generated by the robot.

More recently, a increasing body of work has focused on the application of NLP approaches to HRC specifically. For instance, Cantrell and colleagues demonstrate how a robot can rely on dialog systems to acquire knowledge about new actions from a human [21]. In addition, several studies targeted specific linguistic contents that are typical of collaborative environments. In such scenarios, humans often trigger references to spatial relationships and several methods have been developed to ground language on such constructs [22], [23], [24], [25], [26]. Planning constraints also arise naturally as a way to provide instructions to collaborative robots [27], [28], as well as reward functions [29]. In this work, we present a framework for the robot to learn from demonstrations how to respond to natural language commands. Our system takes advantage of contextual information, but importantly it does not assume previous knowledge on the language, the task, or the grammatical forms used.

III. Material and Methods

A. Experimental Setup

This paper introduces an experiment designed around a human participant and a Baxter collaborative robot engaged in a construction task—more specifically a small-scale chair, developed in prior work [30] and shown in Fig. 1. The chair, depicted in Fig. 2, requires nineteen individual parts to be

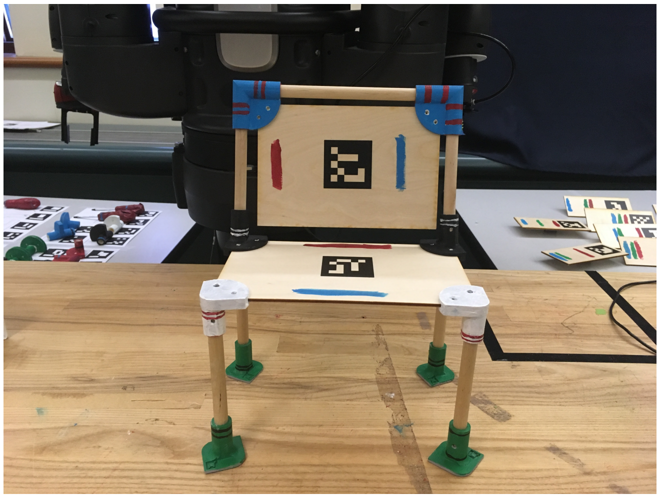


Fig. 2. The application domain the human and the robot are tasked with is the joint construction of a small-scale chair, depicted in figure.

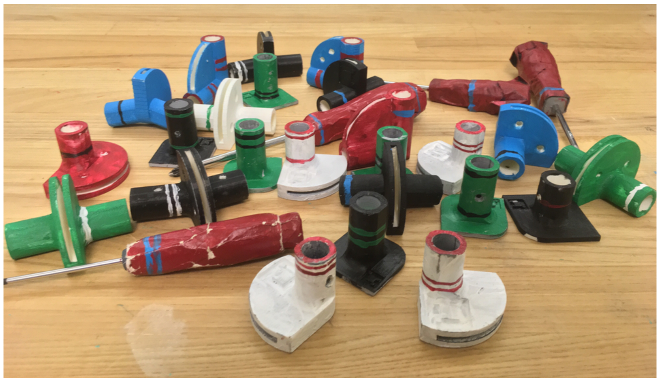


Fig. 3. Detail of the experimental setup. Color patterns can be used to refer to objects and tools, but not unequivocally. For example, the two white pieces in the foreground differ only in the position of their red stripes. Similarly, using purely spatial relationships to refer to objects is difficult due to the large number of objects present.

built: seven dowels that act as legs and supports for the back, a chair seat, a chair back and ten connecting joints that fasten parts together. A single screwdriver is the only tool needed to secure a total of twelve screws.

For the purposes of this work, we confine the interaction to a master–slave configuration, where the robot is expected to provide support to the human upon request—similarly to [31], [4]. More specifically, the parts constituting the model set are placed in two pools of objects (at the right and left side of the robot, see Fig. 1) that can be accessed by the robot exclusively. To successfully perform the task, the participant is requested to ask for constituent parts and tools through speech commands. To facilitate the speech recognition system, we enrolled native English speakers exclusively. Crucially, participants were free to verbally interact with the robot *in any way they preferred*, in order to train the system with as natural interactions as possible. Additionally, a number of 'dummy' objects were integrated into the object pool to add visual noise and increase ambiguity. To this end, another important design decision was to artificially increase the visual complexity of tools and pieces required for the assembly task (Fig. 3). All of the pieces were 3-D printed, novel designs, and lacked clear or otherwise recognizable labels. These pieces were painted with distinct, but overlapping patterns to add more visual noise. Creating a sufficiently ambiguous task was a crucial feature of this experiment, as a way to reproduce realistic environments in which a simple speech model is not sufficient for the robot needs (see Section III-B).

The Baxter robot is provided with a set of basic capabilities, encapsulated into a library of high-level actions originally developed in [2][1]. The perception system is provided by ARuco [32], a library for generation and detection of fiducial markers. Each object in the workspace is provided with an unique marker, which is detected by the end effectors' cameras and then mapped into the robot's 3D operational space via the robot kinematics. In this work, the robot is simply tasked with picking up objects and tools and passing them to its human partner. Additionally, we employ the following software layers: i) a *web interface* to remotely receive feedback from and tele-operate the robot during collection of the training dataset; ii) a *speech-to-text (STT)* software that employs the state-of-the-art Google Cloud STT API [33] to convert verbal commands into text strings; iii) a *feedback channel* visualized on the robot display to provide feedback to the participant about the robot's internal state (see Fig. 1).

B. Data And Learning Algorithm

We propose a flexible approach for incrementally deriving a model of the desired robot's behavior from utterances and context. This is accomplished by maintaining and updating distinct models of the *speech* and the *context* in relation with robot actions, the outputs of which are combined for action selection. For a given utterance and the associated context, each model yields a probability distribution over actions. Our system assumes the independence of the utterance and context observations, given the desired action. For a uniform prior on robot actions, this results in the predicted probability of actions being proportional to the product of the predictions of the speech and utterance models. More formally, for an observed context $c_t \in C$ and an utterance $u_t \in U$, the action $\hat{a}_t$ chosen by the system at time-step t is the action that maximizes:

$$\hat{a_t} = \underset{a \in A_t}{\arg\max} \left[p_{\text{speech}}(a|u_t) \cdot p_{\text{context}}(a|c_t) \right] \quad , \qquad (1)$$

where A_t is the set of feasible actions at time t, which typically excludes objects that are absent from the picking area at a given time.

Here *context* refers to both the presence of objects in the workspace but also the action history of the robot—the sequence of successful actions taken up until the current moment. For the purpose of this experiment we introduce a simple context model that records counts of actions taken

[1] https://github.com/ScazLab/human_robot_collaboration

in any context. We consider contexts represented as action histories: $C = \cup_{m\in\mathbb{N}} A^m$ where A^m is the set of all sequences of m actions. Although the state of all possible contexts is intractable, the system only models here the distribution on actions from observed contexts and assumes a uniform probability on unknown contexts. Because contexts are sequences of successful actions—each sub-sequence prefixing an observed context is itself an observed context—we implement the model by storing the counts of successful occurrences of actions in the tree of observed contexts. For a given, known, context c and $N_{a,c}$ the count of occurrences of action a in context c, the model returns a probability:

$$p_{\text{context}}(a|c) = (1-\varepsilon_{\text{context}})\frac{N_{a,c}}{\sum_{a'\in A} N_{a',c}} + \varepsilon_{\text{context}} \quad , \tag{2}$$

where $\varepsilon_{\text{context}}$ is an exploration parameter (0.15 in the experiment) that accounts for new unexpected actions in known contexts.

For the purposes of this experiment, we use a speech model based on logistic regression to represent $\Pr(a|u_t)$. More precisely, each utterance command u_t is converted by the speech-to-text system and then represented as a bag of n-grams of size one and two. The model for this experiment is based on [34] for both the classifier and the feature extraction. For a given utterance u, represented as a vector x of n-grams counts, the logistic regression model includes a parameter vector θ_a for each action and returns an estimate of the action probability:

$$p_{\text{logistic}}(a|u) = \frac{C}{1+e^{-\theta_a^T x}} \quad , \tag{3}$$

where C is a normalization constant. We then compute:

$$p_{\text{speech}}(a|u) = (1-\varepsilon_{\text{context}}) \cdot p_{\text{logistic}}(a|u) + \varepsilon_{\text{context}} \quad , \tag{4}$$

in which $\varepsilon_{\text{context}}$ accounts for the occurrence of irrelevant speech commands (in practice 0.15 in the experiment).

C. Data Collection Phase

We collected data from recordings of human interactions with a teleoperated robot. All training data was recorded from the chair building task, that each participant built three times, according to three different orderings (denoted as *instructions* A, B, and C in Table I), that were provided through instruction sheets. Each instruction sheet simply includes a list of steps that the participant had to follow. Each step is represented by a picture of the part to ask the robot for and a picture of the current state of the chair being built. We rotated the order in which each participant was asked to build the chair, alternating between ABC, BCA, and CAB.

After receiving the instructions for the task, each participant was familiarized to the process of receiving parts from the robot, which includes pressing a button to trigger release of the part. During the explanation, the experimenters avoided the use of any other word than "part" to refer to the various elements of the assembly, in order not to bias the vocabulary later used by the participants. The instructions specified that participants needed to refer unambiguously to the part they wanted from the robot.

TABLE I

INSTRUCTION SETS USED DURING DATA COLLECTION AND EXPERIMENTAL TRIALS.

A / A'	B / B'	C / C'
foot_1	foot_1	foot_1
leg_1	foot_2	leg_3
screwdriver	foot_3	screwdriver
foot_2	foot_4	back_1
leg_2	leg_1	leg_5
foot_3	screwdriver	top_2
leg_3	leg_2	foot_2
foot_4	leg_3	leg_4
leg_4	leg_4	back_2
front_1	front_1	leg_6
front_3	front_3	top_1
back_1	back_1	leg_7
back_2	back_2	back_2
seat / **leg_5**	seat / **leg_5**	foot_3 / **seat**
leg_5 / **leg_6**	leg_5 / **leg_6**	leg_1 / **foot_4**
leg_6 / **top_1**	leg_6 / **top_1**	front_1 / **leg_1**
top_2 / **top_2**	top_2 / **top_2**	foot_4 / **front_3**
top_1 / **leg_7**	top_1 / **leg_7**	leg_2 / **foot_3**
leg_7 / **back**	leg_7 / **back**	front_3 / **leg_2**
back / **seat**	back / **seat**	seat / **front_1**

During each of the three assemblies, an experimenter was waiting for the participant to formulate an unambiguous request and was then triggering the actions from the robot. However, the participants were led to believe that the robot was operating autonomously in order to elicit the most naturalistic utterances possible. All transcribed utterances (and corresponding robot actions) were collected and used later on to train the system. Importantly, *all* the sentences were collected, without filtering out bad utterances and/or broken requests. In total, 626 pairs of requested objects and actions were collected from 12 participants.

D. Evaluation

In order to evaluate the efficacy of the joint context and speech model system, we deployed the trained model on the robot and conducted autonomous construction trials with 11 participants (please refer to accompanying video for select demonstrations, also available at https://youtu.be/pSdN9NJg_EI). As during data collection, the robot supported the construction of three chairs by retrieving parts based on speech commands issued by participants. In addition to tasks A, B, and C we introduced three corresponding permutations: A', B', C' (here referred to as *prime* tasks). For the prime tasks, only the latter third of the instructions were permuted allowing for the performance of the model to be evaluated in a divergent and unfamiliar context. Participants were provided one of three sets of instructions: $AC'C'$, $BA'A'$, or $CB'B'$. The prime tasks were repeated by each participant in order to gauge how effectively the system was able to learn from the previous trial.

TABLE II

TEN UTTERANCES FROM THE RANDOMLY CHOSEN 'FOOT_4' OBJECT, AS DETECTED BY THE SPEECH-TO-TEXT SYSTEM [33].

	Utterance
1	"Green Leaf shaped green leaf shaped object on your right arm with black pants"
2	"I had the last 3 months and then we're one black lines at the top and wants to bottom"
3	"the leaf shaped object green colored with One Bank on black bun on top and one on the bottom"
4	"thank you can I have another green part that has one black line on the top one black line on the bottom on your right arm"
5	"Baxter can you hand me"
6	"I need the greens Green Park and it has a hole in between the black to Circus"
7	"I want the green part with one black stripe at the top and one black stripe at the bottom"
8	"vodka the Green Bay's peace"
9	"give me the green part with black line on the top and the bottom"
10	"underneath the rectangular blue line on the top and in the middle Back Square private line"

In accordance with the data collection phase, participants constructed the chair in the order specified by the instructions sets. In the event that an incorrect piece was retrieved by the robot, the participants were instructed to press a red button on the robot's end effector resulting in the robot returning the piece to its original location. This was repeated until the robot retrieved the correct piece. The performance of the system was evaluated based on the number of occurrences of such button presses during each trial.

IV. RESULTS

A. Collected data

In this experiment we initially trained a decision process as a classifier on the 20 actions from a little more than 600 samples collected from 12 participants over 3 assemblies per participant. Additionally, further training was acquired online during experimental sessions, with learning persisting across the three trials. Thus we demonstrate the feasibility of training a supportive robot on a limited amount of data, from a relatively unconstrained scenario, and without any assumption on the language used by the participants, outside of the availability of a speech to text system. This setting contrasts sharply with the large amount of data required by typical NLP systems. In particular it brings training to a range feasible for real-world HRC and robotics scenario, where data collection is both hard to constrain and costly.

Table II shows a selection of utterances for one action, chosen randomly from our training set. The table displays the variability of the utterances used to request the same object, as well as the transcription errors from the speech to text system. In particular, several utterances are not correctly transcribed, some are truncated, and some are not associated with the correct object. They illustrate the difficulty of acquiring clean training data from real world scenario, even when using a teleoperated robot.

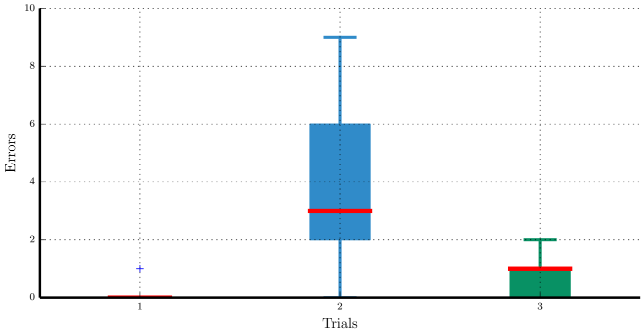


Fig. 4. Errors per trial across participants. Here an error is an incorrect action taken by the robot. A paired t-test revealed a significant decrease in error rate across trials 2 and 3 ($p < 0.001$). This suggests that the system is capable of effectively learning new tasks from limited data.

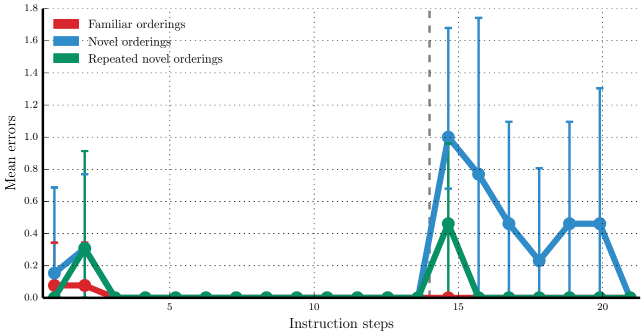


Fig. 5. Errors across instruction steps. The dashed black line indicates the step at which the instructions diverge from those used in the training set for the trails denoted by the blue and green lines.

B. User interaction

Trial 1 utilized one of the three instructions from the training phase (either A, B, or C). Trials 2 and 3 had the participants repeating one of the novel *prime* permutations (either A', B', or C'). The repetition of prime tasks allowed us to evaluate how rapidly and effectively the system was able to learn new tasks.

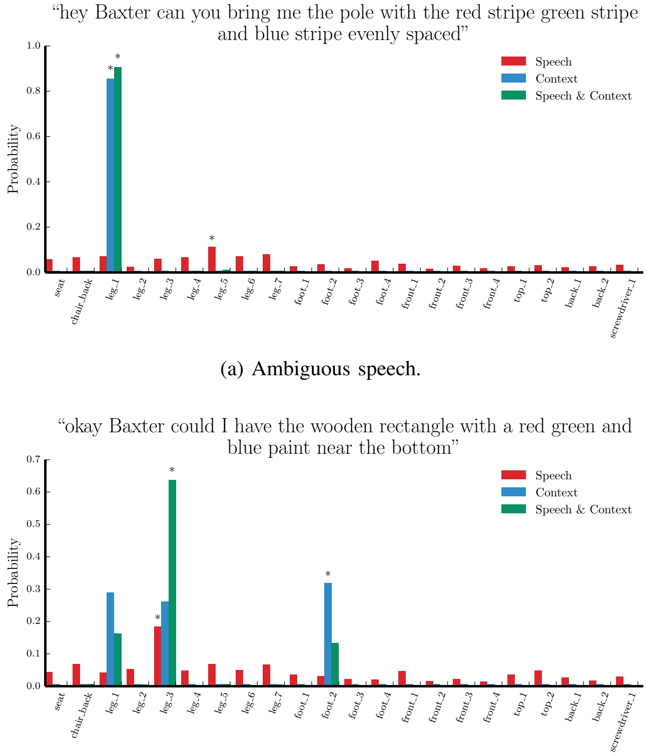
Fig. 4 presents the classification median error rate of the algorithm for the utterances captured from 11 participants across the three trials. No errors were observed in the familiar tasks trial, suggesting that the system robustly learned the structure of the three tasks comprising the trial. While the model performed noticeably worse on the unfamiliar task trial, a paired t-test revealed it's acquired familiarity with the task boosted performance significantly ($p < 0.001$) on the second attempt. In Fig. 5 the effects of learning are made more apparent. Across the unfamiliar tasks trial and the repetition trial, performance is worst at instruction step 14—the step at which the prime instructions diverge from their corresponding originals. However, on the second attempt, there are comparatively fewer errors at step 14, and indeed no errors for subsequent steps, supporting the models ability to rapidly acquire task structures from limited data.

In addition, we compared the error rate of the speech, context, and joint speech and context models in simulation using the data collected from the experimental trials. In order

TABLE III

Average error rate (in parenthesis, standard deviation) per experimental trial for each model.

	Speech	Context	Speech & Context
Familiar	10.1 (2.16)	0.6 (0.48)	**0.0** (0.0)
Unfamiliar	10.8 (2.60)	6.7 (0.64)	**3.6** (1.11)
Repeated unfamiliar	10.3 (2.60)	1.7 (0.64)	**0.9** (0.53)



(a) Ambiguous speech.

(b) Ambiguous context.

Fig. 6. Model predictions from actual participant commands. The quoted text above each figure denotes the command being classified by the three models. The colored bars denote each model's probability distribution over all possible actions. Asterisks above select bars denote the action selected by the corresponding model.

to compare the three models on the data where only one was running, the error rate is the number of robot errors on first attempt. In other words, repetition of the command by the participant after an incorrect action from the robot are discarded. As reported in Table III, the joint speech and context model produced the fewest errors on all trials, supporting efficacy of this approach.

V. Discussion

We present a flexible system capable of building incremental models of speech and context information for producing desired behaviors. In addition, we present results from a baseline HRC experiment that had a participant and an autonomous robot complete an assembly task. For the experimental trials utilizing instructions from the training set, our model correctly classified all of the participants' commands. This is notable given the relatively small number of examples in the training set (approximately 12 of each of the three instruction sets). While performance decreased on unfamiliar tasks, the system's performance increased significantly after only a single exposure. This suggests that the system develops approximate yet adaptable representations of tasks which can be generated quickly, but also deviated from should it be required.

The strength of our system lies in its integration of the outputs of two distinct models, speech and context, for action selection. Figure 6 depicts instructive example model outputs from actual participant commands. When the speech model weights multiple actions equally (Fig. 6a), the context model can break the tie. This is also true in the converse case (Fig. 6b); when multiple actions are weighed ambiguously by the context model, the speech model can induce the correct classification of the command. This also holds when the system is in an unfamiliar context, as is the case when performing an unfamiliar task, and thus must solely rely on the speech model. Having the models compensate for each others' weaknesses, enables to effectively bootstrap one model from the others' learning of the task. Not only does this boost the overall performance on subsequent tasks, but it enables the robot to be trained online as it provides support for its collaborators.

A current weakness of the system is its inability to weight each of the models' predictions by its confidence in said predictions. In some instances this behavior may be desired, for example when the speech model strongly favors one action, but the context model favors another. If the speech command in question was very similar to commands the model had been trained on, then we may want the speech model to override the context model, irrespective of the strength of the context models prediction. This could potentially allow the system to more readily explore unfamiliar contexts, and thus learn new tasks more rapidly. A typical case of this situation, of great interest for future work, is the one of a sentence containing a new word, to be contrasted with the situation of a irrelevant utterance, or one containing a word incorrectly transcribed. In particular, when facing an irrelevant utterance, the system should probably rely on its contextual prediction, while when facing a new word, it might be on the other side relevant to assume that the new word refers to an uncommon action. This latter approach is implemented in pragmatics models (e.g. [16], [20]); it is of great interest for future work to contrast such models in concrete situations like these one.

A limitation of this model is that its effectiveness is constrained to only tasks that are structured. Free-form tasks or tasks that permit no consistent set of approaches to a solution constrain the predictive power of the context model. Additionally, for the purposes of this experiment, words that did not appear in the initial training set were ignored by the speech model and were not learned online. This, however, is not a theoretical limitation of the system but rather a concession made during the implementation of the speech model for simplicity.

VI. Conclusions

In this paper we presented a flexible system capable of incrementally deriving a model of a robot's desired behavior from contextual and speech information. In experimental trials with human participants, our system demonstrated perfect classification rates of commands for task with which it was familiar, and significant performance increases on unfamiliar tasks after one-shot exposures. The minimal training required for robust performance, paired with the simplicity of the system suggests that the system is well suited for HRC, a domain where large quantities of data are difficult to obtain.

Future work will focus on extending this system by dynamically weighting model predictions based on prediction confidence. This should allow the system to mitigate the effects of contextual over-fitting, and allow the system to acquire representations of new tasks more readily.

VII. Acknowledgments

This work was supported by the Office of Naval Research contract number N00014-16-1-2212. We would like to thank Sofya Zeylikman and Kayleigh Bishop for their assistance in designing and running the experimental trials.

References

[1] R. A. Knepper, T. Layton, J. Romanishin, and D. Rus, "IkeaBot: An autonomous multi-robot coordinated furniture assembly system," in *IEEE International Conference on Robotics and Automation*, 2013.

[2] A. Roncone, O. Mangin, and B. Scassellati, "Transparent Role Assignment and Task Allocation in Human Robot Collaboration," *Robotics and Automation (ICRA), IEEE International Conference on*, 2017.

[3] L. Johannsmeier and S. Haddadin, "A Hierarchical Human-Robot Interaction-Planning Framework for Task Allocation in Collaborative Industrial Assembly Processes," *IEEE Robotics and Automation Letters*, vol. 2, no. 1, pp. 41–48, Jan 2017.

[4] B. Hayes and B. Scassellati, "Effective Robot Teammate Behaviors for Supporting Sequential Manipulation Tasks," in *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, 2015.

[5] S. Tellex, P. Thaker, J. Joseph, and N. Roy, "Learning perceptually grounded word meanings from unaligned parallel data," *Machine Learning*, vol. 94, no. 2, pp. 151–167, 2014.

[6] J. Bütepage and D. Kragic, "Human-Robot Collaboration: From Psychology to Social Robotics," *CoRR*, 2017.

[7] B. Wahn, J. Schwandt, M. Krüger, D. Crafa, V. Nunnendorf, and P. König, "Multisensory teamwork: using a tactile or an auditory display to exchange gaze information improves performance in joint visual search," *Ergonomics*, vol. 59, no. 6, pp. 781–795, 2016.

[8] J. Shah and C. Breazeal, "An empirical analysis of team coordination behaviors and action planning with application to human–robot teaming," *The Journal of the Human Factors and Ergonomics Society*, vol. 52, no. 2, pp. 234–245, 2010.

[9] G. Gazdar, "Pragmatics, Implicature, Presuposition and Lógical Form," *Crítica: Revista Hispanoamericana de Filosofía*, vol. 12, no. 35, pp. 113–122, 1980.

[10] T. Iio, M. Shiomi, K. Shinozawa, T. Miyashita, T. Akimoto, and N. Hagita, "Lexical entrainment in human-robot interaction: Can robots entrain human vocabulary?" in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct 2009, pp. 3727–3734.

[11] M. Lohse, K. J. Rohlfing, B. Wrede, and G. Sagerer, "Try something else! When users change their discursive behavior in human-robot interaction," in *2008 IEEE International Conference on Robotics and Automation*, May 2008, pp. 3481–3486.

[12] C. Breazeal, G. Hoffman, and A. Lockerd, "Teaching and working with robots as a collaboration," in *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems-Volume 3*. IEEE Computer Society, 2004, pp. 1030–1037.

[13] G. Briggs and M. Scheutz, "Facilitating Mental Modeling in Collaborative Human-robot Interaction Through Adverbial Cues," in *Proceedings of the SIGDIAL 2011 Conference*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 239–247.

[14] J. Sattar and G. Dudek, "Towards quantitative modeling of task confirmations in human-robot dialog," in *2011 IEEE International Conference on Robotics and Automation*. IEEE, may 2011.

[15] A. St Clair and M. Mataric, "How robot verbal feedback can improve team performance in human-robot task collaborations," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2015, pp. 213–220.

[16] S. Tellex, R. Knepper, A. Li, D. Rus, and N. Roy, "Asking for help using inverse semantics," in *Robotics: Science and Systems*, 2014.

[17] R. Deits, S. Tellex, P. Thaker, D. Simeonov, T. Kollar, and N. Roy, "Clarifying commands with information-theoretic human-robot dialog," *Journal of Human-Robot Interaction*, vol. 2, no. 2, pp. 58–79, 2013.

[18] J. Hirschberg and C. D. Manning, "Advances in natural language processing," *Science*, vol. 349, no. 6245, pp. 261–266, jul 2015.

[19] B. Fransen, V. Morariu, E. Martinson, S. Blisard, M. Marge, S. Thomas, A. Schultz, and D. Perzanowski, "Using vision, acoustics, and natural language for disambiguation," in *Proceeding of the ACM/IEEE international conference on Human-robot interaction (HRI)*. ACM Press, 2007.

[20] S. I. Wang, P. Liang, and C. D. Manning, "Learning Language Games through Interaction," in *Association for Computational Linguistics (ACL)*, 2016.

[21] R. Cantrell, P. Schermerhorn, and M. Scheutz, "Learning actions from human-robot dialogues," in *RO-MAN, 2011 IEEE*. IEEE, 2011, pp. 125–130.

[22] S. Guadarrama, L. Riano, D. Golland, D. Göhring, Y. Jia, D. Klein, P. Abbeel, and T. Darrell, "Grounding spatial relations for human-robot interaction," in *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*. IEEE, 2013, pp. 1640–1647.

[23] T. Kollar, S. Tellex, D. Roy, and N. Roy, "Toward understanding natural language directions," in *Human-Robot Interaction (HRI), 5th ACM/IEEE International Conference on*. IEEE, 2010, pp. 259–266.

[24] S. N. Blisard and M. Skubic, "Modeling spatial referencing language for human-robot interaction," in *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005.* IEEE.

[25] S. Hemachandra, M. R. Walter, S. Tellex, and S. Teller, "Learning spatial-semantic representations from natural language descriptions and scene classifications," in *Robotics and Automation (ICRA), 2014 IEEE International Conference on*. IEEE, 2014, pp. 2623–2630.

[26] 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 *AAAI Conference on Artificial Intelligence*. AAAI Publications, 2011.

[27] T. M. Howard, S. Tellex, and N. Roy, "A natural language planner interface for mobile manipulators," in *Robotics and Automation (ICRA), 2014 IEEE International Conference on*. IEEE, 2014, pp. 6652–6659.

[28] R. Cantrell, J. Benton, K. Talamadupula, S. Kambhampati, P. Schermerhorn, and M. Scheutz, "Tell me when and why to do it! Run-time planner model updates via natural language instruction," in *Human-Robot Interaction (HRI), 2012 7th ACM/IEEE International Conference on*. IEEE, 2012, pp. 471–478.

[29] J. MacGlashan, M. Babes-Vroman, M. desJardins, M. Littman, S. Muresan, S. Squire, S. Tellex, D. Arumugam, and L. Yang, "Grounding English Commands to Reward Functions," in *Proceedings of Robotics: Science and Systems*, Rome, Italy, July 2015.

[30] S. Zeylikman, S. Widder, A. Roncone, O. Mangin, and B. Scassellati, "The HRC model set for human-robot collaboration research," in *Intelligent Robots and Systems (IROS), 2018 IEEE/RSJ International Conference on*. IEEE, Oct 2018.

[31] O. Mangin, A. Roncone, and B. Scassellati, "How to be Helpful? Implementing Supportive Behaviors for Human-Robot Collaboration," *arXiv preprint arXiv:1710.11194*, 2017.

[32] S. Garrido-Jurado, R. Muñoz-Salinas, F. Madrid-Cuevas, and M. Marín-Jiménez, "Automatic generation and detection of highly reliable fiducial markers under occlusion," *Pattern Recognition*, vol. 47, no. 6, pp. 2280 – 2292, 2014.

[33] "Google Cloud Speech API," https://cloud.google.com/speech/docs/, 2018.

[34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.