Learning Interactive Behavior for Service Robots – the Challenge of Mixed-Initiative Interaction

Phoebe Liu, Dylan F. Glas, Takayuki Kanda, *Member, IEEE*, and Hiroshi Ishiguro, *Member, IEEE*

***Abstract*— Learning-by-imitation approaches for developing human-robot interaction logic are relatively new, but they have been gaining popularity in the research community in recent years. Learning interaction logic from human-human interaction data provides several benefits over explicit programming, including a reduced level of effort for interaction design and the ability to capture unconscious, implicit social rules that are difficult to articulate or program. In previous work, we have shown a technique capable of learning behavior logic for a service robot in a shopping scenario, based on non-annotated speech and motion data from human-human example interactions. That approach was effective in reproducing reactive behavior, such as question-answer interactions. In our current work (still in progress), we are focusing on reproducing mixed-initiative interactions which include proactive behavior on the part of the robot. We have collected a much more challenging data set featuring high variability of behavior and proactive behavior in response to backchannel utterances. We are currently investigating techniques for reproducing this mixed-initiative behavior and for adapting the robot's behavior to customers with different needs.**

I. Introduction

As robotic technologies improve, the possibility of service robots in the real world becomes closer to reality. Service robots will need to interact directly with human users, raising a number of difficult challenges. One such challenge is the problem of how to create the overall application logic for interactive robots, including interactive dialog and interactive motion planning.

Although high-level interaction logic might traditionally be programmed manually by an interaction designer, we propose a data-driven technique, in which machine learning techniques could be used to learn application logic through imitation of human behavior. We propose that for situations where large amounts of example human-human interaction data is available, such data-driven approaches could produce more reliable interaction logic and require less effort than manual programming.

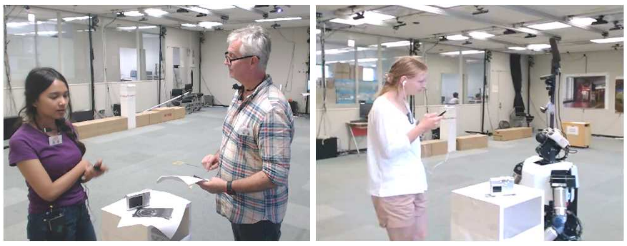


Fig. 1. Overview of the proposed learning technique. Left: Collecting data from example human-human interactions. Right: Reproducing human behavior in a service robot.

In this paper, we will first summarize our previous work on this topic, in which we demonstrated a technique for reproducing interactive behaviors which were primarily user-initiated [1, 2]. We will then present the current state of our work in progress, which examines the challenge of mixed-initiative interaction. We will describe the data set we have collected for this new project and discuss a possible solution for an extension of our technique to mixed-initiative interaction.

II. Related work

A. Learning from Data

In many areas of robotics, machine learning approaches such as learning-by-demonstration are often utilized to learn from a dataset of examples in order to reproduce a demonstrated task, as it is easier for humans, including non-robotic-experts, to input poses, e.g., by moving an arm manually, than to explicitly specify them numerically. Some examples include trajectory following [3, 4] or joint motion replication [5]. Often these approaches are used with low-level inputs such as sensory-motor patterns, rather than cognitive and decision-making skills.

In social robotics as well, machine learning has been used to teach low-level behaviors, for example to mimic gestures and movements [6] and to learn how to direct gaze in response to gestural cues [7]. In one example, pointing and gaze behaviors were recognized in an imitative game using a hidden Markov model [8].

Data-driven dialogue systems have been demonstrated in robots which infer meanings from spoken utterances. Rybski *et al.* developed an algorithm which allowed a human to interact with a robot with a subset of spoken English language in order to train the robot on a new task [9]. Meena *et al.* used a data-driven chunking parser for automatic interpretation of spoken route directions for robot navigation [10].

The focus of our work differs from these other works in that

*This work was supported in part by the JST ERATO Ishiguro Symbiotic Human-Robot Interaction Project and in part by JSPS KAKENHI Grant Number 25240042.

The authors are with the Advanced Telecommunications Research Institute International, Kyoto 619-0288, Japan (e-mail: phoebe@atr.jp; dylan@atr.jp; kanda@atr.jp).

Hiroshi Ishiguro also belongs to the Intelligent Robotics Laboratory, Osaka University, Toyonaka, Japan. Email: ishiguro@sys.es.osaka-u.ac.jp. P. Liu, D.F. Glas, and H. Ishiguro are also with the JST ERATO Ishiguro Symbiotic Human-Robot Interaction Project.

we are trying to reproduce overall high-level interaction logic, rather than specific elements of interaction, based on training examples observed from real human-human interaction, with natural spoken dialogue.

B. Using the crowd for learning

With the advancement of high-precision tracking systems able to monitor real social environments [11, 12], it is becoming possible to collect large amounts of detailed interaction data with little effort. This suggests the possibility of using a "crowdsourcing" approach, like the distributed techniques used over the web to solve complex problems, e.g. users on Amazon's Mechanical Turk helping to annotate images for grasp planning [13].

The use of real human interaction data collected from sensors for learning interactive behaviors has been investigated in numerous works. The robot JAMES was developed to serve drinks in a bar setting, in which a number of supervised (i.e. dialog management) and unsupervised learning techniques (i.e. clustering of social states) have been applied to learn social interaction [14]. In contrast, we propose a completely unsupervised approach for both abstraction and clustering of social states as well as for robot behavior generation

In Young et al.'s work [15] [16], a person provides an example of an interactive locomotion style, which is used to teach the robot to generate interactive locomotive behaviors in real time according to that style. We also propose to use real human interaction to train the robot, but our focus is not only the robot's motion, but its speech as well.

Connectivity to the web has also changed the way interaction data can be collected. The Robot Management System framework was developed to make learning of manipulation and navigation tasks easier by collecting demonstrations from remote users through a browser as a game [17]. The Restaurant Game used annotated crowdsourced data to generate abstracted representation of data to automate game characters [18]. The Mars Escape online game used crowdsourcing to learn robot behaviors [19-21]. The idea was to use a data-driven approach to develop human robot interaction (HRI) behaviors from players of an online collaborative game to provide large amounts of training data and reproduce behaviors in a real autonomous robot.

Our work complements these approaches by considering a crowd-based data collection from sensors in a physical environment, where some new challenges include resolving recognition ambiguities due to sensor noise and natural variation of human behavior.

III. Materials and Methods

In our previous study, we sought to reproduce speech and locomotion behaviors of participants role-playing a shopkeeper in a camera shop scenario [1]. Since this study is an extension of that work, we will summarize the important points of that technique here.

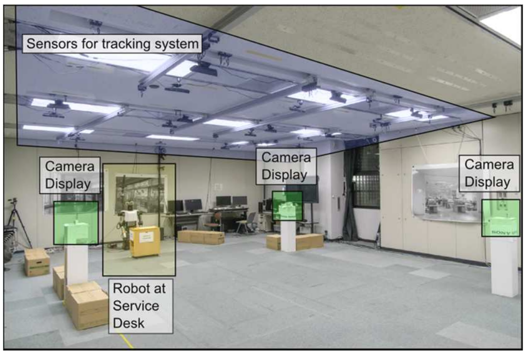


Figure 2. Environment setup for our study, featuring three camera displays. Sensors on the ceiling were used for tracking human position, and smartphones carried by the participants were used to capture speech.

A. Objective

In the near future, we expect that it will be feasible to place sensors in real social environments which can passively observe human social behaviors, such as motion and speech. Such technologies would allow enormous amounts of interaction data to be collected in real interactive environments, such as retail shops, care centers, schools, or homes. By using such data, we believe it should be possible to train robots to perform the socially-interactive duties of service providers in those environments.

The goal of our previous study was to provide a proof-of-concept demonstration that it is possible to learn high-level interaction logic from passive observation of human behavior, in an entirely data-driven way. That is, the technique should not depend on any kind of manual annotation or cleanup of the sensor data. We see this as an important requirement in order to utilize "big data" such as the interaction data described above.

For this study, we chose a camera shop scenario, in which two participants, one representing a shopkeeper and one representing a customer, role-played interactions in a simulated shop environment, in which the shopkeeper provided recommendations and answered the customer's questions about features of three cameras (Canon, Sony, and Panasonic). These interactions involved both dialog and motion, as the participants needed to walk between the camera displays in the shop. The objective was to learn the shopkeeper's speech and motion behavior logic so that the shopkeeper could be replaced with a humanoid robot.

B. Data collection technique

To capture the participants' motion and speech data, we used a human position tracking system to record people's positions in the room (Fig. 2), and we used a set of handheld smartphones for speech recognition.

The position tracking system used data from 16 Microsoft Kinect 1 sensors arranged in rows on the ceiling. Particle filters were used to estimate the position and body orientation of each person in the room based on point cloud data [11].

Speech was captured via a smartphone with a hands-free headset, using the Android speech recognition API to recognize utterances and sending the text to a server via Wi-Fi. Users were required to touch the mobile screen to indicate the beginning and end of their speech. Ideally, we would like to use automatic detection of speech activity and to collect data passively, using sensors mounted in the environment, but reliable technologies to do this are not yet easily available.

Using this data capture system, we collected 178 interactions, including 1194 customer utterances and 1233 shopkeeper utterances.

C. *Learning techniques*

The details of data processing, abstraction, vectorization, and learning are fairly complex, so we will only summarize them here. For a full explanation, please see our journal paper on this work [2].

Our basic approach was to abstract the behaviors of the shopkeeper into a finite set of discrete speech and motion "actions," each of which could be reproduced with a robot. A classifier was then trained such that vectorized representations of customer actions could be used to predict when any one of these discrete robot actions should be executed.

1) *Abstractions*

The first main challenge of this approach was how to reduce the dimensionality of the human behaviors detected by the sensors into a usable and meaningful feature vector. We used several abstraction techniques to achieve this:

- We spatially clustered people's moving and stopped trajectories, to identify a discrete set of typical **stopping locations** (Fig. 3) and **motion trajectories** for each role (customer and shopkeeper). This way, movement could be modeled as simple sequences of moving and stopping, rather than using raw (*x*,*y*) position data. For stopping positions, we used k-means clustering, and for moving trajectories we used k-medoid clustering based on spatiotemporal matching using dynamic time warping.
- We identified common **spatial formations**, such as "face-to-face" and "present object", which correspond to existing HRI proxemics models. By modeling the spatial interaction as a series of transitions between different proxemics formations, the details of exact relative positioning can be computed by the model, reducing the amount of data needed for learning.
- We performed **speech vectorization** of the customer and shopkeeper using common text-processing techniques such as removal of stop words, stemming, enumeration of n-grams, and Latent Semantic Analysis, as well as using a pre-trained model from AlchemyAPI cloud-based service[1] to automatically extract keywords.
- To identify discrete **robot speech actions**, we clustered the shopkeeper utterances using dynamic hierarchical clustering [22]. A total of 233 utterance clusters were generated.

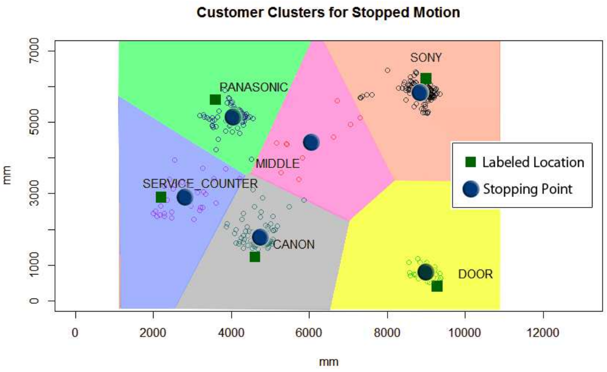


Figure 3. Customer stopping locations for the previous study, identified through unsupervised clustering.

2) *Action Discretization and Learning*

We developed a simple set of rules for discretizing customer and shopkeeper actions in the training data. We defined an "action" to have occurred whenever a participant spoke an utterance and/or changed their moving state (stopped to moving, or vice versa). When two shopkeeper actions were detected consecutively, they were merged into one action according to a set of rules. In this way, every interaction could be modeled as a sequence of customer actions followed by shopkeeper actions. This enables us to train a classifier to predict an appropriate shopkeeper action whenever a customer action is detected.

3) *Learning*

The robot's interaction logic was learned by training a classifier to predict a discrete robot action given an input consisting of a vectorized representation of the customer action. Specifically we used a Naïve Bayesian classifier, and the input vector was defined to contain the following information:

- Text vectorization of the customer speech
- Spatial state of the customer (location, motion origin, and motion target)
- Spatial state of the shopkeeper/robot (location, motion origin, and motion target)
- Spatial formation of the customer and shopkeeper/robot (face-to-face, present product, waiting, or other)

The final input vector was 393 dimensions (346 from text and keywords, and 47 from spatial information), and it was used to predict one of 467 possible robot actions (combinations of utterance clusters and target locations).

Robot actions were defined as follows:

- If the shopkeeper spoke an utterance in the training data, the corresponding **speech cluster ID** was included in the robot action.
- A **target location** for the action is identified. If the shopkeeper is stopped when the action is detected,

[1] http://www.alchemyapi.com

TABLE I. EXAMPLE TRAINING INTERACTION FROM PREVIOUS STUDY

C: Customer, S: Shopkeeper

C: *(Enters shop)*
S: (*Approaches customer*) Hi are you looking for anything in particular today?
C: Yes I would like to… I am looking for a camera with good storage memory.
S: (*Guides to Canon*) Ok the Canon Rebel XTi can hold 10000 photos.
C: Ok, that is very good. What about the price?
S: This camera is $400.
C: I see. Is it heavy?
S: Yes, very heavy.
C: How much?
S: Like, a kilogram.
C: I see, that is very heavy. Well I will think about it. Thank you. (*Leaves shop*)
S: Sure, no problem.

then that stopping location is used. If the shopkeeper is moving, then the motion target location is used.

- “None” was also included as a possible robot action.

During online operation, the predicted robot action was executed in the robot as follows:

- If the predicted action contains an utterance cluster ID, the robot speaks the “typical utterance” for that cluster, chosen by finding the utterance that has the highest average similarity score to other utterances in that cluster.
- If the predicted target location is different from the robot’s current location, the robot moves to the new target location.

D. Performance/evaluation summary from previous work

A comparison study was conducted to evaluate the robot’s performance using our proposed system, compared with a baseline system which did not use techniques such as clustering of utterances or abstraction of interaction states, which we consider to be the concepts at the core of our proposed technique. For the details of this study, please see [2].

The results of this evaluation showed that the proposed system significantly outperformed the baseline system in a variety of metrics, including social appropriateness, consistency of speech and motion, correctness of wording, and an overall evaluation.

Another interesting result showed that the proposed system produced socially-acceptable behaviors 84.8% of the time, whereas automatic speech recognition (ASR) accuracy was only 76.8%. That is, our proposed system was shown to be robust to errors in speech recognition.

TABLE II. EXAMPLE HUMAN-ROBOT INTERACTION FROM PREVIOUS STUDY

C: Customer, R: Robot

C: *(Enters shop)*
R: Hi, is there anything I can help you with today?
C: *(moves to Panasonic)* Yes, this camera, how's the battery life?
R: *(moves to Panasonic)* Long battery life of 9 hours.
C: Ah, I see. And the weight?
R: This camera is 150 grams and it can fit in your pocket.
C: Oh, that's very convenient. Thank you. *(begins walking away)*
R: Ok. *(moves back to service counter)*
C: *(moves to Canon)* How about this one? How is the storage memory?
R: *(while moving to Canon)* Yes, it can store 10000 photos.
C: How much is it?
R: This is $400.
C: Ok. What colors does it come in?
R: Intense grey, red, and brown colors.
C: Actually... *(moves back to Panasonic)*
R: *(moves to Panasonic)*
C: What's the optimal zoom for this camera?
R: 5x optical zoom.
C: Oh, ok. Thank you. *(begins walking to door)*
R: No problem. *(returns to service counter)*

We were quite pleased with the accuracy and lifelikeness of the robot’s behaviors, but that study had some limitations. The interactions in that study were mostly question-answer exchanges, and they included no representation of interaction history. Table I shows an example interaction from that study. Notice that the shopkeeper is always reacting to the customer. Even the shopkeeper’s first utterance can be modeled as a reactive behavior responding to the customer entering the shop. Table II shows an example of a human-robot interaction generated using that technique.

In our current work, we are aiming to enrich the set of learned behaviors to include situations where the robot can generate behaviors proactively, rather than always responding to questions. To achieve this goal, we needed to conduct a new data collection.

IV. PROACTIVE INTERACTION DATA

To capture naturally proactive behavior, we conducted a new data collection, with a single participant playing the role of shopkeeper. Through interviews and trial interactions, we chose a participant with a naturally outgoing personality and a great interest in cameras and photography. Our objective was to try and reproduce the proactive nature of his personality.

A. Data Collection

The data collection was conducted with the same setup as the first data collection, using the same room configuration, position tracking system, and smartphone-based speech recognition application. Three new camera models were chosen for the scenario.

Customer participants were instructed to browse as much or as little as they liked, and they could ask questions about cameras or simply listen to the shopkeeper's recommendations. To create variation in the interactions, customer participants were asked to role-play either "novice" or "advanced" customers and ask questions that would be appropriate for their role. Some camera features were chosen to be more interesting for novice users (color, weight, etc.) and others were more advanced (High-ISO performance, details of the autofocus system, etc.), although they were not explicitly labeled as such.

Customer participants were not given a specific target feature or goal for the interaction, as we were mostly interested in capturing the shopkeeper's proactive sales behavior. All participants were instructed to focus their discussion on the features listed on the camera spec sheet, to minimize the amount of "off-topic" discussion.

We recruited a total of 9 customer participants (8 male, 1 female, average age 34.1), who conducted 12 interactions each (6 as advanced and 6 as novice). The final data set included a total of 2568 shopkeeper utterances and 2299 customer utterances.

B. *Data Properties*

This interaction data differed from that of the previous study in a few ways. First, the shopkeeper's utterances tended to be much longer and more complex, sometimes talking about 2 or 3 topics in one sentence. Second, the shopkeeper often proactively spoke if some silence had elapsed after his last utterance. Third, the customers demonstrated more "backchannel" utterances. For example, a customer might say, "oh, ok," after listening to an explanation, but not ask a follow-up question. In such situations, the shopkeeper in this study often performed proactive behaviors, such as volunteering more information about the current camera or continued his previous explanation.

We performed a preliminary analysis of the customer utterances to identify whether an utterance required a response (such as a question or a request) or did not require a response (such as a backchannel utterance). We found that 527 (22.8%) of the customer's 2299 utterances did not seek a response from the shopkeeper. In these situations, we would expect that the shopkeeper could choose to perform some proactive behavior.

Table III illustrates an example interaction from the new data collection. Notice that after the shopkeeper explains the price, the customer agrees with him ("Oh, very cheap!"), but does not ask a further question. The shopkeeper then volunteers more information regarding the price. Next, after several seconds of silence, the shopkeeper proactively presents more information about a different feature.

Figure 4 also shows an interesting difference from the previous study, in that the stopping locations do not include "service counter". This is because the proactive shopkeeper walked out to meet customers rather than waiting for them to approach him.

These examples show that we can see qualitative differences in both speech and motion data for the new data set. We are currently exploring ways to extend our system to reproduce proactive behaviors like these.

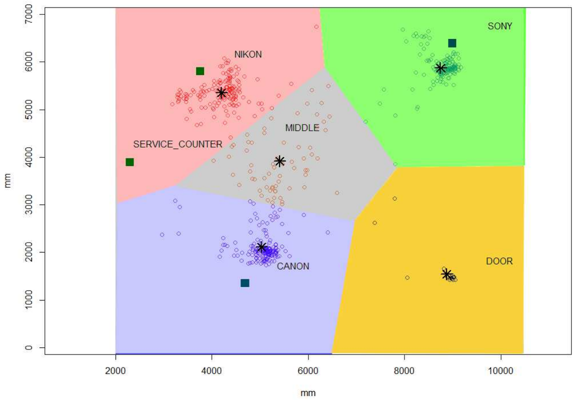


Figure 4. Customer stopping locations from the new data collection. Note that there are only five clusters, as opposed to the six from the previous study.

V. Proposed Techniques

One important consideration in generating proactive behavior is that some level of history representation will be necessary. For example, if a customer is silent or says "ok", then the shopkeeper's next utterance will depend to some degree on the shopkeeper's previous utterance. Other utterances may depend on deeper history, for example, if the shopkeeper needs to present a new feature of a camera without repeating features which have been presented before.

However, it is not clear how history of utterances should be represented. For example, it would be possible to store a Boolean value for each robot action which has been executed previously, but this would not preserve information about sequence. Another option would be to store a full sequence of the last *n* feature vectors, but this could increase the dimensionality to the point where learning useful behaviors would require an unreasonably large amount of data.

A. *Learning adaptive robot behaviors*

There are many techniques which have been developed for learning adaptive robot behaviors, such as goal-directed and habitual robot behaviors through a Bayesian dynamic working memory system [23], or incorporating history in learning for mobile robots [24, 25], we believe this problem is a bit closer to the field of language or dialog learning. In particular, many techniques involving deep neural networks have been developed recently for handling language-related tasks, which are inherently sequential and require some level of history or memory.

Recurrent neural networks are often used for tasks like language processing, and Long Short-Term Memory (LSTM) recurrent neural network techniques are often used for tasks such as word-by-word machine reading, where the meaning of a sentence can only be understood when interpreted in the context of previously encountered words [26].

Some techniques have been developed for generating automated dialog and answering questions, such as the Neural Responding Machine [27]. LSTM networks have also been

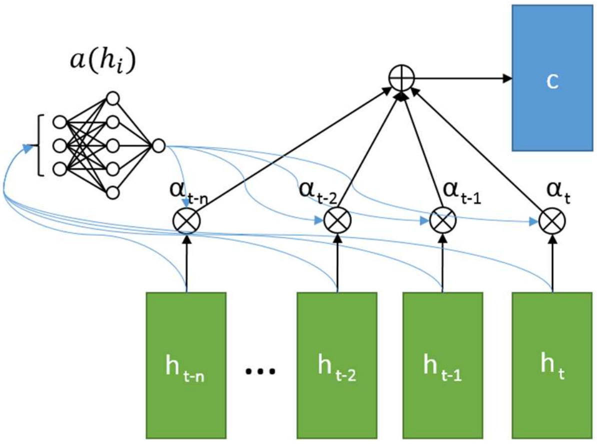


Figure 5. Attention model architecture.

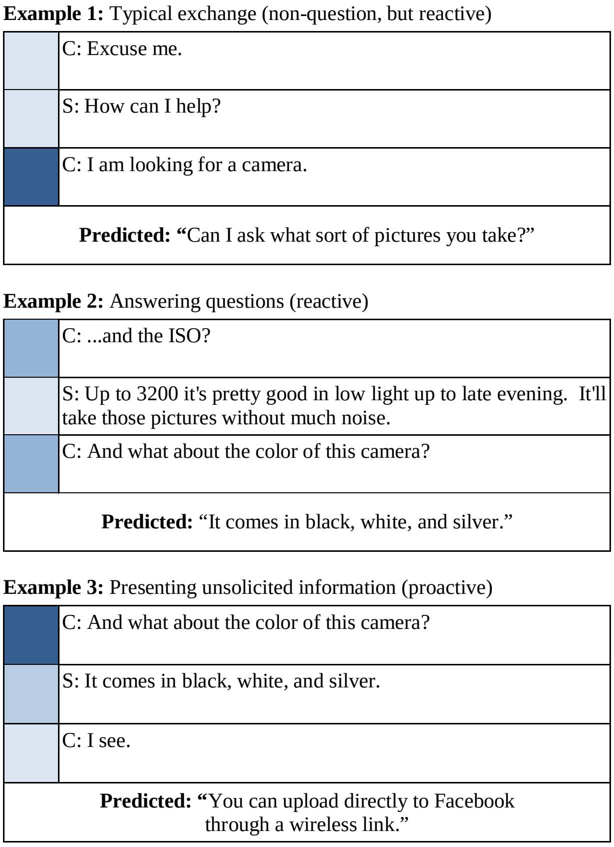
used in conjunction with convolutional neural networks for question-answering tasks[28]. A related technique is the End-to-End memory network, which has been used for tasks like language modeling and question answering [29]. This technique learns which part of an input sequence is important for predicting the answer to a question.

B. Attention Model

One deep-learning technique which we think seems promising for history representation is a structure called an "attention model". Studies based on attention mechanisms exist, such as modeling attentional modulation with flexible scheduling for periodic tasks of a behavior-based robotic system [30, 31]. In this work, we are considering an attention model like that proposed by Raffel and Ellis in [32].

The basic concept behind this attention model is illustrated in Fig. 5. In this network, the output vectors of multilayer perceptrons (h_{t-n} .. h_t) representing the information from the last n time steps (actions in the dataset) are multiplied by weighting factors (α_{t-n} .. α_t). The weighting factors themselves are learned from the vectors (h_{t-n} .. h_t) via a second multilayer perceptron. The weighted vectors are then combined additively to produce an output vector c, which is finally used to predict the robot action. The interesting part of this architecture is that the weighting factors α are determined dynamically, and thus, depending on the context, the system could automatically determine whether it should consider the most recent action or other actions from the interaction history in predicting the next robot action.

For example, in a question-answer situation, we would expect that the most recent customer action (that is, the question), represented by $h_{t,}$ would be the most important contributing factor to the decision process, so α_t should have the highest weight. In a situation where the customer said "I see," perhaps the shopkeeper's last utterance (h_{t-1}) or other utterances in the history would have stronger weight, because knowledge of these history elements would be more helpful in predicting an appropriate proactive behavior to avoid repetition or choose a new utterance relevant to the previous discussion.



Example 1: Typical exchange (non-question, but reactive)

░	C: Excuse me.
░	S: How can I help?
▓	C: I am looking for a camera.
Predicted: "Can I ask what sort of pictures you take?"	

Example 2: Answering questions (reactive)

▒	C: ...and the ISO?
░	S: Up to 3200 it's pretty good in low light up to late evening. It'll take those pictures without much noise.
▒	C: And what about the color of this camera?
Predicted: "It comes in black, white, and silver."	

Example 3: Presenting unsolicited information (proactive)

▓	C: And what about the color of this camera?
▒	S: It comes in black, white, and silver.
░	C: I see.
Predicted: "You can upload directly to Facebook through a wireless link."	

Figure 6. Examples of successful predictions using our attention network technique for a history length of 3. Shaded boxes show the relative weight assigned to each utterance, indicating its importance in predicting the final prediction. Darker shading indicates higher weight.

VI. Preliminary Results

Although this work is still in an early stage, we have implemented a version of this attention model. Reproducing proactive robot behaviors that are robust to noisy sensor data remains part of the challenge for this work. Here we will show some early results showing some examples of the effectiveness (and ineffectiveness) of this technique.

A. Learning System Implementation

To generate the embeddings (h_{t-n} .. h_t) in Fig. 5, we used autoencoders trained on a vectorization of the input text. Text was first processed according to the procedure used in the previous study (removal of stopwords, enumeration of n-grams, keyword identification, and latent semantic analysis), resulting in a 300-dimensional vector. Likewise, we extracted the keywords for each utterances, and processed the keywords in the same way, resulting in an additional 50 dimensions.

This vector was then input to a 4-layer autoencoder with 800 hidden units in each layer, then was trained with a *tanh* activation for each layer, with an output dimensionality of 200. Customer and shopkeeper utterances were embedded into different spaces, and we used a "leaky rectifier" (LReLU) nonlinearity to compute h_t:

$$h_t = LReLU(W_{xh}x_t + b_{xh})$$

The weighting factors were then computed using:

$$a(h_t) = \tanh(W_{hc}h_t + b_{hc})$$

We then added another hidden layer using leaky rectifier nonlinearity to compute the output:

$$y = LReLU(W_{cy}c + b_{cy})$$

The output y is trained against the target value with a cross entropy objective for 10000 epochs. Details of the leaky rectifier technique can be found in [33].

B. Prediction Examples

Figure 6 shows some examples of predictions made by our system. For simplicity of presentation, only utterances are shown, but our intention is to incorporate spatial data as well. These examples were generated by taking a sequence of three utterances from the training data (customer – shopkeeper – customer) and feeding them into the classifier to predict an output shopkeeper utterance.

Example 1 shows a typical exchange which occurs at the beginning of many interactions – when a customer asks for help in selecting a camera, the shopkeeper usually asks what kind of pictures he/she takes. In this example, the attention model has selected the most recent customer utterance ("I am looking for a camera") as the most important in predicting the shopkeeper's response.

Example 2 shows a typical exchange in which a customer asks several questions about features of a camera. In this case, the predictor correctly predicted the answer, but it seemed to consider both the customer's current and previous utterances as being equally important. We have seen many examples where this happens, and we currently have no good explanation of why this occurs.

Example 3 shows a typical example of a situation where the shopkeeper must generate proactive behavior which is not answering a question. In this case, the attention model chooses the customer's *previous* utterance as the most relevant. We hypothesize that this is because the customer's previous question helps to define the set of proactive behaviors which would be appropriate in this context. In this case, the system chooses to present a different feature of the same camera. Since this is a feature of the same camera and since the robot is not repeating itself (presenting the same feature twice), we consider this to be a socially appropriate behavior.

These examples show some successful predictions, but we are not claiming that this is a workable solution yet. In fact, the predictor fails fairly often using our current approach. These examples were chosen because they illustrate the possibility that an attention model such as this could feasibly be a useful tool for incorporating history in the learning of interaction logic. However, we are continuing to develop the system and search for solutions which are robust and generalizable to other scenarios and datasets. By presenting this work at this workshop, we hope to start some discussion about possible approaches that might be useful for this difficult task.

VII. Discussion and Conclusions

A. Discussion

The theme of this workshop is "behavior adaptation, interaction and learning for assistive robotics", and although our chosen retail scenario may differ somewhat from the target applications of assistance for the elderly and disabled, we believe that the principles behind this work are highly relevant.

Regarding adaptation to users, we are endeavoring to create a data-driven technique for learning the subtleties of interaction logic when dealing with a variety of users. In our camera shop scenario, the "novice" and "advanced" users require different kinds of explanation and assistance, and we expect that our system should implicitly learn to provide service consistent with these different sets of user needs.

Though our current method requires interactions to be re-collected when domain knowledge changes (i.e. update in camera price). It would be interesting to investigate solutions to this problem in future work, such as using a data-driven way to update the system with new knowledge, or automatically extract and update a feature specification such as the camera's price.

We believe that a technique like the one we propose could be helpful in learning interaction logic for assistive robots, because the learning could be done directly from caregivers and domain experts, simply by observing them in the process of providing services as they usually do. Eventually, if large amounts of speech, motion, and other types of interaction data could be captured passively in care facilities or hospitals, it might be possible to use a "big data" approach to create very rich interactions in assistive robots, for example to enable conversational robots to learn "active listening" skills to encourage dementia patients to communicate.

B. Conclusion

To put this work into perspective, we understand that it does not seem practical to entirely replace the role of a human interaction designer with a machine learning technique. Realistically speaking, it will always be necessary to have some visibility into a robot's reasoning and to have the ability to debug and improve the robot's behavior.

In an eventual real-world system, we expect that a hybrid approach would be best, combining the strengths of data-driven learning and manual design. The learning component could contribute by collecting the base set of behaviors and conditions, discovering fringe cases which might not be anticipated by a designer, and uncovering rules governed by implicit knowledge of which a designer might not be aware. The manual design component could then be useful for fine-tuning behaviors, correcting errors from noisy data, and extending or updating the robot's behavior set.

To conclude, the problem of learning mixed-initiative interaction logic from data is a difficult one. As we have described in this paper, we believe that the attention model technique shows promise as a flexible way to incorporate interaction history into a data-driven technique for learning interaction logic by imitation. By presenting this work at this workshop, we also hope to gather insights and suggestions from other participants and build upon the knowledge of the human-robot interaction community to develop reusable and generalizable techniques for learning top-level interaction logic for service robots.

ACKNOWLEDGMENT

This work was supported in part by the JST ERATO Ishiguro Symbiotic Human-Robot Interaction Project, and in part by JSPS KAKENHI Grant Number 25240042.

REFERENCES

[1] P. Liu, D. F. Glas, T. Kanda, H. Ishiguro, and N. Hagita, "How to train your robot - teaching service robots to reproduce human social behavior," in Robot and Human Interactive Communication, 2014 RO-MAN: The 23rd IEEE International Symposium on, 2014, pp. 961-968.

[2] P. Liu, D. F. Glas, T. Kanda, H. Ishiguro, and N. Hagita, "Data-driven HRI: Learning social behaviors by example from human-human interaction," IEEE Trans. on Robotics, vol. (to appear), 2016.

[3] M. Nicolescu and M. J. Mataric, "Task learning through imitation and human-robot interaction," Models and mechanisms of imitation and social learning in robots, humans and animals: behavioural, social and communicative dimensions, 2005.

[4] S. P. Chatzis and Y. Demiris, "Nonparametric mixtures of Gaussian processes with power-law behavior," Neural Networks and Learning Systems, IEEE Transactions on, vol. 23, pp. 1862-1871, 2012.

[5] M. Ogino, H. Toichi, Y. Yoshikawa, and M. Asada, "Interaction rule learning with a human partner based on an imitation faculty with a simple visuo-motor mapping," Robotics and Autonomous Systems, vol. 54, pp. 414-418, 2006.

[6] B. M. Scassellati, "Foundations for a Theory of Mind for a Humanoid Robot," Massachusetts Institute of Technology, 2001.

[7] Y. Nagai, "Learning to comprehend deictic gestures in robots and human infants," in Robot and Human Interactive Communication, 2005. ROMAN 2005. IEEE International Workshop on, 2005, pp. 217-222.

[8] S. Calinon and A. Billard, "Teaching a humanoid robot to recognize and reproduce social cues," in Robot and Human Interactive Communication, 2006. ROMAN 2006. The 15th IEEE International Symposium on, 2006, pp. 346-351.

[9] P. E. Rybski, J. Stolarz, K. Yoon, and M. Veloso, "Using dialog and human observations to dictate tasks to a learning robot assistant," Intelligent Service Robotics, vol. 1, pp. 159-167, 2008.

[10] R. Meena, G. Skantze, and J. Gustafson, "A Data-driven Approach to Understanding Spoken Route Directions in Human-Robot Dialogue," in INTERSPEECH, 2012.

[11] D. Brščić, T. Kanda, T. Ikeda, and T. Miyashita, "Person Tracking in Large Public Spaces Using 3-D Range Sensors," Human-Machine Systems, IEEE Transactions on, vol. 43, pp. 522-534, 2013.

[12] W. Yan and D. A. Forsyth, "Learning the behavior of users in a public space through video tracking," in Application of Computer Vision, 2005. WACV/MOTIONS'05 Volume 1. Seventh IEEE Workshops on, 2005, pp. 370-377.

[13] A. Sorokin, D. Berenson, S. S. Srinivasa, and M. Hebert, "People helping robots helping people: Crowdsourcing for grasping novel objects," in Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on, 2010, pp. 2117-2122.

[14] M. E. Foster, S. Keizer, Z. Wang, and O. Lemon, "Machine learning of social states and skills for multi-party human-robot interaction," in Proceedings of the workshop on Machine Learning for Interactive Systems (MLIS 2012), 2012, p. 9.

[15] J. E. Young, E. Sharlin, and T. Igarashi, "Teaching robots style: designing and evaluating style-by-demonstration for interactive robotic locomotion," Human–Computer Interaction, vol. 28, pp. 379-416, 2013.

[16] J. E. Young, T. Igarashi, E. Sharlin, D. Sakamoto, and J. Allen, "Design and evaluation techniques for authoring interactive and stylistic behaviors," ACM Transactions on Interactive Intelligent Systems (TiiS), vol. 3, p. 23, 2014.

[17] R. Toris, D. Kent, and S. Chernova, "The robot management system: A framework for conducting human-robot interaction studies through crowdsourcing," Journal of Human-Robot Interaction, vol. 3, pp. 25-49, 2014.

[18] J. Orkin and D. K. Roy, "Understanding Speech in Interactive Narratives with Crowdsourced Data," in AIIDE, 2012.

[19] C. Breazeal, N. DePalma, J. Orkin, S. Chernova, and M. Jung, "Crowdsourcing Human-Robot Interaction: New Methods and System Evaluation in a Public Environment," Journal of Human-Robot Interaction, vol. 2, pp. 82-111, 2013.

[20] S. Chernova, J. Orkin, and C. Breazeal, "Crowdsourcing HRI through Online Multiplayer Games," presented at the AAAI Fall Symposium Series, 2010.

[21] S. Chernova, N. DePalma, E. Morant, and C. Breazeal, "Crowdsourcing human-robot interaction: Application from virtual to physical worlds," in RO-MAN, 2011 IEEE, 2011, pp. 21-26.

[22] P. Langfelder, B. Zhang, and S. Horvath, "Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R," Bioinformatics, vol. 24, pp. 719-720, 2008.

[23] G. Viejo, M. Khamassi, A. Brovelli, and B. Girard, "Modeling choice and reaction time during arbitrary visuomotor learning through the coordination of adaptive working memory and reinforcement learning," Frontiers in behavioral neuroscience, vol. 9, 2015.

[24] F. Michaud and M. J. Matarić, "Learning from history for behavior-based mobile robots in non-stationary conditions," Machine Learning, vol. 31, pp. 141-167, 1998.

[25] Y. Mohammad and T. Nishdia, "Self-initiated imitation learning. Discovering what to imitate," in Control, Automation and Systems (ICCAS), 2012 12th International Conference on, 2012, pp. 726-732.

[26] J. Cheng, L. Dong, and M. Lapata, "Long short-term memory-networks for machine reading," arXiv preprint arXiv:1601.06733, 2016.

[27] L. Shang, Z. Lu, and H. Li, "Neural responding machine for short-text conversation," arXiv preprint arXiv:1503.02364, 2015.

[28] X. Zhou, B. Hu, Q. Chen, B. Tang, and X. Wang, "Answer sequence learning with neural networks for answer selection in community question answering," arXiv preprint arXiv:1506.06490, 2015.

[29] S. Sukhbaatar, J. Weston, and R. Fergus, "End-to-end memory networks," in Advances in Neural Information Processing Systems, 2015, pp. 2431-2439.

[30] S. Iengo, A. Origlia, M. Staffa, and A. Finzi, "Attentional and emotional regulation in human-robot interaction," in 2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication, 2012, pp. 1135-1140.

[31] E. Burattini, S. Rossi, A. Finzi, and M. Staffa, "Attentional modulation of mutually dependent behaviors," in International Conference on Simulation of Adaptive Behavior, 2010, pp. 283-292.

[32] C. Raffel and D. P. W. Ellis, "Feed-Forward Networks with Attention Can Solve Some Long-Term Memory Problems," arXiv preprint arXiv:1512.08756, 2015.

[33] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in Proc. ICML, 2013, p. 1.