Situated Utterances and Discourse Relations

Matthew Stone[†§], Una Stojnic[‡§] and Ernest Lepore[‡§*]
[†]Computer Science, [‡]Philosophy, [§]Cognitive Science
Rutgers University
matthew.stone@rutgers.edu, ustojnic@eden.rutgers.edu, lepore@ruccs.rutgers.edu

Abstract

Utterances in situated activity are about the world. Theories and systems normally capture this by assuming references must be resolved to real-world entities in utterance understanding. We describe a number of puzzles and problems for this approach, and propose an alternative semantic representation using discourse relations that link utterances to the nonlinguistic context to capture the context-dependent interpretation of situated utterances. Our approach promises better empirical coverage and more straightforward system building. Substantiating these advantages is work in progress.

1 Introduction

People exhibit sophisticated strategies for using language to coordinate their ongoing practical activities (Clark and Krych, 2004). Applications such as human–robot interaction must also also support these strategies (Kruijff et al., 2012). A key question is how to track speakers' reference to the environment. Julia Child's (1), from the "omelette" episode of *The French Chef*, illustrates the issues:

(1) There's your omelette, forming itself in the bottom of the pan.

The accompanying video is a close-up of a stovetop from above, with the omelette Child is cooking front and center. To understand what Child is doing, we need to track the obvious connection between her utterance and *that omelette* in *that pan*.

In this paper, we explore a novel and surprising proposal about how do this: *discourse relations*. Discourse relations are kinds of speech acts, which can connect an utterance to ongoing conversation, establish the coherence of what is said, and allow inference to implicatures and other aspects of the speaker's mental state (Asher and Lascarides, 2003, 2013). Child's (1) we suggest, gets its coherence, in part, as a *report of what's visible in that situation*. Child's reference to the omelette and the pan is a consequence of this interpretation, and so need not and should not be separately represented.

Prior work focuses on symbols that refer directly to the world and placeholders that abstract reference; see Section 2. We introduce our alternative and contrast it with these approaches in Section 3. On the theoretical side, we argue in Section 4, our approach captures meaning more precisely and provides a better account of what's needed to understand and disambiguate situated utterances. On the practical side, we argue in Section 5, our approach provides a more tractable interface between linguistic processing, situated perception, and deliberation.

2 Background

Our work arises out of an interest in combining grounded representations of meaning with coherence approaches to discourse. To appreciate the issues, consider a classic case (Bolt, 1980). A user utters (2), while pointing first at an object in the environment and then at the place it should go.

(2) Put that there.

*This paper is based on work supported by NSF IIS-1017811. We acknowledge the helpful anonymous reviews.

Kaplan (1989) urges us to treat demonstratives in cases such as (2) as *directly referential*. In using them, the speaker gives information about *those very things*. Computationally, this means that the objects of demonstrations should be represented using symbols that are locked onto their referents system-wide. In multimodal interfaces, like Bolt's (1980), the targets are graphical objects and other locations on a computer display, for which appropriate internal representations are readily available. For robotics, the natural strategy is to represent referents using perceptually-grounded symbols anchored in the real world, as do Yu and Ballard (2004) for example.

Speakers *use* many utterances referentially, even if the utterances don't have referential semantics (Kripke, 1977). Systems need to recognize and represent the speaker's referential communicative intentions in those cases too (Allen and Perrault, 1980). Bolt's system, for example, responded to definite descriptions, as in (3), the same way as it did to demonstratives: it moved *those things*.

(3) Put the cruise ship north of the Dominican Republic.

When it's problematic to use grounded symbols, we can use *discourse referents* alongside them. Formally, a discourse referent is just a free variable, but it can be associated with *anchoring constraints* that describe how it is supposed to be linked up with the world (Zeevat, 1999). In practice, anchoring involves representing interpretation in two separate tiers (Luperfoy, 1992). Meaning is represented via variables, the world is represented via suitably grounded symbols, and an evolving assignment of values to variables embodies the system's understanding of the real-world reference of expressions in discourse.

Reconciling grounded reference and discourse anaphora is the job of discourse semantics. We can see this in the planning dialogue of (4), for example.

(4) a. A: Let's put the cruise ship south of Florida.
b. B: That won't fit there.

We need to represent A's utterance as a proposal—a specific kind of problem-solving move that advocates a particular course of action and commits the speaker to following through on it if others concur. The move focuses attention on the entities involved in carrying it out: here, the ship, Florida and the region to its south. Meanwhile, B's response in (4b) is a rejection—a move that offers a negative assessment of the current proposal and commits the speaker against adopting it. Note that B's references with *that* and *there* succeed even if B produces the utterance without any accompanying gesture or demonstration, because the referent has been activated by an earlier mention in a related utterance (Gundel et al., 1993).

The simplest approach to discourse organization is to represent the state of the discourse with an information state (Poesio and Traum, 1997) and associate each move with an appropriate update. For example, we can model utterances such as (2) and (3) as making moves that contribute step-by-step to broader problem-solving activity; see Lochbaum (1998) or Blaylock (2005). Normally, the information state specifies once and for all how each thread of ongoing activity places relevant real-world entities at the center of attention (Grosz and Sidner, 1986; Poesio and Traum, 1997).

We advocate a different approach, based on *discourse coherence* (Kehler, 2002; Asher and Lascarides, 2003). The idea is that discourse is fundamentally composed of *relational* contributions, which establish connections that link each utterance by inference to segments of the preceding conversation. The interpretation of an utterance therefore implicitly refers to the interpretation of some prior discourse and comments on it. On coherence approaches, how an utterance *attaches* to the discourse determines what entities are prominent in interpreting it (Hobbs, 1979). Coherence theory does *not* naturally characterize discourse in terms of state-by-state updates to an overarching model of information and attention. Kehler and colleagues' (5) illustrates what's at stake (Kehler et al., 2008).

(5) Phil tickled Stanley, and Liz poked him.

When we understand the second clause of (5) as a description of a parallel to the first, we prefer to resolve *him* to Stanley. When we understand it to describe its results, we prefer to resolve *him* to Phil. For coherence theory, the two interpretations of the second clause *relate* it to the first, and the relation is what suggests prominent resolutions for its references. The relation in turn structures the discourse into higher level units that shape possibilities for attaching subsequent utterances.

3 Our Proposal

Coherence theories start from the observation that understanding utterances involves recognizing the implicit relationships that connect ideas together. The same is true, we argue, for utterances like (1). Child is not just giving the next step in making an omelette, or giving her audience new information about the principles of cooking. She's describing what's happening on the screen, in terms she expects her audience to confirm for themselves by examining what they see. An interpreter who doesn't recognize this about (1) has failed to understand it.

To sketch the key formal ingredients of our account, we use a simple dynamic semantics (Muskens, 1996) and an expressive ontology of situations and eventualities (Hobbs, 1985; Kratzer, 2002). Dynamic semantics represents meanings as sequences of updates $[\mathbf{v}|\varphi]$ that introduce discourse referents $\mathbf{v}$ and characterize them with conditions φ. We add an update $\langle \pi xc \rangle$ introducing a discourse referent x perceptually grounded in c, and an update $\langle \sigma xs \rangle$ introducing x as the central entity in grounded situation s (provided s does uniquely distinguish one). Where necessary, ∂K marks update K as presupposed.

Situations are parts of the world, capturing particular states of particular objects, perhaps as located in particular spatial regions and changing over particular temporal intervals. Propositions are true in situations; this is important for perceptual, causal and default reasoning. Eventualities, including Davidsonian events, turn out to be situations that exemplify propositions in characteristic ways. We capture discourse coherence by specifying relations among situations; these can be discourse referents for situations introduced by utterances or grounded references to parts of the speech situation.[1]

We capture the interpretation of (2) by specifying the dynamics of discourse referents and their grounded interpretations as in (6), using c_1 for *that* and c_2 for *there*.

(6) $\langle \pi x c_1 \rangle; \langle \pi y c_2 \rangle; [e|command(e), put(e,x,y)]$

Things get more interesting when we factor in coherence. We formalize (1) as in (7).

(7) $[e|Summary(s_0,e)]; \partial[o,p|omelette(o), pan(p)]; [|forming\text{-}self\text{-}in(e,o,p)]$

The discourse relation $Summary(s_0,e)$ captures the interpretive connection between the utterance describing e and what's happening simultaneously on the screen in situation s_0. Like all coherence relations, *Summary* reflects semantic and pragmatic constraints. Semantically, e must be part of s_0. Following Kratzer (2002), this entails that the information describing e is true in s_0. Pragmatically, $Summary(s_0,e)$ holds only if the information describing e provides a good answer about "what's happening" in s_0. A summary appeals to broad, basic categories to provide essential information. We have in mind something like the "vital nuggets of information" needed to answer definition questions (Voorhees, 2003).

For "That's an omelette" we offer (8), which defines the central entity in situation s_0 as an omelette:

(8) $[e|Summary(s_0,e)]; \langle \sigma o s_0 \rangle; [|omelette(e,o)]$

The update $\langle \sigma o s_0 \rangle$ formalizes how the *discourse relation* makes entities prominent for reference, as we observed in (5). Such updates can capture the interpretation of demonstratives when there's no explicit pointing or demonstration in the utterance.

Not all situated utterances offer a *Summary* of an unfolding situation. For example, utterances can offer *Assessments* that invite the audience not to define what's happening but to appraise it. Take "Yummy!" In commenting on the food this way, the speaker expects the audience to join in her appreciation. A formal characterization of *Assessment* would appeal to the semantics of predicates of personal taste and the distinctive pragmatic functions of such judgments, perhaps following Crespo and Fernández (2011). And speakers can also link up questions and instructions to ongoing activity by suitable relations.

Summary and *Assessment* could also be used to formalize the interpretation of successive utterances by relating two described situations. In fact, utterances can relate both to ongoing activity and to previous

[1] Thus we use situations to capture discourse meaning, *not* to formalize events of speech or the common ground as in Poesio and Traum (1997). An alternative approach would follow Zeevat (1999) and Asher and Lascarides (2003) and use labels for DRSs to capture perceptual and discourse content in discourse relations.

discourse. For example, consider (9) and (10), taken from Vi Hart's origami proof of the Pythagorean theorem—a visual narrative much like Child's where utterances describe ongoing events on the screen.[2]

(9) We're just taking advantage of the symmetries of the square for the next step.

(10) This is where you're choosing how long and pointy or short and fat the right triangle is.

Hart uses (9) while folding a square into eight identical segments to explain how to do the folds. Hart uses (10) as she describes the next step of folding, to highlight its result for the proof. Thus, these utterances are linked to the accompanying activity but do not just report what's going on; and they're linked to the ongoing discourse as well. In fact, coherence theory already allows that utterances can bear multiple connections to prior discourse (Asher and Lascarides, 2003). The closest parallel may be that of multimodal communication, where Lascarides and Stone (2009) argue that utterances bear discourse relations both to prior utterances and to simultaneous gesture.

4 Empirical Adequacy

Combining grounded representations with discourse relations, specifically as in (8), makes it possible to give a better characterization of the logical form of demonstrative utterances in otherwise problematic cases. In particular, it captures how speakers and interpreters can rely on the world to disambiguate what they say and to understand one another.

Here's a telling case. It's the beginning of spring, 2012, and Jupiter and Venus are shining brightly very close together—just a few degrees apart—in the evening sky. The speaker has deployed a telescope facing a window over the western sky. When a visitor arrives, the speaker adjusts the telescope, then says, without any further demonstration, either (11) or (12).

(11) That's Jupiter. You can even see four moons.

(12) That's Venus. You can see the crescent.

We (and our informants) find these utterances unproblematic. But the coherence theory is required to get their interpretations right. These are comments on what's visible through the telescope. You can't see four of Jupiter's moons or the crescent of Venus with the naked eye and the speaker isn't suggesting otherwise. Moreover, the coherence relation is what's making it possible for the speaker to refer to Jupiter or Venus as *that*. To comment on the view through the telescope is to evoke whatever entity is centrally imaged in the telescope as a prominent candidate for reference. And nothing else will do. Given the astronomical conjunction, the speaker couldn't distinguish Jupiter from Venus by pointing, nor could the visitor judge which body the telescope was pointed at by the direction of the tube. Letting s_1 name the view through the telescope, we can formalize the key bits of interpretation:

(13) $[e|Summary(s_1,e)];\langle \sigma x s_1\rangle;[|jupiter(e,x)]$

(14) $[e|Summary(s_1,e)];\langle \sigma x s_1\rangle;[|venus(e,x)]$

The representations get the meaning right. More importantly, they explain how the visitor can recover the logical form and understand the speaker's point by recognizing the relationship that makes the speaker's utterance coherent, even though the visitor can't identify which specific body the speaker is referring to until the visitor looks through the telescope for herself. By contrast, if all you had was representations like (6), grounded representations of deixis that made reference explicit, you'd incorrectly predict that there's an ambiguity to resolve in (11) and (12) even *after* you understand them as comments about the view through the scope. You'd have two grounded symbols for bright objects in the western sky, and you'd have to *pick* one as the referent of the speaker's demonstration—or ask for clarification. We take this as strong evidence against the idea that speakers and hearers must coordinate directly on demonstrative referents, a common view in both formal and computational semantics (Neale, 2004; Stone, 2004).

[2] `http://www.youtube.com/watch?v=z61L83wl31E`

5 Cognitive Architecture

Systems with grounded language interpretation have to integrate language and perception. Real robotic understanding systems use a complicated inference process to resolve reference in situated utterances, in order to figure out what thing in the world is a reference of a demonstration, for example (Kruijff et al., 2012). It's not just a matter of selecting the right referent from a set of salient candidates based on linguistic constraints. Like people, robots have cameras with limited fields of view that must be pointed at what they see. So when you do a demonstration for a robot, it has to track the pointing movement to find the possible targets, much as a person would. There is substantial problem solving involved.

The same is true of many other grounded inference tasks involving domain reasoning. We may need the results to calculate the implications of what we hear, or even to select the most likely interpretation. But it's normally prohibitive to try to interleave that problem solving with fine-grained disambiguation, because it requires systems to solve these hard reasoning problems not just for the intended interpretation but for any candidate interpretation it considers.

Representing context dependence via coherence provides an attractive framework to divide interpretation into stages and minimize the problem solving that's necessary to compute logical form. Take (11) and its representation in (13). Here is a formalism that captures the meaning of the utterance while spelling out the further work that will be required to resolve reference. According to (13), when you look through the telescope, you'll find out what the referent of *that* is. Recognizing the logical form of the utterance this way should suffice for understanding. We expect that the discourse relation could be resolved based on shallow constraints on what information counts as a summary. And the system only needs to be tracking the ongoing activity enough to link utterances to relevant situations. It can ground its provisional understanding in perception and action as needed separately.

6 Conclusion and Future Work

We have considered grounded interpretations in coherent discourse, and argued that referential interpretations in cases like (1), (2) and (11) are understood and derived relationally. This requires representations of interpretation that explicitly link discourse entities with grounded symbols and track the heterogeneous prominence that these entities get in virtue of the diverse relationships that utterances can bear to ongoing activity. In brief: we relate our talk to the world around us through suitable discourse relations.

Our approach commits us to representing utterances with specific kinds of interpretive connections to the world. Our characterization of these connections is obviously provisional, and corpus and modeling work is necessary to flesh out the parameters of the approach. We have also suggested that these representations will be useful in designing systems that communicate about the environment where they can perceive and act. Experiments with prototypes are clearly necessary to substantiate this claim.

References

Allen, J. F. and C. R. Perrault (1980). Analyzing intention in utterances. *AIJ 15*(3), 143–178.

Asher, N. and A. Lascarides (2003). *Logics of Conversation*. Cambridge: Cambridge University Press.

Asher, N. and A. Lascarides (2013). Strategic conversation. *Semantics and Pragmatics*.

Blaylock, N. J. (2005). *Towards Tractable Agent-Based Dialogue*. Ph.D. dissertation, Rochester.

Bolt, R. (1980). Put-that-there: Voice and gesture at the graphics interface. *Computer Graphics 14*(3), 262–270.

Clark, H. H. and M. A. Krych (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language 50*, 62–81.

Crespo, I. and R. Fernández (2011). Expressing taste in dialogue. In *SEMDIAL 2011: Proceedings of the 15th Workshop on the Seantics and Pragmatics of Dialogue*, pp. 84–93.

Grosz, B. J. and C. L. Sidner (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics 12*(3), 175–204.

Gundel, J. K., N. Hedberg, and R. Zacharski (1993). Cognitive status and the form of referring expressions in discourse. *Language 69*(2), 274–307.

Hobbs, J. R. (1979). Coherence and coreference. *Cognitive Science 3*(1), 67–90.

Hobbs, J. R. (1985). Ontological promiscuity. In *Proceedings of ACL*, pp. 61–69.

Kaplan, D. (1989). Demonstratives. In J. Almog, J. Perry, and H. Wettstein (Eds.), *Themes from Kaplan*, pp. 481–563. Oxford: Oxford University Press.

Kehler, A. (2002). *Coherence, Reference and the Theory of Grammar*. Stanford: CSLI.

Kehler, A., L. Kertz, H. Rohde, and J. L. Elman (2008). Coherence and coreference revisited. *Journal of Semantics 25*(1), 1–44.

Kratzer, A. (2002). Facts: Particulars or information units? *Linguistics & Philosophy 25*(5–6), 655–670.

Kripke, S. (1977). Speaker's reference and semantic reference. In P. A. French, T. Uehling, Jr., and H. K. Wettstein (Eds.), *Midwest Studies in Philosophy, Volume II*, pp. 255–276. Minneapolis: University of Minnesota Press.

Kruijff, G.-J., M. Janicek, and H. Zender (2012). Situated communication for joint activity in human-robot teams. *IEEE Intelligent Systems 27*(2), 27–35.

Lascarides, A. and M. Stone (2009). Discourse coherence and gesture interpretation. *Gesture 9*(2), 147–180.

Lochbaum, K. E. (1998). A collaborative planning model of intentional structure. *Computational Linguistics 24*(4), 525–572.

Luperfoy, S. (1992). The representation of multimodal user interface dialogues using discourse pegs. In *Proceedings of ACL*, pp. 22–31.

Muskens, R. (1996). Combining Montague semantics and discourse representation. *Linguistics & Philosophy 19*(2), 143–186.

Neale, S. (2004). This, that, and the other. In A. Bezuidenhout and M. Reimer (Eds.), *Descriptions and Beyond*, pp. 68–181. Oxford: Oxford University Press.

Poesio, M. and D. R. Traum (1997). Conversational actions and discourse situations. *Computational Intelligence 13*(3), 309–347.

Stone, M. (2004). Intention, interpretation and the computational structure of language. *Cognitive Science 28*(5), 781–809.

Voorhees, E. M. (2003). Evaluating answers to definition questions. In *Companion Volume of the Proceedings of HLT-NAACL – Short Papers*, pp. 109–111.

Yu, C. and D. H. Ballard (2004). On the integration of grounding language and learning objects. In *Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI)*, pp. 488–494.

Zeevat, H. (1999). Demonstratives in discourse. *Journal of Semantics 16*(4), 279–313.