Annotating Situated Actions in Dialogue

Christopher Tam, Richard Brutti, Kenneth Lai, James Pustejovsky
Department of Computer Science
Brandeis University
Waltham, MA, USA
{christophertam,brutti,klai12,jamesp}@brandeis.edu

Abstract

Actions are critical for interpreting dialogue: they provide context for demonstratives and definite descriptions in discourse, and they continually update the common ground. This paper describes how Abstract Meaning Representation (AMR) can be used to annotate actions in multimodal human-human and human-object interactions. We conduct initial annotations of shared task and first-person point-of-view videos. We show that AMRs can be interpreted by a proxy language, such as VoxML, as executable annotation structures in order to recreate and simulate a series of annotated events.

1 Introduction

In recent years, there is an increasing interest in dialogue systems that interact with humans in a natural and sophisticated manner. ChatGPT (OpenAI, 2022) and other large language models (LLMs) show a remarkable ability to generate fluent responses to textual prompts. However, these systems lack two key capabilities which are necessary for naturalistic interaction. First, they lack the ability to communicate in multiple modalities beyond written language, including gesture, gaze, and facial expression; LLMs, even ones like GPT-4 that accept both text and image input (OpenAI, 2023), are limited to text output. Second, these models do not have a notion of the "world" as such. They do not track actions and objects in an environment, and therefore are unable to perform *situated grounding* (Pustejovsky and Krishnaswamy, 2021).

Much work has addressed the importance of non-linguistic modalities in communication (Cassell et al., 2000; Wahlster, 2006; Foster, 2007; Kopp and Wachsmuth, 2010; Marshall and Hornecker, 2013; Schaffer and Reithinger, 2019). For example, in a spoken sentence "I used this for the sketch", the referent of the demonstrative "this" is unspecified. In conjunction with a gesture, e.g., pointing to a pencil, however, reference resolution and disambiguation are possible.

Less attention has been paid to the role of action in dialogue interpretation. Actions significantly contribute to the multimodal context within which linguistic utterances are made, and thus play a crucial role in understanding and interpreting dialogue. In the previous example, lifting the pencil can also direct attention to it, which is then linked to the demonstrative. Additionally, actions can also serve as antecedents to speech in VP ellipsis constructions, (e.g., "What did you do that for?" after someone slams a door), and as action-based bridging relations, where actions create links between concepts in a narrative (e.g., "I went to the store today", followed by taking fruit out of a grocery bag). Actions can even be referenced directly by participants, such as the case of a child relaying "My brother said 'thumbs up'!" when given permission to play with a favorite toy.

A major aspect of dialogue interpretation is the *common ground*— shared knowledge and beliefs that interlocutors possess about each other and the world (Clark and Brennan, 1991; Stalnaker, 2002; Tomasello and Carpenter, 2007). Conversations between agents introduce the problem of identifying and modifying the common ground (Tellex et al., 2020). Actions can update the common ground in ways that speech and gesture cannot, by adding, modifying, and deleting items within it.

We argue that, given the importance of actions to multimodal NLU and their direct influence on the common ground, it is essential to consider how they may be integrated with language and other communicative modalities in a shared annotation scheme.

In this paper, we review existing action annotation schemes, as well as Abstract Meaning Representation (AMR) (Banarescu et al., 2013). We then describe initial efforts to use AMR to anno-

Proceedings of the 4th International Workshop on Designing Meaning Representations, pages 45–51
June, 2023.

tate actions in video data. We explain how action descriptions made with AMR can be translated to the VoxML interpretation language (Pustejovsky and Krishnaswamy, 2016), where they can be executed in a simulated environment, VoxSim (Krishnaswamy and Pustejovsky, 2016), and then close with a discussion of annotation challenges and future work.

2 Background

2.1 Action Annotation

Action recognition in videos is a prominent research area within computer vision, and numerous datasets have been developed providing lexical descriptions of video content, such as Kinetics (Kay et al., 2017) and MSR-VTT (Xu et al., 2016). To facilitate data-driven learning, many of these datasets consist of trimmed clips, categorized with a coarse-grained label describing the action being performed, such as "making pottery" or "bowling".

However, for the purpose of understanding the interplay between action and other communicative acts, we focus on videos that feature discourse between multiple people, and extend over a period of time, thereby allowing for the annotation of fine-grained actions. Although the Charades dataset (Sigurdsson et al., 2016) only involves single individuals, each clip captures a variety of actions through interval-timestamped captions, from which semantic roles can be inferred. The AVA (Gu et al., 2018) and AVA-Kinetics (Li et al., 2020) datasets provide the spatial information of each action associated with multiple people, though their annotations do not adequately assign semantic roles. VidSitu (Sadhu et al., 2021) excels in capturing actions alongside discourse by using movie datasets, introducing semantic role labeling in addition to coreference and event links.

2.2 Abstract Meaning Representation

AMR is a graph-based representation of the meaning of a sentence in terms of its predicate-argument structure (Banarescu et al., 2013). It was designed to be annotatable by humans, and easily parsed by computers. Several extensions have been put forth by the research community (described below), pointing to AMR's utility and expressiveness. For example, the English language sentence "Put that block there.", would be represented in PENMAN (Matthiessen and Bateman, 1991) notation as follows:

```
(p / put-01
  :ARG0 (y / you)
  :ARG1 (b / block
          :mod (t / that))
  :ARG2 (t2 / there)
  :mode imperative)
```

AMR was designed to represent the propositional content of individual written sentences in text. Various extensions to AMR have been proposed which make it more suitable for representing entire documents or dialogues, even using multiple modalities. First, Multi-sentence AMR (MS-AMR) allows AMR to represent meaning beyond the sentence level (O'Gorman et al., 2018). It augments sentence-level AMRs with implicit roles, and marks coreference and bridging relations between entities and events across AMRs.

AMR does not account for a spoken utterance's illocutionary force or effect on the broader dialogue context. Dialogue-AMR (Bonial et al., 2020) extends AMR to include this information in the form of speech act relations, as well as tense and aspect.

Gesture AMR is a further extension of AMR, that goes beyond the linguistic domain, to cover the semantics of gesture (Brutti et al., 2022). Content-bearing gestures are classified according to a taxonomy of gesture acts, and their meaning is annotated similarly to Dialogue-AMR.

Finally, Spatial AMR adds spatial information to AMR, in the form of spatial rolesets, concepts, and frames (Bonn et al., 2020). Of note, Bonn et al. use Spatial AMR to annotate a corpus of Minecraft dialogues, which include both utterances and textual descriptions of actions, such as *[Builder puts down/picks up a red block at X:0 Y:1 Z:0]*.

In addition to wide community adoption, there are several practical reasons for why we propose the annotation of actions with AMR. Every PropBank sense is associated with a single meaning, providing unambiguous interpretations for the labeled actions. PropBank also provides consistent and interpretable argument structures for semantic role labeling. For modeling multimodal dialogue, the efforts described above to capture natural speech and gesture with AMR extensions allow speech and gesture to be seamlessly linked with AMRs of actions using MS-AMR.

3 Approach

To explore the feasibility of applying AMR to actions, we examine two distinct datasets: the Fibonacci Weights Task dataset (Khebour et al., in

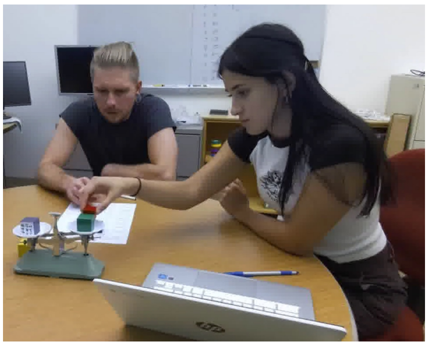

Figure 1: Participant putting a block on a scale.

review), as well as the egocentric Epic Kitchens dataset (Damen et al., 2022). In the examples below, we align observed actions with PropBank senses (Palmer et al., 2003).

3.1 Fibonacci Weights Task

The Weights Task data was designed to elicit teamwork as described in various collaboration frameworks (e.g., PISA (2015); Hesse et al. (2015); Sun et al. (2020)). The task is completed by 2-3 people, and includes blocks, a scale, a worksheet, and a computer with a survey, as seen in Figure 1.

Participants negotiate meaning (and update common ground) via multiple simultaneous modalities. They speak to discuss weights, they gesture to signal the blocks to weigh, and they learn by putting groups of blocks on the scale. The action of putting a block on a scale is annotated as:

```
(p / put-01
  :ARG0 (p1 / participant)
  :ARG1 (b / block)
  :ARG2 (s / scale))
```

Though the actions performed in this dataset are mostly limited to moving and grabbing blocks, they are often prompted by spoken utterances. For instance, an utterance of “let’s try this” followed by the action described by the AMR above is an example of a cataphor, where the word *this* refers to the following action. This phenomenon and others like it can be captured by linking AMR arguments with MS-AMR.

3.2 Epic Kitchens

The Epic Kitchens dataset (Damen et al., 2022) consists of spontaneous first-person recordings of individual participants in kitchens, as in Figure 2. Contrasting with the Weights Task dataset, there is little speech in these videos, but a much wider variety of actions that constantly update the common ground for the viewer. Similar to the description of cooking (text) recipes in Tu et al. (2022), the states of the ingredients and tools are updated by each action. Applying AMR to actions in a scenario like this allows for tracking the progress of the recipe and its components.

An example action annotation for the image in Figure 2 is as follows:

```
(t / transfer-01
   :ARG0 (p / participant)
   :ARG1 (v / vegetables)
   :ARG2 (b / bowl)
   :ARG3 (p1 / pot)
   :instrument (c / chopsticks))
```

The AMR of the action registers the objects from the scene as arguments to the *transfer-01* PropBank predicate. As a direct result of actions like this, the vegetables undergo several transformations during the clip - they are combined, boiled, and eventually eaten. Tracking each entity and the changes they undergo is an interesting issue, motivating the following section.

4 Interpretation

4.1 VoxML as an Interpretation Language

The representation of action with AMR as outlined proves useful in modeling its interactions with speech: both the phenomena of VP ellipsis and anaphoric relations that often occur in spoken language can be resolved with MS-AMR cross-modality coreference chains.

However, AMR alone does not describe how actions affect objects in the common ground, such as their ability to update object locations and cause physical transformations. These changes stem from an associated subevent semantics that can be linked with PropBank predicates. For instance, a human executing PropBank *put-01* would involve a grasping and an ungrasping of a given object, with the end result being the object having moved to a new

Figure 2: Participant transferring vegetables from a pot to a bowl with chopsticks.

$$
\begin{bmatrix}
\textbf{put} \\
\text{LEX} = \begin{bmatrix} \text{PRED} = \textbf{put} \\ \text{TYPE} = \textbf{transition\_event} \end{bmatrix} \\
\text{TYP} = \begin{bmatrix} \text{HEAD} = \textbf{transition} \\ \text{ARGS} = \begin{bmatrix} \text{A}_1 = \textbf{x:agent} \\ \text{A}_2 = \textbf{y:physobj} \\ \text{A}_3 = \textbf{z:location} \end{bmatrix} \\ \text{BODY} = \begin{bmatrix} \text{E}_1 = grasp(x,y) \\ \text{E}_2 = [while((\neg at(y,z) \wedge hold(x,y)), move(x,y)] \\ \text{E}_3 = [at(y,z) \rightarrow ungrasp(x,y)] \end{bmatrix} \end{bmatrix}
\end{bmatrix}
$$

Figure 3: An example VoxML program corresponding to the PropBank predicate *put-01*.

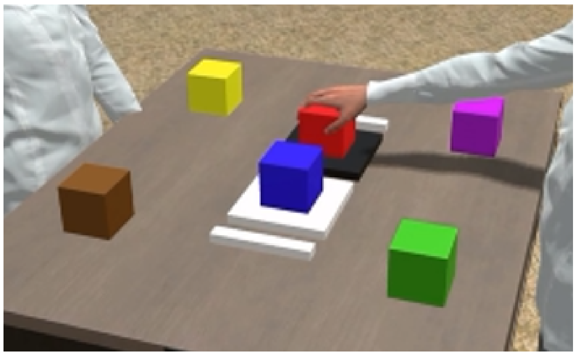

Figure 4: The VoxSim implementation of the Weights Task. At this point in time, two blocks rest on the central scale, one being grasped by a participant.

location. These intermediate subevents are equally valid descriptions of a given action in video, and they can be individually referenced by speech, just as top-level actions can be.

We also note that AMR does not address the *lexical aspect* of its predicates - how they progress over time. To annotate the temporal component of an actions in long videos, we traditionally annotate the timestamps or frame numbers according to when the action begins and ends. However, while some actions suggest a continuous process (e.g., *move*), others are instantaneous results (e.g., *hit*), defined only for a single point in time. We can categorize actions by their lexical aspect in a taxonomy, as either states, atelic (without result) processes, or as telic (with result) achievements and accomplishments (Vendler, 1957).

To encode these semantics, we propose the use of a specification language to enrich these annotations with richer lexical semantics, as provided by Generative Lexicon (Pustejovsky, 2013) and VerbNet (Brown et al., 2022). Such information is encoded directly in VoxML (Pustejovsky and Krishnaswamy, 2016), originally designed as a markup language to describe the semantics of 3D simulations. VoxML consists of a library of concepts called the *voxicon*, where agents and objects are represented in entries called *voxemes*, and action predicates are represented in entries called *programs*. A program outlines a verb's lexical type along with its argument and subevent structure, as shown in Figure 3.

This program is classified as a transition event (telic) as opposed to a state or a process, aligning with the lexical aspect of *put-01*; it continues executing until a specific condition has been met, the result subevent. This characterization is reflected in the program's body, outlining a subevent structure involving grasping and moving the object until the object is finally at location *z*.

Voxemes, on the other hand, encode the affordances of objects given the habitats they reside in (e.g., a cup can only be rolled in a certain orientation), as well as geometric information for spatial reasoning. This specification provides insurance that programs are carried out logically, on the correct arguments in the correct situations.

4.2 AMR to Executable Annotation

The information encoded by VoxML allows it to be modelled in a simulated environment called VoxSim (Krishnaswamy and Pustejovsky, 2016), allowing us to capture and track persistent changes to the common ground. Not only can VoxSim simulate the progression of actions over time, it can also continually track the relations of objects to one another and maintain a history of all events. In our simulation of the Weights Task, displayed in Figure 4, VoxSim maintains the relative locations of each block.

To convert AMR to a format usable by VoxSim, we first require all arguments of AMR annotations to be grounded with specific entities labeled in the world. This can be done by linking every entity node to a string representing the object it refers to in the video. We then find the VoxML program entry that corresponds with the AMR's PropBank predicate, aligning its arguments semantically with that predicate's arguments. A concise executable annotation structure like the following example can then be constructed, where `GreenBlock` and `Table` are proper names assigned to entities in the video:

```
put(GreenBlock,on(Table))
```

Through VoxML, this string can be interpreted as an instruction to execute at a specific timestep defined in the annotation.

5 Discussion

We have described an initial exploration of action annotation within the context of communicative acts in dialogue. By investigating the application of AMR and VoxML, we aim for adequate representations to model the interactions between them, as well as define simulations that can track the evolving common ground. This analysis has highlighted certain challenges associated with annotation and possible directions for future work in designing representations.

5.1 Annotation Challenges

We have discussed how high-level actions can be further broken down into subevents, and how their lexical aspect must be respected. This poses multiple questions for annotation in practice.

The first issue is granularity. As illustrated in Figure 3, a *putting* action can be further broken down into its subevent structure, minimally involving a grabbing motion and a holding period. Other actions, like cutting vegetables, consist of a series of instantaneous slicing events. Other events can be easily annotated but may not considerably affect the state of the world, such as someone blinking.

There are multiple ways to describe a set of actions, and this introduces ambiguity to the annotation problem. To ensure consistency, an annotation environment with multiple annotators should agree on a restricted set of atomic predicates to use, with well-defined descriptions of what events constitute each action instance.

The second issue is temporal. As mentioned in our discussion of lexical aspect, different actions require different descriptions of how they progress through time. While processes and accomplishments are defined by an interval of time, achievements are only defined by a single point. Additionally, in contrast with speech, individuals often perform multiple actions simultaneously, such as when they multitask with both hands. This implies multiple overlapping intervals.

Annotation software like ELAN (Brugman and Russel, 2004) can handle simultaneity by placing intervals on multiple tracks. However, interval annotations alone cannot capture instantaneous events, which must either be omitted, or always placed in the context of an accomplishment event.

5.2 Automation of Action Annotation

Though action annotation is a straightforward process given a well-defined set of predicates, manual AMR annotation is more time-consuming. One approach to the automatic annotation of action AMRs involves first identifying actions in videos, then generating AMRs for those actions. Yang et al. (2022) used the VidSitu dataset (Sadhu et al., 2021) to train models to both identify the verbs in the video and fill in their semantic roles. Given a verb and its arguments, the conversion to AMR is straightforward.

Another possible approach is to generate natural language captions for events in the videos, then parse those captions into AMRs. For example, Xu et al. (2023) developed a modular multimodal model that represents the current state-of-the-art on video captioning on the MSR-VTT dataset (Xu et al., 2016). We can then leverage AMR parsers such as Structured mBART with Maximum Bayes Smatch Ensemble distillation (Lee et al., 2022) to convert those captions to the graph-based structure.

6 Conclusion

In this paper, we argue that representing actions is essential for the proper interpretation of situated dialogues. We describe how AMR can be used to annotate actions in different types of video interactions, and describe the challenges associated with this task. We also show how AMRs can be translated to the VoxML specification language to encode semantic information, allowing for the ability to track changes to the common ground in a simulation environment like VoxSim. In future work, we plan to further develop our annotation methodology, and apply it on a larger scale.

Acknowledgments

This research was supported by the NSF National AI Institute for Student-AI Teaming (iSAT) under grant DRL 2019805. The opinions expressed are those of the authors and do not represent views of the NSF.

References

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, Sofia, Bulgaria. Association for Computational Linguistics.

Claire Bonial, Lucia Donatelli, Mitchell Abrams, Stephanie M. Lukin, Stephen Tratz, Matthew Marge, Ron Artstein, David Traum, and Clare Voss. 2020. Dialogue-AMR: Abstract Meaning Representation for dialogue. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 684–695, Marseille, France. European Language Resources Association.

Julia Bonn, Martha Palmer, Zheng Cai, and Kristin Wright-Bettner. 2020. Spatial AMR: Expanded spatial annotation in the context of a grounded Minecraft corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4883–4892, Marseille, France. European Language Resources Association.

Susan Windisch Brown, Julia Bonn, Ghazaleh Kazeminejad, Annie Zaenen, James Pustejovsky, and Martha Palmer. 2022. Semantic representations for nlp using verbnet and the generative lexicon. *Frontiers in artificial intelligence*, 5.

Hennie Brugman and Albert Russel. 2004. Annotating multi-media/multi-modal resources with ELAN. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Richard Brutti, Lucia Donatelli, Kenneth Lai, and James Pustejovsky. 2022. Abstract Meaning Representation for gesture. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1576–1583, Marseille, France. European Language Resources Association.

Justine Cassell, Joseph Sullivan, Elizabeth Churchill, and Scott Prevost. 2000. *Embodied conversational agents*. MIT Press.

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

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. 2022. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 130:33–55.

Mary Ellen Foster. 2007. Enhancing human-computer interaction with embodied conversational agents. In *International Conference on Universal Access in Human-Computer Interaction*, pages 828–837. Springer.

Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. 2018. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6047–6056.

Friedrich Hesse, Esther Care, Juergen Buder, Kai Sassenberg, and Patrick Griffin. 2015. A framework for teachable collaborative problem solving skills. In *Assessment and teaching of 21st century skills*, pages 37–56. Springer.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Ibrahim Khebour, Richard Brutti, Indrani Dey, Rachel Dickler, Kelsey Sikes, Kenneth Lai, Mariah Bradford, Brittany Cates, Paige Hansen, Changsoo Jung, Brett Wisniewski, C Terpstra, Leanne Hirshfield, Sadhana Puntambekar, Nathaniel Blanchard, James Pustejovsky, and Nikhil Krishnaswamy. in review. When text and speech are not enough: Modeling meaning in situated collaborative tasks.

Stefan Kopp and Ipke Wachsmuth. 2010. *Gesture in embodied communication and human-computer interaction*, volume 5934. Springer.

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.

Young-Suk Lee, Ramón Astudillo, Hoang Thanh Lam, Tahira Naseem, Radu Florian, and Salim Roukos. 2022. Maximum Bayes Smatch ensemble distillation for AMR parsing. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5379–5392, Seattle, United States. Association for Computational Linguistics.

Ang Li, Meghana Thotakuri, David A Ross, João Carreira, Alexander Vostrikov, and Andrew Zisserman. 2020. The ava-kinetics localized human actions video dataset. *arXiv preprint arXiv:2005.00214*.

Paul Marshall and Eva Hornecker. 2013. Theories of embodiment in HCI. *The SAGE Handbook of Digital Technology Research*, 1:144–158.

Christian Matthiessen and John A. Bateman. 1991. *Text generation and systemic-functional linguistics: experiences from English and Japanese*. Burns & Oates.

OpenAI. 2022. Introducing ChatGPT. `https://openai.com/blog/chatgpt`.

OpenAI. 2023. GPT-4 technical report. arXiv.

Tim O'Gorman, Michael Regan, Kira Griffitt, Ulf Hermjakob, Kevin Knight, and Martha Palmer. 2018. Amr beyond the sentence: the multi-sentence amr corpus. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3693–3702.

Martha Palmer, Dan Gildea, and Paul Kingsbury. 2003. The proposition bank: an annotated corpus of semantic roles. *Computational Linguistics*.

OECD PISA. 2015. Assessment and analytical framework: Science. *Reading, Mathematics and Financial Literacy, PISA*.

James Pustejovsky. 2013. Dynamic event structure and habitat theory. In *Proceedings of the 6th International Conference on Generative Approaches to the Lexicon (GL2013)*, pages 1–10.

James Pustejovsky and Nikhil Krishnaswamy. 2016. Voxml: A visualization modeling language. *arXiv preprint arXiv:1610.01508*.

James Pustejovsky and Nikhil Krishnaswamy. 2021. Embodied human computer interaction. *KI-Künstliche Intelligenz*, 35(3):307–327.

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.

Stefan Schaffer and Norbert Reithinger. 2019. Conversation is multimodal: thus conversational user interfaces should be as well. In *Proceedings of the 1st International Conference on Conversational User Interfaces*, pages 1–3.

Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 510–526. Springer.

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

Chen Sun, Valerie J Shute, Angela Stewart, Jade Yonehiro, Nicholas Duran, and Sidney D'Mello. 2020. Towards a generalized competency model of collaborative problem solving. *Computers & Education*, 143:103672.

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.

Michael Tomasello and Malinda Carpenter. 2007. Shared intentionality. *Developmental Science*, 10(1):121–125.

Jingxuan Tu, Eben Holderness, Marco Maru, Simone Conia, Kyeongmin Rim, Kelley Lynch, Richard Brutti, Roberto Navigli, and James Pustejovsky. 2022. Semeval-2022 task 9: R2vq–competence-based multimodal question answering. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1244–1255.

Zeno Vendler. 1957. Verbs and times. *The Philosophical Review*, pages 143–160.

Wolfgang Wahlster. 2006. Dialogue systems go multimodal: The SmartKom experience. In *SmartKom: foundations of multimodal dialogue systems*, pages 3–27. Springer.

Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, Guohai Xu, Ji Zhang, Songfang Huang, Fei Huang, and Jingren Zhou. 2023. mPLUG-2: A modularized multi-modal foundation model across text, image and video. arXiv.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296.

Guang Yang, Manling Li, Jiajie Zhang, Xudong Lin, Shih-Fu Chang, and Heng Ji. 2022. Video event extraction via tracking visual states of arguments.