

# Open Challenges on Generating Referring Expressions for Human-Robot Interaction

Fethiye Irmak Doğan  
KTH Royal Institute of Technology  
Stockholm, Sweden  
fidogan@kth.se

Iolanda Leite  
KTH Royal Institute of Technology  
Stockholm, Sweden  
iolanda@kth.se

## ABSTRACT

Effective verbal communication is crucial in human-robot collaboration. When a robot helps its human partner to complete a task with verbal instructions, referring expressions are commonly employed during the interaction. Despite many studies on generating referring expressions, crucial open challenges still remain for effective interaction. In this work, we discuss some of these challenges (i.e., using contextual information, taking users' perspectives, and handling misinterpretations in an autonomous manner).

## KEYWORDS

Generating Referring Expressions, Context-Dependency, Perspective-Taking, Handling of Misinterpretations

## 1 INTRODUCTION

In a human-robot collaborative task, it is critical that verbal communication between a robot and a human is effective. For instance, when a human assembles furniture, and a robot helps to find the correct pieces, the robot should direct its human partner and describe the target objects effectively. Expressions used for describing objects in terms of their distinguishing features are called referring expressions, and referring expression generation is defined as "choosing the words and phrases to express domain objects" [16].

Generating appropriate referring expressions has the potential of significantly improving human-robot collaboration. It is one of the most studied areas in natural language generation for social robotics because the problem contains a relatively straightforward input and output [16]. Studies on referring expression generation based on primarily rule-based templates or algorithms [21, 35, 36, 38], and recent studies have addressed this problem using learning-based methods [12, 25, 31].

Although referring expression generation has been extensively studied in HRI, there are still open challenges for more efficient generation mechanisms applicable to different tasks. In this paper, after reviewing the state of the art research in this area, we summarize the main open challenges and open further research directions (see Figure 1), which can be summarized as follows:

- Leveraging contextual information to generate referring expressions that facilitate communication
- While generating referring expressions, taking the perspectives of users for an effective collaboration
- Being able to complete the task accurately in dynamic environments by autonomously handling the misinterpretations

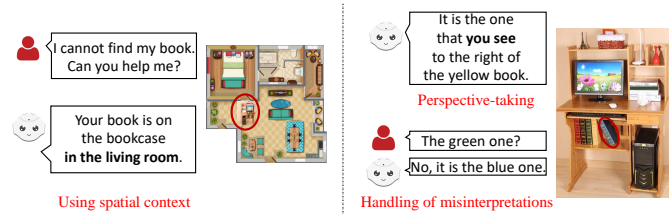


Figure 1: Example interactions for open challenges on generating referring expressions.

## 2 OPEN CHALLENGES

### 2.1 Using Contextual Information

When a robot describes an object to its human partner, it needs to consider the social context (e.g., the person's age and knowledge level about the object) and spatial context (e.g., whether it is likely to find the target object in the existing place or the robot needs to direct the user to another place), and adapt itself accordingly.

Understanding social context is essential to generate comprehensible referring expressions. For instance, if a robot describes a rarely known object to a child, it can be more efficient to use color or shape information of the object instead of solely using the object's name.

To facilitate finding the described object, interpreting spatial context is crucial. For example, if a robot describes an object which is more likely to belong to a kitchen, it can direct the user to the correct place with its referring expressions, e.g., "the object A next to the object B **in the kitchen**". Otherwise, users can waste time by looking for objects in the wrong places, which can affect the user's perception of the interaction and the robot. To address these challenges, recent studies on context modeling in robots have employed deep learning [4, 5, 9, 10]. In these studies, co-occurrences of the objects [10], spatial relations between them [5], and their affordances [4] have been utilized.

Adapting referring expressions concerning context has been focused [17, 18, 20, 33]. In a promising referring expression study, Viethen and Dale [33] showed that human behavior on selecting the content of referring expressions are mostly speaker-dependent. They stated that this speaker-dependency might be correlated to "age, gender, and social or cultural background" of the users, and this is open for further research. Some studies have focused on context-dependent selection of distinguishing features of objects [18, 20, 33] rather than using fixed preference ordering over features [7]. Further, Foster et al. [17] studied the context-sensitive generation of referring expressions on human-robot joint construction tasks. Their context-sensitive expressions include "... this red cube and

screw it ..." where "this red cube" and "it" are called context-sensitive references. Even though these works have important improvements in context-dependency of referring expressions, considering all aspects of social and spatial contexts remains unsolved.

## 2.2 Perspective-Taking

Perspective-taking has been a research topic in a variety of fields (e.g., psychology, cognitive science, robotics, and computer vision). In robotics, it has been studied when the robot comprehends referring expressions [3, 26, 37]. Many of these studies focused on perspective-taking when users employ spatial relations [14, 15, 19, 29, 32]. While generating referring expressions, Magassouba et al. [25] addressed this problem by generating perspective-free expressions, i.e., expressions that don't tie to particular viewpoints.

To generate unambiguous referring expressions, a robot needs to be explicit about from which perspective it refers to the objects, especially while using spatial relations between them. This can be achieved, for example, by clearly stating "the object A that **I** see to the right of the object B" or "the object A that **you** see to the left of the object B" to not to cause any misunderstanding during the interaction.

While the robot is describing objects, taking the perspectives of users and evaluating its impacts have been remained unsolved until very recently. Lately, we have proposed a method for making the first attempt to address this problem. In our recent work [11], we observed a scene from different perspectives and generated an expression from the closest perspective of a user. Further, we evaluated the impact of perspective-taking with regard to different aspects of the interaction (e.g., task efficiency, perception of the task, and the robot). Through a user study, we showed that when the objects are spatially described from the users' perspectives, participants take less time to find the referred objects, find the correct objects more often and consider the task easier.

Although our recent work demonstrates the significance of perspective-taking for effective collaboration, the method we proposed depended on the views of a scene from different perspectives. For a more general solution, i.e., taking different perspectives from a single view, 3D or 6D pose information of the objects can be helpful. However, existing 3D or 6D off the shelf object pose predictors are still generally limited by predicting a few types of objects – mostly vehicles, pedestrians, trees. Therefore, they are not sufficient to use in real-world settings. Further, taking different perspectives in 2D is still an open topic in computer vision. Even though there are novel view synthesis methods [13, 24, 30] that address this problem, existing solutions are still immature to use in real-world HRI.

Perspective-taking is also crucial when there are occluded objects from the views of users. When a robot observes a target object is occluded from the perspective of users, it should inform them that they need to change their viewpoint to see the target object and to achieve the task accurately. To address this problem, 3D cameras can be used, and occlusions can be estimated from the depth information of the objects. Further, if the problem is aimed to be resolved in 2D, studies on single-image depth estimation [6, 22, 23] can be employed to predict the depth of objects.

Although perspective-taking has been addressed while comprehending referring expressions and recently to generate these

expressions, while the robot is describing objects, taking a user's perspectives from a single view and informing the user about occlusions is still open for further research.

## 2.3 Handling of Misinterpretations in an Autonomous Manner

When a robot describes an object to its partner, it needs to cope with misinterpretations of its expressions and clarify them in an autonomous manner. In other words, when environments are highly ambiguous, users might misinterpret the expression of a robot (e.g., they might head towards the wrong objects) or ask for clarification. In these cases, the robot should be able to clarify its expressions in an autonomous manner to accurately describe the target object in dynamically changing real-world environments.

To describe objects autonomously (i.e., generating referring expressions directly from scenes without requiring any prior knowledge about the environment, existing objects, or their configurations), recent studies in robotics have relied on deep learning [12, 25, 31]. In our recent work [12], we have proposed a method to generate spatial referring expressions in a natural and unambiguous manner in real-world environments.

In order to handle misinterpretations while a robot comprehends referring expressions, Shridhar and Hsu [27, 28] have proposed a method grounding by generation. They have used the generation part of their model for asking disambiguating questions during comprehension. Moreover, to handle misinterpretations while a robot refers to objects, Wallbridge et al. [34] have proposed a dynamic method. In this method, there are three different dynamic description categories (i.e., negate, elaborate, positive) to provide user further clarification.

Even though these systems have made some advances to cope with misinterpretations, they are either mainly focused on comprehension and a specific task (i.e., pick and place) [27, 28] or limited by the number of clarification categories [34]. In order to handle misinterpretations while describing objects, existing autonomous referring expression generation methods need to be extended with detecting misinterpretations and providing more flexible clarifications. For this purpose, works on explainability (i.e., building more transparent and understandable models in their prediction-making process [2]), visual question answering (i.e., generating an answer for a given image and a question [1]), or visual dialog (i.e., generating an answer for a given image and a history of a dialog [8]) can be avenues worth exploring.

## 3 CONCLUSION

In this paper, we have focused on open challenges while generating referring expressions. We have suggested that utilizing contextual information and adapting expressions concerning context might contribute to more effective language-based interactions between robots and people. Further, we have stated that being explicit about from which perspective the expression is generated and taking the user perspective might be helpful for an unambiguous and efficient interaction. Finally, we have claimed that the handling of misinterpretations in an autonomous manner is necessary for successfully completing the task in dynamically changing environments.

## REFERENCES

- [1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*. 2425–2433.
- [2] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Benetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (jun 2020), 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- [3] Matt Berlin, Jesse Gray, Andrea Lockerd Thomaz, and Cynthia Breazeal. 2006. Perspective taking: An organizing principle for learning in human-robot interaction. In *AAAI*, Vol. 2. 1444–1450.
- [4] Ilker Bozcan and Sinan Kalkan. 2019. COSMO: contextualized scene modeling with boltzmann machines. *Robotics and Autonomous Systems* 113 (2019), 132–148.
- [5] I. Bozcan, Y. Oymak, İ. Z. Alemdar, and S. Kalkan. 2018. What is (Missing or Wrong) in the Scene? A Hybrid Deep Boltzmann Machine for Contextualized Scene Modeling. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*. 5869–5874. <https://doi.org/10.1109/ICRA.2018.8460828>
- [6] Ayan Chakrabarti, Jingyu Shao, and Greg Shakhnarovich. 2016. Depth from a Single Image by Harmonizing Overcomplete Local Network Predictions. In *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Eds.). Curran Associates, Inc., 2658–2666. <http://papers.nips.cc/paper/6510-depth-from-a-single-image-by-harmonizing-overcomplete-local-network-predictions.pdf>
- [7] Robert Dale and Ehud Reiter. 1995. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive science* 19, 2 (1995), 233–263.
- [8] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, Jose M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual Dialog. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [9] Fethiye Irmak Doğan, Ilker Bozcan, Mehmet Celik, and Sinan Kalkan. 2018. Cinet: A learning based approach to incremental context modeling in robots. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 4641–4646.
- [10] Fethiye Irmak Doğan, Hande Celikkanat, and Sinan Kalkan. 2018. A deep incremental boltzmann machine for modeling context in robots. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2411–2416.
- [11] Fethiye Irmak Doğan, Sarah Gillet, Elizabeth J Carter, and Iolanda Leite. 2020. The impact of adding perspective-taking to spatial referencing during human-robot interaction. *Robotics and Autonomous Systems* (2020), 103654.
- [12] Fethiye Irmak Doğan, Sinan Kalkan, and Iolanda Leite. 2019. Learning to Generate Unambiguous Spatial Referring Expressions for Real-World Environments. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 4992–4999. <https://doi.org/10.1109/IROS40897.2019.8968510>
- [13] John Flynn, Michael Broxton, Paul E. Debevec, Matthew DuVall, Graham Fyfe, Ryan S. Overbeck, Noah Snively, and Richard Tucker. 2019. DeepView: View Synthesis with Learned Gradient Descent. *CoRR* abs/1906.07316 (2019). [arXiv:1906.07316](http://arxiv.org/abs/1906.07316) <http://arxiv.org/abs/1906.07316>
- [14] Terrence Fong, Clayton Kunz, Laura M Hiatt, and Magda Bugajska. 2006. The human-robot interaction operating system. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*. 41–48.
- [15] Terrence Fong, Illah Nourbakhsh, Clayton Kunz, Lorenzo Flückiger, John Schreiner, Robert Ambrose, Robert Burrige, Reid Simmons, Laura Hiatt, Alan Schultz, J. Trafton, Magdalena Bugajska, and Jean Scholtz. 2005. The Peer-to-Peer Human-Robot Interaction Project. In *Space 2005*. Vol. 2. 6750. <https://doi.org/10.2514/6.2005-6750>
- [16] Mary Ellen Foster. 2019. Natural language generation for social robotics: opportunities and challenges. *Philosophical Transactions of the Royal Society B* 374, 1771 (2019), 20180027.
- [17] Mary Ellen Foster, Manuel Giuliani, and Amy Isard. 2014. Task-based evaluation of context-sensitive referring expressions in human-robot dialogue. *Language, Cognition and Neuroscience* 29, 8 (2014), 1018–1034.
- [18] Konstantina Garoufi and Alexander Koller. 2014. Generation of effective referring expressions in situated context. *Language, Cognition and Neuroscience* 29, 8 (2014), 986–1001.
- [19] William G Kennedy, Magdalena D Bugajska, Matthew Marge, William Adams, Benjamin R Fransen, Dennis Perzanowski, Alan C Schultz, and J Gregory Trafton. 2007. Spatial representation and reasoning for human-robot collaboration. In *AAAI*, Vol. 7. 1554–1559.
- [20] Emiel Krahrmer and Mariët Theune. 2002. Efficient context-sensitive generation of referring expressions. *Information sharing: Reference and presupposition in language generation and interpretation* 143 (2002), 223–263.
- [21] Lars Kunze, Tom Williams, Nick Hawes, and Matthias Scheutz. 2017. Spatial Referring Expression Generation for HRI: Algorithms and Evaluation Framework. In *AAAI Fall Symposium on AI and HRI*.
- [22] Jae-Han Lee, Minhyeok Heo, Kyung-Rae Kim, and Chang-Su Kim. 2018. Single-image depth estimation based on fourier domain analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 330–339.
- [23] Fayao Liu, Chunhua Shen, and Guosheng Lin. 2015. Deep convolutional neural fields for depth estimation from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5162–5170.
- [24] Miaomiao Liu, Xuming He, and Mathieu Salzmann. 2018. Geometry-aware Deep Network for Single-Image Novel View Synthesis. *CoRR* abs/1804.06008. [arXiv:1804.06008](http://arxiv.org/abs/1804.06008) <http://arxiv.org/abs/1804.06008>
- [25] Aly Magassouba, Komei Sugiura, and Hisashi Kawai. 2019. Multimodal Attention Branch Network for Perspective-Free Sentence Generation. (2019). [arXiv:cs.CV/1909.05664](http://arxiv.org/abs/1909.05664)
- [26] Amit Kumar Pandey and Rachid Alami. 2010. Mightability maps: A perceptual level decisional framework for co-operative and competitive human-robot interaction. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 5842–5848.
- [27] Mohit Shridhar and David Hsu. 2018. Interactive Visual Grounding of Referring Expressions for Human-Robot Interaction. *CoRR* abs/1806.03831 (2018). [arXiv:1806.03831](http://arxiv.org/abs/1806.03831) <http://arxiv.org/abs/1806.03831>
- [28] Mohit Shridhar, Dixant Mittal, and David Hsu. 2020. INGRESS: Interactive visual grounding of referring expressions. *The International Journal of Robotics Research* (2020), 0278364919897133.
- [29] Emrah Akin Sisbot, Luis F Marin-Urias, Xavier Broquere, Daniel Sidobre, and Rachid Alami. 2010. Synthesizing robot motions adapted to human presence. *International Journal of Social Robotics* 2, 3 (2010), 329–343.
- [30] Shao-Hua Sun, Minyoung Huh, Yuan-Hong Liao, Ning Zhang, and Joseph J Lim. 2018. Multi-view to novel view: Synthesizing novel views with self-learned confidence. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 155–171.
- [31] Mikihito Tanaka, Takayuki Itamochi, Kenichi Narioka, Ikuro Sato, Yoshitaka Ushiku, and Tatsuya Harada. 2019. Generating Easy-to-Understand Referring Expressions for Target Identifications. In *Proceedings of the IEEE International Conference on Computer Vision*. 5794–5803.
- [32] J. G. Trafton, N. L. Cassimatis, M. D. Bugajska, D. P. Brock, F. E. Mintz, and A. C. Schultz. 2005. Enabling Effective Human-robot Interaction Using Perspective-taking in Robots. *Trans. Sys. Man Cyber. Part A* 35, 4 (July 2005), 460–470. <https://doi.org/10.1109/TSMCA.2005.850592>
- [33] Jette Viethen and Robert Dale. 2010. Speaker-dependent variation in content selection for referring expression generation. In *Proceedings of the Australasian Language Technology Association Workshop 2010*. 81–89.
- [34] Christopher David 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* 6 (2019), 67.
- [35] Tom Williams and Matthias Scheutz. 2017. Referring expression generation under uncertainty: Algorithm and evaluation framework. In *INLG*.
- [36] Tom Williams and Matthias Scheutz. 2017. Referring expression generation under uncertainty in integrated robot architectures. In *RSS Workshop on Human-Centered Robotics: Interaction, Physiological Integration and Autonomy*.
- [37] Travis J Wiltshire, Daniel Barber, and Stephen M Fiore. 2013. Towards modeling social-cognitive mechanisms in robots to facilitate human-robot teaming. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 57. SAGE Publications Sage CA: Los Angeles, CA, 1278–1282.
- [38] Hendrik Zender, Geert-Jan M Kruijff, and Ivana Kruijff-Korbayová. 2009. Situated Resolution and Generation of Spatial Referring Expressions for Robotic Assistants. In *IJCAI*.