What A Situated Language-Using Agent Must be Able to Do: A Top-Down Analysis

David Schlangen

Computational Linguistics / Department of Linguistics
University of Potsdam, Germany
david.schlangen@uni-potsdam.de

Abstract

Even in our increasingly text-intensive times, the primary site of language use is situated, co-present interaction. Situated interaction is also the final frontier of Natural Language Processing (NLP), where, compared to the area of text processing, little progress has been made in the past decade, and where a myriad of practical applications is waiting to be unlocked. While the usual approach in the field is to reach, bottom-up, for the ever next "adjacent possible", in this paper I attempt a top-down analysis of what the demands are that unrestricted situated interaction makes on the participating agent, and suggest ways in which this analysis can structure computational models and research on them. Specifically, I discuss representational demands (the building up and application of world model, language model, situation model, discourse model, and agent model) and what I call anchoring processes (incremental processing, incremental learning, conversational grounding, multimodal grounding) that bind the agent to the here, now, and us.

1 Introduction

As Bisk et al. (2020) have noted, NLP as a field is slowly working its way towards ever wider "world scopes", going from modelling corpora to larger collections of text, to collections of text paired with other modalities, to modelling in environments over which the learning agent has some control, currently reaching out to scenarios where other agents need to be modelled as well. It is interesting to note how curiously backwards this would be as a description of the development of a human language user: Humans needs to experience other minds before they can ever begin to experience structured textual information. As a development strategy, the bottom-up methodology reflected in this "widening of scopes" also bears some risks: As Koller (2016) recently argued with respect to distributional semantics, a bottom-up strategy by design moves
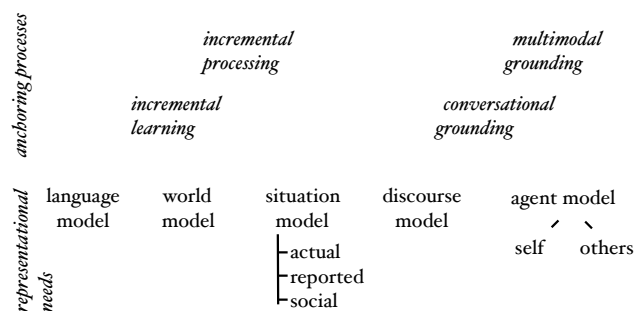

Figure 1: Representational Domains (bottom) and Anchoring Processes (top) Structuring the Situated Agent

from one (relative) success to the next, as the next thing is always the one that is just about possible to do. Without some further guidance, however, this limits the perspective and comes with the risk of getting stuck in local optima. This paper is an attempt to provide such guidance for the field of "embodied social AI", by pulling together some of what is known in the various scientific areas that deal with human verbal interaction, into an abstract description of modelling desiderata. In that sense, the proposal here may serve as a *conceptual benchmark* against which settings, tasks, datasets, and models can be measured in terms of their coverage, and in terms of the costs of the abstractions they make relative to this general model.[1]

Figure 1 shows an outline of the proposal, which the remaining sections will unpack: There are representational demands that the situation puts on the agent—that is, the agent needs to bring some knowledge, and track some information (discussed further in Section 3). The processes with which it handles the interaction (and comes to update this knowledge) also are subjects to some demands, stemming from the fact that the interaction partner is free and independent, but similar (Sections 2, 4).

It is useful to clarify one thing from the out-

---

[1] Note that this way of proceeding is fully compatible with an "empirical approach", insofar as that is used to select the best model, and does not aim to determine the goals as well.

set: What this paper is *not* trying to do is to make any recommendations as to *how* aspects of this model are to be realised (e.g., using symbolic or distributed representation methods; using particular learning algorithms; building in a certain modularisation or modelling monolithically; or using particular decision making algorithms); the intended contribution is an analysis of how the *phenomenon* of situated language use is conceptually structured on a high level, which can then eventually guide the definition of challenges and selection of methods to meet them.

2 Situated Interaction

Here is a (very) high-level, general characterisation of the face-to-face interaction situation: It is a *direct, purposeful encounter of free and independent, but similar agents*. Let us unpack this:

• as *agents*, the participants meet their purposes—and here, specifically, *communicative purposes*—through acting;

• as *free* agents, they cannot be forced, and cannot force the respective other, to do anything, and specifically not to *understand* as intended;

• as *independent* agents, they are individually subject to the same passing of time (while one acts, the other can as well and need not wait); they will also have different histories, including their histories of previous interactions and language use, and will bring different knowledge to the interaction;

• this being a *direct* encounter, the agents must rely on what they can do (produce for the other, receive from the other) with their bodies to create meaning here and now;

• finally, as fundamentally *similar* agents, they can rely on a certain body of shared knowledge and experience, for example in how each parses the shared environment, understands the world, and forms desires, beliefs, and intentions, and, if they are to use language for communication, in how they use language, but where the exact degree of similarity will need to be determined during and through the interaction.

This has consequences: To reach *joint* purposes, the agents need to coordinate, in a process that unfolds continuously in time and which can yield new knowledge, including about how to coordinate, but that also can rest on assumed commonalities.[2]

The next sections will go into the details of what

the situation, thus characterised, demands of the agent.

3 Representational Demands

The central means through with agents in situated interaction meet their purposes is *language* (and a particular one, at that), and hence the agent must come with knowledge of this language, or possess (or represent to itself) what I will call here a **language model**. It is not enough for the agent to be able to produce well-formed strings; rather, the systematic connection to the communicative intentions they express (Grice, 1957) must be modelled as well. As these intentions can concern objects in the world, and to the degree that the model of the language can be presumed to be shared, it is via those that the language can count as *grounded* in the sense that has most currency in the NLP community (Chandu et al., 2021).

Examples like those in (1) below indicate that **world knowledge** also factors into the purpose-directed use of language:

(1)   a.   I couldn't put the coat into the suitcase because it was too small.

          b.   Put the poster up on the wall.

In (1-a), a "Winograd schema" (Levesque et al., 2012) type sentence, information about expectable relative sizes, and in (1-b), knowledge about expected outcomes, is needed to interpret the utterance.[3] Again, underlying the communicative use of this knowledge is an assumption that it is shared.

While subject to possible updates, as we will see, these types of knowledge can be seen as something that the agent brings into the situation. But the situation itself must be understood by the agents, in order to interact in it. The proposed schema splits the **situation model** into three sub-types: A model of the *actual situation* in which the interaction is happening, which would provide not only referents for "poster" and "wall" in a situation in which (1-b) is used, but also potential likely referent for the implicit instrument of the requested action (e.g., perhaps there is a roll of duct tape visible, or a collection of pushpins). For this to work, there is an underlying assumption, which is that the situation will be mostly parsed similarly by the agents, so that it can form the shared basis for assumed

---

[2]This short description places a different focus, but in the broad strokes follows the analysis by Clark (1996).

[3]But see for example Pustejovsky (1991); Murphy (2010) for the notorious difficulties separating linguistic, and in particular lexical knowledge from such more general knowledge.

mutual knowledge (Clark, 1996); repair processes accounting for violations of this assumption will be discussed below.

The discourse of the agents does not always have to be about the actual situation, however. The building up of model of the *reported situation* (van Dijk and Kintsch, 1983), together with world knowledge about the consequences of entering a room, can explain the licensing of the contrast (indicating surprisal) in the following example:

(2)   I saw two people enter the room, but when I followed, the room was empty.

Lastly, the *social situation* also bears on linguistic material: Relative social status, for example, is grammaticalised in many languages (Bender, 2013); and even more generally, the simple fact of who is and who is not party to an interaction determines which linguistic and other behaviour is appropriate (Goffman, 1981; Bohus and Horvitz, 2009).

Next, the **discourse model**, required to keep track of antecedents of anaphoric acts and, more generally, for the determination of *coherence*. In (3), for example, the anaphoric elements *no* (as negating the proposition contained in A's question) and *he* can only be resolved under the assumption that they realise an *answer* (to A's question) and an *explanation* (for the answer), respectively. (See, *inter alia*, Kamp and Reyle (1993); Asher and Lascarides (2003); Ginzburg (2012).)

(3)   A: Is John coming to the party?
        B: No, he's busy.

Finally, there is a large body of work elucidating the role of the **agent model** (representing their *beliefs, desires, and intentions*) in interpreting discourse (Cohen et al., 1990). To again give just one illustrating example, in (4), A must know something about B's likely desires and intentions (to stay awake, or not stay awake) to make sense of their reply.

(4)   A: Do you want some coffee?
        B: It's late.

4   Anchoring Processes

Even if not often taken in its full breadth, *that* the knowledge described above plays a role in situated interaction is presumably not very controversial. The focus of this section is on something that is less
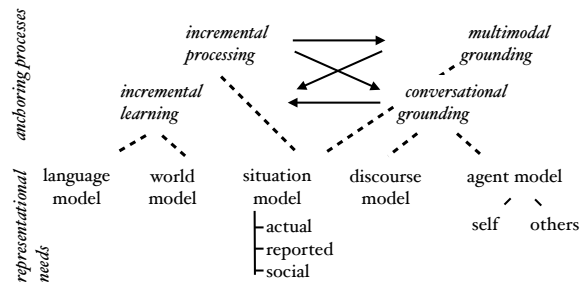

Figure 2: Representations and Processes. Arrows Denote the *enables* Relation, Dotted Lines the *updates* Relation.

often dealt with and brought together, namely, the processes through which the knowledge is applied and updated.

The fact that the agents are *independent* and hence not extrinsically temporally coordinated argues for **incremental processing**, that is, an updating of situation, discourse and agent models that is continual to the observation of the other agent's actions as well as to the agent's own production—this is turn then makes possible the *achievement* of coordination, for example in successful turn-taking (Schlangen and Skantze, 2009).

Only this processing regime then makes available certain devices used in **conversational grounding** (Clark and Brennan, 1991; Clark, 1996)—the process of coordination on what can count as shared knowledge (with respect to either of the models discussed above, most immediately the discourse model)—namely the use of overlapping signals such as "back-channel" behaviours like "uh-hu" or nodding (Yngve, 1970). When understanding is not reached immediately, *repair* can be initiated through clarifications and corrections (Hayashi et al., 2013). In the framework of section 2, this can be understood as the mending and making true of initially overoptimistic assumptions on what was shared.

In some cases, the outcome of the repair process can lead to "localised" learning through the establishment of precedents (Brennan and Clark, 1996), such as for example particular idiosyncratic ways of referring to some object; however, it can also, just like *direct teaching*, lead to a longer term update of language model (e.g., learning new terms) or world model (e.g., learning new facts), in a process of **incremental learning** (Hoppit and Laland, 2013).

Lastly, the multimodal nature of situated interaction (Holler and Levinson, 2019) is accounted for by processes of **multimodal grounding**, which

integrate meaning-making devices such as deictic and iconic gestures (Sowa and Wachsmuth, 2003; Kennington et al., 2013) and facial expressions (Poggi and Pelachaud, 2000). I will also subsume under this header the process of resolving references into the situational context (Roy and Reiter, 2005; Siebert and Schlangen, 2008) by performing the categorisations denoted by the expressions.

To summarise the preceding two sections, Figure 2 again shows the elements of the analysis discussed here, this time with interrelations added.

5 Where We Are, And Where To Go

It should be clear that from the perspective of the analysis detailed above, even the NLP tasks that most seem like they are related to situated interaction are severly limited. Here is not the space for a detailed catalogisation, but we can look at a few examples. "Visual dialog" (Das et al., 2017b), the task of generating a reply to a question about an image, within the context of preceding questions and answers, requires a limited discourse model (the previously asked questions and answer may introduce discourse referents), and a limited form of situation model and multimodal grounding (of the target image), but the strict role asymmetry (questioner and answerer) precludes any need for agent modelling and conversational grounding; its strict turn-by-turn nature and the use of written language abstract away from the questioning agent as independent agent and put control unilaterally in the hand of the model. "Embodied QA" (Das et al., 2017a) and "language & vision navigation" (Anderson et al., 2018), are tasks where in response to a language stimulus actions (in a simulator environment) need to be performed. Hence, these tasks require a more dynamic situation model, but other than that, are not fundamentally different from the visual dialog task (and in fact take away again what little that task requires in terms of discourse model).

Now, tackling a problem by focussing on its parts is a valid strategy, but only if in doing so the whole is kept in mind. In the cases cited above, it seems fair to say that the formulation of the task was driven more by the available modelling method: They basically are tasks that lend themselves to a formulation as sequence-to-sequence problem, and as such are more about transducing the semantics of the stimulus language than they are about situated *interaction* (or interaction at all).

More recently, tasks have been proposed that put more stress on the conversational grounding aspects mentioned above (Ilinykh et al., 2019; Udagawa and Aizawa, 2019; Bara et al., 2021). This is a good start, but in order to systematise these efforts, what is missing is a clearer picture of how the task setting (environment, interaction mode, etc.) determines what a task can even test, and how close it will come to the fuller picture sketched above. Should it turn out that for the richest settings, real interaction with capable language users is required, then ways will have to be found to enable that, and to overcome the batch learning mode that current models are bound to.

6 Related Work

That various kinds of knowledge and update processes are required to model conversational agents is not a new insight. The grandparent of any of those attempts, Winograd's 1972 SHDRLU already made a distinction between language model (in the form of parsing procedures) and situation model, as did the later textbook presentation by Allen (1995). A distinction between conversation situation and reported situation was made by Barwise and Perry (1983); Bratman (1987) and Wooldridge (2000) stressed the importance of modelling agents in terms of their beliefs, desires, and intentions. Allen et al. (2001) were among the first to point out the need for incremental processing. And to conclude this—almost absurdly selective—tour through what is a massive body of work, Traum and Larsson (2003) and Larsson (2002) described a representational system that elegantly interfaced discourse modelling, conversational grounding, and agent modelling. It is not, I want to claim here, that the *analyses* from these papers were wrong; in the light of more recently available methods, what is likely the case is that the realisation of representational demands through manually constructed representation formats and formalism restricted these models, and that this is what our more recent methods can help us overcome.[4]

7 Conclusions

I have argued for a particular analysis of the task of participating in situated interaction, drawing on various literatures. If NLP wants to advance on this phenomenon, I contend, it needs to start to take its complexity seriously, and devise methods and testbeds for tackling it, rather than only invent tasks that fit the available methods.

---

[4]See (Dingemanse et al., 2023) for a very recent, even more wide-ranging argument for the value of looking at situated interaction.

References

James Allen, George Ferguson, and Amanda Stent. 2001. An architecture for more realistic conversational systems. In *Proceedings of the conference on intelligent user interfaces*, Santa Fe, USA.

James C. Allen. 1995. *Natural Language Understanding*, 2nd edition. Benjamin/Cummings, Redwood City, USA.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. 2018. Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR 2018*.

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press.

Cristian-Paul Bara, Sky CH-Wang, and Joyce Chai. 2021. MindCraft: Theory of mind modeling for situated dialogue in collaborative tasks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1112–1125, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jon Barwise and John Perry. 1983. *Situations and Attitudes*. MIT Press, Cambridge, Mass., USA.

E.M. Bender. 2013. *Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language. *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pages 8718–8735.

Dan Bohus and Eric Horvitz. 2009. Models for multi-party engagement in open-world dialog. In *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL '09, pages 225–234.

Michael E. Bratman. 1987. *Intentions, Plans, And Practical Reason*. Harvard University Press, Cambridge, Massachusetts, USA.

Susan E. Brennan and Herbert H. Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6):1482–1493.

Khyathi Raghavi Chandu, Yonatan Bisk, and Alan W Black. 2021. Grounding "Grounding" in NLP. In *Findings of ACL-IJCNLP 2021*.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press, Cambridge.

Herbert H. Clark and Susan E. Brennan. 1991. Grounding in communication. In L. B. Resnick, J. Levine, and S. D. Behrend, editors, *Perspectives on Socially Shared Cognition*, pages 127–149. American Psychological Association Books, Washington D.C., USA.

Philip R. Cohen, Jerry Morgan, and Martha E. Pollack, editors. 1990. *Intentions in Communication*. MIT Press, Cambridge, Mass.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. 2017a. Embodied question answering. *CoRR*, abs/1711.11543.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. 2017b. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Mark Dingemanse, Andreas Liesenfeld, Marlou Rasenberg, Saul Albert, Felix K. Ameka, Abeba Birhane, Dimitris Bolis, Justine Cassell, Rebecca Clift, Elena Cuffari, Hanne De Jaegher, Catarina Dutilh Novaes, N. J. Enfield, Riccardo Fusaroli, Eleni Gregoromichelaki, Edwin Hutchins, Ivana Konvalinka, Damian Milton, Joanna Rączaszek-Leonardi, Vasudevi Reddy, Federico Rossano, David Schlangen, Johanna Seibt, Elizabeth Stokoe, Lucy Suchman, Cordula Vesper, Thalia Wheatley, and Martina Wiltschko. 2023. Beyond single-mindedness: A figure-ground reversal for the cognitive sciences. *Cognitive Science*, 47(1):e13230.

Jonathan Ginzburg. 2012. *The Interactive Stance: Meaning for Conversation*. Oxford University Press, Oxford, UK.

E. Goffman. 1981. *Forms of Talk*. Conduct and Communication. University of Pennsylvania Press, Incorporated.

H.P. Grice. 1957. Meaning. *Philosophical Review*, 66:377–388.

Makoto Hayashi, Geoffrey Raymond, and Jack Sidnell. 2013. *Conversational Repair and Human Understanding*. OAPEN Library. Cambridge University Press.

Judith Holler and Stephen C. Levinson. 2019. Multimodal Language Processing in Human Communication. *Trends in Cognitive Sciences*, pages 1–14.

William Hoppit and Kevin N. Laland. 2013. *Social Learning: An Introduction to Mechanisms, Methods, and Models*. Princeton University Press.

Nikolai Ilinykh, Sina Zarrieß, and David Schlangen. 2019. Meetup! a corpus of joint activity dialogues in a visual environment. In *Proceedings of the 23rd Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2019 / LondonLogue)*, London, UK.

Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic*. Kluwer, Dordrecht.

Casey Kennington, Spyros Kousidis, and David Schlangen. 2013. Interpreting situated dialogue utterances: an update model that uses speech, gaze, and gesture information. In *Proceedings of the SIGDIAL 2013 Conference*, pages 173–182, Metz, France. Association for Computational Linguistics.

Alexander Koller. 2016. Top-down and bottom-up views on success in semantics. Invited talk at the 5th Joint Conference on Lexical and Computational Semantics (*SEM).

Staffan Larsson. 2002. *Issue-based Dialogue Management*. Ph.D. thesis, Göteborg University, Göteborg, Sweden.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The Winograd schema challenge. *Proceedings of the International Workshop on Temporal Representation and Reasoning*, pages 552–561.

M. Lynn Murphy. 2010. *Lexical Meaning*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Isabella Poggi and Catherine Pelachaud. 2000. Facial performatives in a conversational system. In Justine Cassell, Joseph Sullivan, Scott Prevost, and Elizabeth Churchill, editors, *Embodied Conversational Agents*. MIT Press, Cambridge, USA.

James Pustejovsky. 1991. The generative lexicon. *Computational Linguistics*, 17(4):409–441.

Deb Roy and Ehud Reiter. 2005. Connecting language to the world. *Artificial Intelligence*, 167:1–12.

David Schlangen and Gabriel Skantze. 2009. A general, abstract model of incremental dialogue processing. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009)*, pages 710–718, Athens, Greece.

Alexander Siebert and David Schlangen. 2008. A simple method for resolution of definite reference in a shared visual context. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pages 84–87, Columbus, Ohio. Association for Computational Linguistics.

Timo Sowa and Ipke Wachsmuth. 2003. Coverbal Iconic Gestures for Object Descriptions in Virtual Environments: An Empirical Study. In *Gestures: Meaning and Use*, Porto, Portugal.

David R. Traum and Staffan Larsson. 2003. The information state approach to dialogue management. In Ronnie Smith and Jan van Kuppevelt, editors, *Current and New Directions in Discourse and Dialogue*, pages 325–353. Kluwer, Dordrecht, The Netherlands.

Takuma Udagawa and Akiko Aizawa. 2019. A Natural Language Corpus of Common Grounding under Continuous and Partially-Observable Context. In *Proceedings of AAAI*.

T.A. van Dijk and W. Kintsch. 1983. *Strategies of Discourse Comprehension*. Monograph Series. Academic Press.

Terry Winograd. 1972. Procedures as a representation for data in a computer program for understanding natural language. *Cognitive Psychology*, 3(1).

Michael J. Wooldridge. 2000. *Reasoning about rational agents*. Intelligent robots and autonomous agents. MIT Press.

V. H. Yngve. 1970. On getting a word in edgewise. In *Papers from the 6th Regional Meeting*, pages 567–578, Chicago, USA. Chicago Linguistics Society.