A Situated Context Model for Resolution and Generation of Referring Expressions

Hendrik Zender and **Geert-Jan M. Kruijff** and **Ivana Kruijff-Korbayová**
Language Technology Lab, German Research Center for Artificial Intelligence (DFKI)
Saarbrücken, Germany
{zender, gj, ivana.kruijff}@dfki.de

Abstract

The background for this paper is the aim to build robotic assistants that can "naturally" interact with humans. One prerequisite for this is that the robot can correctly identify objects or places a user refers to, and produce comprehensible references itself. As robots typically act in environments that are larger than what is immediately perceivable, the problem arises how to identify the appropriate context, against which to resolve or produce a referring expression (RE). Existing algorithms for generating REs generally bypass this problem by assuming a given context. In this paper, we explicitly address this problem, proposing a method for context determination in large-scale space. We show how it can be applied both for resolving and producing REs.

1 Introduction

The past years have seen an extraordinary increase in research on robotic assistants that help users perform daily chores. Autonomous vacuum cleaners have already found their way into people's homes, but it will still take a while before fully conversational robot "gophers" will assist people in more demanding everyday tasks. Imagine a robot that can deliver objects, and give directions to visitors on a university campus. This robot must be able to verbalize its knowledge in a way that is understandable by humans.

A conversational robot will inevitably face situations in which it needs to refer to an entity (an object, a locality, or even an event) that is located somewhere outside the current scene, as Figure 1 illustrates. There are conceivably many ways in which a robot might refer to things in the world, but many such expressions are unsuitable in most

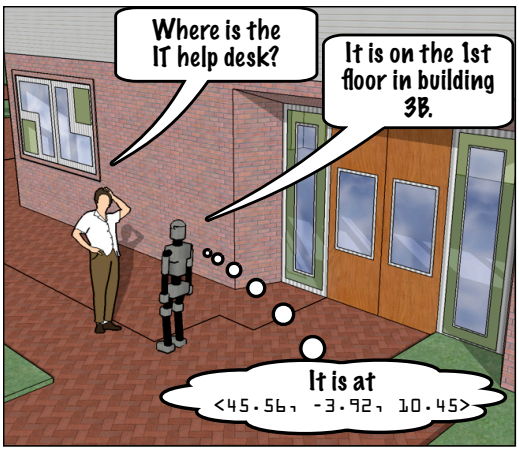

Figure 1: Situated dialogue with a service robot

human-robot dialogues. Consider the following set of examples:

1. "position $P = \langle 45.56, -3.92, 10.45 \rangle$"
2. "Peter's office no. 200 at the end of the corridor on the third floor of the Acme Corp. building 3 in the Acme Corp. complex, 47 Evergreen Terrace, Calisota, Earth, (...)"
3. "the area"

These REs are valid descriptions of their respective referents. Still they fail to achieve their *communicative goal*, which is to specify the right amount of information that the hearer needs to uniquely identify the referent. The next REs *might* serve as more appropriate variants of the previous examples (*in certain contexts!*):

1. "the IT help desk"
2. "Peter's office"
3. "the large hall on the first floor"

The first example highlights a requirement on the knowledge representation to which an algorithm for generating referring expressions (GRE) has access. Although the robot needs a robot-centric representation of its surrounding space that allows it to safely perform actions and navigate its world, it should use human-centric qualitative descriptions when talking about things in the world. We

Proceedings of the 12th European Workshop on Natural Language Generation, pages 126–129,
Athens, Greece, 30 – 31 March 2009.

do not address this issue here, but refer the interested reader to our recent work on multi-layered spatial maps for robots, bridging the gap between robot-centric and human-centric spatial representations (Zender et al., 2008).

The other examples point out another important consideration: how much information does the human need to single out the intended referent among the possible entities that the robot could be referring to? According to the seminal work on GRE by Dale and Reiter (1995), one needs to distinguish whether the intended referent is already in the hearer's *focus of attention* or not. This focus of attention can consist of a local visual scene (visual context) or a shared workspace (spatial context), but also contains recently mentioned entities (dialogue context). If the referent is already part of the current context, the GRE task merely consists of singling it out among the other members of the context, which act as distractors. In this case the generated RE contains *discriminatory* information, e.g. "the red ball" if several kinds of objects with different colors are in the context. If, on the other hand, the referent is not in the hearer's focus of attention, an RE needs to contain what Dale and Reiter call *navigational*, or *attention-directing* information. The example they give is "the black power supply in the equipment rack," where "the equipment rack" is supposed to direct the hearers attention to the rack and its contents.

In the following we propose an approach for context determination and extension that allows a mobile robot to produce and interpret REs to entities outside the current visual context.

2 Background

Most GRE approaches are applied to very limited, visual scenes – so-called *small-scale space*. The domain of such systems is usually a small visual scene, e.g. a number of objects, such as cups and tables, located in the same room), or other closed-context scenarios (Dale and Reiter, 1995; Horacek, 1997; Krahmer and Theune, 2002). Recently, Kelleher and Kruijff (2006) have presented an incremental GRE algorithm for situated dialogue with a robot about a table-top setting, i.e. also about small-scale space. In all these cases, the context set is assumed to be identical to the visual scene that is shared between the interlocutors. The intended referent is thus already in the hearer's *focus of attention*.

In contrast, robots typically act in *large-scale space*, i.e. space "larger than what can be perceived at once" (Kuipers, 1977). They need the ability to understand and produce references to things that are beyond the current visual and spatial context. In any situated dialogue that involves entities beyond the current focus of attention, the task of *extending the context* becomes key.

Paraboni et al. (2007) present an algorithm for *context determination* in hierarchically ordered domains, e.g. a university campus or a document structure. Their approach is mainly targeted at producing textual references to entities in written documents (e.g. figures, tables in book chapters). Consequently they do not address the challenges that arise in physically and perceptually situated dialogues. Still, the approach presents a number of good contributions towards GRE for situated dialogue in large-scale space. An appropriate context, as a subset of the full domain, is determined through Ancestral Search. This search for the intended referent is rooted in the "position of the speaker and the hearer in the domain" (represented as d), a crucial first step towards situatedness. Their approach suffers from the shortcoming that spatial relationships are treated as one-place attributes by their GRE algorithm. For example they transform the spatial containment relation that holds between a room entity and a building entity ("the library in the Cockroft building") into a property of the room entity (BUILDING NAME = COCKROFT) and not a two-place relation (`in(library,Cockroft)`). Thus they avoid recursive calls to the algorithm, which would be needed if the intended referent is related to another entity that needs to be properly referred to.

However, according to Dale and Reiter (1995), these related entities do not necessarily serve as discriminatory information. At least in large-scale space, in contrast to a document structure that is conceivably transparent to a reader, they function as *attention-directing elements* that are introduced to build up *common ground* by incrementally extending the hearer's focus of attention. Moreover, representing some spatial relations as two-place predicates between two entities and some as one-place predicates is an arbitrary decision.

We present an approach for context determination (or *extension*), that imposes less restrictions on its knowledge base, and which can be used as a sub-routine in existing GRE algorithms.

3 Situated Dialogue in Large-Scale Space

Imagine the situation in Figure 1 did not take place somewhere on campus, but rather inside building 3B. Certainly the robot would not have said "the IT help desk is on the 1st floor in building 3B." To avoid confusing the human, an utterance like "the IT help desk is on the 1st floor" would have been appropriate. Likewise, if the IT help desk happened to be located on another site of the university, the robot would have had to identify its location as being "on the 1st floor in building 3B on the new campus." The hierarchical representation of space that people are known to assume (Cohn and Hazarika, 2001), reflects upon the choice of an appropriate context when producing REs.

In the above example the physical and spatial situatedness of the dialogue participants play an important role in determining which related parts of space come into consideration as potential distractors. Another important observation concerns the verbal behavior of humans when talking about remote objects and places during a complex dialogue (i.e. more than just a question and a reply). Consider the following example dialogue:

> Person A: "Where is the exit?"
>
> Person B: "You first go down this corridor. Then you turn right. After a few steps you will see the big glass doors."
>
> Person A: "And the bus station? Is it to the left?"

The dialogue illustrates how utterances become grounded in previously introduced discourse referents, both temporally and spatially. Initially, the physical surroundings of the dialogue partners form the context for anchoring references. As a dialogue unfolds, this point can conceptually move to other locations that have been explicitly introduced. Discourse markers denoting spatial or temporal cohesion (e.g. "then" or "there") can make this move to a new anchor explicit, leading to a "mental tour" through large-scale space.

We propose a general principle of *Topological Abstraction* (TA) for context extension which is rooted in what we will call the *Referential Anchor* a.[1] TA is designed for a multiple abstraction hierarchy (e.g. represented as a lattice structure rather than a simple tree). The Referential Anchor a, corresponding to the current focus of attention, forms the nucleus of the context. In the simple case, a corresponds to the hearer's physical location. As illustrated above, a can also move along the "spatial progression" of the most salient discourse entity during a dialogue. If the intended referent is outside the current context, TA extends the context by incrementally ascending the spatial abstraction hierarchy until the intended referent is an element of the resulting sub-hierarchy, as illustrated in Figure 2. Below we describe two instantiations of the TA principle, a TA algorithm for reference generation (TAA1) and TAA2 for reference resolution.

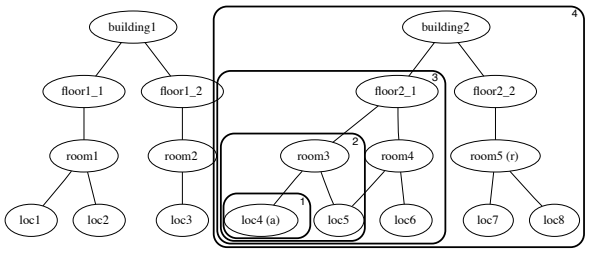

Figure 2: Incremental TA in large-scale space

Context Determination for GRE TAA1 constructs a set of entities dominated by the Referential Anchor a (and a itself). If this set contains the intended referent r, it is taken as the current utterance context set. Else TAA1 moves up one level of abstraction and adds the set of all child nodes to the context set. This loop continues until r is in the context set. At that point TAA1 stops and returns the constructed context set (cf. Algorithm 1).

TAA1 is formulated to be neutral to the kind of GRE algorithm that it is used for. It can be used with the original Incremental Algorithm (Dale and Reiter, 1995), augmented by a recursive call if a relation to another entity is selected as a discriminatory feature. It could in principle also be used with the standard approach to GRE involving relations (Dale and Haddock, 1991), but we agree with Paraboni et al. (2007) that the mutually qualified references that it can produce[2] are not easily resolvable if they pertain to circumstances where a confirmatory search is costly (such as in large-scale space). More recent approaches to avoiding infinite loops when using relations in GRE make use of a graph-based knowledge representation (Krahmer et al., 2003; Croitoru and van Deemter, 2007). TAA1 is compatible with these approaches, as well as with the salience based approach of (Krahmer and Theune, 2002).

[1] similar to Ancestral Search (Paraboni et al., 2007)

[2] An example for such a phenomenon is the expression "the ball on the table" in a context with several tables and several balls, but of which only one is on a table. Humans find such REs natural and easy to resolve in visual scenes.

Algorithm 1 TAA1 (for reference generation)

Require: a = referential anchor; r = intended referent
  Initialize context: $C = \{\}$
  $C = C \cup topologicalChildren(a) \cup \{a\}$
  if $r \in C$ **then**
    return C
  else
    Initialize: $SUPERNODES = \{a\}$
    for each $n \in SUPERNODES$ **do**
      for each $p \in topologicalParents(n)$ **do**
        $SUPERNODES = SUPERNODES \cup \{p\}$
        $C = C \cup topologicalChildren(p)$
      end for
      if $r \in C$ **then**
        return C
      end if
    end for
    return failure
  end if

Algorithm 2 TAA2 (for reference resolution)

Require: a = ref. anchor; $desc(x)$ = description of referent
  Initialize context: $C = \{\}$
  Initialize possible referents: $R = \{\}$
  $C = C \cup topologicalChildren(a) \cup \{a\}$
  $R = desc(x) \cap C$
  if $R \neq \{\}$ **then**
    return R
  else
    Initialize: $SUPERNODES = \{a\}$
    for each $n \in SUPERNODES$ **do**
      for each $p \in topologicalParents(n)$ **do**
        $SUPERNODES = SUPERNODES \cup \{p\}$
        $C = C \cup topologicalChildren(p)$
      end for
      $R = desc(x) \cap C$
      if $R \neq \{\}$ **then**
        return R
      end if
    end for
    return failure
  end if

Resolving References to Elsewhere Analogous to the GRE task, a conversational robot must be able to understand verbal descriptions by its users. In order to avoid overgenerating possible referents, we propose TAA2 (cf. Algorithm 2) which tries to select an appropriate referent from a relevant subset of the full knowledge base. It is initialized with a given semantic representation of the referential expression, $desc(x)$, in a format compatible with the knowledge base. Then, an appropriate entity satisfying this description is searched for in the knowledge base. Similarly to TAA1, the description is first matched against the current *context set* C consisting of a and its child nodes. If this set does not contain any instances that match $desc(x)$, TAA2 increases the context set along the spatial abstraction axis until at least one possible referent can be identified within the context.

4 Conclusions and Future Work

We have presented two algorithms for context determination that can be used both for resolving and generating REs in large-scale space.

We are currently planning a user study to evaluate the performance of the TA algorithms. Another important item for future work is the exact nature of the spatial progression, modeled by "moving" the referential anchor, in a situated dialogue.

Acknowledgments

This work was supported by the EU FP7 ICT Project "CogX" (FP7-ICT-215181).

References

A. G. Cohn and S. M. Hazarika. 2001. Qualitative spatial representation and reasoning: An overview. *Fundamenta Informaticae*, 46:1–29.

M. Croitoru and K. van Deemter. 2007. A conceptual graph approach to the generation of referring expressions. In *Proc. IJCAI-2007*, Hyderabad, India.

R. Dale and N. Haddock. 1991. Generating referring expressions involving relations. In *Proc. of the 5th Meeting of the EACL*, Berlin, Germany, April.

R. Dale and E. Reiter. 1995. Computational interpretations of the Gricean Maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263.

H. Horacek. 1997. An algorithm for generating referential descriptions with flexible interfaces. In *Proc. of the 35th Annual Meeting of the ACL and 8th Conf. of the EACL*, Madrid, Spain.

J. Kelleher and G.-J. Kruijff. 2006. Incremental generation of spatial referring expressions in situated dialogue. In *In Proc. Coling-ACL 06*, Sydney, Australia.

E. Krahmer and M. Theune. 2002. Efficient context-sensitive generation of referring expressions. In K. van Deemter and R.Kibble, editors, *Information Sharing: Givenness and Newness in Language Processing*. CSLI Publications, Stanford, CA, USA.

E. Krahmer, S. van Erk, and A. Verleg. 2003. Graph-based generation of referring expressions. *Computational Linguistics*, 29(1).

B. Kuipers. 1977. *Representing Knowledge of Large-scale Space*. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, USA.

I. Paraboni, K. van Deemter, and J. Masthoff. 2007. Generating referring expressions: Making referents easy to identify. *Computational Linguistics*, 33(2):229–254, June.

H. Zender, O. Martínez Mozos, P. Jensfelt, G.-J. Kruijff, and W. Burgard. 2008. Conceptual spatial representations for indoor mobile robots. *Robotics and Autonomous Systems*, 56(6):493–502, June.