Coordinating spatial perspective in discourse

Simon Dobnik

Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg, Box 200, 405 30 Göteborg
simon.dobnik@gu.se

1. Introduction

Understanding and generating spatial descriptions such as "to the left of" and "above" is crucial for any situated conversational agent such as a robot used in rescue operations. The semantics of spatial descriptions are complex and involve (i) perceptual knowledge obtained from scene geometry (Regier and Carlson, 2001), (ii) world knowledge about the objects involved (Coventry et al., 2001), and (iii) shared knowledge that is established as the common ground in discourse. Dialogue partners coordinate all three types of meaning when describing and interpreting visual scenes.

One example of (iii) is the assignment of perspective or the reference frame (RF). For example, the table may be "to the left of the chair", "to the right of the chair", "behind the chair" or "South of the chair". The RF, may be described linguistically "from your view" or "from there" but in a spontaneous conversation it is frequently omitted. Instead, it is integrated in the content of the conversation as a discourse variable which is applied over several turns and even over several speakers. The RF may also be inferred from the perceptual context if given some configuration of the scene a spatial description is true only in that RF. It follows that when interpreting and generating spatial descriptions humans rely on verification of spatial templates in different RFs which requires considerable computational complexity (Steels and Loetzsch, 2009).

The perspective is grounded by some point in the scene called the *viewpoint* (VPT). There are three ways in which the VPT is set in human languages (Levinson, 2003): (i) *relative RF:* by some third object distinct from the located and reference objects (the speaker, the hearer, the sofa); (ii) *intrinsic RF:* by the reference object itself (the chair); or (iii) *extrinsic RF:* by some global reference point (the North). Sometimes (mostly for route descriptions) a distinction is made between speaker-oriented (egocentric) and external (allocentric) perspective or between route and survey perspective but this is a less specific distinction. The geometric spatial templates are projected within the framework defined by the VPT (Maillat, 2003).

2. Reference frames in conversation

Watson et al. (2004) show experimentally that (i) participants are significantly more likely to use an intrinsic RF after their partner used an intrinsic RF, compared when the partner used a relative RF (with the speaker as the VPT); (ii) participants are significantly more likely to use intrinsic RF when the objects are aligned horizontally (their typical alignment in the world) than when they are aligned vertically; (iii) the alignment of the RFs is not due to the lexical priming caused by using the same preposition. Andonova (2010) shows for the map task that overall partners align with the primed route or survey perspective set by the confederate if priming is consistent – when the confederate changes the perspective only once in the middle of the session. On the other hand, if the confederate regularly alternates between the perspectives their partner has nothing to prime to. The self-assessed spatial ability (using a standardised test) is also important – low ability participants only align with the primed perspective when the switch is from the survey to the route perspective which is otherwise also the most frequently used one.

3. Towards a more natural spatial dialogue

Our interest is to implement these and similar strategies as information state update rules in a dialogue manager such as GoDiS (Larsson, 2002). In such a model each conversational agent must keep a record of their own RF and that of their partner in the common ground. The RFs are updated following perceptual verification and an alignment strategy. The proposal is a move towards a more natural interpretation and generation of projective spatial descriptions in an artificial conversational agent compared to our previous attempt where the RF parameters were not specifically included in the model but some RF knowledge has nonetheless been learned with machine learning. We proceed as follows:

1. Collect a corpus of dialogue interactions containing projective spatial descriptions made in a room scene.
2. Annotate the dialogue utterances with an XML annotation scheme which identifies perceptual states, objects in focus, utterances, turns, speakers, located objects, RFs, VPTs, spatial relations, ref. objects, etc.
3. Replicate the literature findings on the RF usage in our dataset.
4. Repeat the experiments from (1) but where one of the participants is a dialogue manager following an RF strategy. Allow humans conversational partners to rate the performance of the system.
   (a) Always use the relative RF to yourself.
   (b) Always align to the RF used by your partner in the previous turn.
   (c) For each turn select the RF randomly.
   (d) Keep a randomly chosen RF for n turns, then change.

   To prevent over-agreement with the system the evaluators should, ideally, compare pairs of strategies and select the preferred one.

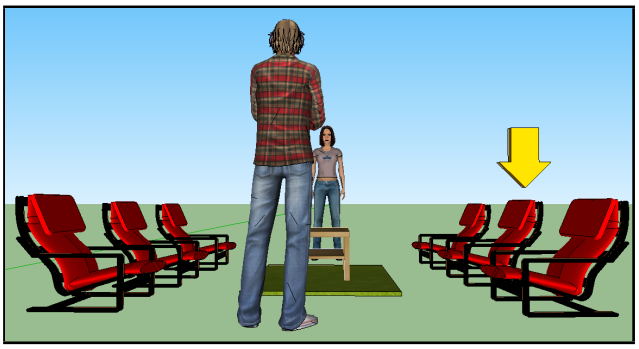

Figure 1: A scene one participant sees during a conversation in the 2nd pilot study. The arrow indicates the object the location of which this participant should describe.

We collect our data and later test the interaction in an online experimental environment specifically developed for this purpose (`http://goo.gl/8KLja`). Participants may create sessions to which they invite other participants and complete them interactively in their own time. During a session each participant sees a 3d generated image of a room containing some furniture. The image also contains two avatars: the one with their back towards the participant is the participant and the one facing the participant from the opposite side of the room is their partner (Figure 1). This is explained to the participants in the instructions and different representations are used to avoid the confusion. The other participant sees the room from the opposite side. The participants communicate via a text chat interface which allows unrestricted entry of text and also logs and partially annotates both the conversation and the perceptual information in the background.

4. Results

By the time of writing this abstract we conducted two pilot studies for which we completed stages 1 to 3 of our plan. In the first pilot study (7 conversations in Slovenian) we used a room with four distinct entities (two participants, a chair and a sofa) arranged around a table in the middle which was placed on a carpet. We instructed the participants to talk about the location of the objects in the scene. Although this method was good in encouraging spontaneous conversations it had two shortcomings: (i) the participants produced less spatial descriptions than desired (11.9 per conversation) as they also discussed their opinions about the objects, etc.; and (ii) they spontaneously took on roles where one was asking questions and the other was giving answers and therefore the conversations included were very few cases of interaction that we were looking for. To overcome the difficulties from the first study we designed a second pilot study (10 conversations in Slovenian) for which we (i) only used one kind of objects (the chairs), (ii) restricted the conversational interaction to pair of turns where in the first turn one participant describes which chair they chose (one is automatically selected for them and marked with an arrow as shown in Figure 1) and then in the second turn their partner selects that chair on their view of the room. The roles are reversed in the next turn. Thus, we get a series of dialogue turns from which we record (i) speaker's strategy for RF choice; (ii) the hearer's understanding of the description. The latter is important as a particular description may be true under more than one RF.

A manual analysis of the data obtained so far confirms that participants align their perspective but only if one participant uses a particular perspective consistently over more than one turn, then the other would follow (priming). Our explanation is that the second speaker assumes that a particular perspective is important in the conversation and should therefore be made part of the common ground. Further we observe that speakers not only align perspectives but also the way the scene is described syntactically. While in the first trials participants may frequently omit the explicit description of perspective and align to the perspective of the other, the structured environment of the second trials forces them to use definitions such as "from your side" nearly all the time even if they are aligned. They may also omit the explicit definition and align to the fact that each participant is describing from its own perspective.

5. Further work

In the time leading to the conference we hope to continue to collect conversations online (in Slovenian, English and Swedish), tag them and integrate them in an automatic agent that will be used for step 4. Note that the method allows us to collect a set of best referring expressions for each object together with all their semantic properties which means that the descriptions can be conveniently applied in generation.

6. References

Elena Andonova. 2010. Aligning spatial perspective in route descriptions. In Christoph Hölscher, Thomas Shipley, Marta Olivetti Belardinelli, John Bateman, and Nora Newcombe, editors, *Spatial Cognition VII*, volume 6222 of *Lecture Notes in Computer Science*, pages 125–138. Springer Berlin, Heidelberg.

Kenny R. Coventry, Mercè Prat-Sala, and Lynn Richards. 2001. The interplay between geometry and function in the apprehension of Over, Under, Above and Below. *Journal of memory and language*, 44(3):376–398.

Staffan Larsson. 2002. *Issue-based Dialogue Management*. Ph.D. thesis, University of Gothenburg.

Stephen C Levinson. 2003. *Space in language and cognition: explorations in cognitive diversity*, volume 5. Cambridge University Press, Cambridge.

Didier Maillat. 2003. *The semantics and pragmatics of directionals: a case study in English and French*. Ph.D. thesis, Committee for Comparative Philology and General Linguistics, University of Oxford, Oxford, UK, May.

Terry Regier and Laura A. Carlson. 2001. Grounding spatial language in perception: an empirical and computational investigation. *Journal of Experimental Psychology: General*, 130(2):273–298.

Luc Steels and Martin Loetzsch. 2009. Perspective alignment in spatial language. In Kenny R. Coventry, Thora Tenbrink, and John. A Bateman, editors, *Spatial Language and Dialogue*. Oxford University Press.

Matthew E. Watson, Martin J. Pickering, and Holly P. Branigan. 2004. Alignment of reference frames in dialogue. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society*, Chicago, USA, August.