A Visually-Aware Conversational Robot Receptionist

Nancie Gunson, Daniel Hernandez Garcia, Weronika Sieińska
Angus Addlesee, **Christian Dondrup**, **Oliver Lemon**
Interaction Lab, School of Mathematical and Computer Sciences
Heriot-Watt University, Edinburgh, Scotland, UK
{n.gunson, d.hernandez_garcia, w.sieinska,
a.addlesee, c.dondrup, o.lemon}@hw.ac.uk

Jose L. Part
Alana AI
Edinburgh, UK
jose@alanaai.com

Yanchao Yu
School of Computing
Edinburgh Napier University, Scotland, UK
Y.Yu@napier.ac.uk

Abstract

Socially Assistive Robots (SARs) have the potential to play an increasingly important role in a variety of contexts including healthcare, but most existing systems have very limited interactive capabilities. We will demonstrate a robot receptionist that not only supports task-based and social dialogue via natural spoken conversation but is also capable of visually grounded dialogue; able to perceive and discuss the shared physical environment (e.g. helping users to locate personal belongings or objects of interest). Task-based dialogues include check-in, navigation and FAQs about facilities, alongside social features such as chit-chat, access to the latest news and a quiz game to play while waiting. We also show how visual context (objects and their spatial relations) can be combined with linguistic representations of dialogue context, to support visual dialogue and question answering. We will demonstrate the system on a humanoid ARI robot, which is being deployed in a hospital reception area.

1 Introduction

Socially Assistive Robots (SARs) are increasingly being explored in contexts ranging from education (Papadopoulos et al., 2020) to healthcare (González-González et al., 2021). It has been noted, however, that despite the success of SARs and spoken dialogue systems in their respective research fields, integration of the two is still rare (Lima et al., 2021) and social robots in general still lack interaction capabilities (Cooper et al., 2020). In a similar fashion, even recent research on combining vision and language has tended to centre around the use of still images (Mostafazadeh et al., 2017; Zhou et al., 2020), with few systems able to support visual dialogue as part of a natural, situated conversation.

Figure 1: Interacting with SPRING-ARI

The SPRING project aims to develop such a system in the form of a robot receptionist for visitors to an eldercare outpatient hospital. In this context the robot must be able to communicate naturally with users on a variety of both functional and social topics, including but not limited to those concerning the shared physical environment. We demonstrate our progress towards this goal, with a multi-modal conversational AI system that is integrated on an ARI robot[1] (Fig. 1) and which combines social

[1] https://pal-robotics-com/robots/ari

Proceedings of the SIGdial 2022 Conference, pages 645–648
Heriot-Watt University, Edinburgh, UK. 07-09, September, 2022 ©2022 Association for Computational Linguistics

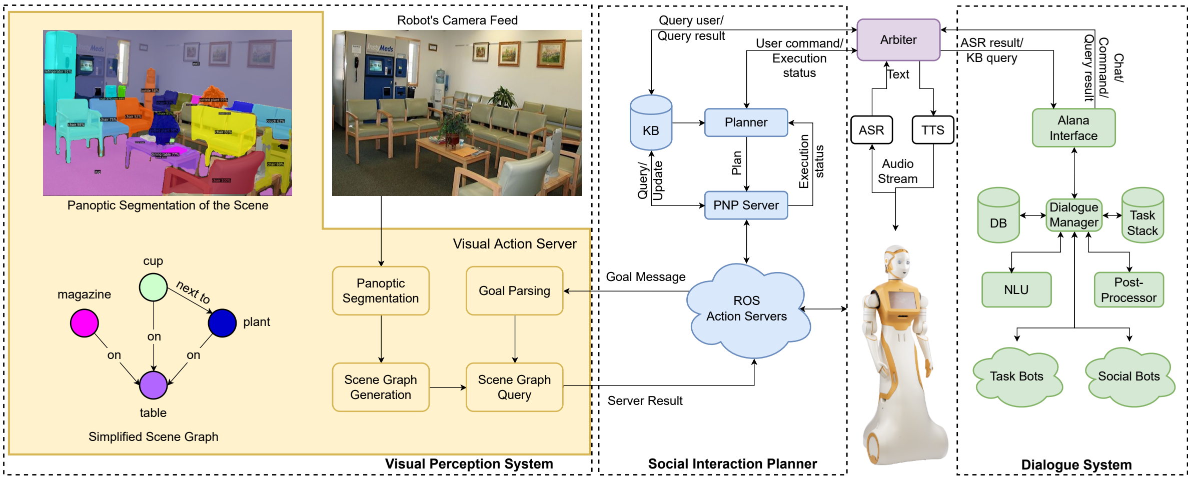

Figure 2: System Architecture. The Social Interaction Planner (blue blocks) interacts with the Dialogue System (green blocks) through an Arbiter module. The Vision module (yellow blocks) is implemented as a ROS Action Server within the planning framework. Based on the user's intent, the Planner can recruit the vision module to respond to questions about the visual scene.

and task-based conversation with visual dialogue regarding navigation and object detection in the shared space. Our system greets visitors, supports them to check-in, answers FAQs and helps users to locate key facilities and objects. It also offers social support/entertainment in the form of chit-chat, a quiz and access to the latest news.

2 System Architecture

The system architecture (Fig. 2), is composed of three main modules; a visual perception system, a dialogue system, and a social interaction planner.

2.1 Visual Perception System

The visual perception system is implemented as a ROS action server and is based on scene segmentation (Wu et al., 2019) from Facebook's Detectron2 framework[2]. From the segmented scene, the goal is to build a scene graph to capture relationships between objects such as location, adjacency, etc.

2.2 Dialogue System

The Dialogue System (Fig. 3) is based on the Alana system (Curry et al., 2018), an ensemble of different bots that compete in parallel to produce a response to user input. There are two types of bot: rule-based bots that can, for example, drive the conversation if it stalls, and express the identity of a virtual 'persona' (e.g. answering questions about the robot's age etc.); and data-driven bots that can retrieve replies from various information sources, e.g., News feeds. The SPRING system retains the rule-based and News bots, supplemented with a number of new, domain-specific bots. *Visual Task Bot* handles visual dialogue within the conversation, converting the user's inferred intent and any entities associated with it to a goal message that is forwarded to the visual action server (Part et al., 2021). *Reception Bot* welcomes visitors, helps them check-in and answers FAQs on, e.g., catering facilities and schedules. *Directions Bot* helps users find key facilities such as the bathrooms and elevator, while *Quiz Bot* is a simple true-or-false game designed to keep users entertained while they wait. The Dialogue Manager decides which response the robot verbalises based on a bot priority list. Automatic Speech Recognition (ASR) on the robot is currently implemented (in English) via Google Cloud[3]. Natural Language Understanding (NLU) is based on the original Alana pipeline, augmented using the RASA framework[4] for the parsing of domain-specific enquiries. Quiz bot employs regex-based intent recognition. Natural Language Generation (NLG) for the majority of bots consists of templates, with only News bot retrieving content from selected online sites. The utterances are voiced on the robot by Acapela's UK English Text-To-Speech voice 'Rachel'[5].

[2] https://github.com/facebookresearch/detectron2

[3] https://cloud.google.com/speech-to-text

[4] https://rasa.com

[5] https://www.acapela-group.com/

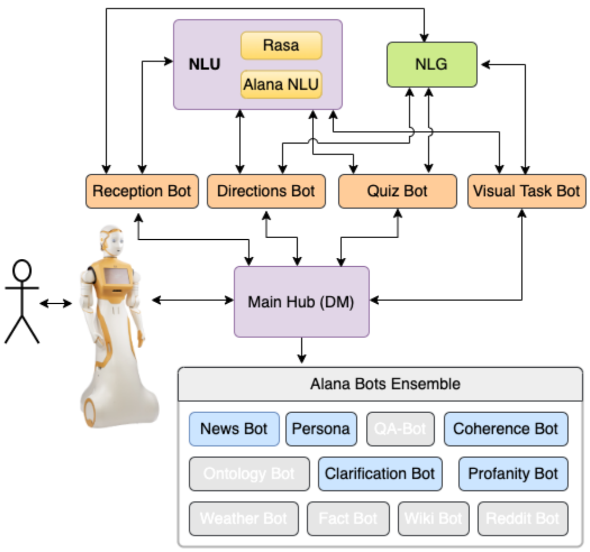

Figure 3: Dialogue System Architecture

2.3 Social Interaction Planner

The Social Interaction Planner interfaces the dialogue, vision systems, and the physical actions of the robot. It creates and executes plan(s) containing dialogue, physical, and perception actions based on the current dialogue context, and is based on the ideas of (Papaioannou et al., 2018; Lemon et al., 2002), enabling multi-threaded task execution and dialogue that is flexible and pausable. As shown in Fig. 2 it comprises several components, with the *Arbiter* managing communication between the dialogue system, the robot, and the planner

The *Planner* is a key component and uses the principle of *recipes and resources* as developed by (Lemon et al., 2002) to eliminate the problems from (re-)planning and concurrent interactive planning and execution. A domain file lists all the possible actions and specifies their types and parameters including their preconditions and effects. Recipes then describe the sequence of actions (i.e. dialogue, physical, or perception) involved in achieving a desired goal. When requested by the Arbiter, these recipes are transformed into Petri-Net Plans (PNP) (Dondrup et al., 2019) and are concurrently executed together with any other plans that may already be running. At run-time, redundant actions whose effects have already been achieved are skipped, or repair actions are executed in cases where an action was unsuccessful. At any time, each action has the ability to communicate with the dialogue system via the Arbiter to allow for clarification or to communicate perception results.

3 The SPRING ARI Robot

ARI is a humanoid robot, designed for use as a socially assistive companion (Cooper et al., 2020). It is 1.65m tall, has a mobile base, a touch-screen on the torso, movable arms and head with LCD eyes that enable gaze behaviour. The version of the robot used here is equipped with several cameras creating a 360º field of view. For audio capture and processing a ReSpeaker Mic v2.0 array, with 4 microphones is mounted on the front of the belly.

4 Example Dialogues

The example dialogue shown in Table 1 illustrates the system's visual dialogue capabilities (in bold), occurring as part of the welcome conversation.

U: Hello, I've got an appointment at 10am.

R: [reception_bot] I can help you with that. May I take the name please?

U: It's Jeanne X.

R: [reception_bot] Pleased to meet you, Jeanne. Is this your first visit?

U: No, I've been before.

R: [reception_bot] OK great, you'll be an expert then! That's you checked in. Please take a seat and a nurse will call you when they're ready. I'm here if you need anything else while you're waiting.

U: Is there something to read while I wait?

R: [visual_task_bot] **I can see a magazine on the table.**

U: Aha thanks.

R: No problem.

Table 1: Task-based Dialogue with Visual Dialogue.

The example dialogue in Table 2 further illustrates the robot's natural language social and task-based capabilities.

U: Excuse me, do you know how much longer I will have to wait?

R: [reception_bot] I expect it won't be long now. A nurse will come and get you. Do you want to play a game to pass the time?

U: OK then, yes please.

R: [quiz_bot] Great, let's play! Please remember, whenever you want to leave the game, just say 'end game'. First question, is Covid-19 on all 7 continents?

U: Nope.

R: Correct. Next question ......

R: [quiz_bot] Congratulations! You won. That was fun. Now, please let me know if you need anything else.

U: I need the bathroom, can you tell me where it is?

R: [directions_bot] To find the nearest bathroom, leave the waiting room and go right towards the door with two round windows. The bathroom will then be on your left.

Table 2: Task-based Dialogue with Social Dialogue.

5 Conclusions and Future Work

The proposed system demonstrates how to successfully weave together a wide range of task-based, social, and visually grounded dialogue and physical actions on an SAR in a receptionist environment. Next steps are to generate the scene graphs automatically by combining data-driven approaches (Zellers et al., 2018; Yang et al., 2018; Zhang et al., 2019; Tang et al., 2020) with prudent use of refining rules. Crucially also, we are working on extending the system to handle multi-party interactions, an active area of research and highly likely to occur in this context.

For the demonstration, we will showcase our system on the ARI robot, inviting attendees to interact with it and experience all the capabilities of the system described in this paper.

Acknowledgements

This research has been funded by the EU H2020 program under grant agreement no. 871245 (http://spring-h2020.eu/).

References

Sara Cooper, Alessandro Di Fava, Carlos Vivas, Luca Marchionni, and Francesco Ferro. 2020. ARI: The Social Assistive Robot and Companion. In *29th IEEE International Conference on Robot and Human Interactive Communication, RO-MAN 2020*, pages 745–751.

Amanda Cercas Curry, Ioannis Papaioannou, Alessandro Suglia, Shubham Agarwal, Igor Shalyminov, Xinnuo Xu, Ondřej Dušek, Arash Eshghi, Ioannis Konstas, Verena Rieser, et al. 2018. Alana v2: Entertaining and Informative Open-Domain Social Dialogue using Ontologies and Entity Linking. *Alexa Prize Proceedings*.

Christian Dondrup, Ioannis Papaioannou, and Oliver Lemon. 2019. Petri Net Machines for Human-Agent Interaction.

Carina Soledad González-González, Verónica Violant-Holz, and Rosa Maria Gil-Iranzo. 2021. Social robots in hospitals: A systematic review. *Applied Sciences*, 11(13).

Oliver Lemon, Alexander Gruenstein, Alexis Battle, and Stanley Peters. 2002. Multi-Tasking and Collaborative Activities in Dialogue Systems. In *Proceedings of the 3rd SIGdial Workshop on Discourse and Dialogue*, pages 113–124.

Maria R. Lima, Maitreyee Wairagkar, Manish Gupta, Ferdinando Rodriguez y Baena, Payam Barnaghi, David J. Sharp, and Ravi Vaidyanathan. 2021. Conversational affective social robots for ageing and dementia support. *IEEE Transactions on Cognitive and Developmental Systems*, pages 1–1.

Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios P. Spithourakis, and Lucy Vanderwende. 2017. Image-Grounded Conversations: Multimodal Context for Natural Question and Response Generation. pages 462–472.

Irena Papadopoulos, Runa Lazzarino, Syed Miah, Tim Weaver, Bernadette Thomas, and Christina Koulouglioti. 2020. A systematic review of the literature regarding socially assistive robots in pre-tertiary education. *Computers Education*, 155:103924.

Ioannis Papaioannou, Christian Dondrup, and Oliver Lemon. 2018. Human-Robot Interaction Requires More Than Slot Filling - Multi-Threaded Dialogue for Collaborative Tasks and Social Conversation. In *Proceedings of the FAIM/ISCA Workshop on Artificial Intelligence for Multimodal Human Robot Interaction*, pages 61–64.

Jose L. Part, Daniel Hernández García, Yanchao Yu, Nancie Gunson, Christian Dondrup, and Oliver Lemon. 2021. Towards Visual Dialogue for Human-Robot Interaction. In *Companion Proceedings of the 16th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 670–672, Boulder, CO, USA.

Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. 2020. Unbiased Scene Graph Generation from Biased Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3713–3722.

Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. 2019. Detectron2. https://github.com/facebookresearch/detectron2.

Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. 2018. Graph R-CNN for Scene Graph Generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 670–685.

Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. 2018. Neural Motifs: Scene Graph Parsing with Global Context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5831–5840.

Ji Zhang, Kevin J. Shih, Ahmed Elgammal, Andrew Tao, and Bryan Catanzaro. 2019. Graphical Contrastive Losses for Scene Graph Parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11535–11543.

Li Zhou, Jianfeng Gao, Di Li, and Heung Yeung Shum. 2020. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93.