

Demonstration of Hospital Receptionist Robot with Extended Hybrid Code Network to Select Responses and Gestures

Eui Jun Hwang, Byeong Kyu Ahn, Bruce A. Macdonald, and Ho Seok Ahn*

Abstract— Task-oriented dialogue system has a vital role in Human-Robot Interaction (HRI). However, it has been developed based on conventional pipeline approach which has several drawbacks; expensive, time-consuming, and so on. Based on this approach, developers manually define a robot's behaviour such as gestures and facial expressions on the corresponding dialogue states. Recently, end-to-end learning of Recurrent Neural Networks (RNNs) is an attractive solution for the dialogue system. In this paper, we proposed a social robot system using end-to-end dialogue system in the context of hospital receptionist. We utilized Hybrid Code Network (HCN) as an end-to-end dialogue system and extended to select both response and gesture using RNN based gesture selector. We evaluate its performance with human users and compare the results with one of the conventional methods. Empirical result shows that the proposed method has benefits in terms of dialogue efficiency, which indicates how efficient users were in performing the given tasks with the help of the robot. Moreover, we achieved the same performance regarding the robot's gesture with the proposed method compared to manually defined gestures.

I. INTRODUCTION

The interface between machine and human is evolving from Command Line Interface (CLI) to Graphic User Interface (GUI) and again to Conversational User Interface (CUI) [1]. CUI is also called dialogue system or chatbot, and it is a primary component in robots as it allows not only the robots to gather information required for tasks but also people to gain easy access to the various services that can be provided. Dialogue systems can be roughly divided into two categories in regards to the intended use case, chat-oriented [2–4] and task-oriented. Chat-oriented dialogue system usually aims to converse with users and provide interesting and reasonable contextually relevant responses. The more extended conversation often indicates better performance in the chat-oriented dialogue system. As an example of this, Amazon held Alexa prize competition [5] to improve the conversational ability of Alexa platform. Task-oriented dialogue system, on the other hand, aims to assist the user to complete a specific task, such as booking a flight, finding a restaurant and so on [6]. A robot with a task-oriented dialogue system can be expected to bring positive impact when it is replaced with a human operator such as a receptionist in terms of reducing time and effort. Thus, a task-oriented dialogue system can be regarded as effective in practice.

E. J. Hwang, B. K. Ahn, Bruce A. Macdonald, and H. S. Ahn* are with Department of Electrical, Computer, and Software Engineering, CARES, University of Auckland, Auckland 1142, New Zealand (phone: +64 9 373 7599; e-mail: ehwa009@aucklanduni.ac.nz, {bk.ahn, b.macdonald, hs.ahn}@auckland.ac.nz). *Corresponding author.

In the existing works, many researchers have developed task-oriented dialogue systems used in HRI. Finite State Machine (FSM) and slot-based method (also called the frame-based method) have been applied with their behaviours such as facial expression, gesture and so on [7–9]. Statistical approach such as Partially Observable Markov Decision Process (POMDP) has been applied to the dialogue system to maintain a distribution over possible dialogue states [10, 11]. Reinforcement learning has been applied to combine chat and task-based dialogue for the dialogue system [12]. However, these systems are based on conventional pipeline approaches and rule-based behaviour selection, which have several common drawbacks. It is often unclear how to define the dialogue state and what history maintain to select system action based on the current dialogue state [13]. Conventional approaches not only make it expensive and time-consuming to deploy but also make it hard to expand to a new domain [14].

Recently, with the success of chit-chat system based on end-to-end trainable neural network models [4, 15], researchers have started exploring end-to-end approaches to solve such difficulties in the pipeline approaches. The end-to-end methods are mainly based on an idea that RNNs can be directly trained on text transcripts of dialogues to represent distributed dialogue representations. With the benefit of RNNs, the end-to-end approaches tend to use a single module to generate a response rather than separate modules as in the pipeline methods. Bordes et al. [16] developed an end-to-end trainable framework for the task-oriented dialogue system using end-to-end memory networks (MemN2N) [17], which consisted of an interference module and a memory component that can be read and written into. In the same line of study, the researchers explored the approach using gated end-to-end memory network [18], query reduction networks [19] and copy-augmented sequence to sequence network [14].

However, according to Williams et al. [13], these purely RNN based approaches were found to lack a general mechanism for injecting domain knowledge. Injecting domain knowledge can be easily solved using a few lines of code, but these previous models required thousands of dialogues to learn such simple behaviours. To solve the limitation, they have introduced a practical RNN based end-to-end framework called HCNs. It combines a domain-specific software and an RNN to learn latent dialogue states and the chosen actions.

We address the research question on how we can apply end-to-end dialogue system to the robot. A different approach may be necessary to build a robot system because end-to-end based dialogue system which calculates the dialogue state with a trained hidden state cannot manually define a robot's

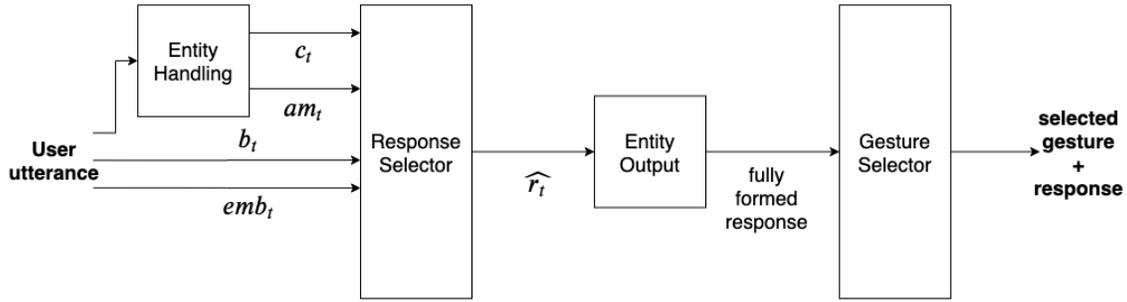


Figure 1. Overview of extended HCN which consists of three components; HCN, knowledge base (KB) and trainable gesture selection. The shaded boxes are trainable components. Bold is an initial input and output respectively.

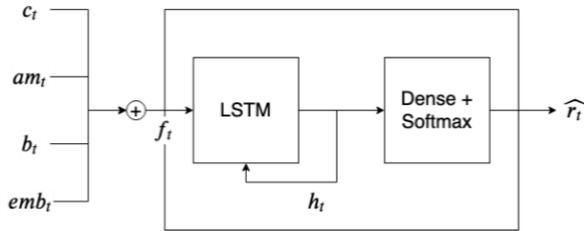


Figure 2. An overview of the response selector.

behaviour such as a gesture, expression and so on. We aim to fill this gap by demonstrating how we build a receptionist robot using the end-to-end dialogue system. There are two aspects of HRI with the end-to-end approach; 1) how end-to-end dialogue system is applied to the robot and 2) how we make the robot express its behaviour. For the sake of this, we propose trainable components to generate not only the response but robot to execute its gestures. Note that we only focus on generating robot’s gesture as a first attempt to HRI with end-to-end approach. We utilized Hybrid Code Network (HCN) and extended to produce a response with selected gesture. We applied Recurrent Neural Network (RNN) to select the robot’s gesture depending on the system response generated from HCN.

This paper is organized as follow. Section II presents our extended HCN for HRI. Section III describes system components of our receptionist robot. Section IV presents an experiment of receptionist robot with real users. Section V summarises and concludes this paper.

I. EXTENDED HCN FOR RECEPTIONIST ROBOT

A. Overall Structure

As shown in Figure 1, the extended HCN consists of two components; HCN dialogue application and trainable gesture selector. The details of each components will be presented in the following paragraph.

Entity handling module, which is a part of domain-specific components, identifies entities in the user’s utterance and maintain the identified entities. Based on the entities identified, it produces a corresponding action mask am_t that prevents non-permitted actions, and context feature c_t at time t . Response selection module determines the response \hat{r}_t based on a concatenated feature vector f_t of the current embed user utterance emb_t , bag of words b_t , current context c_t and action mask am_t . Entity output, which is also a part of

domain-specific components, substitutes the selected response template from the response selector with identified entities.

B. Entity Handling

Entity handling module identifies and maintains entities in user’s utterance such as patient name, address, time and location in our case. For example, if an utterance “where is the bathroom?” is given, the module extracts ‘bathroom’ as a location entity. Then later, if new entity value identified from the user’s utterance, the old entity is replaced with new identified entity. The module generates action mask am_t and context feature vector c_t at time t as a part of input feature in the response selector.

C. Response Selector

Response selector is responsible for computing latent dialogue state and selecting response \hat{r}_t at the time step t . It consists of an LSTM, dense and softmax layers. To determine the response \hat{r}_t , the LSTM is fed with a concatenated feature vector $f_{(t-1)}$ of embed user utterance emb_{t-1} , a bag of words b_{t-1} , current context c_{t-1} and action mask am_{t-1} until the $(t-1)^{th}$ step. At the last step t , the LSTM receive feature vector f_t and generate a 11-dimensional probability distribution for each of the 11 action templates through the dense and softmax layer. In our domain, total of 11 domain-specific response templates were generated as shown in Figure 2.

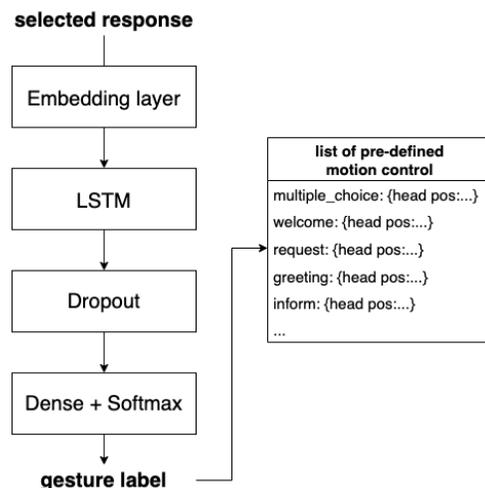


Figure 3. An overview of trainable gesture selector. It consists of 5 different layers; embedding, RNN, dense, dropout and softmax layers.

D. Entity Output

This module is responsible for generating fully-formed response based on the response templates from the response selector. For example, if an action template ‘api_call location <location>’ selected the previous module, then entity output fills stored entity as ‘api_call location *bathroom*’.

E. Gesture Selector

The trainable gesture selector is based on the idea of intent classification works [20]–[22]. It is built up with an embedding, LSTM, dense, dropout and softmax layers as shown in Figure 3. Once the dialogue system generates a response, each word in the response is tokenized and converted into a vector using the embedding layer. Then it is fed to RNN sequentially, and the dense and softmax layers output the probability of the gesture labels, which is matched with one of the pre-defined actual motion controls of a robot. Additionally, we add a dropout layer between dense and softmax layers to prevent overfitting of our model.

II. RECEPTIONIST ROBOT SYSTEM

Figure 4 shows the overall structure of our proposed receptionist robot system. It consists of four components; a sensory perception, extended HCN, Social Human-Robot Interaction (SHRI) framework [23] and robot platform. Automatic Speech Recognition (ASR) and Face detection modules work as sensory perception for the robot. Extended HCN has the responsibility for generating a robot’s response with a gesture that the robot should take. SHRI framework [23] generate robot’s non-verbal capabilities such as turn-taking, gaze, emotion, gesture, posture and so on. All components have been developed for the Robot Operating System (ROS). To test our system, we used NAO [24], which is a humanoid robot that is widely used in research and educational purpose as shown in Figure 4.

A. Sensory Perception

Speech is recognized using a commercial ASR provided by Google Cloud Speech [25]. This system enables developers to convert audio to text by applying neural network models in an easy-to-use API. Face detection is a vision module in which enable the robot to detect ROI(region of interest) of face in front of him. For this, we implement the face detection module with pre-built face detection module in the NAO robot

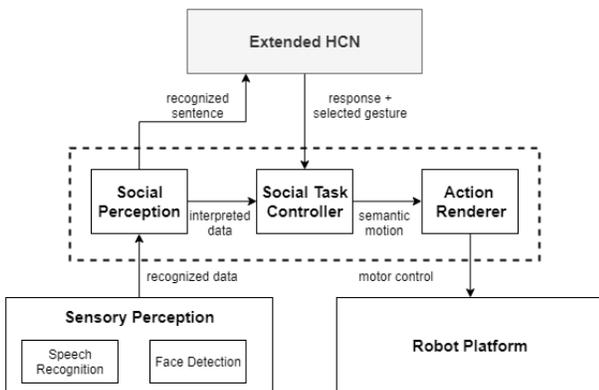


Figure 2. Overall structure of receptionist robot system. It consists of a speech recognition, face detection, extended HCN, Social Human-Robot Interaction (SHRI) framework and robot platform.

platform.

B. Extended HCN

This module is responsible for inferring what response is required and what gesture is required based on the user’s utterance. It depends on RNN to perform its operation. Its only input is a single utterance, and its output is a system response with a selected gesture. The details of this module are described in Section II.

To train HCN, we synthetically generated 250 dialogue datasets [26] which consist of four different tasks; ask for prescription, checking in for doctor’s appointment, asking waiting time, asking location of bathroom) in the same way as a work presented in [16]. To train gesture selector, we extracted a total of 19,671 sentences with eight corresponding labels (multiple choice, welcome, request, greeting, inform, confirm answer, thanks, closing) from dialogue corpus provided by Microsoft dialogue challenge [27]. We selected an LSTM [28] for the RNN. To build our dialogue application for practical purpose, the pre-trained model is loaded into the system. User utterance, as text input from ASR module, is used for input for HCN. We made the system re-prompt user if the highest probability of system action or the confidence of the ASR is less than 50%.

C. SHRI framework

To integrate with robot platform, we used SHRI framework [23], which is the modularized human-robot software separating robot’s social behaviour and domain tasks (dialogue system in our case). It can be separated into domain tasks, which is responsible for controlling the execution flow of scenario, and social behaviour as a framework so that developers reduce their effort to implement robot’s non-verbal capabilities such as turn-taking, gazing, its emotions, gestures, posture and so on.

The framework is composed of three major components; social perception, social task controller and action renderer as shown in Figure 4. Social perception interprets the situation based on the output of sensory perception modules. For example, audio-visual saliency is continuously evaluated, the user’s turn-taking intention is inferred, and cognitive and emotional state of interaction participants are estimated. Social task controller accomplishes action requested by the domain task. It uses tag information such as saying, gazing, pointing, facial expressions and so on. For example, if domain task request “<sm=tag:greeting> Hello. My name is Silbot”. Then, it generates a portable motion command format called semantic motion. Action renderer, which is a robot dependent component, is responsible for executing actual motor control. It interprets the semantic motion generated from the social task controller module.

III. EXPERIMENT

The human-robot dialogue system was evaluated via a user study in which human subjects interacted with the NAO acting autonomously using the system described above; all interactions were in English.

A. Experimental Environments

1) Physical Setup

During each session, the experimenter seats in front of the

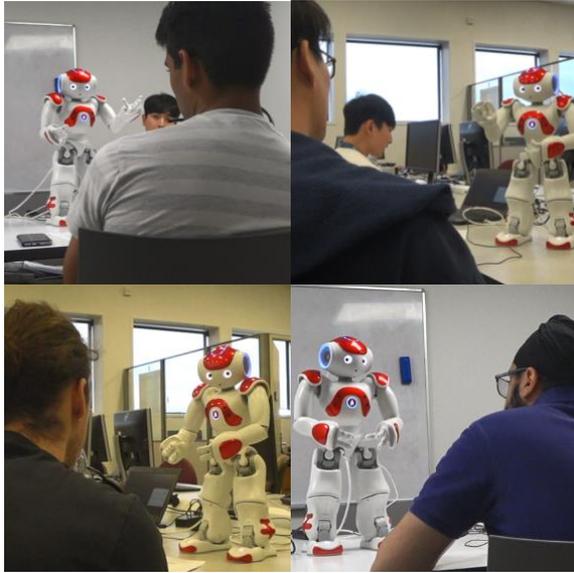


Figure 3. User study with receptionist robot NAO.

robot and the experimenter seat next to the robot to be available in case the participant would need any help as shown in Figure 4. The scene was recorded from participant’s perspective focusing on the robot. The duration of each session did not exceed 20 minutes.

2) Experimental Design

To explore benefit of our proposed system, two conditions are compared; robot with the proposed system, and robot with conventional method. We used slot-based method for which is workhorse of the conventional pipeline dialogue system [16], with having manually defined gesture in each response as a baseline system. Slot-based method predefines the structure of a dialog state as a set of slots to be filled during the dialogue.

To implement the baseline system, we used Google’s DialogFlow [29] and here, we have manually defined seven intents (welcoming, prescriptions, checking-in, silence, location, farewell, and fall back) and four different slots (name, address, time and location) to make an API call to query the knowledge base (KB). A between-subject design was used to compare the two conditions. Thus, different participants are assigned to different conditions.

We collected a range of objective measures from the log files and video recording of the interactions. We considered two metrics for objective measures which were used in PARADISE framework [29, 30]; dialogue efficiency and quality. Dialogue efficiency was evaluated using elapsed time, the number of tasks completed, and the number of utterances made by the user and robot’s during a session. Dialogue quality indicates the number of system behaviour that affects the quality of the resulting dialogue. This was evaluated using the number of time-outs and re-prompt and the confidence of ASR. Moreover, we also explore subjective measures. For this, we asked participants to fill in a questionnaire to analyse their perception of the robot. The questionnaire includes user’s overall rating and the God speed test [32], which is a measurement tool for HRI with five key concepts;

anthropomorphism, animacy, likeability, perceived intelligence and perceived safety.

3) Experimental Scenario

An experimental scenario was designed to show how well the robot works as a receptionist robot. Participants were asked to complete the given tasks as many as possible (ask for prescription, checking in for doctor’s appointment, asking waiting time, asking location of bathroom).

4) Participants Setting

Every participant was asked to imagine that they were entering a hospital that they had never been to before where the robot was installed in the reception area interacting with the patients. Before starting the experiment, they were asked to use natural language spontaneously. Moreover, they were provided with hints on how better communication with the robot. For example, “please wait for your turn to speak” and “please keep in mind that the robot only listens to you while its eyes turn blue”.

B. Experimental Result

20 people (7 males, 13 females) agreed to participate in our study, ranging in age from 19 to 30 ($M=24.1$, $SD=3.4$). M denotes mean value and SD denotes standard deviation. Participants did not receive any financial compensation, and most of them were students with little or no previous experience in interaction with a robot.

Table I shows the result of the experiment that participants were successful in performing the given tasks in both conditions with the average number of completed tasks being 3.8 ($SD=0.42$) in the proposed method and 3.9 ($SD=0.32$) in the baseline. The number of completed tasks was not significantly different between the two conditions (one-tailed T-test, $p = 0.24$). However, there is a significant difference in the number of the user turns, and robot turns, and elapsed time (one-tailed T-test, $p = 0.028$, 0.03 and 0.003 , respectively). More specifically, the user with the proposed method has an average of 1.4 turns less than the baseline method. Moreover, the robot with the proposed method has an average of 1.8 turns less, and the users have an average of 38.4 seconds shorter interaction time. In terms of dialogue quality, the result of data analysis shows that the number of time out, re-prompts and speech recognition all have no significant difference (one

TABLE I. RESULTS OF CONVERSATIONAL EFFICIENCY AND DIALOGUE QUALITY IN TWO CONDITIONS. M DENOTES MEAN VALUE AND SD DENOTES STANDARD DEVIATION

Measure	Baseline	Proposed method
Number of tasks	$M=3.9$, $SD=0.32$	$M=3.8$, $SD=0.42$
Number of user turns	$M=13.4$, $SD=1.57$	$M=12.0$, $SD=1.49$
Number of robot turns	$M=15$, $SD=2.31$	$M=13.2$, $SD=1.61$
Elapsed time(sec)	$M=243.1$, $SD=29.17$	$M=204.7$, $SD=24.22$
Number of time out	$M=0.7$, $SD=0.67$	$M=0.2$, $SD=0.42$
Number of re-prompts	$M=0.8$, $SD=0.91$	$M=1.3$, $SD=1.42$
Speech recognition	$M=93.53$, $SD=2.24$	$M=93.39$, $SD=3.20$
User satisfaction	$M=3.72$, $SD=0.55$	$M=3.88$, $SD=0.43$
perception	$M=3.79$, $SD=0.55$	$M=3.73$, $SD=0.32$

U1: Hi there.
R1: Hi there, my name is Nao, the receptionist robot. How may I help you?
U2: I'm Sam Smith, and I want to check when my doctor's appointment was.
R2: Thanks Sam, and what is your address?
U3: 12 Green Street.
R3: Okay Sam, what time is your appointment?
U4: I believe it's at 4 p.m.
R4: And which doctor are you seeing?
U5: I'm actually not too sure
R5: No problem Sam Smith, I can see ...
U6: <Silence>
R6: Is there anything else I can help you with?
U7: Yeah, I wanted to know if my prescription is ready
R7: I'm sorry Sam Smith, your doctor has not yet written your prescription
U8: Cool thanks, I wanted to know how long the appointment would be approximately
R8: You are next to see doctor Jones, he will be around 5 more minutes.
U9: Could you tell me where the bathroom is
R9: The bathroom is located down the hall, second door on the right.
U10: <Silence>
R10: I hope you have a nice day Sam

Figure 4. An example of dialogue between the proposed system and real user.

U1: Hi there.
R1: Hi there, my name is Nao, the receptionist robot. How may I help you?
U2: I'm Sam Smith and I have a doctor's appointment at 4 p.m.
R3: Thanks Sam, and what is your address?
U4: 12 Green Street
R4: Okay Sam, What time is your appointment?
U5: It's 4 p.m.
R5: And which doctor are you seeing?
U6: I can't remember the doctor's name.
R6: No problem Sam, I can see that you...
U7: Where's the bathroom?
R7: the bathroom is located down the hall...
U8: Can I get my prescription please?
R8: Okay, can you please tell me your name?
U9: Sam Smith.
R9: Thanks Sam, and what is your address?
U10: 12 Green Street.
R10: I'm sorry Sam, your doctor has not...
U11: do you know how long you'll be
R11: You are next to see doctor...
U12: <silence>
R12: is there anything else I can help you?
U13: No thank you.
R13: I hope you have a nice day.

Figure 5. An example of dialogue between the baseline system and real user.

tailed T-test, $p = 0.07, 0.18$ and 0.32 , respectively). However, we found that the proposed method was more likely to re-prompt an average 0.5 more.

We also analysed the result of robot perception and user satisfaction questionnaire to find the acceptability of our proposed system compared to the baseline. The validity of the used questionnaire was tested by measuring its internal consistency with Cronbach's α , which was equal to 0.89 (good consistency). Based on this value, we assume that our participants in the given context interpreted the robot characteristics, provided in the questionnaire, in an expected way. We averaged the 5-point Likert scale of the questionnaire we collected. As a result, we could not find any significant difference between two methods; trainable gesture selector

and conventional method which manually defines gesture in the response (one tailed T-test, $p = 0.24$ and 0.39 , respectively).

C. Discussion

Our evaluation shows that both conditions show similar performance during the experiment with real users. However, the proposed system shows better results in terms of dialogue efficiency. To gain more insight into this experiment, we performed a detailed analysis of the dialogue log files and recording, and it revealed that the proposed model tends to comprehend the dialogue context than the baseline system. Figure 6 shows an example of dialogue between the proposed system and the user. In the same way as work presented in [30], the whole dialogue can be divided into four different sub-dialogues which is same as given tasks; check-in (U1~R5), collect prescription (U7~R7), ask waiting time (U8~R8), and ask location (U9~R9). "<Silence>" means that user did not provide any utterance. It shows that the collected information such as name and address carry over to the different tasks naturally (from check-in to collect prescription in this case) and may notice conversation is going to be over (U10~R10).

On the other hands, the baseline system requests the information (U8~R10) that has already been collected from the previous task (U1~R6) as shown in Figure 7. We found that the baseline system is not suitable for complex conversation in terms of dialogue efficiency. Of course, slot-based method can store entity values and hand it over the next intents, but it can be required a lot of handcrafted works and there may be missing works that can lead to conversation failures. On the side of robot's gesture, it can be said our proposed system is better than the baseline system due to no handcrafted works were required to define the robot's gesture according to the response.

IV. CONCLUSION AND FUTURE WORK

We presented and evaluated our autonomous hospital receptionist robot, using an end-to-end approach to generate response and gesture. For this, we extended HCN to select not only a response but also a proper gesture based on the generated response. We experimented with real users. We found that our proposed system has an advantage in terms of dialogue efficiency, which indicate how efficient users were in achieving the given tasks with the receptionist robot. Moreover, participants found no difference between the proposed system and baseline system in terms of the robot's perception. It means that no more handcrafted works are required to define the robot's gesture according to the robot's response.

In future work, it will be possible to make several improvements to extend the realm of possibility for the receptionist robot with the end-to-end approach. The dialogue system, it was limited to a rather small domain. Tests on others, perhaps broader, the domain would be needed to see the resulting scale. On the side of the robot's behaviour, we have tested how the users feel when using the receptionist robot. However, the robot is rather equipped with minimal features at this stage; verbal interaction with the robot's gesture and gaze would be good examples. To improve the user experience, we can extend our work to more diverse

forms such as robot's expression, voice pitch and so on.

ACKNOWLEDGMENT

This work was supported by the Technology Innovation Program (10077553, Development of Social Robot Intelligence for Social Human-Robot Interaction of Service Robots) funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea).

REFERENCES

- [1] KaKao, "developing service for AI era," *LinkedIn SlideShare*, 2018. [Online]. Available: <https://www.slideshare.net/ifkakao/ai-113145560>. [Accessed: 11-Sep-2018].
- [2] Z. Yan *et al.*, "DocChat: An Information Retrieval Approach for Chatbot Engines Using Unstructured Documents," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 2016, pp. 516–525, doi: 10.18653/v1/p16-1049.
- [3] R. E. Banchs and H. Li, "IRIS: a Chat-oriented Dialogue System based on the Vector Space Model," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, 2012, pp. 37–42.
- [4] O. Vinyals and Q. Le, "A Neural Conversational Model," *CoRR*, vol. abs/1506.0, 2015, doi: 10.1210/jc.2006-0173.
- [5] A. Ram *et al.*, "Conversational AI: The Science Behind the Alexa Prize," in *arXiv preprint arXiv:1801.03604*, 2018.
- [6] D. McTear, Michael and Callejas, Zoraida and Griol, *The conversational interface: Talking to smart devices*. Springer, 2016.
- [7] F. Faber, "The Humanoid Museum Tour Guide Robotinho," in *Robot and Human Interactive Communication, 2009. RO-MAN 2009. The 18th IEEE International Symposium on*, 2009, pp. 891–896.
- [8] T. Fong, C. Thorpe, and C. Baur, "Collaboration, Dialogue, and Human-Robot Interaction," in *Proceedings of the 10th International Symposium of Robotics Research, Lorne, Victoria, Australia*, 2001.
- [9] D. Spiliotopoulos and C. D. Spyropoulos, "Human-Robot Interaction Based On Spoken Natural Language Dialogue," in *Proceedings of the European workshop on service and humanoid robots*, 2001, pp. 25–27.
- [10] L. Lucignano, F. Cutugno, S. Rossi, and A. Finzi, "A dialogue system for multimodal human-robot interaction," in *Proceedings of the 15th ACM on International conference on multimodal interaction*, 2013, pp. 197–204, doi: 10.1145/2522848.2522873.
- [11] R. Stiefelhagen *et al.*, "Enabling multimodal human-robot interaction for the Karlsruhe humanoid robot," *IEEE Trans. Robot.*, vol. 23, no. 5, pp. 840–851, 2007, doi: 10.1109/TRO.2007.907484.
- [12] I. Papaioannou, C. Dondrup, J. Novikova, and O. Lemon, "Hybrid chat and task dialogue for more engaging HRI using reinforcement learning," in *RO-MAN 2017 - 26th IEEE International Symposium on Robot and Human Interactive Communication*, 2017, pp. 593–598, doi: 10.1109/ROMAN.2017.8172363.
- [13] J. D. Williams, K. Asadi, and G. Zweig, "Hybrid Code Networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2017, pp. 665–677, doi: 10.18653/v1/P17-1062.
- [14] M. Eric and C. D. Manning, "A Copy-Augmented Sequence-to-Sequence Architecture Gives Good Performance on Task-Oriented Dialogue," in *the Association for Computational Linguistics*, 2017, vol. 2, pp. 468–473.
- [15] L. Shang, Z. Lu, and H. Li, "Neural Responding Machine for Short-Text Conversation," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2015, pp. 1577–1586, doi: 10.3115/v1/P15-1152.
- [16] A. Bordes, Y.-L. Boureau, and J. Weston, "Learning end-to-end goal-oriented dialog," in *ICLR 2017*, 2017.
- [17] S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus, "End-To-End Memory Networks," in *Advances in neural information processing systems. 2015*, 2015.
- [18] F. Liu and J. Perez, "Gated end-to-end memory networks," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 2017, vol. 1, pp. 1–10, doi: 10.1038/nature20101.
- [19] M. Seo, S. Min, A. Farhadi, and H. Hajishirzi, "Query-reduction networks for question answering," in *International Conference on Learning Representations*, 2017.
- [20] P. Zhou *et al.*, "Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2016, pp. 207–212, doi: 10.18653/v1/P16-2034.
- [21] N. Karpov, A. Demidovskij, and A. Malafeev, "Development of a Model to Predict Intention Using Deep Learning," in *Proceedings of the 6th International Conference "Analysis of Images, Social Networks and Texts" (AIST 2017)*, 2017, pp. 27–29.
- [22] H. Purohit, G. Dong, V. Shalin, K. Thirunarayan, and A. Sheth, "Intent classification of short-text on social media," *Proc. - 2015 IEEE Int. Conf. Smart City, SmartCity 2015, Held Jointly with 8th IEEE Int. Conf. Soc. Comput. Networking, Soc. 2015, 5th IEEE Int. Conf. Sustain. Comput. Commun.*, pp. 222–228, 2015, doi: 10.1109/SmartCity.2015.75.
- [23] M. Jang, J. Kim, and B. K. Ahn, "A software framework design for social human-robot interaction," in *2015 12th International Conference on Ubiquitous Robots and Ambient Intelligence, URAI 2015*, 2015, pp. 411–412, doi: 10.1109/URAI.2015.7358887.
- [24] V. Hugel *et al.*, "Mechatronic design of NAO humanoid," in *IEEE International Conference on Robotics and Automation*, 2009, pp. 769–774, doi: 10.1109/robot.2009.5152516.
- [25] Google, "Google Cloud Speech." [Online]. Available: <https://cloud.google.com/speech-to-text/>. [Accessed: 30-Sep-2018].
- [26] E. J. Hwang, B. A. MacDonald, and H. S. Ahn, "End-to-End Dialogue System with Multi Languages for Hospital Receptionist Robot," *2019 16th Int. Conf. Ubiquitous Robot. UR 2019*, pp. 278–283, 2019, doi: 10.1109/URAI.2019.8768694.
- [27] X. Li, Y. Wang, S. Sun, S. Panda, J. Liu, and J. Gao, "Microsoft dialogue challenge: building end-to-end task-completion dialogue systems," in *arxiv.org*, 2018.
- [28] S. Hochreiter and J. Unger Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.
- [29] Google, "Dialogflow." [Online]. Available: <https://dialogflow.com/>. [Accessed: 28-Oct-2018].
- [30] M. A. Walker, D. J. Litman, C. A. Kamm, and A. Abella, "PARADISE: A Framework for Evaluating Spoken Dialogue Agents," in *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, 1997, pp. 271–280.
- [31] M. A. Walker, "An Application of Reinforcement Learning to Dialogue Strategy Selection in a Spoken Dialogue System for Email," *J. Artif. Intell. Res.*, vol. 12, p. 387, 2000.
- [32] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *Int. J. Soc. Robot.*, vol. 1, no. 1, pp. 71–81, 2009, doi: 10.1007/s12369-008-0001-3.