Bridging the Gap between Language and Action

Tokunaga Takenobu[1], Koyama Tomofumi[1],
Saito Suguru[2], and Okumura Manabu[2]

[1] Department of Computer Science, Tokyo Institute of Technology
Tokyo Meguro Ôookayama 2-12-1, Japan 152-8552
{take@cl,tomoshit@img}.cs.titech.ac.jp
[2] Precision and Intelligence Laboratory, Tokyo Institute of Technology
Yokohama Midori Nagatsuta 4259, Japan 226-8503
{suguru,oku}@pi.titech.ac.jp

Abstract. When communicating with animated agents in a virtual space through natural language dialogue, it is necessary to deal with vagueness of language. To deal with vagueness, in particular vagueness of spatial relation, this paper proposes a new representation of locations. The representation is designed to have bilateral character, symbolic and numeric, in order to bridge the gap between the symbolic system (language processing) and the continuous system (animation generation). Through the implementation of a prototype system, the effectiveness of the proposed representation is evaluated.

1 Introduction

Research of animated agents capable of interacting with humans through natural language has drawn much attention in recent years [1–3]. When communicating with animated agents in a virtual space, it is necessary to deal with vagueness of language as well as ambiguity. Vagueness and ambiguity of language are similar but different concepts.

The following short conversation between a human (H) and a virtual agent (A) highlights the contrast between vagueness and ambiguity.

H: Do you see a ball in front of the desk?
A: Yes.
H: Put it on the desk.

In the third utterance, the pronoun "it" could refer to one of the objects mentioned in the preceding utterance, "a ball" or "the desk". So there is ambiguity in reference. Solving this kind of ambiguity has been studied for many years as anaphora resolution [5].

This example includes vagueness as well. When putting the ball on the desk, a location to place the ball should be decided. There is no explicit mention of the location on the desk where the ball to be placed. It is just mentioned as "*on* the desk". In contrast with the reference ambiguity, there is, in principle, infinite choices of the location. When we interact with virtual agents through natural

T. Rist et al. (Eds.): IVA 2003, LNAI 2792, pp. 127-135, 2003.

language, such kind of vagueness is inevitable. In particular, vagueness of spatial relations could be a crucial obstacle for autonomous agents, because the agent cannot perform a proper action without dealing with the vagueness.

As this example shows, solving ambiguity is a process of choosing a correct one from discrete and categorical choices, which has an affinity to the symbolic nature of language. On the other hand, solving vagueness involves finding a plausible point or area in continuous space, which is incompatible with the symbolic system. This would be one of the main reasons that vagueness has not drawn much attention in past natural language processing research.

Most of the past natural language dialogue systems worked in discrete space where every relation among objects and locations are described in terms of symbols. The discrete space has an affinity to conventional symbol-based planning which plays a crucial role in realizing intelligent agents. When moving from discrete space to continuous space, however, symbolic planning faces difficulty of vagueness. If it treats every location as a symbol, the number of symbols could be infinite in theory. To avoid this problem, Shinyama et al. proposed to use composite lambda functions to delay computation of locations [11]. However, since the result of computation are the coordinate values of a single location, their method does not deal with vagueness in the strict sense.

The above speculation tells us that the representation of locations and spatial relations for the virtual agents needs to have both symbolic and numeric character. With such representation, bridging the gap between a symbolic system (language) and a continuous system (action) could be achieved. Olivier et al. also proposed a similar idea in which the representation had both qualitative and quantitative properties [9]. However, their motivation was the visualization of spatial description and did not consider its use in a more dynamic environment. Horswill proposed a framework in which all logical variables are directly grounded on visual information in the real world [7]. It is not clear if this framework is applicable to general linguistic expressions. Our research is motivated to explore the spatial representation satisfying these requirements for intelligent agents.

The structure of the paper is as follows. Section 2 describes an overview of our prototype system with its architecture. Section 3 proposes the Space object which fulfills the above requirements. In Section 4, we show an example of how the Space object behaves in the planning process. Finally, Section 5 concludes the paper and looks at the future work.

2 System Architecture

To achieve the above goal, we are developing a prototype system $\mathcal{K}2$ as a test bed to evaluate our idea. Fig. 1 shows a screen shot of $\mathcal{K}2$. There are two agents and several objects (colored balls and desks) in a virtual world. Through speech input, a user can command the agents to manipulate the objects. The current system accepts simple Japanese utterances with anaphoric and elliptical expressions, such as "Walk to the desk.", "Further". The size of the lexicon is about

100 words. The agent's behavior and the subsequent changes in the virtual world are presented to the user in terms of a three-dimensional animation.

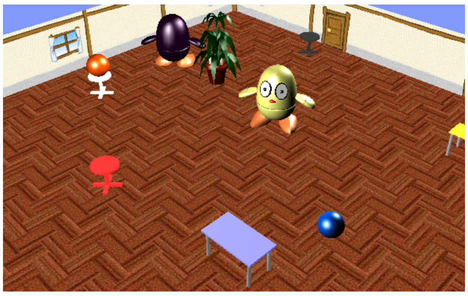


Fig. 1. A screen shot of $\mathcal{K}\mathit{2}$

Fig. 2 illustrates the architecture of the $\mathcal{K}\mathit{2}$ system. The speech recognition module receives the user's speech input and generates a sequence of words. The syntactic and semantic analysis modules analyze the word sequence to extract a case frame. At this stage, not all case slots are necessarily filled, because of ellipses in the utterance. Even in cases there is no ellipsis, instances of objects are not identified at this stage. Resolving ellipses and anaphora, and identifying the instances in the world are performed by the discourse analysis module.

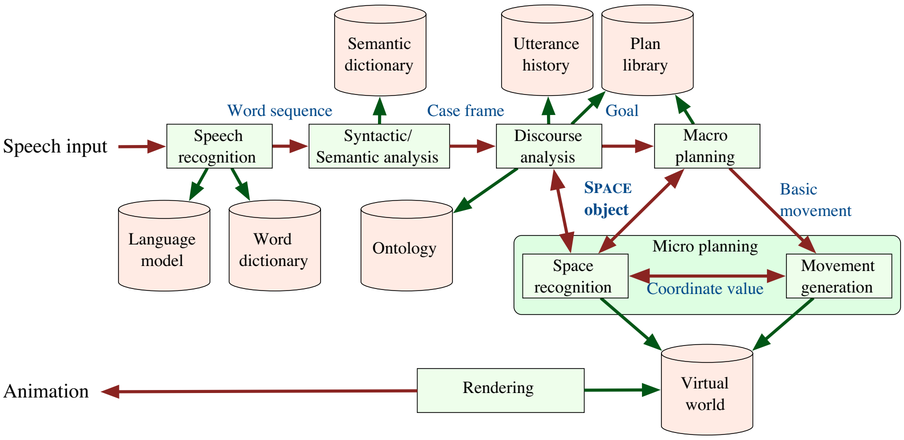


Fig. 2. The system architecture of $\mathcal{K}\mathit{2}$

The discourse analysis module extracts the user's goal as well and hands it over to the planning modules which build a plan to generate the appropriate animation. In other words, the planning modules translate the user's goal into animation data. However, the properties of these two ends are very different and straightforward translation is rather difficult. The user's goal is represented in term of symbols, while the animation data is a sequence of numeric values. To bridge this gap, we take a two-stage approach – macro and micro planning.

The macro planner adopts a conventional planning framework like STRIPS [4], that is, given a goal, it generates a sequence of predefined primitive operators. In this case, the planner generates a sequence of basic movements of the agent. For instance, the goal "on(ball#1, desk#2)" would be satisfied by a sequence of basic movements, "go near to the ball", "pick up the ball" and "put the ball on the desk". It is generally difficult to define a set of basic movements, since it depends on the application domain. One approach to this issue is described in [12].

During the macro planning, the planner happens to need to know the physical properties of objects, such as their size, location and so on. For example, to pick up a ball, the agent first needs to move to the location at which he can reach the ball. In this planning process, the distance between the ball and the agent needs to be calculated. This sort of information is represented in terms of coordinate values of the virtual space and they are handled by the micro planner.

The results of micro planning update the virtual world database, and the update reflects the output animation through the rendering module.

3 The Space Object

To interface the macro and micro planning, we propose the Space object to represent a location in the virtual space, with its bilateral character; symbolic and numeric. To realize such bilateral character, the following two requirements arise for the Space object describing a location.

R1. It can be an argument of logical functions.
R2. It can represent plausibility of a location.

The requirement **R1.** comes from the macro planner side. The macro planner uses plan operators described in terms of logical forms, in which a location is described such as InFrontOf(Obj). Such representation needs to be an argument of another logical function. From the viewpoint of the macro planner, the Space object is designed to behave as a symbolic object by referring to its unique identifier.

The requirement **R2.** comes from the micro planner side. A location could have vagueness and the most plausible place changes depending on the situation. Therefore it should be treated as a certain region rather than a single point. To fulfill this requirement, we adopt the idea of the potential model proposed by Yamada et al. [13], in which a potential function maps a location to its plausibility. Vagueness of a location is naturally realized as a potential function embedded in the Space object.

We design the potential function f to satisfy the following two conditions.

C1. It is differentiable throughout the domain.
C2. It moves range 0 to 1.

When the most plausible point is required by the micro planner for generating the animation, the point is calculated by using the potential function with the Steepest Descent Method (SDM). The condition **C1.** is necessary to adopt the SDM.

In the condition **C2.**, the point with value 1 is defined as the most plausible location and that with value 0 is the least plausible one. This definition makes it possible to translate the logical AND operation on the Space objects to the product of the potential function values of the objects. The NOT operation is defined as $(1-f)$ and the OR operation can be derived from the combination of the AND and NOT operations. As this example shows, relations between objects and locations are represented as symbols in the macro planner, and as composition of potential functions in the micro planner.

Currently, we have defined the potential functions for the following spatial concepts:

- relations "front", "back", "left", "right", "on", "between"
- a place occupied by an object
- a place close to an object

The parameters of a potential function are derived from the size and shape of objects.

We can also use the Space object to represent more complex spatial constraints such as a reachableByHand(Agent) location. In this case the potential function reflects the stress on the agent's arm.

4 Example of Planning with the Space Object

This section describes how the Space object plays a role as a mediator between the symbolic and continuous system through the example introduced in Section 1.

When an utterance "Do you see a ball in front of the desk?" is given in the situation shown in Fig. 1, the discourse analysis module identifies an instance of "a ball" in the following steps.

```
(1) space#1 := new inFrontOf(desk#1, viewpoint#1, MIRROR)
(2) list#1 := space#1.findObjects()
(3) ball#1 := list#1.getFirstMatch(kindOf(BALL))
```

In step (1), an instance of Space is created as an instance of the class inFrontOf. The constructor of inFrontOf takes three arguments; the reference object, the viewpoint and the axis order [3] [6]. There have been several studies on the classification of spatial reference [6, 10, 8]. In this paper, we follow Herskovits's formulation [6] due to its simplicity, in which a reference frame is determined in terms of the above three parameters.

[3] There are two types of axis order, basic and mirror. In the basic order, the axes are ordered clockwise around the origin as "front", "right", "back" and "left". In the mirror order, however, the order is "front", "left", "back" and "right". The mirror order is used when the speaker faces an object.

To interpret a speaker's utterance correctly, it is necessary to identify the reference frame which the speaker used. However, there is no decisive method to accomplish this, since various factors, such as object properties, preceding discourse context and psychological factors are involved. The current prototype system adopts a naive algorithm to determine the reference frame based on heuristic rules. In this paper, we focus on the calculation of potential functions given a reference frame.

Suppose the parameters of inFrontOf have been resolved in the preceding steps, and the discourse analysis module chose the axis mirror order and the orientation of the axis based on the viewpoint as the light-colored arrows in Fig. 3. While the desk has four potential direction, (1) through (4), only one of them can be the "front" axis of the desk. The closest one to the viewpoint-based "front" axis is chosen as the "front" of the desk. In this example, (1) is chosen. Then, the parameters of potential function f corresponding to "front" are set as shown in Fig. 4.

The potential function f is defined as given in equation (1). The first Gaussian factor expresses the expanse of the potential on both sides of the "front" axis, the second sigmoid factor reduces the potential of the "back" side region. Since these two factors satisfy the conditions described in Section 3, f also satisfies them.

$$f(d_1, d_2) = \exp^{\frac{-d_2^2}{b^2(d_1^2 - \frac{l_2}{2} + \frac{l_1}{2})^2}} \times \frac{1}{1 + \exp^{-ad_1}} \qquad (1)$$

d_1 : value of the "front" axis
d_2 : value of the "left-right" axis
l_1, l_2 : maximum length of the reference object along axes d_1, d_2
a, b : coefficient

In step (2), the method matchObjects() returns a list of objects located in the potential field of space#1 shown in Fig. 5. The objects in the list is sorted in descending order of the potential value of its location.

In step (3), the most plausible object satisfying the type constraint (BALL) is selected by the method getFirstMatch().

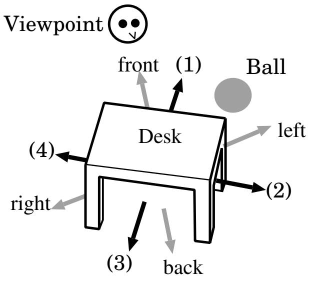


Fig. 3. Adjustment of axis

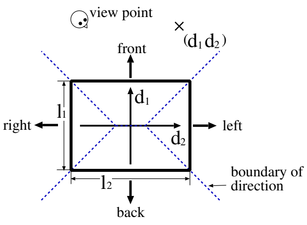


Fig. 4. Parameters of f

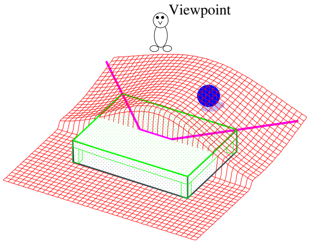


Fig. 5. Potential field of f

When receiving the next utterance “Put it on the desk.”, the discourse analysis module resolves the referent of the pronoun “it” and extract the user’s goal. The macro planner constructs a plan to satisfy the goal as follows:

(1) walk(inFrontOf(ball#1, viewpoint#1, MIRROR) AND
    reachableByHand(ball#1) AND NOT(occupied(ball#1)))
(2) grasp(ball#1)
(3) put(ball#1,on(desk#1, viewpoint#1, MIRROR)) [4]

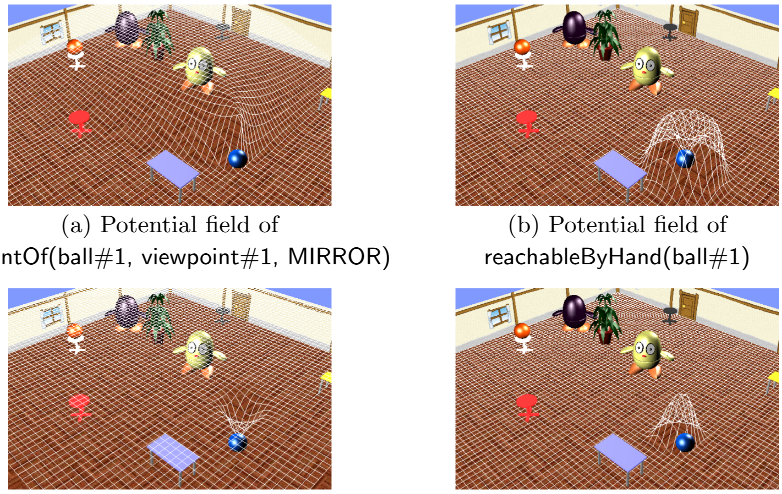
Walk, grasp and put are defined as basic movements. They are handed over to the micro planner one by one.

The movement walk takes a Space object representing its destination as an argument. In this example, the conjunction of three Space objects is given as the argument. The potential function of the resultant Space is calculated by multiplying the values of corresponding three potential functions at each point. Fig. 6 illustrates the three potential fields (a) through (c) and the resultant field (d). The agent walks to the location which has the maximum potential value with respect to the field (d).

After moving to the specified location, the movement grasp is performed to grab the ball#1. This movement should succeed because the agent is guaranteed to be at the location from which the ball is reachable.

When putting the ball on the desk, the micro planner looks for a space on the desk which no other object occupies by composing the potential functions similar to the walk step.

As this example illustrates, the Space object effectively plays a role as a mediator between the macro and micro planning.



(a) Potential field of inFrontOf(ball#1, viewpoint#1, MIRROR)

(b) Potential field of reachableByHand(ball#1)

(c) Potential field of NOT(occupied(ball#1))

(d) Composition of (a) – (c)

Fig. 6. Composition of potential fields

[4] Actually, further constraints are necessary to ensure enough room for the ball.

5 Conclusion

This paper proposed a representation of a location in the virtual world. The proposed representation, the SPACE object is designed to have bilateral character in order to bridge the gap between the symbolic system (language processing) and the continuous system (animation generation). Through the implementation of the prototype system $\mathcal{K}2$ in which a user can interact with animated agents in the virtual world, we found the SPACE object is a promising candidate to deal with vagueness of language.

Our future research plan includes increasing the number of spatial relations and utilizing potential fields for path planning in the micro planner. For example, introducing a potential field decreasing the value along with the orthogonal direction of the wall makes it possible to deal with an expression like "Walk along the wall to the big desk.". In addition, more principled algorithm to disambiguate the reference frame should be incorporated into the system.

References

1. N. I. Badler, M. S. Palmer, and R. Bindinganavale. Animation control for realtime visual humans. *Communication of the ACM*, 42(8):65–73, 1999.
2. R. Bindinganavale, W. Schuler, J. Allbeck, N. Badler, A. Joshi, and M. Palmer. Dynamically altering agent behaviors using natural language instructions. In *Autonomous Agents 2000*, pages 293–300, 2000.
3. J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, editors. *Embodied Conversational Agents*. The MIT Press, 2000.
4. R. E. Fikes. STRIPS: A new approach to the application of theorem problem solving. *Artificial Intelligence*, 2:189–208, 1971.
5. B. J. Grosz, A. K. Joshi, and P. Weinstein. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–226, 1995.
6. A. Herskovits. *Language and Spatial Cognition. An Interdisciplinary Study of the Prepositions in English*. Cambridge University Press, 1986.
7. I. D. Horswill. Visual routines and visual search. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, August 1995.
8. W.J.M. Levelt. *Speaking: From Intention to Articulation*. The MIT Press, 1989.
9. P. Olivier, T. Maeda, and J. Tsujii. Automatic depiction of spatial descriptions. In *AAAI 94*, pages 1405–1410, 1994.
10. G. Retsz-Schmidt. Various views on spatial prepositions. *AI Magazine*, 9(2):95–105, 1988.
11. Y. Shinyama, T. Tokunaga, and H. Tanaka. Processing of 3-D spatial relations for virtual agents acting on natural language instructions. In *the Second Workshop on Intelligent Virtual Agents*, pages 67–78, 1999.
12. T. Tokunaga, M. Okumura, S. Saitô, and H. Tanaka. Constructing a lexicon of action. In *the 3rd International Conference on Language Resources and Evaluation (LREC)*, pages 172–175, 2002.
13. A. Yamada, T. Nishida, and S. Doshita. Figuring out most plausible interpretation from spatial description. In *the 12th International Conference on Computational Linguistics (COLING)*, pages 764–769, 1988.