# Directions Robot: In-the-Wild Experiences and Lessons Learned

Dan Bohus
Microsoft Research
One Microsoft Way
Redmond, WA, 98052
+1-425-706-5880
dbohus@microsoft.com

Chit W. Saw
Microsoft Research
One Microsoft Way
Redmond, WA, 98052
+1-425-722-6430
Nick.Saw@microsoft.com

Eric Horvitz
Microsoft Research
One Microsoft Way
Redmond, WA, 98052
+1-425-706-2127
horvitz@microsoft.com

## ABSTRACT
We introduce *Directions Robot*, a system we have fielded for studying open-world human-robot interaction. The system brings together models for situated spoken language interaction with directions-generation and a gesturing humanoid robot. We describe the perceptual, interaction, and output generation competencies of this system. We then discuss experiences and lessons drawn from data collected in an initial *in-the-wild* deployment, and highlight several challenges with managing engagement, providing directions, and handling out-of-domain queries that arise in open-world, multiparty settings.

## Categories and Subject Descriptors
H.1.2 **[Models and Principles]**: User/Machine System – *Human Information Processing*; H.5.2 **[Information Interfaces and Presentation]** User Interfaces – *Natural Language*; I.4.8 [**Scene Analysis**]: Tracking, Sensor Fusion

## Keywords
Human-robot interaction; directions giving; engagement.

## 1. INTRODUCTION
We are pursuing computational systems that can interact with people in a natural manner with spoken language in the *open world*. Conducting effective dialog in physically situated settings requires the handling of multiple nuances and uncertainties that cannot be easily identified or emulated in the laboratory. In open environments, people may come and go, as individuals or as groups, and complex and evolving social relationships may be in play. The focus of attention is often shared, and people may intermittently attend to and interact with a system, while engaging in conversations with others, or glancing at the screen of a smartphone or other attractors in their field of view. There is uncertainty about intentions and goals, the sources, targets, and meanings of utterances, and the rationale for patterns of movement and attention over time.

Learning about and testing the real-world interactive competencies of such open-world dialog systems requires situating them *in the wild*—*e.g.*, placing them in a workspace or location where people can engage with them at will. Here, we introduce *Directions Robot*, a humanoid robot that gives directions inside a building, and provides a research testbed for open-world human-robot interaction. We describe the system and the key components it harnesses for perception, reasoning, learning, path-generation, dialog, and gesturing. We then discuss lessons learned from an initial live deployment, with a focus on conversational engagement, handling out-of-domain queries, and providing directions.

## 2. RELATED WORK
Several research projects have investigated the problem of providing directions in natural language, and with embodied conversational agents. A framework and architecture for generating natural language directions are presented in [1]. In [2], a representation of locations is used for generating descriptive directions. The MACK [3, 4, 5], and later NUMACK [6] efforts have explored the use of embodied conversational agents to give directions on a college campus. These systems have been used to investigate topics such as the synchronized generation of gaze, gesture, and speech [3, 5], face-to-face verbal and non-verbal grounding [4], and models for generating iconic gestures from communicative intentions [6]. The Virtual Guide [7] is an embodied conversational agent that provides natural language directions in a virtual 3D environment and uses a corpus-based approach for generating beat gestures [8]. Recently, the GIVE challenge has fostered research on generating instructions for guiding users within 3D environments [9].

Directions-giving tasks have also been investigated in the human-robot interaction community. In [10], a model that integrates utterances, gestures and timing is presented and evaluated for a directions-giving task with the Robovie robot. The experiments explore the influence of gestures and highlight the importance of timing in the directions-giving task. Other researchers [11] have studied the use of spatial language when giving directions to robots.

A comparison of the use of different embodiments for performing a directions-giving task found no significant differences in the effectiveness of using a physical robot versus a virtual avatar [12].

**Appears in:** *Alessio Lomuscio, Paul Scerri, Ana Bazzan, and Michael Huhns (eds.), Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2014), May 5-9, 2014, Paris, France.*

However, participants interacting with the virtual agent performed more complementary gestures. The study also found that listener-perspective gestures led to a reduction of errors in a directions retelling task. Another investigation [13] of listener- and speaker-perspective gestures, with a human directions giver, found no significant differences in their effectiveness.

Directions giving remains an active area of research with numerous open questions. Although we frame our studies around a directions task, our primary motivation and goals in developing the Directions Robot are broader. We report on the development of the system as a testbed for studying human-robot interaction in open-world settings. The work follows on the trajectory of several other research efforts, where robots or embodied conversational agents have been deployed in real-world environments—in an effort to identify and address the challenges of performing for long periods of time as competent social actors, *i.a.* [14, 15, 16, 17]. We discuss lessons learned from an initial deployment of the Directions Robot, which highlighted both expected and unforeseen challenges, and we outline directions for future work.

# 3. SYSTEM

The Directions Robot interacts using natural language with one or multiple participants and provides directions to people's offices, conference rooms, and other public areas inside our building such as the kitchen, cafeteria, and bathrooms. The system couples robotics hardware with an existing platform for multimodal, physically situated spoken language interaction and a mapping and directions-generation framework. An example interaction with the robot is presented in Figure 2; videos are available at [18].

## 3.1 Hardware setup

We used a NAO robot from Aldebaran Robotics [19], shown in Figure 1. Its small humanoid form factor and high degree of maneuverability make it well suited for giving directions, and, more specifically, for producing gestural outputs. Its reduced size and clean design foster a positive attitude and curiosity on the part of people that encounter it on a busy hallway.

While the NAO provides on-board sensors and computation, given the complexity of the task at hand and the associated processing needs, we use separate, fixed sensors and a multicore desktop computer to control the robot. The robot stands in a fixed position on a table, with an external wide-angle camera and microphone array placed above its head, as seen at the top right of Figure 1. A Flea3 PointGrey camera with a wide-angle lens (140° field-of-view), captures video at a resolution of 2048x1080 pixels. The microphone array in a Kinect sensor is used to collect audio and sound source localization data. Audio output is produced on the computer and is rendered via external speakers placed behind the robot, which enables acoustic echo cancellation on the audio input.

## 3.2 System components

The software infrastructure for the robot subsumes components for making inferences from audio-visual signals (*e.g.*, face tracking, speech recognition, and scene analysis) and combines them with interaction planning, decision making, output generation, and with a mapping and directions-generation component. A high-level diagram of the system components is shown in Figure 1. Next, we describe them in more detail.

### *3.2.1 Vision*

Interaction in physically situated settings hinges critically on the ability to accurately perceive and track actors in the environment. The Directions Robot tracks people using a face detector that recognizes faces in frontal and side orientations, coupled with a mean-shift tracker that seeks the corresponding image patches for each face in the next frame. To increase robustness, we implemented several additional heuristic mechanisms, including one for resolving occlusions, and for aligning and shrinking tracked faces based on the head crowns detected in the foreground mask. Furthermore, we trained a face-confidence model that assesses the correctness of each tracked face at every frame. Faces with confidence below a threshold are discarded.

### *3.2.2 Speech recognition*

We use the recognition engine distributed with the Kinect for Windows SDK to perform speech recognition on the audio collected from the Kinect microphone array.

The speech recognizer is configured with dialog-state—specific grammars: depending on the interaction context, a particular subset of grammar rules is enabled. Some of the rules were manually authored to capture expected user questions and responses at different points in the dialog. In addition, a subset of rules that model the names of people, places, and office numbers were automatically generated from existing office directory and building map information. In principle, the automatic generation of the building-specific grammar rules should enable easy relocation of the Directions Robot to other buildings.

### *3.2.3 Conversational scene analysis*

Information from the speech recognition and vision components is used to analyze the scene in real-time and model communicative processes such as engagement and turn taking.

**Engagement**. Engagement is the process by which participants initiate, maintain, and break their connection during interactions that they jointly undertake [20, 21]. Human-human engagement is a mixed-initiative, highly coordinated process, regulated by signals across multiple modalities, including proxemics, head and hand gestures, mutual gaze and speech. Modeling engagement in multi-party open-world settings is a challenging problem: people come and go, and initiate, break and interleave interactions with the robot and with each other at will. Within groups, the focus of attention is divided between the robot and the other participants.

Directions Robot uses an engagement model introduced in [22], which makes real-time inferences about the engagement *state, actions,* and *intentions* of each detected actor in the scene. The engagement state captures whether an actor is *engaged* or *not-engaged* in an interaction with the robot. In contrast, the engagement intention represents whether an actor *desires* to be engaged or not. Finally, the engagement action models whether an actor is performing a *maintaining* versus a *disengaging* action (if the actor is currently engaged), or an *initiating* action versus *no-action* (if not currently engaged).

In the initial implementation of this engagement model, the probability of engagement actions is assessed via a soft, heuristic rule. The rule fuses information about the centrality of the face in the scene, the size of the face, and the number of times the face was detected in a frontal orientation, and aims to model whether or not actors are close by, in front of the robot, in an *f-formation* [23] with it (the natural configuration that two or more people assume when they coordinate with each other.) If so, we assume they are performing an *initiating* action (if they are not already engaged) or a *maintaining* action (if they are engaged). The default model also assumes that an actor *intends to be engaged* in a conversation if and only if an initiating or maintaining action is performed. In effect, we make an assumption that intentions are the same as actions. Finally, the state is updated based on the joint actions of the actor

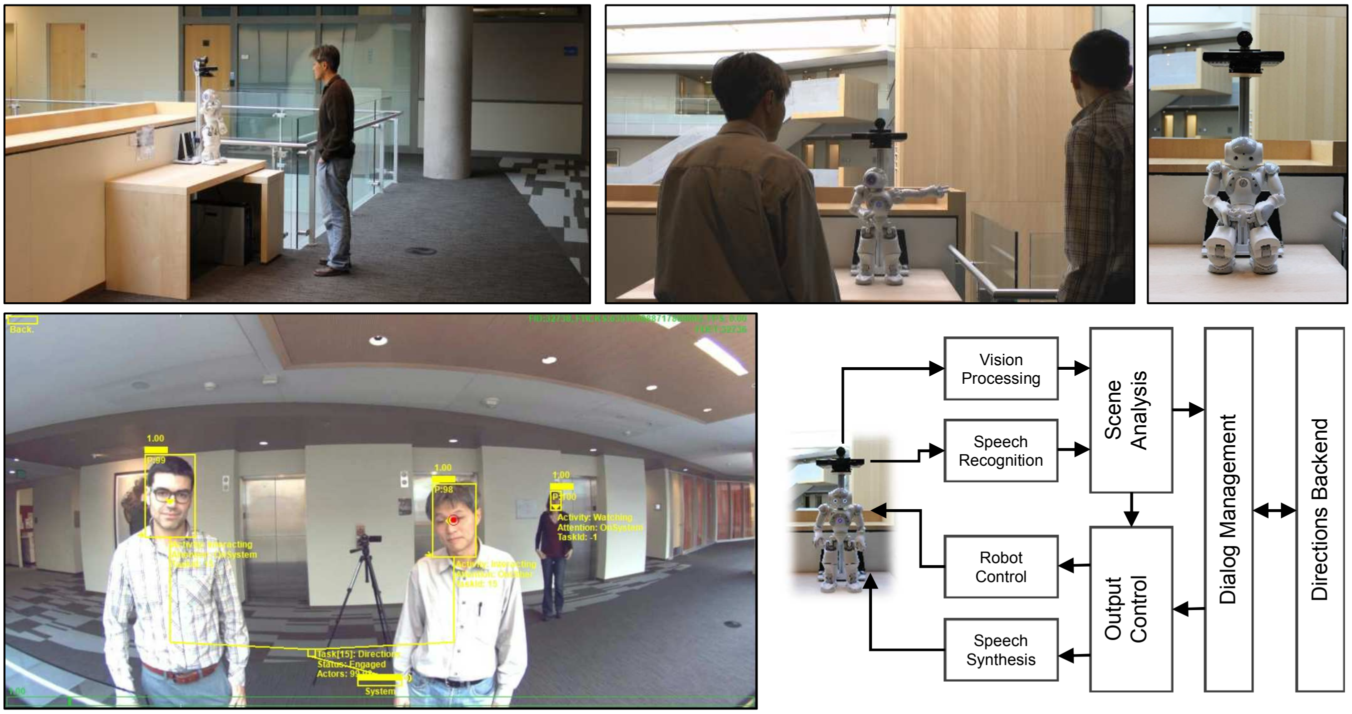


**Figure 1. Directions Robot, from top-left to bottom-right: Robot interacting with one participant; robot interacting with two participants; robot in offline position; real-time robot scene analysis; system diagram.**

and the robot; for instance, to transition from *not-engaged* to *engaged*, both the actor and the robot have to be performing an *initiating* action.

Based on the inferred state, actions and intentions, the robot selects its own engagement actions. With the default policy, the robot initiates a new engagement, *i.e.* starts a new conversation, when it detects that the probability that an actor intends to engage exceeds a preset 0.8 threshold (except if the actor had just disengaged, in which case the robot waits for a few seconds to elapse before a new engagement is started). The system adds actors to an existing engagement if the probability of their engagement intention exceeds a preset 0.8 threshold, until a maximum of three actors are engaged. In the example displayed in Figure 1, the two participants in front are in an f-formation and engaged with the robot, while the third one is not. The system disengages with actors when their *not-engaged* intention exceeds a 0.8 probability. When the last engaged actor becomes disengaged, the conversation is terminated.

The robot's engagement actions are rendered into coordinated gaze, gesture, and speech behaviors. Initiation with the first actor is performed with a verbal greeting "*Hi!*" accompanied by a gesture. The final disengagement action (towards the last disengaging actor) is contextualized based on the dialog state. If disengagement happens close to the beginning of the dialog, the robot terminates the interaction without any speech or gesture: this mostly covers cases when the engagement was in fact incorrectly initiated as a person goes by. If the engaged participant(s) leave in the middle of the dialog, the robot stops speaking and triggers a verbal "*Well, catch you later then.*" communicating its surprise at the early user-initiated disengagement. Finally, if the engagement terminates after directions are given, a simple salutation such as *"Bye-bye!"* is performed, accompanied by a hand-waving gesture.

**Turn taking**. Once engagement is established, the robot must coordinate with the participants on the production of spoken signals in a process known as turn taking. The Directions Robot manages turn taking based on a model described in [24]. The model uses information from the microphone array sound source localizer, coupled with additional heuristics based on several observations, including duration of utterances, level of understanding, and visual focus of attention to track who is talking to whom at every point in time and to infer the floor control actions of each participant in the conversation (*e.g.* who is releasing the floor to whom, who is trying to take the floor, *etc.*). Based on these inferences, the model makes floor control decisions, and signals to the dialog manager when it should take the floor and generate outputs. The turn-taking model enables a diversity of behaviors in multiparty interaction settings.

### *3.2.4 Dialog management*

A finite-state based dialog manager controls the conversation. A sample dialog is shown in Figure 2. The interaction begins with the Directions Robot greeting the user and, once the user is engaged, the robot asks if they need directions. Assuming the user responds affirmatively, the robot asks who or what they are looking for. If the recognized response includes a room number, place, or person within the grammar, spoken directions to the destination are generated, coupled with appropriate arm and hand gestures. The robot can answer multiple sequential requests for directions. Finally, if the user remains engaged after receiving directions and makes no further request, the robot asks if they want anything else and decides whether to continue or terminate the interaction.

Speech recognition is challenging with a far-field microphone array in open environments. Given background noise (including from the robot's motors), as well as acoustic, language modeling, and out-of-domain challenges, a fair number of misunderstandings and non-understandings arise. To mitigate such situations, the dialog manager uses error handling mechanisms such as asking for clarification or confirmation, *e.g. "Pardon me?"*, *"You're looking for the cafeteria, correct?"* Rarely, when consecutive errors occur repeatedly, the dialog manager determines that it is futile to continue the interaction. In this case, the robot apologizes for not

**Robot:** **Hi! Do you need directions?**
User: Yes.
**Robot:** **What room number can I direct you to?**
User: I'm looking for Conference Room 3800.
**Robot:** **To get to Conference Room 3800, go to the end of this hallway. Turn right and keep walking down the hallway for a bit. Conference Room 3800 will be the 1st room on your right.**
User: Also, where can I find John Doe?
**Robot:** **John Doe is in office number 4120. To get there, take the elevator up to the 4th floor. Turn right as you exit the elevator and walk to the end of this hallway. John's office will be in that wing of the building.**
User: Thank you.
**Robot:** **No problem. Can I help you find anything else?**
User: No.
Robot: Well, it was nice talking with you. Goodbye.

**Figure 2. Sample dialog with Directions Robot.**

being able to understand, and tries to gracefully end the interaction by inviting the user(s) to stop by later.

### *3.2.5 Directions backend*

The Directions Robot uses a building map framework [25] which represents the building floor plan overlaid with markups describing the traversable route segments and connectors to valid endpoints. Directions from the robot to other locations in the building are computed using an A* search with a Euclidean distance metric. The resulting path consists of a sequence of segments describing the distance and direction of travel, with references to the underlying route segments, connectors or places traversed.

Natural language navigation instructions, along with synchronized robot gestures are generated based on the computed path, as illustrated in Figure 3. For each segment, a spoken phrase and gesture may be generated to describe the navigation step action implied by the distance and direction of the segment. At the beginning of the navigation sequence, deictic gestures are used, coupled with spoken references, *e.g.* "*Go to the end of this hallway*" to provide a clear initial trajectory (see Figure 3). For subsequent navigation steps, the robot switches to route-perspective gestures indicating orientation at the current point on the path, *e.g.* "*Turn right and keep walking down the hallway.*"

Information about traversed or passed places, and their relationship to path segments, is used to augment the description of navigation steps. For instance, if a path segment traverses only a short portion of the hallway, the qualifier *"for a bit"* may be inserted, as in the example shown in Figure 3. In addition, underlying map features may be used to inform the relative position of the destination, for example by providing the ordinal, type and side of the destination place, *e.g. "the first room on your right"*. To avoid generating long and verbose directions and to increase naturalness, we apply a number of simplification rules. When the path contains a single segment, an absolute reference and pointing gesture to the destination is generated, *e.g. "The restroom is just down this hallway on your right"*. Insignificant or short path segments are skipped (step 2 in Figure 3), and coincident path segments are merged prior to generating directions (step 3 in Figure 3). If the path is still overly long, we short-circuit the directions after a maximum number of navigation steps and provide a summary for the remainder of the path, *e.g.* "*Room 1234 will be in that wing of the building,*" coupled with a deictic gesture. If the destination is on a different floor, the final pointing gesture is performed with an

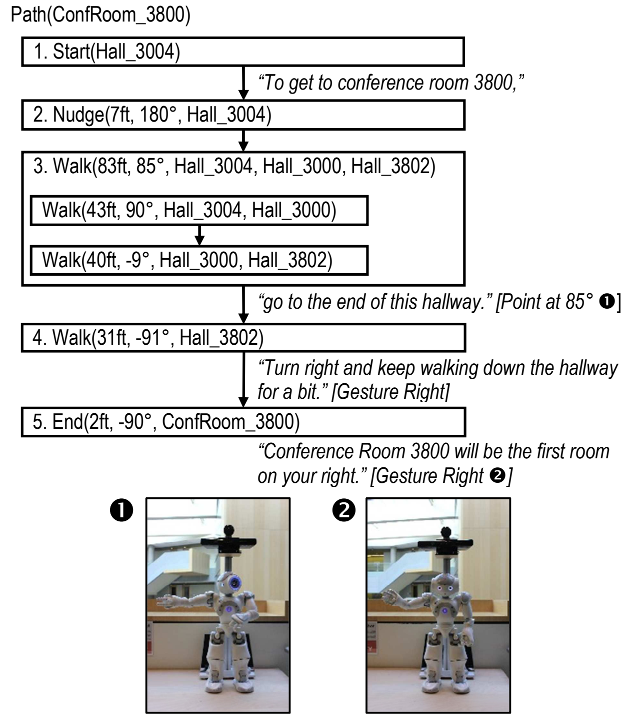


**Figure 3. Above: Path conversion to directions**
**Below: ❶ deictic gesture, ❷ route-relative gesture**

elevation angle, indicating the straight-line direction to the destination in both the horizontal and vertical planes.

### *3.2.6 Output planning and behavioral control*

The output planning component receives high-level semantic outputs from the dialog manager and renders them into speech coordinated with gestures on the robot. For each output, a plan is constructed, consisting of spoken utterances and gestures. The execution of the output plan is governed by an output behavior, which takes into account scene analysis information, *e.g.* whether the robot currently has the floor or is talking. Other low-level robot behaviors, such as gazing towards people that walk by when it is not engaged in a conversation, are controlled from this layer.

### *3.2.7 Robot control*

The robot is controlled via a set of behaviors activated by signals received from the output planner. For example, the robot moves its head to follow people when a new actor is detected in the scene, providing a strong signal of attention and awareness. The position of the actor in the scene and an estimate of distance from the robot based on face size are translated into head joint angles (yaw and pitch) that are sent to the robot and updated periodically, causing the robot to appear to be actively looking at the actor. Robot-relative gestures are performed by a gesture behavior which selects from a set of pre-authored timelines defining robot joint angle key frames, and sends them to the robot. For absolute, directional pointing gestures, the behavior computes at runtime the key frame angles based on the horizontal angle and elevation as specified in the output plan, which may also define the duration of the gesture.

An expression behavior controls the LEDs on the robot to indicate its engagement state. When idle, LEDs on the robot's body are dimmed and during interactions their intensities are raised. In addition, to prevent the robot from appearing too stiff and lifeless, a behavior periodically performs small joint motions and activates the eye LEDs to simulate blinking.

# 4. DEPLOYMENT

We deployed the Directions Robot in a space close to the elevators on a floor of our building (Figure 1). The traffic in this area includes people with offices on the floor, as well as visitors who come to see people on the floor and to attend meetings in various conference rooms. Numerous visitors are unfamiliar with the surroundings.

The robot is setup on a table located directly ahead of the elevators, and interactions occur at eye level (Figure 1). Signs about the robot are posted inside the elevators, on approach corridors, and near the robot, and alert people to the presence and basic operation of the robot, as well as disclose that the system is recording for research purposes. The robot operates during business hours: it is manually started each morning, and each evening it turns off automatically and returns to a resting position (Figure 1).

The robot was deployed for several weeks, including multiple development and testing cycles. Below, we discuss lessons learned by observing interactions with the robot over a period of two weeks. During the first week, it employed the heuristic engagement model described above. In the second week a machine-learned model was used to infer engagement actions. In total, the robot initiated 565 interactions. We eliminated one interaction that was terminated due to a fatal error, and 14 interactions with one of the authors. In addition, we excluded from the second week an atypical sequence of 51 engagements with an employee's small child whose height, small face, and location at the edge of the table frequently led to incorrect engagement decisions being made in quick succession.

# 5. DISCUSSION

The open-world deployment made salient several interesting challenges. The robot frequently has to grapple with multiparty situations. An analysis of interactions from the second week shows that groups of two or more (up to six) people were detected in 57% of interactions. The presence of multiple people leads sometimes to parallel, side conversations, and a divided focus of attention, where people attend to and communicate with others before, during, and after their engagement with the system. This leads to important challenges in correctly managing engagement. It also increases the challenges of accurate turn taking and speech recognition, affecting the system's performance. For example, natural breaks in a conversation with the robot to speak to others about the robot (of a nature that might be assumed to be easily detected in a social engagement where all actors are people) can be confusing to the robot and lead to inappropriate actions. Also, people sometimes ask the robot about topics outside its expertise on employees, locations, and directions, highlighting the need for the graceful handling of out-of-domain requests. Visitors also may react to the novelty of the system and its engaging and intelligent appearance with curiosity and tests of its abilities, as well as humorous utterances. We see the need for distinguishing between genuine requests for directions versus interactions driven by curiosity and playfulness, for developing more social awareness, and engaging in friendly chit-chat. Below, we discuss some of these observations in more detail, and highlight directions for future work.

## 5.1 Engagement

During the first week, the robot used the heuristic engagement model we described in subsection 3.2.3, with minor tuning to the camera properties and to the specific location of the robot. A back-off rule was used in situations where the face tracker lost an already engaged actor: the actor was temporarily persisted, but the engagement action probability was decreased over time, at a constant rate, from the last known value toward zero.

During this first week, 18% of the 249 engagements were incorrectly started by the robot. Most of the initiation errors occurred because the heuristic overestimated the probability of engagement when people passed by on a close trajectory to the robot. In addition, 8% of the 203 correctly initiated engagements were incorrectly terminated by the robot prematurely. These were based in speech recognition errors (the system misunderstood that the user wanted to terminate the interaction when in fact they did not), face tracking problems, and incorrect predictions by the engagement action heuristic.

Previous work [14, 26] has shown that additional features about the visual focus of attention and the temporal dynamics of location and attention are important in estimating engagement actions and intentions. In an effort to improve performance, we constructed a predictive model via supervised learning to predict at every frame whether a given actor is in an f-formation with the robot. Labels were acquired by manual annotation of three days of data, collected during the first week (2358 actor traces, about 537K frames) with the help of a professional tagger who was contracted to watch logged videos and use a multimodal annotation tool that we developed. For learning, we used features describing the location and size of the actor's face and the face-frontal indicator from the tracker. To model temporal dynamics, we constructed temporal features over these streams, including measures of slope, mean, standard deviation over windows of 0.25, 0.5, 1, 2, 4, and 8 seconds. Additional features, such as the total number of actors in the frame and the time since the actor was first observed were included. A logistic regression model was trained to predict the existence of an f-formation. In a batch evaluation, it obtained a frame-based classification error rate of 4.9% (in a 5-fold cross-validation), a significant lift over the majority baseline of 22.2%.

During the second week of the observational study, we used the trained model. The model backed-off to the original heuristic when features were missing *e.g.*, on the first tick some derived features like slope, standard deviation, etc. are not available, and used the same lost faces back-off method described earlier. With the new model, 6% of the 250 interactions were incorrectly initiated. This represents a sizable reduction from the first week. However, we saw an increase in the number of incorrect (early) terminations of engagement to 22%. We note that the incorrect initiations of engagement observed during the first week generally did not appear to be costly: the robot says "*Hi!*" to people passing close-by. In contrast, the incorrect, abrupt disengagements observed in the second week with the use of the data-driven model appeared to be significantly more costly. In these cases, the robot assumed that the participant was disengaging, interrupted itself, and terminated the interaction right away, sometimes much to the surprise of the participant. These disengagements happened as the engagement model predicted with confidence greater than 0.8 that an actor had left the f-formation with the robot. We found that this often occurred as participants turned temporarily towards other engaged participants or bystanders, or performed quick body or head motions, such as when laughing heartily or shifting positions during the interaction. In fact, the rate of incorrect disengagements was twice as large when two or more people were present, compared to when only one person was present: 28% versus 14%.

The data analysis we present focuses on inspecting interactions initiated by the robot. As such, we do not identify potential missed opportunities for engagement, or finer-grain issues related to the timing of engagement decisions, *i.e.,* how early or late these decisions are made, as people approach or depart from the system. Nevertheless, a closer look at the errors identified in this analysis brings to the fore a number of insights, highlighting the limitations of our current inference and decision making mechanisms. We discuss these below.

### 5.1.1 Engagement inferences

**Better visual tracking**. A number of incorrect disengagements were related to errors in the face tracker. We found that fast motions, including the turning of body and head by participants, where their faces did not remain in a frontal pose sometimes led to a loss of tracking of faces, compounding the challenges of inferring engagement. We expect that improvements in the lower-level tracking infrastructure will lead to further reductions in engagement and interaction errors.

**Beyond proximity and attention.** Often, the estimated probability of f-formation dropped sharply when an actor simply turned their head and focus of attention towards other bystanders, or to look in the direction the robot is pointing. While loss of attention and head motion can indicate a loss of f-formation, the data highlights the need to take more context into account when making inferences about engagement. Engagement is really not just about the actor, and proxemics and attention, but also about the interaction between the actor, robot and others. First, we must include the robot in the equation: features that characterize the robot's state, *e.g.* is the robot pointing somewhere, could provide additional relevant context for engagement inferences. Second, including additional features about the state of the conversation would be informative. For example, more robust inferences might take into consideration whether the robot is talking, waiting for the participant, giving directions, or asking a question. Finally, the wider context, including the presence of other bystanders and their behaviors, could be taken into account. We observed nuanced differences in behaviors with groups of different sizes, and depending on whether or not bystanders were part of the same group as the engaged participants. We believe models that reason jointly about all actors present, their group relationships, and engagements hold promise for further gains in the accuracy of inferences. Finally, additional sensory data (*e.g.* skeletal tracking) and more sophisticated base-level classifiers may lead to further improvements.

**Engagement intentions versus engagement actions.** In some of the incorrect disengagements participants turned and broke the f-formation, but did so temporarily, only to turn back moments later. This indicates that equating engagement intentions with actions, or with being in an f-formation, as our current implementation did, is problematic, especially in multiparty settings. Actors might briefly break an f-formation but still have an intention to maintain engagement. A better approach might be to train machine learning models that directly predict intentions, and take into account broader contextual factors, such as the presence of others, whether side conversations are going on, the history and rate of *connection events* [27] between the participant and the robot.

### 5.1.2 Engagement decisions

While better inferences may lead to better decisions, we believe an even more important limitation highlighted by the data has its roots in the framing of the decision making of the current system. Shifts in the decision making approach, and in the design of the action space could address the observed problems with engagement, as well as enable new scenarios.

**Decision making.** The robot used a myopic approach, where it made decisions to engage or disengage by comparing the probability of the user's estimated intention at every frame, against a preset 0.8 threshold. We have seen that this creates significant problems, especially in conjunction with the intention-same-as-action assumption. To a degree, these poor decisions are also a result of the choice of threshold, which did not reflect well the high cost of a decision to incorrectly terminate an engagement, as opposed to incorrectly continuing and reassessing at the next frame.

The disengagement problem reflects a wait-versus-act tradeoff, where the robot must balance the costs of inappropriately disengaging against the costs of continuing the interaction while participants are leaving. Less myopic approaches that reason about how the belief over the engagement intention might evolve in the future, and about different costs of incorrectly disengaging or incorrectly maintaining at different points in the interaction may lead to better outcomes. In reasoning about the future, the robot may attempt to distinguish between temporary disengagements, in which case it could suspend the conversation momentarily, and permanent ones, in which case the robot could terminate the engagement and bid goodbye.

**Engagement initiation.** We have seen that using a machine learned model that predicts whether actors are in an f-formation helps to eliminate some of the incorrect initiations of engagement. The robot lets the user initiate the engagement and responds to the initiation. Another situation however that is interesting to consider is creating robot-initiated engagements at a distance. For instance, if the robot detects that a person is lost, *e.g.* they are hesitant in their motion, glance back-and-forth, look at their smartphones, *etc.,* the robot could proactively initiate an engagement at a distance: *"Excuse me, can I help you find something?"* Another scenario we would like to support is that of a fly-by user-initiated engagement, where a person in a hurry quickly quips *"3042?"* as they go by, and the robot responds: *"Second room on the left!,"* understanding the situation of a user needing help in stream with getting to a location. This could be accomplished by broadening the scope of recognized engagement actions to also include user utterances, even if spoken from a distance and while the actor is in motion.

The data also highlights the importance of leveraging short- and long-term memories of interactions in determining how to start each conversation. Currently, the robot begins each conversation anew, as if it is seeing the participant for the first time. However, sometimes interactions happen in bursts, where one person disengages, and then he or she, or someone else in their group, initiates another engagement. In these cases, contextualizing the engagement in an appropriate manner, *e.g. "Oh, is there something else?"* if directions were previously given, or *"Oh, I thought you were leaving"* if no directions were given, would help create more natural interactions and mitigate prior incorrect disengagements.

**Managing groups.** For the majority of interactions, groups of people were present. We believe more sophisticated policies can be designed that consider when bystanders enter or leave an existing interaction, and generally, how the robot manages the group of actors and the f-formation. For instance, the robot may act to identify groups and bystanders, *e.g. "Are you all together?"*, *"Are they with you?,"* or to encourage bystanders to approach or to indicate that they are not part of the group: *"You guys can come closer if you want,"* or to stand back: *"Okay, I let me work with two people at a time here, can you guys step back for a second?"*. Controlling the number of participants in an engagement may help improve performance, as speech recognition accuracy is affected by distance, overlapping speech, and background noise.

**Termination.** The data also highlights the value of custom-tailoring the system's output during disengagement actions to the situational context. For instance, in some cases, as the robot began giving directions, the human actor started walking in the direction indicated. In this case, the immediate (mid-word) ending of the robot's spoken directions, followed by *"Well, I'll catch you later* *then"* is inopportune. The system should continue for a bit longer, or encourage the user to stay: *"Wait, wait, I'm not done."* We plan to investigate further improvements in the speech and gesture generation components to allow for re-planning, and continuously taking the scene context into account.

## 5.2 Task performance

Next, we discuss some observations regarding the overall task performance and the contents of the interactions.

### *5.2.1 In-domain interactions*

We inspected the 235 interactions correctly started by the robot during the second week and determined when users asked for directions in the known robot domain, and whether the robot provided the corresponding directions *by the end of the interaction*. In 59% of them, one or more questions for directions in the building were asked. The total number of questions asked was 154, and for 83% of these questions, the correct answer was given by the end of the interaction. We are encouraged by the high completion rate. At the same time, we observe that speech recognition problems and misunderstandings are still quite frequent, and sometimes users have to repeat the question multiple times to be understood. In future work, we plan to investigate challenges with speech recognition in this open-world setting.

Given the novelty of the robot, especially following a recent deployment, numerous interactions are driven by curiosity rather than a genuine need for directions. We estimated whether the in-domain questions stemmed from a real need for directions or whether people were simply testing the robot. Discrimination of these intentions is sometimes difficult, and we aimed for a conservative estimate of need-based interactions: if there was an indication that people might be testing, *e.g.* repeated interactions, people not walking after the directions were provided, asking for well-known locations like cafeteria or elevators, *etc.*, we marked the interaction as a test. According to this analysis, 23% of the 138 interactions with in-domain questions asked (14% of the 235 correctly initiated interactions) were genuine, need-based engagements. While the presence of a novel robot attracts attention and curiosity-based interactions, it is also serving a real need. Furthermore, the robot correctly gave directions in 100% of these need based-interactions.

### *5.2.2 Out-of-domain utterances*

About a quarter of the interactions (22%) contained at least one out-of-domain query or assertion addressed to the robot. An important category of our-of-domain utterances includes social commentary or questions and attempts at chit-chat, *e.g.* "*What's your name?" "You're cute", "What is your favorite food in the cafeteria?"* With the current grammar-based recognition approach, such utterances at best lead to non-understandings, and a *"Pardon me?"* from the robot. At worst, out-of-domain utterances are misunderstood as in-domain directions queries. The data highlights the need to develop the robot's capabilities in this social dimension. We anticipate that as the robot becomes more socially competent, people will also engage it in ever more sophisticated ways.

Other frequent out-of-domain utterances include requests that are on-topic per people's understanding of the system's competencies. These requests include queries about people not in our building, about other buildings on campus, or activity-based location queries such as *"Where can I eat?"* In future work, we plan to investigate methods for detecting out-of-scope queries, classify them by topic, and handle them in a more appropriate manner.

## 5.3 Directions giving

We have not yet conducted a qualitative or quantitative assessment of the effectiveness of the robot's directions. However, we made several observations about challenges that came to the fore over the development and testing of the Directions Robot, highlighting opportunities for future work.

### *5.3.1 More effective directions*

We faced the challenge of communicating long paths, resulting in directions that are difficult for the listener to understand and retain. Taking into account not only the Euclidean distance, but also the cognitive complexity of generated directions when scoring paths, *e.g.* number of turns, *etc.* could help improve the comprehensibility and memorability of directions. For long, complicated directions, the robot might change its strategy and simply point to the location of the destination "as the crow flies," and allow the user to use the persistent visualization of the pointed-out location to navigate on their own. So far, we have made limited use of auxiliary map information, mainly to identify the ordinal location of a destination room on a hallway. However, landmarks could also serve as useful discriminators or navigational waypoints, *e.g. "turn left at the corridor just after the copy room"*. Other features such as accessibility of alternate routes could also be used to better tailor directions to the users' needs.

We also seek to leverage the dialog management capabilities of the system to chunk [6] directions and make the process more interactive. For example, the system may begin by first providing summary directions, then establish which additional details (if any) to provide to the user. In this back-and-forth, the robot must continuously assess levels of grounding, and tailor its instructions appropriately. It might elicit and listen for backchannels, head nods, or other confirmatory signals at natural points where users would communicate comprehension. The analyses could be further informed by inferences about user state and intentions, *e.g.* is the user in a hurry or confused?

### *5.3.2 Gestures and natural language*

The deictic gestures used to indicate the initial trajectory or the general direction of the destination provide strong value. Future work could explore combining these with additional gestures to convey multiple navigation steps in a single composite gesture. So far, we adopted a simple approach that maps path segments directly to sentence parts, with random selection of word and phrase alternates where possible to provide some variability. The majority of the gestures used were pre-authored on fixed timelines.

We plan also to investigate algorithms for runtime generation of speech and gesture that also take into account system state and information from the scene analysis. Variables such as the actors' location in the scene, their focus of attention, levels of grounding, the previous gestures performed, *etc.*, may all be used to influence the timing, surface realization, and parameters of the speech and gesture. For example, the system can look in the direction that it is gesturing about, but recurrently bring its attention back to the user with eye contact and a brief pause to check for understanding—or to appear to do that while giving the user a chance to comprehend. The ultimate goal is to generate coordinated outputs that are congruent with the situation at hand.

# 6. CONCLUSION

We have described *Directions Robot*, a platform for investigating the challenges of open-world human-robot interaction. An initial deployment in the wild has highlighted several challenges when the system interacts in dynamic, multiparty environments, where people come and go, and interleave their interactions with the system and with each other. These include challenges with creating, initiating, and terminating conversational engagements in an appropriate manner, with out-of-domain utterances, social interaction, and directions giving.

The lessons and insights gained from this initial deployment help shape future research. We plan to investigate decision-theoretic approaches to managing wait-versus-act tradeoffs related to engagement decisions, in conjunction with expanded inferences that take into account groups, relationships, and the broader social context of the interactions. Future work on engagement will also include the development and real-time control of a richer repertoire of actions, such as temporary suspensions and reprisals. Beyond managing engagement, we plan to explore the design of extended social awareness and competencies, as well as more effective directions via gesture and speech.

## 7. ACKNOWLEDGMENTS

We thank Rebecca Hanson for assistance with data annotation and Anne Loomis Thompson for contributions to the project.

## 8. REFERENCES

[1] Dale, R., Geldof, S., and Prost, J., 2005. Using Natural Language Generation in Automatic Route Description, *in Journal of Research and Practice in Information Technology*, Vol. 37, No. 1, February 2005.

[2] Look, G., Kottahachchi, B., Laddaga, R. and Shrobe, H., 2005. A Location Representation for Generating Descriptive Walking Directions, *in Proc. of IUI'2005*, pp. 122-129.

[3] Cassell, J., Stocky, T., Bickmore, T., Gao, Y., Nakano, Y., Ryokai, K., Tversky, D., Vaucelle, C., and Vilhjalmsson, H. 2002. MACK: Media lab Autonomous Conversational Kiosk, *in Proc. of Imagina'02*, Monte Carlo.

[4] Nakano, Y., Reinstein, G., Stocky, T., and Cassell, J., 2003. Towards a Model of Face-to-Face Grounding, *in Proc. of ACL'03*, Sapporo, Japan.

[5] Stocky, T., and Cassell, J., 2002. Shared Reality: Spatial Intelligence in Intuitive User Interfaces, *in Proc. of IUI'2002,* San Francisco, CA.

[6] Cassell, J., Kopp, S., Tepper, P.A., Ferriman, K., and Striegnitz, K., 2007. Trading Spaces: How Humans and Humanoids use Speech and Gesture to Give Directions, *in Conversational Informatics: An Engineering Approach* (ed T. Nishida), John Wiley & Sons, Ltd, Chichester, UK.

[7] Theune, M., Hofs, D., and Kessel, M., 2007. The Virtual Guide: A Direction Giving Embodied Conversational Agent, *in Proceedings of Interspeech 2007,* Antwerp, Belgium.

[8] Theune, M., and Brandhorst, C.J., 2010. To Beat or Not to Beat: Beat Gestures in Directions Giving, *in Proc. of the 8th International Conference on Gesture in Embodied Communication and Human-Computer Interaction.*

[9] Byron, D., Koller, A., Oberlander, J., Stoia, L., and Striegnitz, K., 2007. Generating Instructions in Virtual Environments (GIVE): A challenge and evaluation test-bed for NLG, *In Proceedings of the Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation*, Arlington, VA.

[10] Okuno, Y., Kanda, T., Imai, M., Ishiguro, H., and Hagita, N., 2009. Providing Route Directions: Design of Robot's Utterance, Gesture and Timing, *in Proc. of HRI'09*, San Diego, CA.

[11] Marge, M., and Rudnicky, A.I., 2010. Comparing Spoken Language Route Instructions for Robots across Environment Representations, *in Proc. of SIGdial'2010*, Tokyo, Japan

[12] Hasegawa, D., Cassell, J., and Araki, K. 2010. The Role of Embodiment and Perspective in Direction-Giving Systems, *In Proc. of Dialog with Robots AAAI Fall Symposium*, Arlington, VA.

[13] Evers, M., Theune, M., and Karreman, J. 2007. Which way to turn? Guide orientation in virtual way finding, *in Proceedings of Workshop on Embodied Language Processing*, Prague, Czech Republic.

[14] Michalowski, M., Sabanovic, S., and Simmons, S., 2006. A Spatial Model of Engagement for a Social Robot. *in Proc. of AMC'2006*, Istanbul, Turkey.

[15] Sabanovic, S., Michalowski, M., and Simmons, R., 2006. Robots in the Wild: Observing Human-Robot Social Interaction Outside the Lab, *in Proc. of AMC'2006*, Istanbul, Turkey.

[16] Gockley, R., Bruce, A., Forlizzi, J., Michalowski, M., Mundell, A., Rosenthal, S., Sellner, B., Simmons, R., Snipes, K., Schultz, A.C., and Wang, J., 2005. Designing Robots for Long-Term Social Interaction, *in Proc. of IROS'2005*, Edmonton, Canada.

[17] Rosenthal, S., Biswas, J., and Veloso, M., 2010. An Effective Personal Mobile Robot Agent Through Symbiotic Human-Robot Interaction, *in Proc. of AAMAS'10*, Toronto, Canada.

[18] http://research.microsoft.com/~dbohus/videos/aamas14.wmv

[19] Gouaillier, D., Hugel, V., Blazevic, P., Kilner, C., Monceaux, J., Lafourcade, P., Marnier, B., Serre, J., Maisonnier, B. 2009. Mechatronic design of NAO humanoid. *In Proceedings of ICRA'09*, Kobe, Japan.

[20] Sidner, C.L., Kidd, C.D., Lee, C., and Lesh, N. 2004. Where to Look: A Study of Human-Robot Engagement, In *Proceedings of IUI'04*, pp. 78-84, Madeira, Portugal.

[21] Sidner, C.L., Lee, C., Kidd, C.D., Lesh, N., and Rich, C. 2005. Explorations in engagement for humans and robots, *Artificial Intelligence*, 166 (1-2), pp. 140-164.

[22] Bohus, D., and Horvitz, E., 2009. Models for Multiparty Engagement in Open-World Dialog, *in Proc. of SIGdial'2009*, London, UK.

[23] Kendon, A. 1990. Spatial organization in social encounters: the F-formation system, *Conducting Interaction: Patterns of behavior in focused encounters*, Studies in International Sociolinguistics, Cambridge University Press.

[24] Bohus, D., and Horvitz, E., 2011. Multiparty Turn Taking in Situated Dialog: Study, Lessons, and Directions, *in Proc. of SIGdial'2011*, Portland, OR.

[25] Shafer, S., 2009. A Framework for Creating and Using Maps of Privately Owned Spaces, *in Proc. of LOCA'2009*, Tokyo, Japan.

[26] Bohus, D., and Horvitz, E., 2009. Learning to Predict Engagement with a Spoken Dialog System in Open-World Settings, *in Proc. of SIGdial'2009*, London, UK.

[27] Rich, C., Ponsler, B., Holroyd, A., Sidner, C., 2010. Recognizing Engagement in Human-Robot Interaction, *in Proc. of HRI'10*, Osaka, Japan.