Situated UMR for Multimodal Interactions

Kenneth Lai[1], Richard Brutti[1], Lucia Donatelli[2], James Pustejovsky[1]
[1]Department of Computer Science
Brandeis University, USA
[2]Department of Language Science and Technology
Saarland University, Germany
{klai12,brutti,jamesp}@brandeis.edu
donatelli@coli.uni-saarland.de

Abstract

We discuss the requirements on a meaning representation language for annotating both verbal and non-verbal communicative acts in multimodal interactions, as it impacts both the deployment of annotation efforts and the development of corpora reflecting these phenomena. We argue that Uniform Meaning Representation (UMR) can be naturally extended to capture multiple communicative channels in discourse, including both language and gesture, while also encoding the annotation of referential grounding in situated contexts.

1 Introduction and Related Work

As multimodal interactive systems become more common and sophisticated, there are increasing expectations that they will approximate interactions with other humans. Human-computer interaction (HCI) and human-robot interaction (HRI) involve communicating intentions, goals, and attitudes through multiple modalities beyond language, including gesture, gaze, and situational awareness. With this interest comes a need for capturing and representing the data that encodes these different modalities during such interactions.

Any representation suitable to this task should, at a minimum, both accommodate the structure and content of the different modalities, as well as facilitate alignment and binding across them. However, it is also important to 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.

Presently, there are few meaning representation languages for situated (dialogue) interactions, that are both adequately expressive of the content and compact enough for corpus development. There have been several annotation efforts utilizing Abstract Meaning Representation (AMR) (Banarescu et al., 2013). Advantages of AMR include its relative simplicity, ease of annotation, and available corpora. AMR has been expanded and applied to multi-sentence settings (O'Gorman et al., 2018), and to task-oriented dialogues (Bonial et al., 2020).

More recently an extension of AMR, Uniform Meaning Representation (UMR) has been developed to be scalable, accomodate cross-linguistic diversity, and support lexical and logical inference (Van Gysel et al., 2021). To this end, UMR incorporates aspect, scope, temporal and modal dependencies, as well as inter-sentential coreference.

We argue that we can combine multimodal elements in a single representation for alignment and grounded meaning. Specifically, we believe that an enriched version of UMR, which we call *Situated UMR (SUMR)*, is an ideal representational format to this end. This allows for an immediate referencing for deictic, pronominal, and underspecified expressions, as well as a spatial "registration" for objects in the discourse (and common ground).

A variety of corpora exist that seek to capture language and dialogue in a situated environment. However, existing annotation schemes often fall short of capturing true situated *meaning*, instead annotating distinct channels separately with little guidance as to how these channels interact to create emergent meaning (Krishnaswamy and Pustejovsky, 2019). The SCOUT corpus is an example of a situated, but *unimodal*, dataset (Bonial et al., 2020). It introduces Dialogue-AMR to extend and enrich AMR in support of HRI, in a navigation setting. The EGGNOG corpus (Wang et al., 2017) is comprised of video of two participants working on shared tasks. It is annotated with the annotator-inferred *intent* of the gestures, as well as their morphology

Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue, September 20–22, 2021, Potsdam / The Internet.

(physical description). However, the annotation is a label with no grounding. Objects introduced in one intent are not available in the next; objects are *identified* but not referenced or registered.

2 Common Ground in Situated UMR

For the present discussion, we focus on the semantics of integrated multimodal expressions in the context of task-oriented dialogues. We assume the model presented in Pustejovsky and Krishnaswamy (2021) and Krishnaswamy and Pustejovsky (2021), where a *common ground structure* (CGS) integrates both intermodal expressions in the discourse and the situational anchoring to objects perceived and referenced in the context. An agent's communicative act, C_a, is a tuple of expressions from the diverse modalities involved (e.g., speech S, gesture G). The CGS embeds C_a within a monad identifying: **A** the communicating agents; **B**, the salient shared belief space; **P**, the objects and relations that are jointly perceived in the environment; and $\mathcal{E}$, the agents' joint embedding space. Here we focus on a communication in relation to the perceived context.

Consider Figure 1, where a multimodal command aligns the linguistic utterance, "*That move there*" with an ACTION-RESULT gesture sequence of "*Point Action Point*". This example illustrates two kinds of gestures: (a) establishing a reference; and (b) depicting an action-object pair (Kendon, 2004; Lascarides and Stone, 2009).[1]

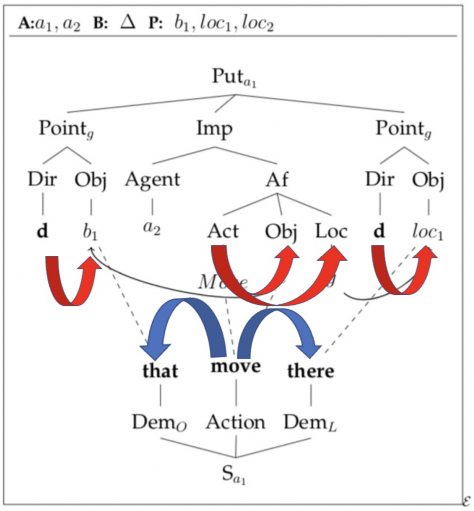

Figure 1: Intermodal alignment between linguistic and gesture dependency structures

Given these assumptions, we introduce a multimodal extension of UMR we call *Situated UMR (SUMR)*, that allows for the representation of both multiple channels of communication, as well as the perceptual (object and situational) awareness present to an agent in the common ground.

```
(a) (c / cgs
      :agent (a / agent)
      :agent (a2 / agent)
      :perception (b / block)
      :perception (l / location)
      :perception (l2 / location))
(b) (s1c2 / command-00
      :ARG0 a1
      :ARG1 (c3 / communicative-act
             :gesture (g / gesture-unit
                        :op1 (d / deixis
                              :DIR (v / vector)
                              :OBJ b)
                        :op2 (a3 / action
                              :ACT (m / move-01)
                              :OBJ (i / implicit-role
                                    :op1 "moved")
                              :LOC (i2 / implicit-role
                                    :op1 "destination"))
                        :op3 (d2 / deixis
                              :DIR (v2 / vector))
                              :OBJ l))
             :speech (m2 / move-01
                       :mode imperative
                       :ARG0 (i3 / implicit-role
                              :op1 "mover")
                       :ARG1 (t / that)
                       :ARG2 (t2 / there))
      :ARG2 a2)
(c) (s1 / sentence
      :coref ((a2 :same-entity i3)
              (b :same-entity i)
              (b :same-entity t)
              (l :same-entity i2)
              (l :same-entity t2)))
```

Figure 2: Example SUMR corresponding to the communicative act in Figure 1

The example SUMR in Figure 2 has three parts. First, in (a), the agents and perceived objects are listed in the CGS (in the example, **B** and $\mathcal{E}$ are omitted for brevity, but can be included). For each communicative act, we have a sentence-level UMR representation (b), with the gesture and speech modalities labeled. We assume the dialogue act annotation from Bonial et al. (2020). As with sentences in text and discourse, gestural expressions can also be sequenced; and, as in multi-sentence AMR, their corresponding individual representations can capture the object coreference inherent in the discourse (O'Gorman et al., 2018). This is captured in the document-level representation (c). Details of the alignment of the speech and gesture expressions are beyond the scope of this poster.

As a platform for multimodal situated dialogue annotation, we believe that SUMR has some attractive properties. It is adequately expressive at both utterance and dialogue levels, while easily accommodating the dependency structures inherent in gestural expressions. Further, the native reentrancy facilitates both the linking between modalities and situational grounding to contextual bindings.

[1] We adopt the gesture grammar developed in (Pustejovsky and Krishnaswamy, 2021).

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.

Claire Bonial, Lucia Donatelli, Mitchell Abrams, Stephanie 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.

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

Nikhil Krishnaswamy and James Pustejovsky. 2019. Generating a novel dataset of multimodal referring expressions. In *Proceedings of the 13th International Conference on Computational Semantics-Short Papers*, pages 44–51.

Nikhil Krishnaswamy and James Pustejovsky. 2021. The role of embodiment and simulation in evaluating hci: Experiments and evaluation. In *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management. Human Body, Motion and Behavior*, pages 220–232, Cham. Springer International Publishing.

Alex Lascarides and Matthew Stone. 2009. A formal semantic analysis of gesture. *Journal of Semantics*, page ffp004.

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, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

J Pustejovsky and N Krishnaswamy. 2021. Embodied human computer interaction. *Künstliche Intelligenz*.

Jens E. L. Van Gysel, Meagan Vigus, Jayeol Chun, Kenneth Lai, Sarah Moeller, Jiarui Yao, Tim O'Gorman, Andrew Cowell, William Croft, Chu-Ren Huang, Jan Hajič, James H. Martin, Stephan Oepen, Martha Palmer, James Pustejovsky, Rosa Vallejos, and Nianwen Xue. 2021. Designing a uniform meaning representation for natural language processing. *KI-Künstliche Intelligenz*, pages 1–18.

Isaac Wang, Pradyumna Narayana, Dhruva Patil, Gururaj Mulay, Rahul Bangar, Bruce Draper, Ross Beveridge, and Jaime Ruiz. 2017. EGGNOG: A continuous, multi-modal data set of naturally occurring gestures with ground truth labels. In *To appear in the Proceedings of the 12th IEEE International Conference on Automatic Face & Gesture Recognition*.