A Hybrid Neural Network and Virtual Reality System for Spatial Language Processing

Guillermina C. Martinez[1], Angelo Cangelosi[1], Kenny R. Coventry[2]

[1]School of Computing and [2]Department of Psychology
University of Plymouth
Drake Circus, Plymouth PL4 8AA, UK
guille@usitmail.com, acangelosi@plymouth.ac.uk, kcoventry@plymouth.ac.uk

Abstract

This paper describes a neural network model for the study of spatial language. It deals with both geometric and functional variables, which have been shown to play an important role in the comprehension of spatial prepositions. The network is integrated with a virtual reality interface for the direct manipulation of geometric and functional factors. The training uses experimental stimuli and data. Results show that the networks reach low training and generalization errors. Cluster analyses of hidden activation show that stimuli primarily group according to extra-geometrical variables.

1 Introduction

The aim of this work is to develop a hybrid neural network (NN) and virtual reality (VR) system for the study of spatial language and cognition. It will also be tested as a prototype natural language interface for virtual environments.

Spatial language, and in particular the use and understanding of spatial terms such as *over*, *above*, *under*, and *below*, has proven to be an important experimental field for the investigation of cognition [3,13]. The use of an expression involving a spatial preposition in English conveys to a hearer where one object (figure) is located in relation to a reference object (ground). Understanding the meaning of spatial prepositions is of particular importance in semantics as they are among the set of closed class terms which are generally regarded as having the role of acting as organizing structure for further conceptual material [14]. Recently, both experimental research and computational models have investigated the use of spatial prepositions, and their role in spatial cognition.

1.1 Psychological Literature on Spatial Language and Function

In the experimental psychological literature it has been shown that both geometric (e.g., the distance between two objects and their relative orientation) and extra-geometric variables (e.g., the function of an object and its size and shape) play an important role in the comprehension of spatial prepositions.

Traditionally, geometric constructs have been invoked to underpin prepositions' lexical entries (e.g., [10,11]). For example, in the sentence, "The pear is *in* the bowl," the figure (the pear) is located in the region described by the prepositional phrase "*in* the bowl", with the spatial relation expressed by *in* corresponding to "contained interior to the reference object."

Clearly, while geometry is important in the use and comprehension of spatial prepositions, other extra-geometric variables need to be invoked in order to account for use and comprehension. For example the expression, *the man is at the piano*, implies that the man is playing the piano, not just that he is in close proximity to it. There have been a number of empirical demonstrations showing that extra-geometric factors play an important role in the use and comprehension of spatial prepositions. Functional relations have been postulated as key components underlying the meaning of the spatial prepositions *in*, *on* and *at* [1, 3, 4].

Functional relations have to do with how objects interact with each other, and what the functions of objects are. For example, with *in*, Garrod and Sanford [7] and Coventry [3] propose that the lexical entry is: in [functional containment - in is appropriate if the ground is conceived of as fulfilling its containment function]. Whether or not in is appropriate depends on a number of factors which determine whether the container is fulfilling its function. Empirical evidence for the importance of this functional analysis has been forthcoming for topological prepositions.

It has also recently been shown that prepositions are influenced differentially by geometric and extra-geometric variables. Coventry, Prat-Sala and Richards [5] found that the comprehension of *over* and *under* was more influenced by function than *above* and *below*, while the comprehension of *above* and *below* was better predicted by geometry than *over* and *under*. In addition, effects of extra-geometric variables have been shown to influence use and comprehension even when the prototypical geometric constraint holds. For example, they found that appropriateness ratings of expressions such as *the umbrella*

is over the man to describe a picture of a man holding an umbrella were reduced when rain was depicted as falling on the man even when the umbrella was depicted directly above the man's head [5].

1.2 Neural Network Models of Spatial Language

There is some computational work that has modeled the acquisition and use of spatial terms using neural networks with a psychologically and linguistically plausible approach. Harris [9] used neural networks to model the polysemy of the preposition *over*, that is the fact that the term *over* appears to have many different senses, such as "being above", "up", "across", etc. Harris's model used feedforward neural networks trained through back propagation to learn to associate the correct meaning of *over* with different sentences. All input sentences contained the term *over* to relate the position of a figure object with respect to a ground object. After learning the correct mapping of the meanings of *over*, the activity of some of the hidden units auto-organizes in a way that units become sensitive to certain features of the object set used in the training sentences. There are units whose activation distinguishes between objects which are or are not normally in contact with a surface, and other units that are sensitive to the size and shape of the objects.

The model introduces the problem of polysemy and openness of the meaning of some spatial terms [9]. It shows the emergence of the role of object-knowledge effects for spatial language using auto-organization systems, such as neural networks. However, this work lacks any reference to the role of geometrical features in the learning and use of spatial prepositions. The encoding of input in only linguistic terms does not allow any processing of geometrical properties between objects. The neural network model is subject to the problem of symbol grounding in cognitively plausible models [8].

Terry Regier [12] has proposed a computational model for spatial prepositions using a method called "constrained connectionism" [6]. The model is trained on the use of various spatial prepositions for static (e.g. *over* and *above*) and moving (e.g. *through*) objects, and makes explicit use of the processing of geometrical information. The model consists of a complex neural network in which the units' layers and connection patterns are structured according to neuropsychological and cognitive evidence; only a few units are based on unstructured parallel distributed processing. An image of two objects (ground and figure) is input to the lower layer of the network. Then the image goes through several levels of geometrical processing. The output units, corresponding to spatial prepositions, are activated according to the geometrical position of the figure object with respect to the central ground. Regier [12] tested this model for various cognitive and cross-linguistic spatial language phenomena. For example, the model proved suitable for reproducing the experimental data of Logan & Sadler's [11] spatial templates for the prepositions *over, above, under* and *below.*

The Regier model, even though it is able to reproduce many of the experimental and cross-linguistic data on the use and learning of spatial terms, has the limitations of relying only on geometrical-based processing and only deals with abstract objects. The network uses different geometrical indices, such as the center of mass between the two objects, their minimal distance, and the overlapping of their shapes. Although the use of these geometric components does allow the system to deal with change over time, no other information is extracted and used, such as that of the objects' functionality.

Recently, a new computational model for spatial language has been proposed by Regier & Carlson [13]. This does not use connectionist techniques. It is based both on attentional factors on the processing of geometrical features of abstract objects.

2 Method

The prototype of a hybrid NN and VR system has been developed. The NN learns to use spatial prepositions in response to input stimuli describing geometrical and functional relationships between two objects. The NN module is integrated with a VR interface, where a user can directly manipulate geometric and extra-geometric factors. This system can be used as an experimental tool for spatial language and for natural language interfacing in VR environments.

2.1 Neural Network

The NN architecture consists of a multi-layer perceptron. The input layer receives information about a visual scene depicting specific spatial configurations of objects. The output units activate the correct spatial preposition(s) describing the scene. The network has four output units, respectively for the prepositions *over*, *above*, *under* and *below*. The activation of each unit corresponds to the level of agreement for the use of a specific term. After training, the activation must correspond to the subjective ratings collected in experimental studies. The hidden layer contains five units, a number sufficient for the network to learn the training data. The number of input units varies according to the explicit/implicit encoding of some of the properties of the objects and the scene.

The training and testing task utilize the stimuli and data from an experiment on the role of functional factors in the rating of the spatial prepositions *over/above/under/below* (experiment 2 in [5]). In this study, subjects used a 7-point Likert scale to rate the use of the four spatial prepositions for 72 scenes. A scene always depicted a man holding/wearing an object (e.g. umbrella, visor) to protect himself from another object (e.g., rain, spray). In this experiment four independent variables were manipulated: ORIENTATION of the protecting object (3 levels: an umbrella

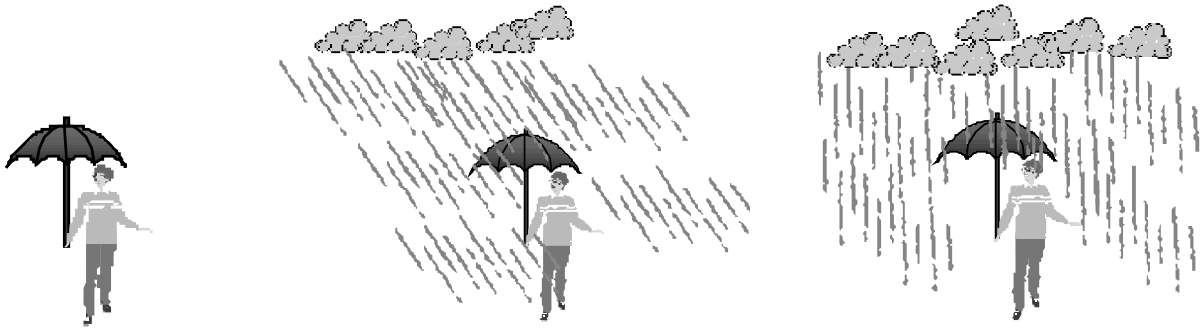

Figure 1: Examples of experimental conditions in the second experiment of Coventry et al. [5]. The three scenes differ in the level of variable FUNCTION. In the control condition (left) there is not rain, in the non-functional condition (center) the umbrella does not protect the man from the rain, and in the functional condition (right) the umbrella is fulfilling its function of protection the man from the rain.

can be rotated at 90, 45, and 0 degrees) FUNCTION fulfillment of protection from the rain (3 levels: yes, no, control), APPROPRIATENESS of object for protection function, e.g. umbrella or suitcase (2 levels: yes, no) and OBJECT type (4 levels). This results in 72 experimental scenes/conditions. An example of three scenes is presented in Figure 1. The scenes differ in the level of the variable FUNCTION.

Three network architectures are used. They only differ in the number of input units and the way input scenes are encoded. The five hidden units and the four output units are the same in all networks (Figure 2).

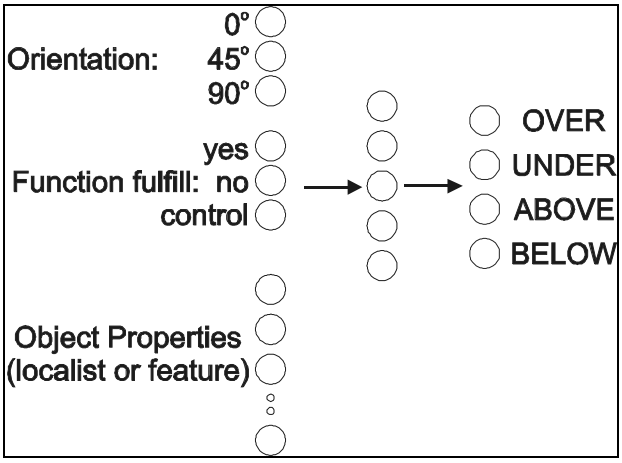

Figure 2: Neural network architecture

Network A: Localist experiment encoding

In this network, the number of input units exactly reflects the number and levels of the four experimental variables. This architecture has a total of 12 localist input units. We use the term localist to indicate that for each variable only one unit is active.

Three input units are used to encode the three levels of ORIENTATION of the protecting object. Three localist units are used for the three levels of the FUNCTION independent variable. Two units encode the levels of APPROPRIATENESS, and four units the types of OBJECT.

Network B: Localist Object Encoding

This network does not have an explicit representation of the object appropriateness, because eight localist units are used to represent all objects. There are also three localist units for ORIENTATION and three for FUNCTION. This architecture has a total of 14 input units.

Network C: Feature-based Object Encoding

In this network the objects are encoded according to their geometrical and functional features. Each object is represented using eight feature-based units. Three units encode the dimension of the object in the three dimensions (x, y, z) and three encode the major shape components (hemispherical, conical, cuboid). Two units refer to the lexicalized function of the object (i.e. APPROPRIATENESS). For example, the object umbrella is encoded as x=1, y=1, z=.67, hemispherical=1, conical=0, cuboid=0, appropriate=1, inappropriate=0.

There are three localist units for ORIENTATION and three for FUNCTION. This architecture has a total of 14 input units.

Training

A standard error backpropagation algorithm was used, with a learning rate of .01, momentum of .9 and 10000 epochs. Of the total of 72 scenes, 71 were used for each training epoch, and 1 for the generalization test. The training of each network type A/B/C was replicated ten times, by varying the initial random weights and the stimulus randomly taken out for the generalization test.

The subjects' mean ratings for the use of the four prepositions were normalized in the range 0-1 and were used as teaching input for the backpropagation training.

2.2 Virtual Reality Environment

The VR module consists of an interface for the manipulation of 3D objects in the scene. For example, in the umbrella scene there are three objects that the user can manipulate: the man, the protecting object (e.g. umbrella or

suitcase), and the rain. For the protecting objects, the user can edit some of their features, such as the size and rotation. The program starts by showing an almost full-screen window with eleven buttons and displays a man with his right hand up. This man is rotated 60 degrees around his Y-axis. The user can then display/hide an object and edit its features. Once all the attributes are ready, the user can click on the "NNAnswer" button to ask the NN module to provide the rating for the four prepositions (Figure 3).

Figure 3: Interface of the VR system. The user can choose the protecting object to display and edit its features. After the NN processes the scene, the ratings for the four spatial prepositions are shown in the bottom right corner of the interface.

This VR module was developed in Java using Borland's Builder Java3D library. Through the Java3D API is possible to create simple virtual reality worlds. The Java program also controlled the communication with the NN module running in Mathlab.

3 Results

3.1 Training and generalization

The training task was relatively easy to learn for a multiplayer perceptron, mainly due to the limited set of training data (71 training stimuli). The final error for all different architectures resulted in an average SSE 0.05. The networks were also able to generalize well to the stimulus taken out from the training set. The average generalization error for all architectures was 0.04. Table 1 reports the detailed average errors for each architecture. The results are similar in the three conditions, with a tendency for the feature-based object encoding network to reach lower training error.

The whole VR and NN system was also successfully tested. After manipulating the properties of objects in the VR interface, the network produced the correct rating for each preposition that were passed back to the VR interface and shown to the user.

Table 1: Average training and generalization errors for the three network architectures.

SSE error	*Net A: Localist experiment*	*Net B: Localist object*	*Net C: Feature object*
Training	0.051	0.055	0.046
Generalization	0.041	0.046	0.044

3.2 Analysis of Internal Representations

To understand the way geometrical and extra-geometrical factors are processed by the networks, a cluster analysis of the hidden activation was performed. This informs us about the major criteria used by the network to perform the spatial language task. A greater distance between clusters indicates which variables are used first to process (i.e. separate) stimuli and experimental conditions.

For each of the three network architectures, we chose the five out of the ten replications with the best learning performance. The connection weights of the fifteen selected networks after epoch 10000 were used to calculate the hidden activation. The activation values of the five hidden units for each of the 72 input scenes were saved and used to perform a cluster analysis. Subsequently, we studied the cluster diagrams to identify the order in which some functional and/or geometrical factors are used to separate clusters of experimental scenes. Although there was variability between the five cluster analyses of each architecture, it was possible to identify some common clustering strategies for each condition.

Diagrams of network A

With the experiment encoding architecture there are three diagrams that share the use of common and consistent clustering criteria. In these networks, clusters are created early according to a geometrical factor, i.e. the ORIENTATION variable. The first divisions group input scenes according to the degree of rotation (0, 45, 90) of the protecting object. The second consistent clustering criterion groups scenes according to the type of objects falling on the man (e.g. rain or spray). In the fourth diagram, the early clustering criteria are a mix of the FUNCTION fulfillment and the ORIENTATION variables. The fifth diagram does not have an identifiable clustering criterion.

Diagrams of network B

In the five diagrams for the architecture with localist object encoding, the early divisions into clusters are determined by the variables ORIENTATION and by that of the falling object. There is not clear and consistent prioritization of these two factors.

Diagrams of network C

The condition with feature-based encoding of objects has four diagrams that share the same clustering criteria. The

first factor determining the early clusters is the APPROPRIATENESS of objects for the protection function. Secondly, the clusters are then subdivided according to the type of falling objects. Thirdly, scenes group into clusters that have similar dimensions or shape components. Figure 4 shows a cluster diagram for this condition. A major difference between this condition and the other two is that up to the last level of clustering the appropriate and inappropriate objects are always kept separate. In networks A and B only at the level of the final clusters the two objects are separated. Finally, one cluster out of five uses an unclear and inconsistent grouping strategy.

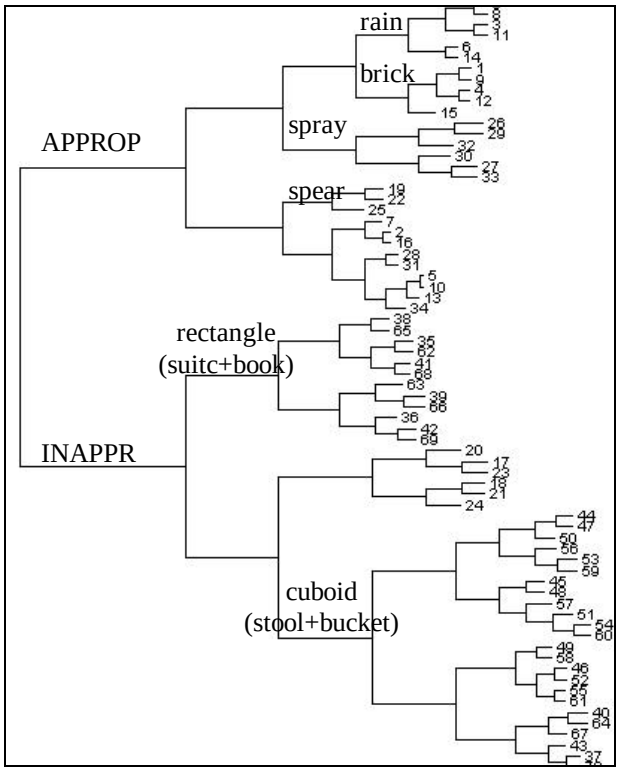

Figure 4: Cluster analysis diagram of the hidden units' activation of a network of condition C (feature-based encoding). Input stimuli first group according to their function (i.e., approriateness/inappropriateness), and subsequently according to the type of falling objects and the similarity of the shape components. Pure geometrical factors such as the orientation of the protecting object are ignored in the early stages of processing.

Overall, the results of hidden activation clustering show that with architectures using localist encodings (networks A and B), geometrical factors such as the orientation of the protecting object prevail. When an explicit encoding of extra-geometrical factors is used, as with architecture C, the stimuli tend to primarily group according to variables related to the function of objects. Most of these extra-geometrical variables, such as the object's lexical functional appropriateness and its size, have been proven to greatly affect the use and comprehension of spatial terms [2]. Therefore, the explicit encoding of objects' extra-geometrical properties (e.g. through feature-based input unit of network C) and its subsequent effect on the network processing strategies seem to more adequately reflect the phenomena observed in experimental subjects. This better match between the network and experimental data favors the use of such a type of architecture for the further development of a computational model of spatial language and cognition.

4 Conclusion

This hybrid NN and VR system allowed us to model the effects of functional and geometrical factors on the comprehension of spatial prepositions. Moreover, it provides a prototype NLP interface for interactive VR applications.

Further research is being conducted in order to develop a psychologically plausible neural network model for the processing of spatial language. The current prototype model shows the importance of explicitly encoding and inputting the extra-geometrical features of objects, as well as their geometrical properties. However, the use of a pre-defined set of functional features and its distributed and explicit encoding in the input units is not yet satisfactory. A computational model of spatial language and cognition should be able to derive, on-demand, and use the right set of properties that are salient to the scene and its context. This is the direction that we are following in our on-going research.

References

[1] Carlson-Radvansky, L.A. & Radvansky, G.A. (1996). The influence of functional relations on spatial term selection. *Psychological Science*, 7(1), 56-60.

[2] Coventry, K.R. (in submission). Spatial prepositions and the instantiation of object knowledge: the case of 'over', 'under', 'above' and 'below'.

[3] Coventry, K.R. (1998). Spatial prepositions, functional relations and lexical specification. In P. Olivier and K. Gapp (Eds.), *The Representation and Processing of Spatial Expressions*, pp247-262. Lawrence Erlbaum Associates.

[5] Coventry. K.R., Carmichael, R. & Garrod, S.C. (1994). Spatial prepositions, functional relations and task requirements. *Journal of Semantics, 11,* 289-309.

[4] Coventry, K.R., Prat-Sala, M. & Richards, L. (2001). The interplay between geometry and function in the comprehension of 'over', 'under', 'above' and 'below'. *Journal of Memory and Language*, 44, 376-398.

[6] Feldman J., Fanty M. & Goddard N. (1988). Computing with structured neural networks. *IEEE Computer*, 21, 91-104.

[7] Garrod, S.C. & Sanford, A.J. (1989). Discourse models as interfaces between language and the spatial world. *Journal of Semantics*, 6, 147-160.

[8] Harnad S. (1990). The Symbol Grounding Problem. *Physica D*, 42, 335-346

[9] Harris C. (1990). Connectionism and cognitive linguistics. *Connection Science*, 2(l), 7-33.

[10] Herskovits, A. (1986). *Language and spatial cognition.* Cambridge University Press.

[11] Logan, G.D. & Sadler, D.D. (1996). A computational analysis of the apprehension of spatial relations. In P. Bloom, M. A. Peterson, L. Nadel, & M. Garrett (Eds.), *Language and Space*, pp 493-529. Cambridge, MA: MIT Press.

[12] Regier, T. (1996). *The human semantic potential: Spatial language and constrained connectionism.* Cambridge, MA: MIT Press.

[13] Regier, T. & Carlson, L.A. (in press). Grounding spatial language in perception: An empirical and computational investigation. *Journal of Experimental Psychology: General.*

[14] Talmy, L. (1983). How language structures space. In H. Pick & L. Acredolo (Eds.), *Spatial orientation: Theory, research and application* (pp. 225-282). New York: Plenum.