Role Switching in Task-Oriented Multimodal Human-Robot Collaboration

Natawut Monaikul[1*], Bahareh Abbasi[2*], Zhanibek Rysbek[3*], Barbara Di Eugenio[1], and Miloš Žefran[3]

***Abstract*— In a collaborative task and the interaction that accompanies it, the participants often take on distinct roles, and dynamically switch the roles as the task requires. A domestic assistive robot thus needs to have similar capabilities. Using our previously proposed Multimodal Interaction Manager (MIM) framework, this paper investigates how role switching for a robot can be implemented. It identifies a set of primitive subtasks that encode common interaction patterns observed in our data corpus and that can be used to easily construct complex task models. It also describes an implementation on the NAO robot that, together with our original work, demonstrates that the robot can take on different roles. We provide a detailed analysis of the performance of the system and discuss the challenges that arise when switching roles in human-robot interactions.**

I. Introduction

There is a growing need for domestic robots that can support the independent living of individuals who may require some assistance in performing various activities of daily living (ADLs). Many of the interactions that would take place between the human and the robot in these scenarios are inherently multimodal, involving the production and the understanding of language and physical actions from both participants. Such an assistive robot must therefore be equipped with a multimodal interaction manager that can process input from multiple sensors and generate appropriate responses while helping its human partner complete a particular task successfully.

Moreover, in a collaborative task, either participant asks questions, gives instructions, or performs physical actions at various points throughout the task depending on factors such as ability, environment, and engagement. For example, a pervasive kind of task that occurs during many ADLs is what we call the *Find* task, which occurs when people need objects that are not immediately accessible and whose specific location may be unknown to at least one of them; the partners will collaborate on finding them and retrieving them. For example, when two friends are cooking together in a kitchen, one person may ask the partner to get them a specific tool; this could result in following that request with an instruction to look in a certain location, the partner requesting further information about the object to find, or the partner choosing to search a nearby location. A domestic assistive robot participating in such collaborative tasks should likewise be able to engagein these dialogues while providing physical assistance in completing the task.

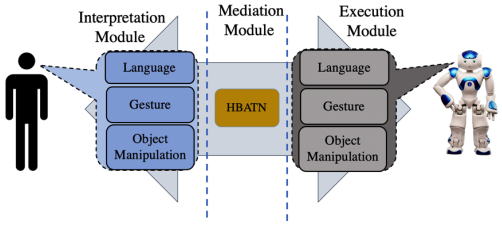

Fig. 1: Multimodal HRI.

In our previous work [1], we proposed a framework for multimodal human-robot interaction (HRI), and in particular for a robot to participate in the *Find* task. The development of this Multimodal Interaction Manager (MIM) was based on a corpus of human-human interactions between elderly individuals (elder role: ELD) and nursing students (helper role: HEL) assisting in ADLs [2], in which 32% of utterances were part of a *Find* task. In these interactions, ELD would typically provide instructions and information, while HEL would ask questions and perform actions manipulating the environment. We devised a new model, a Hierarchical Bipartite Action-Transition Network (HBATN), that we used to represent the task as an interconnected set of networks that modeled both participants simultaneously, providing a means for interpreting actions and planning subsequent actions. The feasibility of an implementation of our MIM in a robot (Baxter) was demonstrated through a user study in which the robot, playing the role of HEL, would interact with a human participant playing the role of ELD and act in order to determine and locate the target object.

In the general *Find* task, however, it is not necessarily the case that only one participant gives instructions, nor that only one participant seeks information and clarification. We envision an assistive robot that is able to assume different roles as needed throughout the task. While many frameworks for assistive task-oriented robots have been proposed, none directly address the challenges of role-switching during multimodal interaction, which is

*First three authors contributed equally to this work.

[1]N. Monaikul and B. Di Eugenio are with the Natural Language Processing Lab, Computer Science Department, University of Illinois at Chicago, Chicago, IL 60607, USA.

[2]B. Abbasi is with the Computer Science Department, California State University Channel Islands, Camarillo, CA 93012, USA.

[3]Z. Rysbek and Miloš Žefran are with the Robotics Lab, Electrical and Computer Engineering Department, University of Illinois at Chicago, Chicago, IL 60607, USA.

This work has been supported by the National Science Foundation grants IIS-1705058 and CMMI-1762924.

crucial to developing a fully collaborative robot. In this work we demonstrate that our framework proposed in [1] allows such role switching. We show that by studying a scenario in which the roles are switched: the robot is the information giver rather than seeker. To further show the flexibility of our framework we also use a different robot (NAO). The new implementation – taken together with our previous implementation – shows that our approach is platform-independent and allows the robot to perform the required actions in either role. We also present a refined HBATN that represents the *Find* task in terms of primitive subtasks that are better suited for learning and reuse in other collaborative tasks, as well as for switching the roles during the task. We report the results of a preliminary user study evaluating this new implementation, and we discuss the insights this study provides on the challenges brought and important considerations called for by the role-switching in the context of HRI.

II. Related Work

Domestic service robots require information from multiple modalities to maintain active interaction with a human user during a collaborative task. Partners can communicate through spoken utterances and/or physical actions such as gestures or object manipulations. Moreover, such domestic robots need a planning and execution framework for generating multimodal responses given observations throughout the interaction.

In literature, the effect of adding human-like nonverbal behaviors to robots on the overall interaction has also been extensively studied. In particular, it was shown that they can improve the user's experience and promote engagement [3]–[6]. However, these studies rarely involve a multimodal task-oriented collaboration between a human and a robot in which the robot can autonomously manipulate the environment. Therefore, the question of how the robot should plan for and communicate during a collaborative task remains largely unanswered.

From a task representation point of view, one common approach is a hierarchical representation in which the task is decomposed into smaller subtasks. An example of this approach are Hierarchical Task Networks (HTNs) [7], [8]. These methods are well suited for extracting and teaching the subtasks to a robot using human demonstrations [9]–[11]. Further, [12] uses the hierarchical task representation to extract policies for navigating dialogic interactions.

Markov Decision Processes (MDPs) and Partially Observable MDPs (POMDPs)are two widely used methods for robot planning in collaborative tasks [13]. Their weakness is that they are computationally expensive [14], [15] and may not be solvable for large observation or state spaces [13]. Reinforcement learningis another alternative which is extensively used for modeling dialogue systems and extracting the dialogue policy [12], [16] when sufficient training data is available.

Role switching for robots has been previously studied in the multi-robot/multi-agent literature [17]–[20], where it is necessary for the robot to autonomously choose one of several possible roles on the team to best contribute towards achieving a goal. Another domain where this problem has been studied is collaborative manipulation, where it is common to distinguish between a leader and a follower behavior [21]–[23]. None of these works involves language and turn-taking that are the focus of our work.

Role sensitivity has also been explored in collaborative dialogues such as solving a problem together [24] or connecting a caller to someone else [25]. These studies analyzed human-human dialogues to investigate how initiative switches during the conversation to bring both participants to the goal. It was found that either participant can initiate the common grounding of information, whether it is in the form of a question requesting information or a statement that voluntarily offers information. These studies reveal that the role that one plays in a dialogue affects the types of utterances one produces.

This was also shown to be the case in our corpus of *Find* tasks, in which the types of utterances greatly differ between roles; These analyses suggest that developing systems that attempt to recognize a human partner's intent or initiative in dialogue may benefit from considering the role the partner is playing.

III. Multimodal Interaction Manager Framework

We have previously proposed a framework for multimodal HRI in collaborative tasks [1]. The Multimodal Interaction Manager (MIM) described therein contains three main components: the Interpretation, Mediation, and Execution Modules (Fig. 1). The Interpretation Module processes the sensory input. The Mediation Module, which includes the HBATN, is responsible for planning the robot's next action, which is then sent to the Execution Module to actually perform the action. This section presents a refined HBATN that further abstracts our previous task representation and provides even greater flexibility.

A. Primitive Subtasks

The structure of the HBATN was developed through an analysis of the ELDERLY-AT-HOME corpus [2], a corpus of human-human multimodal interaction. As a result, the *Find* task was formulated as a set of subtasks with the goal of identifying two main unknowns: O, the target object, and L, the location of the object. The four main subtasks – *Det(O_T)* (determining the desired object type), *Det(L)* (determining a potential location to check), *Open(L)* (opening the location), and *Det(O)* (determining the actual object) – are each modeled as Action-Transition Networks (AcTNets). An AcTNet is a bipartite graph representing the states of both participants (ELD and HEL) and their possible multimodal actions, which are defined as vectors consisting of linguistic

Move	Actor	Utterance	DA	H-O	Pointing	Subtask
1	ELD	Can you help me look for a cup?	Query-yn	-	-	Estab(O_t)
2	HEL	Where should I look?	Query-w	-	-	Estab(L)
3	ELD	Try that drawer	Instruct	-	d3	Estab(L)
4	HEL	-	-	Open(d3)	-	Open(L)
5	HEL	I don't see any cups in here	State-n	-	-	Finish(L)
6	ELD	How about that one?	Instruct	-	d1	Estab(L)
7	HEL	Here?	Check	Touch(d1)	-	Verify(L)
8	ELD	Yeah, there!	Reply-y	-	-	Verify(L)
9	HEL	-	-	Open(d1)	-	Open(L)
10	HEL	You want this spoon?	Query-yn	Takeout(spoon)	-	Verify(O)
11	ELD	That's perfect, thank you.	Reply-y	-	-	Verify(O)

Fig. 2: A sample human-human multimodal interaction in *Find* task along with annotations.

features (the *dialogue act* (DA) [2] of the utterance and object or location words) and physical features (pointing gestures or *haptic-ostensive* (H-O) actions). The HBATN encompassing these AcTNets then allows a robot to not only infer the state of its partner, but also to plan its next action accordingly.

An example of such an interaction is given in Fig. 2. A further analysis of the corpus and the previously proposed HBATN revealed that each subtask can be decomposed into finer subtasks, which we call *primitive subtasks*. In particular, we found that *Det(O_T)* and *Det(L)* both consist of establishing the object or location (*Estab*), potentially followed by verification (*Verify*) or follow-up questions specifying more information (*Spec*), and that *Det(O)* consists of confirming the presence or absence of the desired object (*Finish*) in the current location or verifying a physical object with the partner.

Each of these primitive subtasks can then be represented as simpler AcTNets, as shown in Fig. 3. Note that we adopt similar illustrative conventions in this diagram as in [1]. Specifically, blue nodes represent the *seeker* – the participant who is seeking information or manipulating the environment (HEL in the corpus) – while red nodes represent the *giver* – the participant who is providing answers or giving instructions (ELD in the corpus). Nodes with dotted lines are optional states; for example, to initiate *Estab(L)*, the seeker may ask the giver where to look, or the giver may ask the seeker to look in a specific location.

The arrows in Fig. 3 connecting subtasks and primitive subtasks represent the general flow of the interaction. For example, an object type has to be established before verification or specification of the object type can occur (connecting *Estab(O_T)* to *Verify(O_T)* and *Spec(O_T)*), and when it has been determined that a location does not contain the target object, a new location must be determined (connecting *Det(O)* back to *Det(L)*). Furthermore, subtasks or primitive subtasks may be repeated in succession if, for example, the object type was not properly established (e.g., the seeker did not hear the giver's answer), or a wrong location was verified, so another location has to be verified.

A formal analysis showed that not every path in the previous HBATN can be generated by the refined HBATN. However, we found that the paths that could not be reproduced were not meaningful in an interaction during a collaborative task, so we argue that our refined model is in fact superior. One advantage of representing the task as a HBATN is that since each subtask is its own AcTNet, at the end of each subtask, either participant may initiate the next subtask. This means that the model does not require strictly alternating turns throughout the task and, more importantly, that there are clear points at which roles may seamlessly switch. Another advantage of the task decomposition is that the subtasks can be reused as parts of other complex multimodal collaborative tasks that require common grounding of some information.

B. Automatic Primitive Subtask Segmentation

One limitation of our HBATN and AcTNet approach to task and subtask modeling is that the topologies were derived from the data through manual inspection. By introducing simpler, more primitive AcTNets, the structures can be more easily learned. To this end, we developed a classifier to determine, given a turn in an interaction, which primitive subtask the participant is in. Such a classifier not only helps a robot infer its partner's state in the task, it also provides a means for automatically annotating a large set of multimodal interaction data, which could then be used to learn the structures of those subtask AcTNets.

To create the training data for this classifier, we follow established practice in computational dialogue processing, in which a portion of the corpus is annotated by multiple annotators to establish the validity of the annotation via metrics of inter-coder agreement. Two annotators manually labeled the turns in 84 *Find* task interactions from the ELDERLY-AT-HOME corpus (a total of 1,101 moves) with one of our primitive subtasks. To measure inter-coder agreement on assigning primitive subtask labels, 9 interactions (127 total turns) were independently annotated by both annotators prior to the 84, and relatively high agreement was found using Cohen's kappa ($\kappa = 0.7769$). After some discussion on labeling strategies, a second set of 10 trials with 168 total turns was again independently annotated and shown to have improved agreement ($\kappa = 0.8010$).

We extracted three main types of features from each move: features from the previous move (its label, DA tag, actor, and number of consecutive preceding moves with the same label), the current move (its DA tag, actor, number of tokens, turn number, 50-dimensional GloVe word embeddings [26] averaged over the entire utterance, and if any object words, location words, pointing gestures, or H-O actions were used), and the state of the task (whether or not the object type and location have been determined). These task state features were extracted heuristically using the history of DA tags, primitive subtask labels, and mentions of object and location words. For example, if a previous turn uses the *Instruct* DA tag with an object reference in the *Estab(O_T)* subtask, then the object type has been determined, but if this is followed by a turn using the *Reply-n* DA tag in the

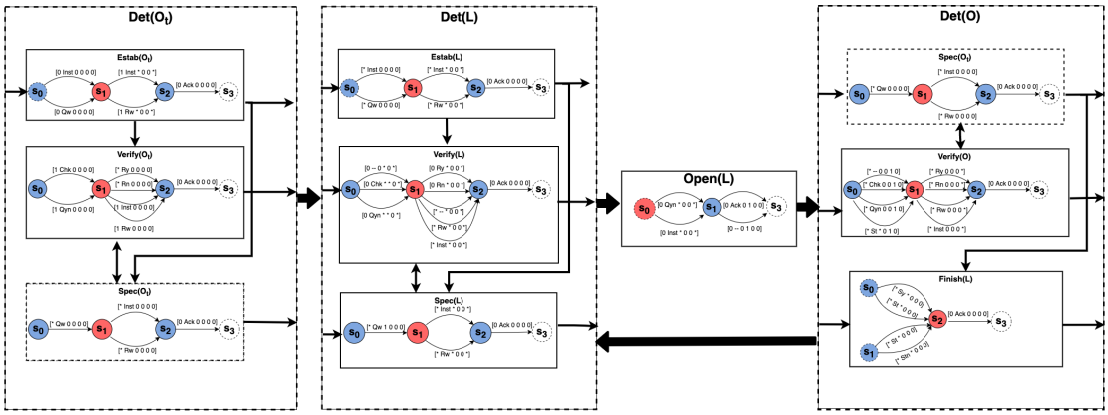

Fig. 3: The hierarchical subtask model decomposed into multiple primitive subtasks – each primitive subtask is an Action-Transition Network that simultaneously models both agents in a concise way.

	CRF	MLP	LR
human-human data	85.3	88.9	88.5
human-robot data	87.6	60	61.1

TABLE I: Subtask classification accuracy of our three classifiers on different datasets.

Verify(O_T) subtask, then the object type has not been determined.

We then trained and tested three different classifiers on the annotated corpus using 5-fold cross-validation: logistic regression (LR), multi-layer perceptron (MLP), and conditional random field (CRF). These results are given in the first row of Table I. Because the results were similar across the three classifiers, with LR and MLP performing slightly better than CRF, we also evaluated the classifiers on the human-robot interaction data collected in our previous work [1]. For this, the classifiers were trained on the full corpus, the same annotators labeled the new interaction data (185 total turns) for primitive subtasks, and the classifiers were tested on this new labeled data. These results are given in the second row of Table I. The observed decrease in performance for LR and MLP may be attributable to imperfect speech-to-text transcriptions in the human-robot data, resulting in extracted features that do not reflect the actual move. However, this did not affect the performance of CRF, which may be due to its ability to better learn patterns in the sequence of subtask labels within a trial.

Overall, the CRF results show the potential for automatically annotating multimodal interaction data for our primitive subtasks, which could then be used to learn the topologies of each subtask by extracting consecutive sequences of moves belonging to the same subtask and using well-established techniques for learning, for example, Markov models. The trained CRF can also be used in implementations of the MIM to help infer the state of the human partner by comparing the observed human action with all possible actions in the HBATN with preference given to those in the predicted subtask.

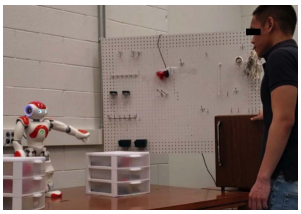

Fig. 4: A snapshot of a trial in which NAO and a human partner interact to find an object.

IV. Experimental Evaluation

In [1], we described an implementation of the MIM to evaluate the efficacy of the HBATN. To investigate the feasibility of switching roles, we again implemented the MIM with the refined HBATN. We chose to use the NAO robot to show the platform-independence of our framework. We then ran a preliminary user study in which participants, acting as the seeker, interacted with NAO, acting as the giver, in the *Find* task.

A. Experimental Setup

NAO was placed on a table to be at around chest-level with the participants. Two storage units, each with three drawers, and a cabinet were placed in front of NAO with handles facing towards the participants. These were the possible locations in which the four target objects (two balls and two cups of different colors) could be hidden. A Kinect sensor was placed on the table, and a camera was mounted off to the side. Participants stood facing NAO and spoke into a hands-free microphone (see Fig. 4).

B. Implementation

To realize the full MIM, we implemented several components to recognize and understand multimodal human actions (the Interpretation Module) and generate robot actions (the Execution Module, see Figure 5). We employed the Google Cloud Speech-to-Text API for speech recognition. Pointing gesture recognition is accomplished by using the skeleton-tracking feature of the Kinect – thresholds on the yaw, pitch, and roll angles of the vector

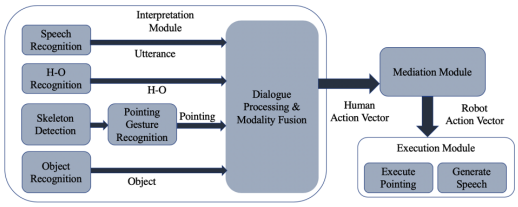

Fig. 5: The subcomponents of the implemented modules of our MIM framework.

from the participant's right shoulder to their right hand were set to identify if and to which of the three locations the participant is pointing.

To recognize H-O actions, we classify each frame from the camera as one of the H-O actions of interest: open, touch, takeout, or no action. We fine-tuned MobileNet-v2 [27] using 36 minutes of recorded and annotated videos (over 21k frames) of three experimenters performing the various H-O actions (95% validation accuracy). For object recognition, skeleton information from the Kinect is used to create a bounding box around the participant's hand, and this image is sent to a fine-tuned ResNet-50 [28] (96% validation accuracy) to predict which, if any, of the objects of interest is being grasped.

The final component in the Interpretation Module, Dialogue Processing & Modality Fusion, performs DA and subtask classification and combines its results and inputs into an action vector for the Mediation Module. Subtask classification was performed using the CRF classifier described in Section III-B. We built a DA classifier with the features described in [2] except the textual features (e.g., part-of-speech tags, dependency parse tree information), which were replaced with features from BERT, a neural language model that has been shown to produce state-of-the-art results on many NLP tasks [29]. Features were obtained by inputting the current and previous utterances and taking the vector output of the special classifier token used in BERT. This updated DA classifier was validated on the ELDERLY-AT-HOME corpus, resulting in an accuracy of 69%, comparable to that reported in [2]. Speech generation and pointing gestures in the Execution Module used NAO's built-in components. Utterances were formed using templates that were created for each possible action vector in the HBATN, and pointing gestures were performed by having NAO move its arm to predefined positions.

C. User Study

A preliminary user study was carried out to evaluate the performance of our HBATN when the robot acts as the giver. Six participants were recruited for this experiment. Each participant was asked to interact with NAO to determine what object it would like (which was randomly decided internally at the start of each trial) and to help it find that object. Each participant performed 5 to 6 trials, resulting in 33 total trials.

D. Evaluation Procedure and Results

We measured the performance of each component of the Interpretation Module by calculating overall accuracy. Speech recognizers are typically evaluated using the word error rate (WER), which is essentially the word-level edit distance between the recognized utterance and the actual utterance, but we report the complementary measure (the percentage of words correctly transcribed) here for consistency, which we call the speech-to-text (STT) accuracy.

To evaluate the quality of the interactions themselves, we report the percentage of successful trials – trials in which NAO acknowledged that the participant had located the correct object – as well as the percentage of human turns that resulted in a *non-understanding* [30]. We define two categories of non-understandings: human turns for which the robot's response is to ask for repetition, and human turns consisting of a question or instruction for which the robot's response does not answer the question or follow the instruction. To further investigate the components which may be contributing to the occurrence of non-understandings, we also report the percentage of non-understandings in which the various components make mistakes, including the HBATN model itself (i.e., if an action is not accounted for in the model).

The overall results of the user study are reported in Table II, and the breakdown of non-understanding results are given in Table III. We see that a majority of the trials were successfully completed with about two-thirds of all human requests being properly understood. The vast majority of non-understandings occur with an error in subtask classification. While the DA classifier feeds directly into the subtask classifier, only about half of the subtask classification errors in non-understandings occur with DA classification errors.

V. Discussion

Overall, participants achieved a very high success rate with NAO in locating the desired object. However, we observed multiple instances of frustrated utterances from the participant or unexpected responses from NAO throughout the interactions despite the eventual completion. What are the main causes of these unexpected responses, and how do the participants adapt to these shortcomings to still complete the task? We investigate these questions to shed light on the challenges that arise when switching roles in HRIs.

The high co-occurrence rate of non-understandings and subtask classification errors suggests that the subtask classifier was the greatest contributor to producing these unexpected responses. Although the subtask classifier performed with high accuracy on the Baxter data, it could not achieve similar results in the NAO experiment. This suggests that the change in role had a significant impact on subtask classification accuracy. An analysis of both types of HRIs revealed that when the robot acts as the seeker, it would typically be the initiator of new

Avg. # Moves	Successful Trials	Non-Understandings	STT Accuracy	DA Accuracy	H-O Accuracy	Pointing Accuracy	Subtask Accuracy
19	84.8%	32.6%	84.8%	57%	83.1%	96%	49.3%

TABLE II: The primary results of our preliminary user study using our implemented MIM.

DA Failure	Speech Failure	H-O Failure	Pointing Failure	Subtask Failure	DA & Subtask Failure	Model Failure
55.5%	43.4%	22.2%	2%	92.9%	51.5%	14.1%

TABLE III: Percentage of non-understandings in which errors in each component occur.

subtasks by asking questions to gather the information, whereas when the robot acts as the giver, the human would typically ask the questions. When the robot asks questions, thereby choosing the subtask, the participant is likely to respond appropriately, remaining in the same subtask, which can be a strong indicator for the classifier since the previous subtask label is one of the features. However, when the human can change the current subtask, the classifier may be less able to predict the label since there is more variation in what subtask will follow.

Another case in which it was difficult for the classifier to detect the initiation of a new subtask is when the participant would utter two sentences that encompass two subtasks, such as "I don't see it there. Where should I look next?" This example can be decomposed into two actions, one belonging to *Finish(L)* and the other to *Estab(L)*, and would be annotated as such in the corpus. However, the speech recognizer cannot recognize sentence boundaries, so this utterance is treated as one action that is unlike others in the corpus, making it hard to classify.

The performance of the DA classifier was also unexpectedly lower than prior evaluations. While errors in the DA classifier did not co-occur with non-understandings in the user study as much as errors in the subtask classifier, the DA classifier does play a direct role in subtask classification, as its predicted labels are features in the subtask classifier, and the heuristic rules for extracting some other features in the subtask classifier depend on DA tags. Thus, the propagation of error from the DA classifier may have contributed to non-understandings.

The classification accuracy of the DA classifier in this experiment contrasts significantly with the accuracy of those reported in [1], [2], despite being trained on the same corpus. One possible reason for this discrepancy is the use of BERT features over the linguistic features of those reported classifiers. To investigate this possibility, we ran the DA classifier used in our previous experiment with Baxter, which uses the same linguistic features as described in [2], on the data collected from the present experiment. It was found that the previous classifier performed at 61.6% accuracy on the NAO data, only marginally better than the current classifier, which suggests that the change in features is not the main cause.

A more likely explanation is again in the role switch itself. In the Baxter experiment, the DA classifier tagged ELD utterances, while in the NAO experiment, the classifier tags HEL utterances. The ELDERLY-AT-HOME corpus was collected from 15 elderly participants, but only two nursing student helpers; a classifier trained on this data may then be able to account for greater variation in ELD utterances at test time, but not in HEL utterances. Moreover, the distribution of DA tags in the corpus for each role reported in [2] shows a clear imbalance, both between the roles and within the roles.

Another contributor to non-understandings is the HBATN itself, which became apparent when the participant would ask questions that NAO had not been programmed to answer, such as "Can you see me?". These questions do not appear in the corpus, nor are they encoded in any primitive subtask AcTNet. Since the participant is the seeker, there is much more variation in the questions that they can ask, many of which could not be accounted for *a priori*.

Our user study provides important insight into the challenges of switching roles in a robot performing a collaborative multimodal task with a human partner. While our DA and subtask classifiers had high accuracy when evaluated on our Baxter data, in which the participant was the giver, their performances degraded significantly when applied to the NAO experiment, in which the participant was the seeker. This change in role brought about a significant change in the human actions that need to be recognized and responded to by the robot, including subtask-initiating actions and actions combining multiple DAs and subtasks. We also note that the trajectories in these interactions may be a significant departure from those in human-human interactions due to errors in the system. Specifically, when a non-understanding occurs between a human and a robot, the robot may ask for repetition or provide an unexpected response, which may lead to frustrating loops, if the human persists in asking their question, or not, if the human ignores the response and continues onward to complete the task (both of which were observed in our user study). Between two humans, however, when non-understandings occur, unexpected responses are typically not simply ignored and can typically be resolved within a few turns [30]. The difference in frequency and resolution strategies of non-understandings between two humans versus a human and a robot creates a testing environment that is significantly different from the training one. A major consideration in building these collaborative robots then is the data on which the different modules are trained – it is not necessarily the case that human-human interaction data

will translate well into human-robot interactions.

VI. Conclusion

Building a robotic assistant that can perform collaborative multimodal tasks requires consideration of the role of the robot. Participants in such tasks often assume distinct roles that may switch at various points throughout the interaction, and so a truly collaborative robot must be equipped with the ability to perform the actions expected of each role. We built upon our previous framework for multimodal human-robot interactions and showed that it can allow for these role switches. We also showed that by implementing such a framework, we can explore a largely unanswered question in the field of human-robot interaction – particularly, how changing roles impacts the dialogic interaction between the human and the robot, as well as the components needed to effectively understand the human in a new context.

References

[1] B. Abbasi, N. Monaikul, Z. Rysbek, B. Di Eugenio, and M. Žefran, "A multimodal human-robot interaction manager for assistive robots," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 6756–6762.

[2] L. Chen, M. Javaid, B. Di Eugenio, and M. Žefran, "The roles and recognition of haptic-ostensive actions in collaborative multimodal human–human dialogues," *Computer Speech & Language*, vol. 34, no. 1, pp. 201–231, 2015.

[3] C.-M. Huang and B. Mutlu, "Learning-based modeling of multimodal behaviors for humanlike robots," in *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '14. ACM, 2014.

[4] J. K. Lee and C. Breazeal, "Human social response toward humanoid robot's head and facial features," in *CHI'10 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2010, pp. 4237–4242.

[5] M. Salem, F. Eyssel, K. Rohlfing, S. Kopp, and F. Joublin, "To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability," *International Journal of Social Robotics*, vol. 5, no. 3, pp. 313–323, 2013.

[6] J. Hemminghaus and S. Kopp, "Towards adaptive social behavior generation for assistive robots using reinforcement learning," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '17. ACM, 2017, pp. 332–340.

[7] K. Erol, J. Hendler, and D. S. Nau, "HTN planning: Complexity and expressivity," in *AAAI*, vol. 94, 1994, pp. 1123–1128.

[8] S. J. Russell and P. Norvig, *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,, 2016.

[9] B. Nooraei, C. Rich, and C. L. Sidner, "A real-time architecture for embodied conversational agents: beyond turn-taking," *ACHI*, vol. 14, pp. 381–388, 2014.

[10] A. Mohseni-Kabir, C. Rich, S. Chernova, C. L. Sidner, and D. Miller, "Interactive hierarchical task learning from a single demonstration," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, 2015, pp. 205–212.

[11] A. S. Clair, C. Saldanha, A. Boteanu, and S. Chernova, "Interactive hierarchical task learning via crowdsourcing for robot adaptability," in *Refereed workshop Planning for Human-Robot Interaction: Shared Autonomy and Collaborative Robotics at Robotics: Science and Systems, Ann Arbor, Michigan. RSS*, 2016.

[12] B. Peng, X. Li, L. Li, J. Gao, A. Celikyilmaz, S. Lee, and K.-F. Wong, "Composite task-completion dialogue policy learning via hierarchical deep reinforcement learning," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017.

[13] N. Gopalan and S. Tellex, "Modeling and solving human-robot collaborative tasks using pomdps," in *RSS Workshop on Model Learning for Human-Robot Communication*, 2015.

[14] C. H. Papadimitriou and J. N. Tsitsiklis, "The complexity of markov decision processes," *Mathematics of operations research*, vol. 12, no. 3, pp. 441–450, 1987.

[15] C. Boutilier, "Planning, learning and coordination in multiagent decision processes," in *Proceedings of the 6th conference on Theoretical aspects of rationality and knowledge*. Morgan Kaufmann Publishers Inc., 1996, pp. 195–210.

[16] M. Gašić, D. Kim, P. Tsiakoulis, and S. Young, "Distributed dialogue policies for multi-domain statistical dialogue management," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5371–5375.

[17] A. Freedy, O. Sert, E. Freedy, J. McDonough, G. Weltman, M. Tambe, T. Gupta, W. Grayson, and P. Cabrera, "Multiagent Adjustable Autonomy Framework (MAAF) for multi-robot, multi-human teams," in *2008 International Symposium on Collaborative Technologies and Systems*, May 2008, pp. 498–505.

[18] E. Martinson and R. Arkin, "Learning to role-switch in multi-robot systems," in *2003 IEEE International Conference on Robotics and Automation (Cat. No.03CH37422)*, vol. 2, Sep. 2003, pp. 2727–2734 vol.2.

[19] H. C.-H. Hsu and A. Liu, "A Flexible Architecture for Navigation Control of a Mobile Robot," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 37, no. 3, pp. 310–318, May 2007.

[20] C. McMillen and M. Veloso, "Distributed, Play-Based Role Assignment for Robot Teams in Dynamic Environments," in *Distributed Autonomous Robotic Systems 7*, M. Gini and R. Voyles, Eds. Tokyo: Springer Japan, 2006, pp. 145–154.

[21] A. Bussy, P. Gergondet, A. Kheddar, F. Keith, and A. Crosnier, "Proactive behavior of a humanoid robot in a haptic transportation task with a human partner," in *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*, Sep. 2012, pp. 962–967.

[22] P. Evrard and A. Kheddar, "Homotopy-based controller for physical human-robot interaction," in *RO-MAN 2009 - The 18th IEEE International Symposium on Robot and Human Interactive Communication*, Sep. 2009, pp. 1–6.

[23] A. Thobbi, Y. Gu, and W. Sheng, "Using human motion estimation for human-robot cooperative manipulation," in *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Sep. 2011, pp. 2873–2878.

[24] J. Chu-Carroll and M. K. Brown, "Tracking initiative in collaborative dialogue interactions," in *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, 1997, pp. 262–270.

[25] C. Balkanski and M. Hurault-Plantet, "Cooperative requests and replies in a collaborative dialogue model," *International Journal of Human-Computer Studies*, vol. 53, pp. 915–968, 2000.

[26] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[27] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[30] G. Hirst, S. McRoy, P. Heeman, P. Edmonds, and D. Horton, "Repairing conversational misunderstandings and non-understandings," *Speech Communication*, vol. 15, no. 3-4, pp. 213–229, 1994.