Abstract

The paper claims that spatial representation and reasoning (SRR) in robotic vision is closely related to SRR in natural language processing (NLP) because the two areas share similar issues in knowledge representation; because the way SRR is expressed by natural language cannot help influencing researchers in their developments of artificial systems of SRR; and because natural language interfaces (NLI) of a special kind are becoming increasingly essential for any type of computer system including robotic vision systems. The paper introduces a frame-based knowledge representation system for SRR, reviews NLP research on SRR, outlines SRR in natural language, and, finally, focuses on SRR in NLIs for robotic vision systems. It is demonstrated that:

- SRR in a robotic vision system is seriously affected by SRR in natural language.
- SRR in natural language is a complex, multifaceted, and fuzzy phenomenon.
- Sublanguage- and meaning-based NLIs to robotic vision systems can easily accommodate all the SRR requirements of those systems.
The paper focuses on those problems of space management in natural language processing which pertain to robotics, especially to robotic vision. The relevance of NLP-related spatial research to robotics is determined by the following three considerations (in the increasing order of importance):

- Knowledge-based robotic systems and NLP systems face similar problems of knowledge representation and reasoning, including spatial representation and reasoning (SRR), and the solutions found in one of the areas may be transportable to the other.

- As the research in robotics progresses and more systems become operational on the shop floor or in a military environment, increasingly more users unfamiliar with the system design and command language will need to access them; this is a situation in which robust and non-brittle interfaces become crucial, and natural language interfaces (NLI) with a reasonably unconstrained input within a specific realistic domain are the most efficient ones for the following reasons:
  - it is natural for humans to represent, and reason about, space in natural language.
  - computer systems have to have spatial knowledge represented in frames, predicate calculus formulas or another notation that facilitates mechanical inference-making.
  - NLIs will thus provide a link between the natural language and the language of spatial representation and reasoning (SRR).
  - NLIs are becoming less costly and more feasible as the research in NLP progresses in the semantic/pragmatic direction.

- While intricate systems of geometric and/or topographic representations can, and need to, be developed artificially for robotic systems, the researchers tend to be guided by the way spatial references are made in their language.

So far, the research devoted to acquiring, representing, understanding and reasoning about spatial relations has been largely confined to the areas of vision and robotics. Much of the work in SRR concentrated on path-finding and navigation for autonomous mobile devices (cf. Meystel, 1985 for a survey). There has been a number of interesting representation (e.g., Agin, 1972; Kuipers, 1978; Davis, 1981), acquisition (e.g., Davis, 1984) and reasoning (e.g., Forbus, 1981; McDermott and Davis, 1984) efforts in the area of shape descriptions, scene descriptions, path finding, etc.

Section 1 of the paper introduces a frame-based knowledge-representation system optimal for NLP systems in general and NLIs in particular, and quite familiar to those involved in knowledge-based robotic vision. Section 2 contains a brief survey of research on SRR in NLP. Section 3 deals with the way natural languages deal with space. Finally, Section 4 discusses the optimal approach to the development of SRR modules of NLIs for robotic systems and demonstrates the adequacy of the frame-based representation for SRR information required in IMASes and other robotic vision systems.

1. Frame-Based Representation in NLP and NLIs

Frame-based knowledge representation is used in many state-of-the-art NLP systems. TRANSLATOR, a knowledge-based machine translation system (see, for instance, Nirenburg et al., 1987), uses a system of this type to represent the interlingua (IL), which is the result of a
comprehensive semantic/pragmatic, syntactic, and morphological analysis of the input text in a source language, e.g., English. The information thus recorded can then be expressed in a target language in the process of generation. If the target language is another natural language, e.g., German, the system performs machine translation. But the IL information can be used for information retrieval, automatic text summarization, etc.

What follows is the IL representation of an English sentence, *Data such as the above, that are stored more or less permanently in a computer, we term a database* by TRANSLATOR's analyzer.

(object
  (id object1)
  (is-token-of data) *
  (subworld computerworld) *
  (quantifier (type all) (scope (and clause1 clause2))))

(object
  (id object2)
  (is-token-of computer)
  (subworld computerworld)
  (quantifier any))

(object
  (id object3)
  (is-token-of database)
  (subworld computerworld))

(state
  (id state1)
  (is-token-of be-equivalent)
  (phase static)
  (patient1 object1)
  (patient2 (antecedent-of above))
  (time always)
  (space none)
  (subworld computerworld))
(state
  (id state2)
  (is-token-of in)
  (phase static)
  (patient1 object1)
  (patient2 object2)
  (time always)
  (space none)
  (subworld computerworld))

(state
  (id state3)
  (is-token-of be-a-name-of)
  (phase static)
  (patient1 object3)
  (patient2 object1)
  (time always)
  (space none)
  (subworld computerworld))

(clause
  (id clause1)
  (discourse-structure (+expan clause1 clause3))
  (event state1)
  (focus state1.patient2)
  (modality conditional)
  (subworld computerworld)
  (time always)
  (space none))

(clause
  (id clause2)
  (discourse-structure (+expan clause2 clause3))
  (event state2)
  (focus time)
  (modality conditional)
  (subworld computerworld)
  (time always)
  (space (in object1 object2)))

(clause
  (id clause3)
  (discourse-structure none)
  (event state3)
  (focus object3)
  (modality real)
  (subworld computerworld)
  (time always)
  (space none))
It is clear that there is hardly any provision for SRR in this system besides characterizing one state as IN. The space property slot remains uncharacterized throughout, and nevertheless, this system works quite well for TRANSLATOR. The reason for that is that many NLP systems (see Raskin, 1986) can get away with very little SRR. This is also the reason why NLP has not paid enough attention to SRR. It will be demonstrated in Section 4 that the representation expands easily to accommodate as much SRR information as necessary.

2. A Brief Survey of NLP Research on SRR

Winograd (1972) introduced a measure of spatial reasoning - thus, for instance, the robotic hand knew that nothing could be stacked on top of a pyramid or on top of a cube on which a pyramid had been stacked earlier and not subsequently removed. This was achieved with the help of a simple and straightforward formalism involving the categories of shape, location, and size (in our notation):

('space'
  ('shape' (Round Pointed Rectangular))
  ('location' (first-coordinate) (second-coordinate) (third-coordinate))
  ('size' (first-dimension) (second-dimension) (third-dimension)))

For size and location, Winograd used a 3D coordinate system with coordinates ranging from 0 to 1200 on all three dimensions. Thus,

('shape' Rectangular)
('location' (100 200 300))
('size' (400 500 600))

would describe a block whose front lower left-hand corner is at 100, 200, 300, and whose three dimensions are 400, 500, 600.

The robotic hand had no sensors whatsoever, and the recognition was simulated by inputting the initial parameters in symbolic form.

Winograd's effort did not, however, lead to much additional research on SRR in NLP. Moreover, neither the spatial module nor the entire system was ever scaled up. The relative lack of interest in space is even more surprising if one considers how seriously treatment of time has been approached
in the NLP community over the past several years (see, for instance, Allen, 1984, Steedman and Moens, 1987).

Research by Waltz and his associates at the University of Illinois (e.g., Waltz, 1981; Boggess, 1978) was directed at the design of spatial representations for the use of natural language question-answering systems as part of their effort to relate verbal descriptions to visual representations and, more generally, to adequately analyze language descriptions of the physical world. Such representations support transitivity judgments (i.e., given as input the sequence "the goldfish is in the goldfish bowl; the goldfish bowl is on the shelf; the shelf is on the desk; the desk is in the room," the system was able to answer in the affirmative the query "Is the goldfish in the room?")

The knowledge-representation system used in TRANSLATOR already provides for more SRR than illustrated in Section 1. Its current properties set includes two SRR-related frames, 'size' and 'shape': (see Nirenburg et al. 1987, p. 97)

However, in order to be more complete (if the served domain requires) the properties should include frames for 'position' and 'distance' as well. The nature of the value-sets for these frames will be addressed in Section 4 while the issue of motion-related meaning, an important aspect of SRR, is touched upon in the following section.

3. SRR in Natural Language

Research on space representation in natural language has focused primarily on space adverbials, typically realized as prepositional phrases, such as:

(1) Tom is in bed

(2) John walked into the trap

(3) The tree is by the house

Talmy (e.g., 1983) came up with the most complete representation of motion, an important aspect of SRR, to date, and it has been followed up to some extent by Herskovits (1987). Oriented primarily at linguistic description, the framework can be rendered in the frame notation roughly as follows:

\[
\text{(motion)}
\quad \text{('figure' \ object)}
\quad \text{('motion' \ action)}
\quad \text{('manner' \ action \ property)}
\quad \text{('path' \ location \ property \ | \ direction \ property)}
\quad \text{('ground' \ object)}
\]

Thus, in

(4) The bottle was floating in the cove

the figure is 'bottle,' the motion and manner are expressed by 'float,' the path is 'in,' and the figure is 'cove.' Generally and typically, the figure and ground are expressed by nouns, the motion by verbs, the manner by adverbs (often incorporated into the verb as in the example), and the path is indicated
by a preposition.

Even this one aspect of SRR as expressed in natural language clearly demonstrates its complexity. The other problems of NL-SRR are ambiguity, fuzziness, and extreme context sensitivity.

Most space prepositions double up in natural languages as temporal and abstract, such as:

(5) in the house in a minute in a hurry
(6) on the bed on completion on the way
(7) at the market at noon at ease

In conjunction with the ambiguity of other words, there are many sentences which are ambiguous, with one SRR-related meaning and the other(s) not, such as:

(8) Jim cut through the hedges

For NLP, this type of ambiguity is no different from any other and is treated in the same way. There can also be cases of SRR anaphora which need to be resolved. Thus, 'here' can mean 'on this chair,' 'at this table,' 'in this room,' 'in this house,' 'in this town,' etc.

More significantly, there is a considerable amount of indeterminacy, or fuzziness, in the way natural language express SRR in comparison with its precise geometry and topography. Thus, 'in the house' determines only a range of positions within the structure, and the actual object (or figure, in Talmy's terms) so located can be situated on any floor, in any room, and anywhere within each room. Moreover, 'in the house' or 'in the box' are different from 'in the city' because a city is 2D rather than 3D. Similarly, the geometric designations in natural languages are very crude and radically different from their scientific counterparts.

The context-sensitivity of SRR-related expressions in natural languages can be exemplified by the following examples:

(9) The piano is in the house

(10) Three people live in the house

(11) Real love can be found in the house

The same locative expression 'in the house' has a truly SRR-related meaning in the first sentence, partially SRR-related meaning in the second, and almost entirely non-SRR-related meaning in the third. This can be demonstrated by the different inferences from the three sentences. While the piano is confined to the house, the people who live in it are not, and the love is associated with the people and not with any physical structure.

Understanding, representation and generation of SRR in natural language is an important functionality for a number of knowledge-based systems, including intelligent (natural language, graphics, and mixed language/graphics) interfaces to vision and robotic and other reasoning systems, on the one hand, and the various text understanding, skimming and generation systems, on the other.
The resolution of ambiguity, anaphora, semantic complexity stemming from context-sensitivity have to be dealt with in both kinds systems but it is the former which seem to be more affected by fuzziness.

4. SRR in NLP

Extracting meaning from text, is a *sine qua non* of any NLP system, and the only way to make this complex task feasible is by adopting the sublanguage approach. Because all computer systems deal with a very constrained domain in manufacturing, communication, or military environment, most of the ambiguity issues simply do not arise. Thus, a specific domain would deal EITHER with physical motion OR with conversation, so that 'cutting through the hedges' would be immediately reduced to only one of its meaning. It has been demonstrated elsewhere (Nirenburg and Raskin, 1987) that the word 'operator' which can have up to 7 distinct meanings in English as a whole keeps only 1.5 of them in the English sublanguage of computer science.

SRR in NLIs for computer systems, in general, and for robotic vision, in particular, should - and easily can - be flexible enough to accommodate more or less information depending on the nature of the domain and the task. A robust NLI for a typical Navigator-type program for an IMAS should be able to interpret accurately a fuzzy command such as:

(12) *Fetch the big wrench from the shelfcase!*

It should be able to analyze 'fetch' roughly as:

1. 'move' from your current position to any spot near the shelfcase in question
2. look for the wrench, shelf after shelf, until you find it
3. grasp the wrench
4. bring it to me

It is important here to understand the precise "division of labor" between the NLI and the robotic vision system *per se*. The former will identify the robot as the addressee of the command; it will also provide a SRR frame, whose slots will be filled by the robotic vision system. For a typical navigating task, the frame will probably be similar to the ones introduced by Kuipers in a different context (1978). In the frame notation, such a frame will look roughly as follows:

(`'motion' go-to`

    (`'from' location1 [the robot's current position]`
    (`'to' location2 [near the operator]`
    (`'via' location3 [shelfcase, where the wrench is]`
    (`'purpose' move (wrench location2 location3)`)`

The semantic analysis of 'the big wrench' would typically involve:

- the recognition of a wrench-type shape
- the discovery of several such objects
- the comparison of their relative sizes
- the ability to choose the biggest one
Similarly, ‘the shelfcase’ involves the recognition of a unique shelfcase-shaped object and the realization of its internal structure, i.e., that it consists of more than one shape. The actual IMAS domain will determine what shapes populate the subworld, and the complexity semantic analysis in the NLI will be determined by this information. Such an analysis will be always much simpler than in "the whole world."

Shape recognition is not an NLI problem. However, there is an interesting development in contemporary semantics which can make the translation from natural language into the command language of a robotic vision system easier. Shape can be described geometrically, and various methods have been proposed for that (see, for instance, Kuipers, 1978; McDermott and Davis, 1984). It can also be described as related to a prototypical shape, as in the analysis above. Semantic analysis can be also performed in terms of "prototype semantics" (see, for instance, Lakoff, 1987, and references there), a theory which attempts to determine the meaning of each word in terms of its prototype and a distance from it.

An NLI to an IMAS does not have to contain commands triggering navigation programs or include route-planning, obstacle- and collision-avoidance, etc. If ‘fetch’ means ‘go from A to B,’ such triggering occurs automatically. Similarly, the fuzziness of SRR in natural language means simply that the interface must be able to translate the fuzzy spatial expression at the input into the grid-, map- or coordinate-related notation used in the IMAS. Whether the range usually expressed by a locative in natural language is translated into a range or a random value within this range in the IMAS is a matter of convenience and design of the robotic vision system itself. This problem is simply a particular aspect of a larger issue addressed in part by Waltz and his associates (see Section 2) of generating a "scene" from its verbal description which always underdetermines reality.

Conclusion

The main points of the paper are:

- SRR in a robotic vision system is seriously affected by SRR in natural language.
- SRR in natural language is a complex, multifaceted, and fuzzy phenomenon.
- Sublanguage- and meaning-based NLIs to robotic vision systems can easily accommodate all the SRR requirements of those systems.

Acknowledgments

The authors would like to thank John Brolio for the many useful discussions of the subject.
References


Forbus, K. 1981. A study of qualitative and geometric knowledge in reasoning about motion. TR 615. MIT AI Lab.


