Learning Relational Correlations

Michael Gasser (Gasser@Indiana.Edu)
Department of Computer Science; Indiana University
Bloomington, IN 47405 USA

Eliana Colunga (EColunga@Indiana.Edu)
Department of Computer Science; Indiana University
Bloomington, IN 47405 USA

Abstract

A conventional view of object categories is that they represent correlations among sets of object features. In this paper we present an analogous view of relational categories, the Micro-Relation Theory. On this view, relational categories such as ON and HIT are built up out of correlations among primitive relational features, which we call **micro-relations**. The process of learning relational categories involves three phases, the learning of the micro-relations within object dimensions, the learning of correlations between the micro-relations across dimensions, and the generalization from absolute to relative relations within dimensions. This paper focuses on the first two phases. We describe an experiment demonstrating the first phase of relational learning and a neural network simulation of the experiment. We conclude with a discussion of future work on the second and third phases of relational learning predicted by the theory.

Grounding Object Categories

For an animal, an object is a cognitive achievement, the outcome of a process that segments sensory/perceptual input into regions as it attends to the sensory/perceptual dimensions that have proven useful in making predictions, dimensions such as color, texture, size, and (the multiple dimensions that make up) shape. An object can be seen in part as the co-occurrence of a set of values on these dimensions, that is, as a feature vector.

Two sorts of generalizations about objects are possible. One singles out a range of values along a single object dimension, treating all objects with that feature as belonging to a one category for some purpose. Such generalizations are often realized in natural language as adjectives (Gasser & Smith, 1998). For example, the object category RED groups together all objects with a particular range of values on the COLOR dimension. The generalization is that there is a set of objects of this type. Typically more useful is a second type of generalization, based on the discovery that certain features co-occur regularly. Categories in this second sense are bundles of correlations of values along different object dimensions. We will refer to these as **object feature correlations**; they usually take the form of nouns in natural language. For example, the category APPLE is characterized by a particular shape, size, taste, smell, and texture. Categories of this type are valuable because of the inferences (predictions) they permit; given a subset of the associated features, the system can predict values on the other, correlating dimensions.

This view of objects and object categories is straightforward to implement within a simple connectionist model. Each processing unit represents a range of values along a particular dimension. A pattern of activation across the units, that is, a vector of activation values, represents an object. That is, rather than being represented by an atomic symbol, an object is **distributed** across a set of **micro-features**. Categories can be learned through unsupervised Hebbian learning, which strengthens the weights on connections between units which are co-activated. The weights then represent correlations across dimensions, the basis of correlational categories such as APPLE. Thus the matrix of weights after learning encompasses all correlational categories, and each category can be seen as a subset of the units which are mutually excitatory. Activation of some of these units causes others to be activated or inhibited, representing inference or prediction.

Grounding Relational Categories

Now let's consider how we can similarly view relations in terms of co-occurring features and correlations. Starting again with object feature dimensions, a (binary) relation instance is the co-occurrence of **pairs** of feature values on the object dimensions. Note that some dimensions which may not be relevant for the identification of the objects because they tend not to be stable *are* relevant for relations, in particular, the location of the objects.

As with objects, there appear to be two sorts of relational categories. One type is defined over a single dimension. Analogous to RED for objects, a particular **pair** of values for two objects along a single dimension may characterize a class of relation instances in the world. For example, in some environment it may be the case that red and blue objects tend to occur together. The co-occurrence of red with blue objects is an example of a primitive **micro-relation**. Because a micro-relation is relational, it already involves a correlation, a correlation between values in two ranges along a single dimension such as COLOR or SIZE. Thus we will also refer to micro-relations as **simple relational correlations**. Note that unlike primitive object features, micro-relations are not normally labeled. There is no word in English, for example, for the situation in which a red and a blue object co-occur. Instead labels are often applied to an elaboration of a micro-relation, a **relative relational correlation**. For such a correlation, the object dimension in question must be ordinal (for example, SIZE or DARKNESS rather than COLOR or SHAPE), and the relation applies to multiple pairs of values across the dimension. Examples are DARKER and SAME SIZE.

Analogous to object feature correlations, we have a second type of relational category: correlations between relational features on different dimensions. Thus a LOCATION relational feature may co-occur with a SIZE relational feature, for example, if an object of a particular size is on top an object of another size. We refer to these as **complex relational correlations**. When such relational feature correlations remain specific to particular values or ranges of values on the two dimensions, they normally do not have associated linguistic labels. But when the relations along one or both dimensions become relative (HIGHER, SMALLER), we often do. Thus for the relation we call *sunset* in English several dimensions seem to be relevant, including the movement of the sun with respect to the horizon, the changing color and apparent shape of the sun, and the changing color of the western sky. An extremely important class of categories involving complex relational correlations consists of relational terms, such as the word *on*. These represent correlations between syntactic patterns, in particular the relative position of the noun phrase arguments of the relational term and semantic dimensions such as the relative position of the referents of the noun phrases. Thus the spatial relation ON correlates with the syntactic pattern associated with the English preposition *on*. Note that for complex relational correlations, it is necessary to specify which of the arguments in one relation corresponds to which of the arguments in the other relation. Thus it is the first of the two noun phrases in a phrase like *the book on the table* which refers to the upper object and the second which refers to the lower. (This is the main way in which *on* differs from *under*.)

The next section spells out our claims about the three phases in the learning of relations and describes a connectionist implementation of the model.

Micro-Relation Theory

Phases in Learning Relations

Our main claim is that relations are built up out of micro-relations, associations between specific features on two object dimensions, and that the most important and easy-to-learn relations involve complex relational correlations (between micro-relations). There are three phases on the way to full-blown relations.

1. The micro-relations themselves must be built up. In the neural network implementation of the model each is represented by a unit with initially weak weights from the units representing the two object features. These weights are strengthened if the unit is activated in response to input patterns containing the associated object values.
2. Once the micro-relation unit is sufficiently activated by object feature input, it can be associated with another micro-relation unit on another dimension, representing a complex relational correlation.
3. Once the system comes to explicitly represent ordering within an object feature dimension such as LENGTH, it becomes possible to learn relative relational correlations such as LONGER.

Phase 1 must precede phase 2 because complex relational correlations are formed micro-relations; hence the micro-relation units must be sufficiently activated for the weight that represent these correlations to be learned. Phase 1 should also precede phase 3 because there is nothing preventing relational learning from beginning even before the dimensions themselves have been figured out.

The Architecture of Relational Learning

Representing relation instances requires a way of distinguishing the different objects from one another, that is, a way of binding together the features associated with a given object. The **binding problem**, in one form or another, has surfaced in many forms in recent years, and a number of connectionist solutions have been proposed (Hummel & Biederman, 1992; Hummel & Holyoak, 1997; Shastri & Ajjanagadde, 1993). Most of these solutions, including the one we proposed in earlier versions of this model (Colunga & Gasser, 1998; Gasser & Colunga, 2000), make use of a dimension in addition to activation which characterizes network processing units, with synchronization along this dimension representing the binding of units. Here we propose a simpler solution, one that makes use of copies of dimensions. The idea is to treat relative position in space or time as a special dimension, one that maps directly onto hardware. This requires a relatively large number of units, but the visual system already utilizes a similar approach in deploying multiple feature detectors of a particular type (for example, motion in particular direction) that are specific to particular regions within the visual field.

Multiple copies of object feature units alone do not solve the problem of where relations come from, however. There is still the need for some sort of segmentation mechanism, a process which can "find" objects in sensory input. A full-blown account of how this happens is beyond the scope of our model. We assume that the process involves two sorts of **micro-relation units**, those that tend to respond to inputs from features of a single object (**sameness units**) and those that tend to respond to inputs from features of different objects (**difference units**). Each micro-relation unit multiplies, rather than adds, the inputs along the connections from the two objects features that it joins, similar to the "sigma-pi" units introduced by Rumelhart, Hinton, and Williams (1986).

Thus with respect to these inputs, it behaves like something like an AND unit. Micro-relation units also excite or inhibit each other to the extent that they represent consistent or inconsistent parsings of the scene.

Micro-relations (simple relational correlations) take the form of difference micro-relation units. The model begins with one of these for each combination of ranges of values on each dimension, but the connections to the relation object dimension units are initially weak, and all of the connections between micro-relation units begin with 0 weights. Each time a micro-relation is coactivated with its related object feature units, the weights on the multiplicative connections are strengthened. Each time two connected micro-relation units are co-activated, the connection between them is strengthened.

The details of our proposal, in particular how it implements the binding of relation roles in complex cases like the meanings of relational terms and how relative relational correlations are handled, are beyond the scope of this paper.

Learning

Units are connected to one another in an extension of a continuous Hopfield network. All connections are symmetric. The learning algorithm is an unsupervised variant of Contrastive Hebbian Learning (Movellan, 1990). There are two phases to learning for each pattern, a "positive" and a "negative" phase. In the positive phase, a training pattern is first clamped on a set of input units. Next the network is allowed to settle, and for each weight Hebbian learning is performed. That is, each weight is incremented by an amount which is proportional to the activations of the two connected units. Then, in the negative phase, the input units are unclamped, a small amount of noise is injected into the network, and the network is allowed to settle again. Next *anti-Hebbian* learning is performed for each weight; each weight is *decremented* by an amount proportional to the product of the two units' activations. The negative phase functions to eliminate spurious attractors in the network, providing a solution the problem of the lack of negative evidence: the network is punished for producing patterns that do not occur in the training set. When the positive and negative phase weight changes cancel each other out, learning has been successful. That is, for each pattern, the network settles to the same states when the input units are clamped and when they are then immediately unclamped.

Predictions

Micro-Relation Theory, as implemented in a network of the type described, predicts that people should be sensitive to relational correlations in input patterns. Presented with a set of training patterns embodying relational correlations, subjects should later accept patterns agreeing with the correlations and reject those violating them. In addition, the theory predicts that relational correlations should begin as absolute (between specific object features) rather than relative.

We developed the micro-relational architecture because of our interest in the learning of relational terms in natural language, perhaps the best example of relational correlations in human behavior. While it is clear that people do learn the meanings of relational terms, it is difficult to test the specific predictions of the theory in this complex domain. For this reason, we began with a much simpler task, described in the next section. While there has been research of this sort on unsupervised correlational learning (Billman & Knutson, 1996), to our knowledge it has not addressed the learning of relational correlations.

Experiment

The goal of this experiment is to explore to what extent people are sensitive to relational correlations. To do this we presented subjects with a simple unsupervised learning task. Subjects were shown instances of parent-child alien pairs from a fictitious planet. The members of these pairs represent the two arguments in a relation. The instances follow three "rules" realized as relational correlations between dimensions characterizing the parent and the child. We wanted to know (1) whether people could learn what made a pair of aliens a good example of a parent-child pair and if so, (2) what it was that they were actually learning.

Method

Subjects 10 undergraduates participated in this experiment.

Stimuli The familiarization stimuli consisted of 128 computer-generated pictures of parent-child pairs. The stimuli included four "species" of aliens. The aliens within a species all had the same basic body shape but varied along the dimensions of size, darkness, body shape, body hue, and eye color. We used four values along each of these dimensions. The parent-child pairs followed the following "rules":

1. The child was always at least as big as the parent. This represents a relative simple relational correlation. Each specific combination of size values constitutes an (absolute) simple relational correlation.
2. The parent was always at least as dark as the child. This is also a relative relational correlation with a simple relational correlation for each combination of values.
3. The size of the child matched the darkness level of the parent. This takes the form of specific combinations of parent and child values along both dimensions. Each of these constitutes an (absolute) complex relational correlation. The rule itself would be a relative complex relational correlation.

For each of the first two rules, there were 10 possible specific combinations of size or darkness values. For the third rule, both size and darkness values were relevant, and there were 20 possible combinations of these.

The testing stimuli consisted of 72 computer-generated pictures of parent-child pairs. Half of the test pictures followed all of the above rules: half of these were completely familiar patterns (“valid”), and the other half were generalizations of the rules (“darkness-size”) in which values on the distracter dimensions differed from those presented during training but the darkness and size values were combinations that had been presented. The other half violated the rules in one way or another:

1. All three rules were violated (“wrong”).
2. The darkness rule was violated (“size”).
3. The size rule was violated (“darkness”).
4. The darkness-size complex relational correlations were violated (“darkness & size”); that is, for a given pair, the pair of darkness values and the pair of size values did occur during training, but the combination of the two pairs did not.

Procedure At the start of the experiment the subjects were told: “The following are scenes representing animals on Planet X. Each scene shows a parent-child pair of some species on Planet X. The child is always on the left. Study the scenes carefully, taking as much time as you need to look at each one. You will be tested on the pictures later.” During this familiarization phase, the scenes were presented on a computer screen and the subject was allowed to look at a scene for as long as desired before going on to the next.

After the familiarization phase, subjects were told: “For each of the following pictures, hit ‘y’ if they represent a parent-child pair and ‘n’ if they don’t.” Again, they were allowed to look at each picture for as long as they wanted before judging it.

Results and Discussion

Figure 1 shows the mean proportion of each type of test item accepted as a parent-child pair by the subjects.

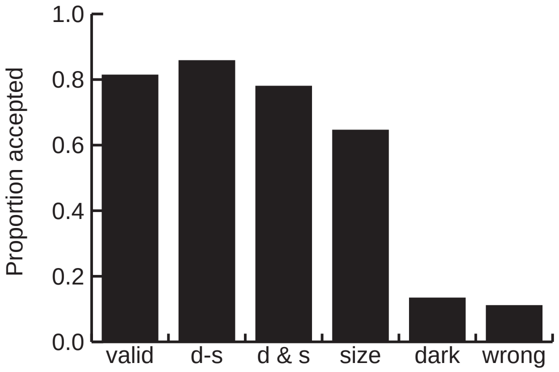

Figure 1: Results of experiment. Mean proportion of each type of test item accepted as a parent-child pair by the subjects.

Subjects accepted at significantly greater than chance frequency the valid ($p < .0001$) and darkness-size ($p = .007$) patterns and rejected at significantly greater than chance frequency the darkness ($p = 03$) and wrong ($p = .004$) patterns. Subjects were also significantly more likely to accept the darkness-size patterns than either the size ($p < .0155$) or the darkness ($p < .0001$) patterns.

The answer to our first question, then, is that people can learn what it is that makes a good parent-child pair. The subjects generalized over the training patterns, readily accepting novel patterns (darkness-size) which differed from the training patterns on non-correlating dimensions. The answer to our second question is that people are sensitive at least to the simple relational correlations in the data. The advantage of the darkness & size patterns over the darkness, size, or wrong patterns shows that they have learned these correlations. That is, in terms of our account, they have reached the first phase of relational learning. The subjects also preferred patterns obeying the complex relational correlations (darkness-size) to patterns which matched the training patterns on *both* size and darkness but failed to obey the complex correlations (darkness & size), though this difference was not significant. Thus there is as yet no evidence that the subjects have achieved the second phase of relational learning.

Simulation

To simulate the results of the experiment, we trained a relational network on a simplified version of the subjects’ task. Input layers were divided into CHILD and PARENT groups; that is, we assumed that segmentation of the input had already taken place. Within each of these groups there were separate layers of units for the two correlating dimensions (size and darkness) and one of the distracter dimensions (body hue). These layers had a single unit for each possible value on the relevant dimension.

For each input dimension there was an associated hidden layer of micro-relation units. Each of these units was joined by multiplicative connections to two object feature units, one each for the child and the parent. These units represent potential simple relational correlations. These weights were initialized at a constant, small value (0.04). All of the micro-relation units were also joined to each other by ordinary additive connections initialized with weights of 0.0. It is these connections that have the potential to represent complex relational correlations. The architecture of the network is shown in Figure 2.

On each training trial, one of the 56 combinations of child and parent darkness, size, and hue that the subjects were trained on was presented to the network. Input units representing the child and parent values on the three dimensions for the training pattern were clamped on, and all of the other input units were clamped off. Then the units in the relational layers were allowed to settle, and the weights were updated using Hebbian learning. Next the input units were unclamped, a small amount of noise was injected into the network, and the network was allowed to settle again. Finally the weights were updated using anti-Hebbian learning.

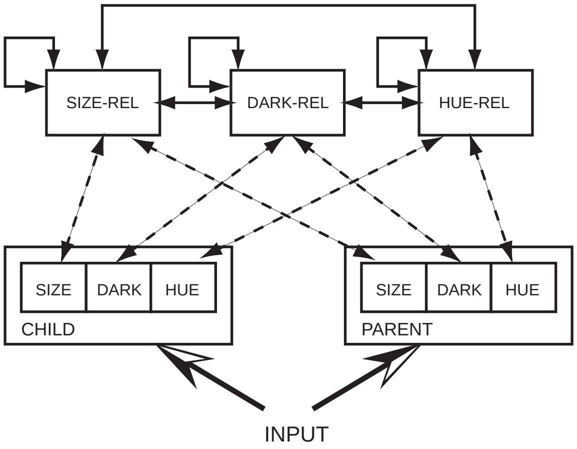

Figure 2: Architecture of simulation network. Rectangles indicate layers of units, arrows connections between layers. There are input layers for each of three dimensions, for both parent and child. Hidden layers consist of micro-relation units. Micro-relation units are connected locally to pairs of object feature units in a given dimension, one for child, one for parent. These connections are multiplicative (indicated by the dashed lines). Each micro-relation unit is also connected to all other relation units by conventional additive connections.

We tested the network and examined the weights following 1 and 10 epochs of training patterns. Following 1 epoch, the multiplicative weights into those relation units representing the simple relational correlations embodied in the training patterns had already clearly increased while the others were near zero. Following 10 epochs of training, we examined the weights connecting the relation units. For each of the (absolute) complex relational correlations between size and darkness in the training patterns, the network had learned a positive weight between the corresponding relation units. All of the other weights connecting relation units had become negative or very close to 0. Thus the network seems to have learned all of the absolute complex relational correlations and to have learned negative weights (during anti-Hebbian learning) which inhibit other combinations of values.

The network was tested on the same set of patterns as the subjects (except that since the network had only one distracter dimension, there were fewer patterns that generalized over the rules).

To measure the extent to which the network treated a pattern as acceptable, we followed a procedure similar to that followed during training. We first clamped the test pattern on the input units, just as during training, and allowed the relational layers to settle. We recorded the activations of all of the units in the network at this point. Next we unclamped the input units and allowed the network to settle again. We calculated the Euclidian distance between the final vector of activations of all units (input and hidden) and that recorded after the clamped phase. For patterns that the network accepts, there should be relatively little change in activation. For patterns that the network treats as unfamiliar, activation should change during the unclamped phase as the network alters the input pattern in the direction of more familiar patterns. To make the network performance comparable to the subjects' data, we subtracted each activation change from an estimate of the maximum possible change in activation (the largest change observed over the test patterns before training). We will refer to this measure as "clamped-unclamped similarity".

Figure 3 shows the mean performance of two separate networks[1] We combined the valid and darkness-size patterns because there were not enough patterns in the darkness-size set to compare the two. We show darkness, but not size, since the two are completely analogous.

Like the subjects, the network "prefers" patterns that agree with both size and darkness simple relational correlations to those which agree on neither or on only one of the two dimensions. In fact, from the observed weights and the data in the graph, we see that these correlations, representing the first phase of relational learning, were learned with a single pass through the patterns. The main difference following additional training is in the significantly increased preference for patterns obeying the (absolute) complex relational correlations (darkness-size). As in our account of relational learning, the mastery of the simple relational correlations is followed by mastery of the complex relational correlations.

Of course it is not surprising that the network treats the patterns it was trained on as more familiar than patterns which differ from the training patterns. The main points of the simulation have been to show that the network exhibits two of the phases of learning that we posited and that it is not just the values on each dimension that matter but the combination of values.

Discussion

Both the subjects in our experiment and the relational network respond to absolute simple relational correlations in unsupervised learning. But note that the model predicts a stronger effect than we found among the subjects. The network preferred patterns obeying the complex relational correlations to those in which both of the dimension-specific rules were obeyed but the complex relational correlations were violated. We believe that this behavior will emerge in the subjects with more training; at least one of our subjects did exhibit this advantage for the patterns obeying the complex correlations.

A further prediction of the theory is that relational categories start out highly specific, that *relative* relational correlations come later. This prediction is not tested directly in the experiment we reported. To test this, we will need to work with dimensions that are less familiar to subjects, dimensions for which they have not already learned an ordering of the values.

The solution to the binding problem that we offer in

[1] Both networks started with the same weights, but because units are selected randomly during network settling, performance varied somewhat from one network to another.

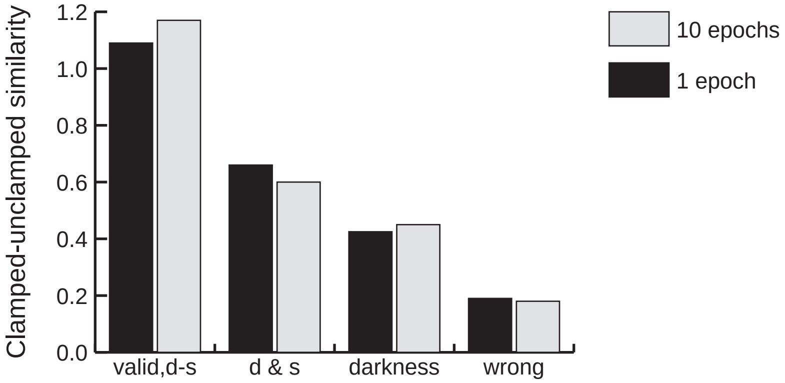

Figure 3: Results of simulation. Similarity between activations following clamping of test patterns and immediate unclamping at two points during training.

this paper also applies to another sort of binding, the binding of a variable with its value. While this may seem quite different from the binding of the features of an object with one another, the Micro-Relation Theory also offers an account of behavior that seems to require explicit variables. Like the roles in a relation, the variables in a rule on this view are implicit in the pattern of activation across a set of micro-relation units representing the primitive relations of sameness and difference in object features. Thus our theory may offer a unified account of a wide range of basic behaviors.

Conclusions

Human cognition is deeply relational, yet we lack a clear picture of how relations emerge in the first years of life and how we acquire new ones later on. Like object categories, relational categories are grounded in experience. Like object categories, they are presumably built up out of more basic stuff. In this paper, we have argued that basic relation stuff is quite similar to basic object stuff; it takes the form of distributed patterns across simple processing units. We have argued that a correlational, associationist account of the learning of relations is possible. We believe this is a first step towards an understanding of where relations come from.

References

Billman, D. & Knutson, J. (1996). Unsupervised concept learning and value systematicity: a complex whole aids learning the parts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 458–475.

Colunga, E. & Gasser, M. (1998). Linguistic relativity and word acquisition: a computational approach. *Annual Conference of the Cognitive Science Society*, *20*, 244–249.

Gasser, M. & Colunga, E. (2000). Babies, variables, and relational correlations. *Annual Conference of the Cognitive Science Society*, *22*, 160–165.

Gasser, M. & Smith, L. B. (1998). Learning nouns and adjectives: a connectionist account. *Language and Cognitive Processes*, *13*, 269–306.

Hummel, J. E. & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review*, *99*, 480–517.

Hummel, J. E. & Holyoak, K. J. (1997). Distributed representation of structure: a theory of analogical access and mapping. *Psychological Review*, *104*, 427–466.

Movellan, J. (1990). Contrastive Hebbian learning in the continuous Hopfield model. In Touretzky, D., Elman, J., Sejnowski, T., & Hinton, G. (Eds.), *Proceedings of the 1990 Connectionist Models Summer School*, pp. 10–17. Morgan Kaufmann, San Mateo, CA.

Rumelhart, D. E., Hinton, G., & Williams, R. (1986). Learning internal representations by error propagation. In Rumelhart, D. E. & McClelland, J. L. (Eds.), *Parallel Distributed Processing, Volume 1*, pp. 318–364. MIT Press, Cambridge, MA.

Shastri, L. & Ajjanagadde, V. (1993). From simple associations so systematic reasoning: a connectionist representation of rules, variables, and dynamic bindings using temporal synchrony. *Behavioral and Brain Sciences*, *16*, 417–494.