Commonsense-enhanced Natural Language Generation for Human-Robot Interaction

Dimitra Gkatzia
School of Computing
Edinburgh Napier University
Edinburgh, UK
d.gkatzia@napier.ac.uk

ABSTRACT

Commonsense is vital for human communication, as it allows us to make inferences without explicitly mentioning the context. Equipping robots with commonsense knowledge would lead to better communication between humans and robots and will allow robots to be introduced in real-world environments. However, this is an extremely hard task due to the complex interdisciplinary nature of the problem, which spans across several fields including natural language generation, reasoning, computer vision and robotics. Addressing this challenge will unlock a plethora of opportunities for assistive and care robotics, service robotics and novel educational and training applications, to tackle immediate challenges such as caring for the elderly population, upscale skills, automate tasks and increase productivity. This paper proposes the *Robot-Commonsense* challenge that goes beyond traditional multi-modal interaction (vision, deictic gestures, language, gaze) and focuses on incorporating commonsense knowledge to enhance human-robot interaction.

1 INTRODUCTION

As robots are set to leave the lab environments and be introduced in public and domestic spaces, the need for seamless integration and efficient communication with humans is vital. When humans interact with their environment, they rely on their commonsense knowledge, unspoken assumptions about spatial relations of visible and invisible objects (such as objects in cupboards or containers), facts and social conventions [4]. Consider, for instance, the conversation below:

(1.1) Person A: Is there a screwdriver?
Person B: Maybe, let me get the toolbox.

In this conversation, Person A asks for a tool that is not visible at that moment. Person B knows that in order to answer this question, they will have to check a toolbox because this is where tools are normally stored. Although this type of knowledge seems trivial for humans, robots and other natural language interfaces still do not possess this ability.

Commonsense can appear in various forms, including but not limited to:

- *Empathy*: a robot/system should react in empathetic ways by reasoning about the human users' mental states based on the events the user has experienced, without the users explicitly stating how they are feeling.
- *Inference on intends and reactions beyond empathy*: a robot should understand that a situation will have an effect on a human which might cause a reaction. This commonsense reasoning can be derived from short texts as in [11], or by visual observations.
- *Ambiguity resolution* as a result of implicit knowledge and underspecification. For instance, co-reference resolution has been studied as a commonsense problem by [12].
- *Reasoning about actions*: for instance, handling objects is done through commonsense [1].
- *Reasoning about situations or relations between situations*: Recent work has looked into generating explanations of relations between situations or events (described in texts) that have taken place in the recent past (e.g. [10]).
- *Interpretation of natural language instructions*: Commonsense can enhance the robots' ability to comprehend incomplete natural language instructions by utilising environmental context to fill in missing information [3].
- *Reasoning about visual objects* that goes beyond their visual properties: for instance, a robot should be able to understand that a full box is heavy as opposed to an empty box, or how an object can be used, which goes beyond object recognition.

The last topic can be seen as a natural next step of the very well known Referring Expression Generation (REG) task. REG refers to a family of methods that aim to generate descriptions of visual objects in natural language, so that the hearer of the referring expression can uniquely identify the described object [8]. The obvious relationship between the physical environment and REG has contributed in making REG the most studied Natural Language Generation (NLG) task in the context of Human-Robot Interaction (see for instance [5, 13]). Previous works have focused on developing models and algorithms that describe physical objects based on their visual properties, such as appearance, colour and shape. The *utility* and non-visual features of physical objects, which can be refer to through commonsense knowledge rather than perception, have not received much attention. In this paper, we argue that commonsense-enhanced NLG is central for HRI, as it will allow robots to naturally communicate with humans in real-world dynamic environments.

2 OPEN CHALLENGES

This paper describes three challenges that can be immediately addressed by imitating human commonsense: (1) describing unknown objects or other entities; (2) reasoning about objects' utility beyond their appearance; (3) inference of invisible objects or parts due to obscured visibility (e.g. objects inside containers, or internal structure of objects etc.).

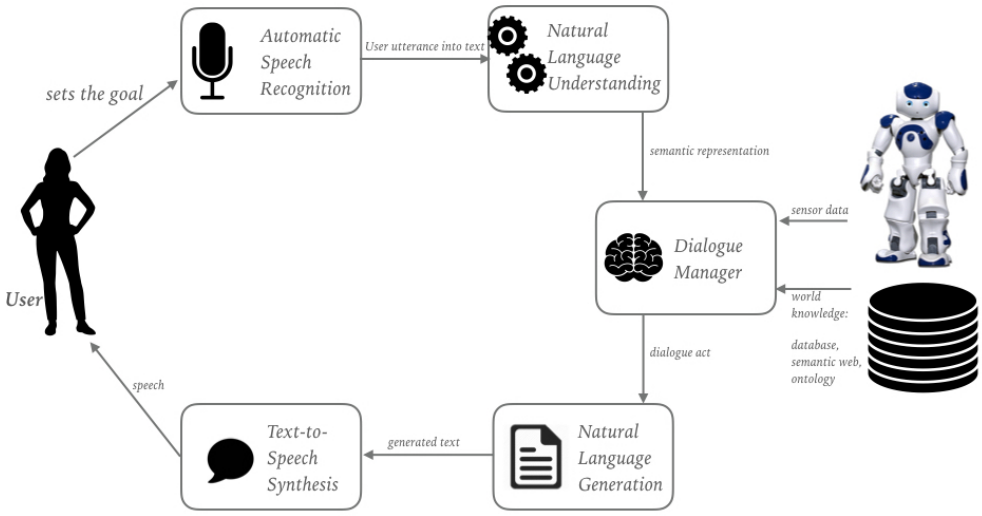

Figure 1: Proposed system architecture for commonsense-enhanced NLG for Human-Robot Interaction.

2.1 Unknown objects and entities

In Human-Robot Interaction scenarios, where humans and robots need to communicate and collaborate to perform tasks, successfully referring to unknown objects or entities in the environment is of vital importance. Traditional REG approaches [6] assume perfect input, i.e. perfect representation of the environment, the objects therein and the objects' properties. In real-world scenarios however, the environment is dynamic - changes are introduced through various means: (1) objects are movable and can be transferred to new locations; (2) the point of view of a robot and a human can change while navigating through the environment; (3) new known objects can be introduced; (4) new unknown objects can be introduced which will be impossible to ground. Humans effectively refer to objects in a variety of ways, even objects that have never encountered before. For instance, unknown objects can be described through their *known* parts [7]. For instance, a tricycle can be described as a bike with three wheels. This human ability of successfully describing unknown entities and objects as well as choosing what attributes to mention is a result of commonsense ability. Endowing robots with this ability will enable communication when computer vision systems fail to recognise objects successfully and will enable life-long learning, which will be particularly useful in situations where limited data are available for learning [2].

2.2 Reasoning about object utility - beyond visual descriptions

In addition to successful descriptions of objects and other entities, reasoning about the objects' utility will lead to natural communication between humans and robots. Robot's ability to infer how objects could be used is useful for human-robot collaboration tasks, when the goal is to achieve tasks by interacting with the environment. This is under-researched task in NLG which aims at imitating abilities already present in humans, for instance a human knows that a mug can be used for measuring cooking ingredients or that a box can be used to store objects. However, a robot or more generally an artificial agent will not know this property unless it is clearly described in a domain ontology or knowledge base.

2.3 Inference of objects that are not visible

Finally, referring to non-visible object parts (e.g. we know that a wardrobe will contain clothes although we cannot see them) is an important challenge for situated human-robot interaction. Exploiting commonsense knowledge available from other sources such as wiki data to enhance Natural Language Generation can be a step toward this direction.

3 PROPOSED CHALLENGE

Based on the aforementioned open challenges and inspired by [9], this paper proposes the *Robot-Commonsense* challenge as an extension of the home move task [9], where one or multiple humans need to collaborate with a robot to pack objects. The robot should observe the humans and make inferences about their intentions, recognise objects and recommend the best way to pack given the space as well as the specific objects' attributes, such as whether they are fragile, whether they can act as containers themselves etc.

The proposed task requires utilising multi-modal information, commonsense knowledge as well as situated dialogue management as shown in Figure 1. The robot must be able to recognise objects and actions, understand and participate in communication situations, both explicitly, for instance when the human addresses directly the robot, but also implicitly, for instance when the human points to an object. In addition, the robot must be able to plan and recommend actions to humans. This task does not propose handling of objects, although that could be the natural next step for Natural Language Generation for Human-Robot Interaction.

ACKNOWLEDGMENTS

The author is supported by the EPSRC grant CiViL: EP/T014598/1.

REFERENCES

[1] Paola Ardón, Èric Pairet, Ron Petrick, Subramanian Ramamoorthy, and Katrin Lohan. 2019. Reasoning on Grasp-Action Affordances. In *Proceedings of the 20th Annual Conference on Towards Autonomous Robotic Systems (TAROS)*.

[2] Jekaterina Belakova and Dimitra Gkatzia. 2018. Learning from limited datasets: Implications for Natural Language Generation and Human-Robot Interaction. In *Proceedings of the Workshop on Natural Language Generation for Human–Robot Interaction*.

[3] Haonan Chen, Hao Tan, Alan Kuntz, Mohit Bansal, and Ron Alterovitz. 2019. Enabling Robots to Understand Incomplete Natural Language Instructions Using Commonsense Reasoning. *CoRR* abs/1904.12907 (2019).

[4] Suzanne Eggins and Diana Slade. 2004. *Analysing: Casual Conversation*.

[5] R. Fang, M. Doering, and J. Y. Chai. 2015. Embodied Collaborative Referring Expression Generation in Situated Human-Robot Interaction. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*.

[6] Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in Natural Language Generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research (JAIR)* 61 (2018), 65–170.

[7] Dimitra Gkatzia, Verena Rieser, Phil Bartie, and William Mackaness. 2015. From the Virtual to the RealWorld: Referring to Objects in Real-World Spatial Scenes. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[8] Emiel Krahmer and Kees van Deemter. 2012. Computational Generation of Referring Expressions: A Survey. *Comput. Linguist.* , Vol. 38, 1 (2012), 173–218.

[9] Séverin Lemaignan, Mathieu Warnier, E. Akin Sisbot, Aurélie Clodic, and Rachid Alami. 2017. Artificial cognition for social human–robot interaction: An implementation. *Artificial Intelligence* 247 (2017), 45 – 69. Special Issue on AI and Robotics.

[10] Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain Yourself! Leveraging Language Models for Commonsense Reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

[11] Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi. 2018. Event2Mind: Commonsense Inference on Events, Intents, and Reactions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*.

[12] Paul Trichelair, Ali Emami, Adam Trischler, Kaheer Suleman, and Jackie Chi Kit Cheung. 2019. How Reasonable are Common-Sense Reasoning Tasks: A Case-Study on the Winograd Schema Challenge and SWAG. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

[13] Christopher D. Wallbridge, Séverin Lemaignan, Emmanuel Senft, and Tony Belpaeme. 2019. Generating Spatial Referring Expressions in a Social Robot: Dynamic vs. Non-ambiguous. *Frontiers in Robotics and AI* , Vol. 6 (2019), 67.