Grounding Meaning Representation for Situated Reasoning

Nikhil Krishnaswamy
Department of Computer Science
Colorado State University
Fort Collins, CO USA
nkrishna@colostate.edu

James Pustejovsky
Department of Computer Science
Brandeis University
Waltham, MA USA
jamesp@brandeis.edu

1 Tutorial Description

As natural language technology becomes ever-present in everyday life, people will expect artificial agents to understand language use as humans do. Nevertheless, most advanced neural AI systems fail at some types of interactions that are trivial for humans (e.g., ask a smart system "What am I pointing at?"). One critical aspect of human language understanding is *situated reasoning*, where inferences make reference to the local context, perceptual surroundings, and contextual groundings from the interaction. In this **cutting-edge** tutorial, we bring to the NLP/CL community a synthesis of multimodal grounding and meaning representation techniques with formal and computational models of *embodied reasoning*. We will discuss existing approaches to multimodal language grounding and meaning representations, discuss the kind of information each method captures and their relative suitability to situated reasoning tasks, and demonstrate how to construct agents that conduct situated reasoning by embodying a simulated environment. In doing so, these agents also represent their human interlocutor(s) within the simulation, and are represented through their virtual embodiment in the real world, enabling true bidirectional communication with a computer using multiple modalities.

"Grounding" in much of the NLP literature involves linking linguistic expressions to information expressed in another modality, often images or video (Yatskar et al., 2016; Li et al., 2019). Examples include linking semantic roles to entities in an image, or joint linguistic-visual attention between a caption and an image or video. Efforts have also focused on creating common meaning representation formalisms for linguistic data that are known to be relatively expressive, easy to annotate, and extensible to accommodate linguistic diversity, scale, and support inference, e.g., Copestake et al. (2005); Banarescu et al. (2013); Cooper and Ginzburg (2015); Pustejovsky et al. (2019); Lai et al. (2021).

Robust human-computer interactions and human-robot interactions will require representations with all these features, that encode the different modalities in use in such an interaction and ground them to the shared environment, enabling bidirectional, symmetric communication, and shared reference. Central to such situated meaning is the recognition and interpretation of gesture in the common ground (Holler and Wilkin, 2009; Alahverdzhieva et al., 2018).

Certain problems in human-to-human communication cannot be solved without *situated reasoning*, meaning they cannot be adequately addressed with ungrounded meaning representation or cross-modal linking of instances alone. Examples include grounding an object and then reasoning with it ("Pick up *this* box. Put it *there*."), referring to a previously-established concept or instance that was never explicitly introduced into the dialogue, underspecification of deixis, and in general, dynamic updating of context through perceptual, linguistic, action, or self-announcement. Without *both* a representation framework and mechanism for grounding references and inferences to the environment, such problems may well remain out of reach for NLP.

An appropriate representation should accommodate both the structure and content of different modalities, as well as facilitate alignment and binding across them. However, it must also distinguish between alignment across channels in a multimodal dialogue (language, gesture, gaze), and the situated grounding of an expression to the local environment, be it objects in a situated context, an image, or a formal registration in a database. Therefore, such a meaning representation should also have the basic facility for situated grounding; i.e., explicit mention of object and situational state in context.

To date there has been interest in creating meaning representations that capture multimodality, and in multimodal corpora that capture language use in a situated environment (e.g., Chen et al.

Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Tutorial Abstracts, pages 22–27
November 20-21, 2022.

(2019)), yet the two have been largely distinct. We will demonstrate how to bring these together into grounded meaning representations that capture language, gesture, object, and event semantics that can be used to not only represent situated meaning, but drive situated *reasoning* in embodied agents that occupy a three-dimensional environment.

There have been recent *ACL tutorials on meaning representations (Lopez and Gilroy, 2018; Koller et al., 2019), on common-sense reasoning (Sap et al., 2020), and on common ground and multimodality (Alikhani and Stone, 2020). To our knowledge this is the first time these three areas have been brought together with situated, grounded reasoning for an NLP/CL audience.

This tutorial will cover the most pressing problems in situated reasoning: namely, those requiring both multimodal grounding of expressions, as well as contextual reasoning with this information. Three example areas we will cover are:

"Make Me Another" Grounding an underspecified item or concept to previous elements of a dialogue requires an understanding of both what is salient in context, and of what elements of that item or concept are relevant to the situation inhabited by the interlocutors (Schlangen and Skantze, 2011). For example, if someone is cooking a stack of pancakes for someone else, and the diner says "make me *another*," a human would likely infer a reference to a single pancake, not the whole stack. The computational mechanisms for representing the elements of the environment and making this inference are richly involved. Addressing this problem and similar ones is an important part of building agents that respond to queries and requests in ways that are situationally appropriate.

Underspecification of Deixis The referent of a deixis may be ambiguous, though it naturally grounds to an object if one is available (Alahverdzhieva and Lascarides, 2011). Adding demonstratives like "this" or "there" naturally selects for objects vs. locations, and we will present models to capture these joint gesture-language semantics (Alahverdzhieva and Lascarides, 2010). Even coupling gesture and language may be insufficient for reasoning. "Pick up *this one* and put *it* · *there*," plus deixis, singles out a liftable object in the embedding space, and a possible ambiguity. If *there* refers to a location, then the command can be fully grounded in space. But, if the referent of the deixis is an object, additional reasoning must be conducted vis-à-vis what part of the object accommodates both the action *put* and the denotatum of *this one*. The many possible interpretations lead to rich reasoning strategies in situated space.

Dynamic Updating of Context Through Announcement When a participant in a dialogue sees, says, does, or realizes something new, the external and/or internal world changes for the participants, along with the capabilities for reasoning over the situation. For example, someone can verbally or gesturally announce an intent or provide information; perceptually demonstrate that something is present or absent; visibly act on a request or command; and personally realize something based on the current context. Each of these requires situated grounding and reasoning within those worlds.

1.1 Outline

This tutorial comprises 4 45-minute parts. 1) We will first present existing approaches to multimodal grounding, in the form of cross-modal linking (Yatskar et al., 2016; Yang et al., 2016; Sadhu et al., 2021) or linguistic-visual attention (Antol et al., 2015; Shih et al., 2016; Zhu et al., 2018; Sood et al., 2020) along with datasets that exist for this purpose (e.g., Kontogiorgos et al., 2018; Chen et al., 2019; Shridhar et al., 2020), and 2) common approaches to structured meaning representation (Copestake et al., 2005; Banarescu et al., 2013; Cooper and Ginzburg, 2015). 3) We will describe the forumation of *common ground* as a data structure of the information associated with a state in a dialogue or discourse (Clark et al., 1983; Stalnaker, 2002), and how it can be used to ground elements like gestures and situations to meaning representations (Lascarides and Stone, 2009; Alahverdzhieva et al., 2018). Each section will focus the material with regard how the discussed frameworks treat the grounding and reasoning questions from Sec. 1.

4) Finally we will present some of own work, including the grounded modeling language VoxML (Pustejovsky and Krishnaswamy, 2016) and a demonstration of building agents capable of situated reasoning in *VoxWorld* (Krishnaswamy and Pustejovsky, 2016; Pustejovsky and Krishnaswamy, 2021), a platform built on VoxML for developing embodied agent behaviors. We will provide a starter scene with an agent who can act upon the world, and discuss the computational and modeling considerations that go into developing

distinct types of agents, such as virtual collaborative assistants (Krishnaswamy et al., 2017), mobile robots (Krajovic et al., 2020; Tellex et al., 2020), and self-guided exploratory agents (Tan et al., 2019; Pustejovsky and Krishnaswamy, 2022), comparing our own framework to others'.

Technical requirements We have no special hardware requirements for this tutorial except for a projector or display screen.

Distribution of materials We plan to make all tutorial materials fully available to the community.

2 Target and Expected Audience

This tutorial will be of interest to both researchers in meaning representation, and in multimodal NLP and grounding, particularly those interested in both theoretical and data-driven approaches to language grounding and those interested in treating automated reasoning as more than just a pure machine learning problem. The diverse approaches to linguistic grounding of situated meaning have also provoked significant interest from the robotics community. Given the increased interest in interactive agents and grounding for robotics at in the *ACL community, as witnessed by the recent creation of Language Grounding to Vision, Robotics, and Beyond tracks at most *ACL venues, this tutorial, that synthesizes various approaches to situated conversation and interaction will be a timely way to bring these two communities closer. We expect this tutorial will draw an audience of roughly 30-45.

2.1 Requisite Background

This tutorial will be self-contained. However, to get the most out of this tutorial, attendees will want to be familiar with both theoretical and machine-learning approaches to semantics. Familiarity with common meaning representation frameworks, such as abstract meaning representation (Banarescu et al., 2013) or minimal recursion semantics (Copestake et al., 2005), is desirable, as is familiarity with multimodal language and vision techniques, such as VQA or image captioning (Antol et al., 2015; Shih et al., 2016). Participants will be invited to "code along" for the last part of the tutorial if they so desire, for which knowledge of C# and the Unity game engine will be advantageous but *not* prerequisite.

3 Breadth and Reading List

This tutorial draws on a wealth of both theory and applied research in multimodal semantics, including not only the central meaning representation work mentioned above (Copestake et al., 2005; Banarescu et al., 2013; Cooper and Ginzburg, 2015; Pustejovsky et al., 2019), but also gesture semantics and situated dialogue (Kendon, 2004; Lascarides and Stone, 2006, 2009; Kelleher and Kruijff, 2006), and qualitative spatiotemporal reasoning (Freksa, 1991; Forbus et al., 1991; Zimmermann and Freksa, 1996; Cohn and Renz, 2008). We bring these diverse areas together in the modeling language VoxML (Pustejovsky and Krishnaswamy, 2016), and this tutorial will demonstrate how to exploit the strengths of both meaning representations and data-driven multimodal methods to create agents that reason with vision, language, action, and gesture about the environments they inhabit and share with human beings. Suggested reading is below:

- L. Banarescu, et al. (2013). Abstract meaning representation for sembanking. In Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, pages 178–186. https://aclanthology.org/W13-2322.pdf
- A. Copestake, et al. (2005). Minimal recursion semantics: An introduction. Research on Language and Computation, 3(2):281–332. https://doi.org/10.1007/s11168-006-6327-9
- R. Cooper and J. Ginzburg. (2015). Type theory with records for natural language semantics. The Handbook of Contemporary Semantic Theory, pages 375–407. https://doi.org/10.1002/9781118882139.ch12
- A. Lascarides and M. Stone. (2009). A formal semantic analysis of gesture. Journal of Semantics, 26(4), 393-449. https://homepages.inf.ed.ac.uk/alex/papers/gesture_formal.pdf
- K. Alahverdzhieva, et al. Aligning speech and co-speech gesture in a constraint-based grammar. https://jlm.ipipan.waw.pl/index.php/JLM/article/view/167/179
- The qualitative spatial dynamics of motion in language. J. Pustejovsky, and J. L. Moszkowicz. (2011). Spatial Cognition & Computation 11, no. 1 (2011): 15-44. http://www.cs-135.org/wp-content/uploads/2017/12/SCC-2011.pdf

- J. Pustejovsky and N. Krishnaswamy (2021). Situated Meaning in Multimodal Dialogue: Human-Robot and Human-Computer Interactions, in TAL Volume 61 issue 3, pp 17-41. https://www.atala.org/sites/default/files/TAL-61-3-1_Pustejovsky.pdf

- J. Pustejovsky and N. Krishnaswamy (2021). Embodied Human Computer Interaction, Künstliche Intelligenz, Springer. https://doi.org/10.1007/s13218-021-00727-5

4 Instructors

Nikhil Krishnaswamy is Assistant Professor of Computer Science at Colorado State University and director of the Situated Grounding and Natural Language Lab (www.signallab.ai). He received his Ph.D. from Brandeis University in 2017. His primary research is in situated grounding and natural language semantics, using computational, formal, and simulation methods to study how language works and how humans use it. He is the co-creator of VoxML. He has taught courses on machine learning and NLP, previously taught at EACL 2017 (with J. Pustejovsky), and he will be co-teaching (also with J. Pustejovsky) at ESSLLI 2022 on multimodal semantics of affordances and actions. He has routinely received positive feedback as an instructor, including "always willing to engage in in-depth discussions regarding class material," "his understanding of the subject matter is phenomenal," "my favorite course this semester," and "he clearly spends a lot of time making his lectures engaging." He has served on the PC for ACL, EACL, NAACL, EMNLP, AAAI, AACL, etc. Email: nkrishna@colostate.edu, Website: https://www.nikhilkrishnaswamy.com.

James Pustejovsky is the TJX Feldberg Chair in Computer Science at Brandeis University, where he is also Chair of the Linguistics Program, Chair of the Computational Linguistics M.S. Program, and Director of the Lab for Linguistics and Computation. He received his B.S. from MIT and his Ph.D. from UMass Amherst. He has worked on computational and lexical semantics for 25 years and is chief developer of Generative Lexicon Theory; the TARSQI platform for temporal reasoning in language; TimeML and ISO-TimeML, a recently adopted ISO standard for temporal information in language; the recently adopted standard ISO-Space, a specification for spatial information in language; and the co-creator of the VoxML modeling framework for linguistic expressions and interactions as multimodal simulations VoxML (co-created with N. Krishnaswamy), enables real-time communication between humans and computers or robots for joint tasks, utilizing speech, gesture, gaze, and action. He is currently working with robotics researchers in HRI to allow the VoxML platform to act as both a dialogue management system as well as a simulation environment that reveals realtime epistemic state and perceptual input to a computational agent. Email: jamesp@brandeis.edu, Website: https://www.pusto.com.

5 Diversity

Situated reasoning and grounding inherently crosses language boundaries. Language grounding in English can be compared to language grounding a low-resourced language by way of a situated model. From a research perspective these are important questions to answer, to explore how different languages represent the same environment or situation. Therefore situated reasoning is an important potential way to broaden the linguistic diversity of NLP, and we hope the meaning representation component of this tutorial may inspire broadening meaning representations to more languages yet.

The instructors are junior and senior faculty, respectively, established in the NLP community. We actively recruit women and underrepresented minorities to our respective research groups, and plan to promote this tutorial to an international and diverse audience. We are experienced instructors in a hybrid format, and we will accommodate and promote remote attendance to broaden participation.

6 Ethics Statement

Computational agents that reason situationally necessarily require sight and hearing, and come with concomitant ethical issues regarding computer vision and speech recognition. In the course of this tutorial, we will discuss many of the considerations surrounding user privacy and storing user data (or, in the case of our own research, explicitly *not* doing that (Wang et al., 2017)). We will also discuss adapting speech recognition models to user diversity as part of the multimodal grounding section (Krishnaswamy and Alalyani, 2021).

Real-time, situated reasoning requires smaller, lightweight models. While we use large models where necessary, our use of meaning representations to guide search within multimodal grounding tasks provides a way to accomplish this task with less computational overhead and resource use.

References

Katya Alahverdzhieva and Alex Lascarides. 2010. Analysing speech and co-speech gesture in constraint-based grammars. In *The Proceedings of the 17th International Conference on Head-Driven Phrase Structure Grammar*, pages 6–26. Citeseer.

Katya Alahverdzhieva and Alex Lascarides. 2011. An hpsg approach to synchronous speech and deixis. In *Proceedings of the 18th international conference on head-driven phase structure grammar (HPSG)*, pages 6–24. CSLI Publications.

Katya Alahverdzhieva, Alex Lascarides, and Dan Flickinger. 2018. Aligning speech and co-speech gesture in a constraint-based grammar. *Journal of Language Modelling*, 5(3):421–464.

Malihe Alikhani and Matthew Stone. 2020. Achieving common ground in multi-modal dialogue. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 10–15.

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*, pages 2425–2433.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 178–186.

Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. 2019. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12538–12547.

Herbert H Clark, Robert Schreuder, and Samuel Buttrick. 1983. Common ground at the understanding of demonstrative reference. *Journal of verbal learning and verbal behavior*, 22(2):245–258.

Anthony G Cohn and Jochen Renz. 2008. Qualitative spatial representation and reasoning. *Foundations of Artificial Intelligence*, 3:551–596.

Robin Cooper and Jonathan Ginzburg. 2015. Type theory with records for natural language semantics. *The handbook of contemporary semantic theory*, pages 375–407.

Ann Copestake, Dan Flickinger, Carl Pollard, and Ivan A Sag. 2005. Minimal recursion semantics: An introduction. *Research on language and computation*, 3(2):281–332.

Kenneth D Forbus, Paul Nielsen, and Boi Faltings. 1991. Qualitative spatial reasoning: The clock project. *Artificial Intelligence*, 51(1-3):417–471.

Christian Freksa. 1991. Qualitative spatial reasoning. In *Cognitive and linguistic aspects of geographic space*, pages 361–372. Springer.

Judith Holler and Katie Wilkin. 2009. Communicating common ground: How mutually shared knowledge influences speech and gesture in a narrative task. *Language and Cognitive Processes*, 24(2):267–289.

John Kelleher and Geert-Jan M Kruijff. 2006. Incremental generation of spatial referring expressions in situated dialog. In *Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the association for computational linguistics*, pages 1041–1048.

Adam Kendon. 2004. *Gesture: Visible action as utterance*. Cambridge University Press.

Alexander Koller, Stephan Oepen, and Weiwei Sun. 2019. Graph-based meaning representations: Design and processing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 6–11.

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 *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Katherine Krajovic, Nikhil Krishnaswamy, Nathaniel J Dimick, R Pito Salas, and James Pustejovsky. 2020. Situated multimodal control of a mobile robot: Navigation through a virtual environment. *arXiv preprint arXiv:2007.09053*.

Nikhil Krishnaswamy and Nada Alalyani. 2021. Embodied multimodal agents to bridge the understanding gap. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 41–46.

Nikhil Krishnaswamy, Pradyumna Narayana, Isaac Wang, Kyeongmin Rim, Rahul Bangar, Dhruva Patil, Gururaj Mulay, Ross Beveridge, Jaime Ruiz, Bruce Draper, et al. 2017. Communicating and acting: Understanding gesture in simulation semantics. In *IWCS 2017—12th International Conference on Computational Semantics—Short papers*.

Nikhil Krishnaswamy and James Pustejovsky. 2016. Voxsim: A visual platform for modeling motion language. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 54–58.

Kenneth Lai, Richard Brutti, Lucia Donatelli, and James Pustejovsky. 2021. Situated umr for multimodal interactions. In *Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue*.

Alex Lascarides and Matthew Stone. 2006. *Formal semantics for iconic gesture*. Universität Potsdam.

Alex Lascarides and Matthew Stone. 2009. A formal semantic analysis of gesture. *Journal of Semantics*, 26(4):393–449.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Adam Lopez and Sorcha Gilroy. 2018. Graph formalisms for meaning representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*.

James Pustejovsky and Nikhil Krishnaswamy. 2016. Voxml: A visualization modeling language. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4606–4613.

James Pustejovsky and Nikhil Krishnaswamy. 2021. Embodied human computer interaction. *KI-Künstliche Intelligenz*, pages 1–21.

James Pustejovsky and Nikhil Krishnaswamy. 2022. Multimodal semantics for affordances and actiona. In *International Conference on Human-Computer Interaction*. Springer.

James Pustejovsky, Ken Lai, and Nianwen Xue. 2019. Modeling quantification and scope in abstract meaning representations. In *Proceedings of the first international workshop on designing meaning representations*, pages 28–33.

Arka Sadhu, Tanmay Gupta, Mark Yatskar, Ram Nevatia, and Aniruddha Kembhavi. 2021. Visual semantic role labeling for video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5589–5600.

Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi, and Dan Roth. 2020. Introductory tutorial: Commonsense reasoning for natural language processing. *Association for Computational Linguistics (ACL 2020): Tutorial Abstracts*, 27.

David Schlangen and Gabriel Skantze. 2011. A general, abstract model of incremental dialogue processing. *Dialogue & Discourse*, 2(1):83–111.

Kevin J Shih, Saurabh Singh, and Derek Hoiem. 2016. Where to look: Focus regions for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4613–4621.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10740–10749.

Ekta Sood, Simon Tannert, Philipp Müller, and Andreas Bulling. 2020. Improving natural language processing tasks with human gaze-guided neural attention. *Advances in Neural Information Processing Systems*, 33:6327–6341.

Robert Stalnaker. 2002. Common ground. *Linguistics and philosophy*, 25(5/6):701–721.

Hao Tan, Licheng Yu, and Mohit Bansal. 2019. Learning to navigate unseen environments: Back translation with environmental dropout. 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 2610–2621.

Stefanie Tellex, Nakul Gopalan, Hadas Kress-Gazit, and Cynthia Matuszek. 2020. Robots that use language. *Annual Review of Control, Robotics, and Autonomous Systems*, 3:25–55.

Isaac Wang, Mohtadi Ben Fraj, Pradyumna Narayana, Dhruva Patil, Gururaj Mulay, Rahul Bangar, J Ross Beveridge, Bruce A Draper, and Jaime Ruiz. 2017. Eggnog: A continuous, multi-modal data set of naturally occurring gestures with ground truth labels. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 414–421. IEEE.

Shaohua Yang, Qiaozi Gao, Changsong Liu, Caiming Xiong, Song-Chun Zhu, and Joyce Chai. 2016. Grounded semantic role labeling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 149–159.

Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. 2016. Situation recognition: Visual semantic role labeling for image understanding. *In Proceedings of the Conference of Computer Vision and Pattern Recognition (CVPR)*.

Junnan Zhu, Haoran Li, Tianshang Liu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2018. Msmo: Multimodal summarization with multimodal output. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 4154–4164.

Kai Zimmermann and Christian Freksa. 1996. Qualitative spatial reasoning using orientation, distance, and path knowledge. *Applied intelligence*, 6(1):49–58.