Tutorial: Achieving Common Ground in Multi-modal Dialogue

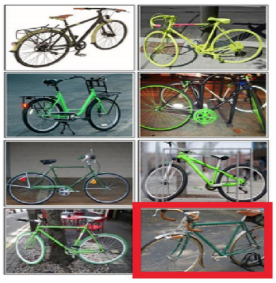
Malihe Alikhani
Computer Science
Rutgers University
malihe@pitt.edu

Matthew Stone
Computer Science
Rutgers University
matthew.stone@rutgers.edu

1 Description

All communication aims at achieving common ground (grounding): interlocutors can work together effectively only with mutual beliefs about what the state of the world is, about what their goals are, and about how they plan to make their goals a reality (Clark et al., 1991). Computational dialogue research, in particular, has a long history of influential work on how implemented systems can achieve common ground with human users, from formal results on grounding actions in conversation (Traum, 1994) to machine learning results on how best to fold confirmation actions into dialogue flow (Levin et al., 1998; Walker, 2000). Such classic results, however, offer scant guidance to the design of grounding modules and behaviors in cutting-edge systems, which increasingly combine multiple communication modalities, address complex tasks, and include the possibility for lightweight practical action interleaved with communication. This tutorial is premised on the idea that it's time to revisit work on grounding in human–human conversation, particularly Brennan's general and important characterization of grounding as seeking and providing evidence of mutual understanding (Brennan, 1990), in light of the opportunities and challenges of multi-modal settings such as human–robot interaction.

In this tutorial, we focus on three main topic areas: 1) grounding in human-human communication; 2) grounding in dialogue systems; and 3) grounding in multi-modal interactive systems, including image-oriented conversations and human-robot interactions. We highlight a number of achievements of recent computational research in coordinating complex content, show how these results lead to rich and challenging opportunities for doing grounding in more flexible and powerful ways, and canvass relevant insights from the literature on human–human conversation. We expect that the tutorial will be of interest to researchers in dialogue systems, computational semantics and cognitive modeling, and hope that it will catalyze research and system building that more directly explores the creative, strategic ways conversational agents might be able to seek and offer evidence about their understanding of their interlocutors.



A: A green bike with tan handlebars. B: Got it (Manuvinakurike et al., 2017)

A: The green cup is called Bill. B: Ok, the green cup is Bill. [point to the inferred object] (Liu and Chai, 2015)

Figure 1: Examples of the generation and interpretation of grounded referring expressions in multimodal interactive settings. Grounding is making sure that the listener understands what the speaker said.

Grounding in human-human communication. Clark et al. (1991) argued that communication is accomplished in two phases. In the presentation phase, the speaker presents signals intended to specify the content of the contributions. In the second phase, the participants work together to establish mutual beliefs that serve the purposes of the conversation. The two phases together constitute a unit of communication–*contributions*. Clark and Krych (2004) show how this model applies to coordinated action, while Stone and Stojnić (2015) applies the model to text-and-video presentations.

Coherence is key.

Grounding in dialogue systems. Computer systems achieve grounding mechanistically by ensuring they get attention and feedback from their users, tracking user state, and planning actions with reinforcement learning to resolve problematic situations. We will review techniques for maintaining engagement (Sidner et al., 2005; Bohus and Horvitz, 2014; Foster et al., 2017) and problems that arises in joint attention (Kontogiorgos et al., 2018) and turn taking such as incremental interpretation (DeVault and Stone, 2004; DeVault et al., 2011), ambiguity resolution (DeVault and Stone, 2009) and learning flexible dialogue management policies (Henderson et al., 2005). Similar questions have been studied in the context of instruction games (Perera et al., 2018; Thomason et al., 2019; Suhr and Artzi, 2018), and interactive tutoring systems (Yu et al., 2016; Wiggins et al., 2019).

Grounding in multi-modal systems. Multimodal systems offer the ability to use signals such as nodding, certain hand gestures and gazing at a speaker to communicate meaning and contribute to establishing common ground (Mavridis, 2015). However, multi-modal grounding is more than just using pointing to clarify. Multi-modal systems have diverse opportunities to demonstrate understanding. For example, recent work has aimed to bridge vision, interactive learning, and natural language understanding through language learning tasks based on natural images (Zhang et al., 2018; Kazemzadeh et al., 2014; De Vries et al., 2017a; Kim et al., 2020). The work on visual dialogue games (Geman et al., 2015) brings new resources and models for generating referring expression for referents in images (Suhr et al., 2019; Shekhar et al., 2018), visually grounded spoken language communication (Roy, 2002; Gkatzia et al., 2015), and captioning (Levinboim et al., 2019; Alikhani and Stone, 2019), which can be used creatively to demonstrate how a system understand a user. Figure 1 shows two examples of models that understand and generate referring expressions in multimodal settings.

Similarly, robots can demostrate how they understand a task by carring it out—in research on interactive task learning in human-robot interaction (Zarrieß and Schlangen, 2018; Carlmeyer et al., 2018) as well as embodied agents performing interactive tasks (Gordon et al., 2018; Das et al., 2018) in physically simulated environments (Anderson et al., 2018; Tan and Bansal, 2018) often drawing on the successes of deep learning and reinforcement learning (Branavan et al., 2009; Liu and Chai, 2015). A lesson that can be learned from this line of research is that one main factor that affects grounding is the choice of medium of communication. Thus, researchers have developed different techniques and methods for data collection and modeling of multimodal communication (Alikhani et al., 2019; Novikova et al., 2016). Figure 2 shows two example resources that were put together using crowdsourcing and virtual reality systems. We will discuss the strengths and shortcomings of these methods.

Show me a restaurant by the river, serving pasta/Italian food, highly rated and expensive, not child-friendly, located near Cafe Adriatic. (Novikova et al., 2016)

Crystal Island, an interactive narrative-centered virtual learning environment (Rowe et al., 2008)

Figure 2: Content and medium affect grounding. This figure shows two examples of interactive multimodal dialogue systems.

We pay special attention to non-verbal grounding in languages beyond English, including German (Han and Schlangen, 2018), Swedish (Kontogiorgos, 2017), Japanese (Endrass et al., 2013; Nakano et al., 2003), French (Lemaignan and Alami, 2013; Steels, 2001), Italian (Borghi and Cangelosi, 2014; Taylor et al., 1986), Spanish (Kery et al., 2019), Russian (Janda, 1988), and American sign language (Emmorey and Casey, 1995). These investigations often describe important language-dependent characteristics and cultural differences in studying non-verbal grounding.

Grounding in end-to-end language & vision systems. With current advances in neural mod-

elling and the availability of large pretrained models in language and vision, multi-modal interaction often is enabled by neural end-to-end architectures with multimodal encodings, e.g. by answering questions abut visual scenes (Antol et al., 2015; Das et al., 2017). It is argued that these shared representations help to ground word meanings. In this tutorial, we will discuss how this type of lexical grounding relates to grounding in dialogue from a theoretical perspective (Larsson, 2018), as well as within different interactive application scenarios – ranging from interactively identifying an object (De Vries et al., 2017b) to dialogue-based learning of word meanings (Yu et al., 2016). We then critically review existing datasets and shared tasks and showcase some of the shortcomings of current vision and language models, e.g. (Agarwal et al., 2018). In contrast to previous ACL tutorials on Multimodal Learning and Reasoning, we will concentrate on identifying different grounding phenomena as identified in the first part of this tutorial.

2 Outline

We begin by discussing grounding in human-human communication (∼20 min). After that, we discuss the role of grounding in spoken dialogue systems (∼30 min) and visually grounded interactions including grounding visual explanations in images and multimodal language grounding for human-robot collaboration (∼90 min). We then survey methods for developing and testing multimodal systems to study non-verbal grounding (∼20 min). We follow this by describing common solution concepts and barrier problems that cross application domains and interaction types (∼20 min).

3 Prerequisites and reading list

The tutorial will be self-contained. For further readings, we recommend the following publications that are central to the non-verbal grounding framework as of late 2019:

1. Grounding in communication, Herb Clark and Susan Brennan. (Clark et al., 1991)
2. Meaning and demonstration by Una Stojnic and Matthew Stone (Stone and Stojnić, 2015)
3. Using Reinforcement Learning to Model Incrementality in a Fast-Paced Dialogue Game, Ramesh Manuvinakurike, David DeVault and Kallirroi Georgila. (Manuvinakurike et al., 2017)
4. Language to Action: Towards Interactive Task Learning with Physical Agents, Joyce Y. Chai by Joyce Y. Chai et al.(Chai et al., 2018)
5. It's Not What You Do, It's How You Do It: Grounding Uncertainty for a Simple Robot, Julian Hough and David Schlangen. (Hough and Schlangen, 2017)
6. Learning Effective Multimodal Dialogue Strategies from Wizard-of-Oz Data: Bootstrapping and Evaluation rieser-lemon by Verena Rieser and Oliver Lemon. (Rieser and Lemon, 2008)
7. A survey of nonverbal signaling methods for non-humanoid robots by Elizabeth Cha et al. (Cha et al., 2018)
8. The Devil is in the Details: A Magnifying Glass for the GuessWhich Visual Dialogue Game by Alberto Testoni et al. (Testoni et al., 2019)

4 Authors

Malihe Alikhani is an assistant professor of computer science in the School of Computing and Information at the University of Pittsburgh. Her research aims at teaching machines to understand and generate multimodal communication. She is the recipient of the fellowship award for excellence in computation and data sciences from Rutgers Discovery Informatics Institute in 2018 and the Anita Berg student fellowship in 2019. Before joining Rutgers, she was a lecturer and an adjunct professor of Mathematics and Statistics for a year at San Diego State University and San Diego Mesa College. She has served as the program committee of ACL, NAACL, EMNLP, AAAI, ICRL, ICMI, and INLG and is currently the associate editor of the Mental Note Journal. email: ma1195@cs.rutgers.edu, webpage: www.malihealikhani.com

Matthew Stone is professor and chair in the Department of Computer Science at Rutgers University; he holds a joint appointment in the Rutgers Center for Cognitive Science. His research focuses on discourse, dialogue and natural language generation; he is particularly interested in leveraging semantics to make interactive systems easier to build and more human-like in their behavior. He was program co-chair for NAACL 2007,

general co-chair for SIGDIAL 2014. He has also served as program co-chair for INLG and IWCS, as an information officer for SIGSEM, and on the editorial board for Computational Linguistics. email: mdstone@cs.rutgers.edu, website: www.cs.rutgers.edu/~mdstone/

Acknowledgments

Preparation of this tutorial was supported in part by the DATA-INSPIRE Institute at Rutgers http://robotics.cs.rutgers.edu/data-inspire/ under NSF HDR TRIPODS award CCF-1934924. We gratefully acknowledge the effort of Professor Verena Rieser of Heriot-Watt University, who discussed the tutorial with us extensively but was ultimately unable to participate due to the disruption of COVID-19.

References

Shubham Agarwal, Ondřej Dušek, Ioannis Konstas, and Verena Rieser. 2018. Improving context modelling in multimodal dialogue generation. In *Proceedings of the 11th International Conference on Natural Language Generation*.

Malihe Alikhani, Sreyasi Nag Chowdhury, Gerard de Melo, and Matthew Stone. 2019. Cite: A corpus of image-text discourse relations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 570–575.

Malihe Alikhani and Matthew Stone. 2019. caption as a coherence relation: Evidence and implications. In *Proceedings of the Second Workshop on Shortcomings in Vision and Language*, pages 58–67.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3674–3683.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *ICCV*.

Dan Bohus and Eric Horvitz. 2014. Managing human-robot engagement with forecasts and... um... hesitations. In *Proceedings of the 16th international conference on multimodal interaction*. ACM.

Anna M Borghi and Angelo Cangelosi. 2014. Action and language integration: From humans to cognitive robots. *Topics in cognitive science*, 6(3):344–358.

Satchuthananthavale RK Branavan, Harr Chen, Luke S Zettlemoyer, and Regina Barzilay. 2009. Reinforcement learning for mapping instructions to actions. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL*, pages 82–90.

Susan E. Brennan. 1990. *Seeking and Providing Evidence for Mutual Understanding*. Ph.D. thesis, Stanford University.

Birte Carlmeyer, Simon Betz, Petra Wagner, Britta Wrede, and David Schlangen. 2018. The hesitating robot-implementation and first impressions. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*.

Elizabeth Cha, Yunkyung Kim, Terrence Fong, Maja J Mataric, et al. 2018. A survey of nonverbal signaling methods for non-humanoid robots. *Foundations and Trends® in Robotics*, 6(4):211–323.

Joyce Y Chai, Qiaozi Gao, Lanbo She, Shaohua Yang, Sari Saba-Sadiya, and Guangyue Xu. 2018. Language to action: Towards interactive task learning with physical agents. In *AAMAS*, page 6.

Herbert H Clark, Susan E Brennan, et al. 1991. Grounding in communication. *Perspectives on socially shared cognition*, 13(1991):127–149.

Herbert H. Clark and Meredyth A. Krych. 2004. Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, 50:62–81.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. 2018. Embodied question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2054–2063.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *CVPR*.

Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017a. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017b. Guesswhat?! visual object discovery through multi-modal dialogue. In *CVPR*.

David DeVault, Kenji Sagae, and David Traum. 2011. Incremental interpretation and prediction of utterance meaning for interactive dialogue. *Dialogue & Discourse*, 2(1):143–170.

David DeVault and Matthew Stone. 2004. Interpreting vague utterances in context. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1247. Association for Computational Linguistics.

David DeVault and Matthew Stone. 2009. Learning to interpret utterances using dialogue history. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*.

Karen Emmorey and Shannon Casey. 1995. A comparison of spatial language in english & american sign language. *Sign Language Studies*, 88(1):255–288.

Birgit Endrass, Elisabeth André, Matthias Rehm, and Yukiko Nakano. 2013. Investigating culture-related aspects of behavior for virtual characters. *Autonomous Agents and Multi-Agent Systems*.

Mary Ellen Foster, Andre Gaschler, and Manuel Giuliani. 2017. Automatically classifying user engagement for dynamic multi-party human–robot interaction. *International Journal of Social Robotics*.

Donald Geman, Stuart Geman, Neil Hallonquist, and Laurent Younes. 2015. Visual turing test for computer vision systems. *Proceedings of the National Academy of Sciences*, 112(12):3618–3623.

Dimitra Gkatzia, Amanda Cercas Curry, Verena Rieser, and Oliver Lemon. 2015. A game-based setup for data collection and task-based evaluation of uncertain information presentation. In *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG)*.

Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. 2018. Iqa: Visual question answering in interactive environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4089–4098.

Ting Han and David Schlangen. 2018. A corpus of natural multimodal spatial scene descriptions. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

James Henderson, Oliver Lemon, and Kallirroi Georgila. 2005. Hybrid reinforcement/supervised learning for dialogue policies from communicator data. In *IJCAI workshop on knowledge and reasoning in practical dialogue systems*, pages 68–75. Citeseer.

Julian Hough and David Schlangen. 2017. It's not what you do, it's how you do it: Grounding uncertainty for a simple robot. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, pages 274–282. ACM.

Laura A Janda. 1988. The mapping of elements of cognitive space onto grammatical relations: An example from russian verbal prefixation. *Topics in cognitive linguistics*, 50:327–343.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798.

Caroline Kery, Francis Ferraro, and Cynthia Matuszek. 2019. ¿ es un plátano? exploring the application of a physically grounded language acquisition system to spanish. In *Proceedings of the Combined Workshop on Spatial Language Understanding (SpLU) and Grounded Communication for Robotics (RoboNLP)*, pages 7–17.

Hyounghun Kim, Hao Tan, and Mohit Bansal. 2020. Modality-balanced models for visual dialogue.

Dimosthenis Kontogiorgos. 2017. Multimodal language grounding for improved human-robot collaboration: exploring spatial semantic representations in the shared space of attention. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 660–664. ACM.

Dimosthenis Kontogiorgos, Vanya Avramova, Simon Alexanderson, Patrik Jonell, Catharine Oertel, Jonas Beskow, Gabriel Skantze, and Joakim Gustafson. 2018. A multimodal corpus for mutual gaze and joint attention in multiparty situated interaction. In *International Conference on Language Resources and Evaluation (LREC 2018)*.

Staffan Larsson. 2018. Grounding as a side-effect of grounding. *Topics in Cognitive Science*, 10(2):389–408.

Séverin Lemaignan and Rachid Alami. 2013. talking to my robot: From knowledge grounding to dialogue processing. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 409–409. IEEE.

E. Levin, R. Pieraccini, and W. Eckert. 1998. Using markov decision process for learning dialogue strategies. In *Proceedings International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Tomer Levinboim, Ashish Thapliyal, Piyush Sharma, and Radu Soricut. 2019. Quality estimation for image captions based on large-scale human evaluations. *arXiv preprint arXiv:1909.03396*.

Changsong Liu and Joyce Yue Chai. 2015. Learning to mediate perceptual differences in situated human-robot dialogue. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Ramesh Manuvinakurike, David DeVault, and Kallirroi Georgila. 2017. Using reinforcement learning to model incrementality in a fast-paced dialogue game. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 331–341.

Nikolaos Mavridis. 2015. A review of verbal and non-verbal human–robot interactive communication. *Robotics and Autonomous Systems*, 63:22–35.

Yukiko I Nakano, Gabe Reinstein, Tom Stocky, and Justine Cassell. 2003. Towards a model of face-to-face grounding. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 553–561. Association for Computational Linguistics.

Jekaterina Novikova, Oliver Lemon, and Verena Rieser. 2016. Crowd-sourcing nlg data: Pictures elicit better data. *arXiv preprint arXiv:1608.00339*.

Ian Perera, James Allen, Choh Man Teng, and Lucian Galescu. 2018. A situated dialogue system for learning structural concepts in blocks world. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 89–98.

Verena Rieser and Oliver Lemon. 2008. Learning effective multimodal dialogue strategies from Wizard-of-Oz data: Bootstrapping and evaluation. In *Proceedings of ACL-08: HLT*, pages 638–646, Columbus, Ohio. Association for Computational Linguistics.

Jonathan P Rowe, Eun Young Ha, and James C Lester. 2008. Archetype-driven character dialogue generation for interactive narrative. In *International Workshop on Intelligent Virtual Agents*. Springer.

Deb Roy. 2002. A trainable visually-grounded spoken language generation system. In *Proceedings of the international conference of spoken language processing*. Citeseer.

Ravi Shekhar, Aashish Venkatesh, Tim Baumgärtner, Elia Bruni, Barbara Plank, Raffaella Bernardi, and Raquel Fernández. 2018. Beyond task success: A closer look at jointly learning to see, ask, and guesswhat.

Candace L. Sidner, Christopher Lee, Cory D. Kidd, Neal Lesh, and Charles Rich. 2005. Explorations in engagement for humans and robots. *Artificial Intelligence*, 166(1):140 – 164.

Luc Steels. 2001. Language games for autonomous robots. *IEEE Intelligent systems*, 16(5):16–22.

Matthew Stone and Una Stojnić. 2015. Meaning and demonstration. *Review of Philosophy and Psychology*, 6:69–97.

Alane Suhr and Yoav Artzi. 2018. Situated mapping of sequential instructions to actions with single-step reward observation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2072–2082.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Hao Tan and Mohit Bansal. 2018. Source-target inference models for spatial instruction understanding. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

John Taylor et al. 1986. *Contrasting prepositional categories: English and Italian*. Linguistic Agency University of Duisburg.

Alberto Testoni, Ravi Shekhar, Raquel Fernández, and Raffaella Bernardi. 2019. The devil is in the details: A magnifying glass for the guesswhich visual dialogue game.

Jesse Thomason, Aishwarya Padmakumar, Jivko Sinapov, Nick Walker, Yuqian Jiang, Harel Yedidsion, Justin Hart, Peter Stone, and Raymond J Mooney. 2019. Improving grounded natural language understanding through human-robot dialog. *arXiv preprint arXiv:1903.00122*.

David R Traum. 1994. A computational theory of grounding in natural language conversation. Technical report, Rochester Univ NY Dept of Computer Science.

Marilyn A. Walker. 2000. An application of reinforcement learning to dialogue strategy selection in a spoken dialogue system for email. *J. Artif. Intell. Res. (JAIR)*, 12:387–416.

Joseph B Wiggins, Mayank Kulkarni, Wookhee Min, Kristy Elizabeth Boyer, Bradford Mott, Eric Wiebe, and James Lester. 2019. Take the initiative: Mixed initiative dialogue policies for pedagogical agents in game-based learning environments. In *International Conference on Artificial Intelligence in Education*. Springer.

Yanchao Yu, Arash Eshghi, and Oliver Lemon. 2016. Interactively learning visually grounded word meanings from a human tutor. In *Proceedings of the 5th Workshop on Vision and Language*.

Sina Zarrieß and David Schlangen. 2018. Being data-driven is not enough: Revisiting interactive instruction giving as a challenge for nlg. In *Proceedings of the Workshop on NLG for Human–Robot Interaction*, pages 27–31.

Hanwang Zhang, Yulei Niu, and Shih-Fu Chang. 2018. Grounding referring expressions in images by variational context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4158–4166.