In _Proceedings of the 15th Annual Meeting of the Cognitive Science Society_, (pp.729-734), 1993.

Modeling Property Intercorrelations in Conceptual Memory

Ken McRae
Department of Psychology
University of Rochester
Rochester, NY 14627
kenm@psych.rochester.edu

Virginia R. de Sa
Department of Computer Science
University of Rochester
Rochester, NY 14627
desa@cs.rochester.edu

Mark S. Seidenberg
Neuroscience Program
University of Southern California
Los Angeles, CA 90089
marks@neuro.usc.edu

Abstract

Behavioral experiments have demonstrated that people encode knowledge of correlations among semantic properties of entities and that this knowledge influences performance on semantic tasks (McRae, 1992; McRae, de Sa, & Seidenberg, 1993). Independently, in connectionist theory, it has been claimed that relationships among semantic properties may provide structure that is required for the relatively arbitrary mapping from word form to word meaning (Hinton & Shallice, 1991). We explored these issues by implementing a modified Hopfield network (1982, 1984) to simulate the computation from word form to meaning. The model was used as a vehicle for developing explanations for the role played by correlated properties in determining short interval semantic priming effects and in determining the ease with which a property is verified as part of a concept. Simulations of the priming and property verification experiments of McRae (1992) are reported. It is concluded that correlations among properties encoded in conceptual memory play a key role in the dynamics of the computation of word meaning. Furthermore, a model in which property intercorrelations are central to forming basins of attraction corresponding to concepts may provide important insights into lexical memory.

Introduction

An important question in cognitive science involves how entities such as dogs and chairs are represented in memory. In the work reported in this article, we explored the role of correlations among properties in conceptual memory. Behavioral experiments described in McRae (1992) demonstrated that people encode correlations among properties of real-world entities. Below, we present a connectionist model that was used to investigate a theory in which property intercorrelations play a key role in the computation of concepts from words. Simulations of the behavioral phenomena described in McRae are presented. A more complete description of the experiments and simulations is available in McRae et al. (1993).

The human empirical studies reported in McRae (1992) investigated an important and basic aspect of people's knowledge of real-world entities, namely, correlations among properties (e.g., *has fur* and *has a tail* are correlated in entities in the world). The studies demonstrated a clear influence of correlated properties on people's performance in semantic tasks. Specifically, Experiment 1 showed that the speed with which people indicate that a property (e.g., *hunted by people*) is reasonably true of an entity (e.g., DEER) can be predicted from how strongly the property is correlated with other properties of that entity. In Experiment 2, a short interval semantic priming task, similarity defined in terms of overlap of individual and correlated properties was used to predict priming effects. Similarity in terms of individual properties predicted priming effects for artifacts (things that are made by humans), but similarity in terms of correlated property pairs predicted priming effects for biological kinds (living things). The results of these studies as well as others from the artificial concepts literature (e.g., Medin et al., 1982; Younger & Cohen, 1983) indicate that conceptual memory includes information about property intercorrelations.

Connectionist models provide a natural vehicle to study the notion that stored property intercorrelations are central to concepts. A connectionist model that uses a correlational learning rule (e.g., a Hopfield network, 1982, 1984) can naturally be used to encode correlations among the properties of entities in its environment. Furthermore, this type of model may shed light on the processes by which property intercorrelations influence people's performance on semantic tasks. Our goal was to use an explicit computational model to investigate a theory in which encoded knowledge of the correlational structure of semantic space plays a critical role in computing concepts from words.

There is further independent motivation for hypothesizing that encoded correlations may be important for computing word meaning. The mapping from word form to meaning contains few regularities, particularly at the monomorphemic level. Thus, words that have similar pronunciations or spelling do **not** tend to have similar meanings (e.g., MAT, RAT, THAT, HAT, CAT, and SAT). As pointed out by Hinton and Shallice (1991), it is difficult for a network with distributed representations to map similar inputs (e.g., the orthographic or phonological forms of

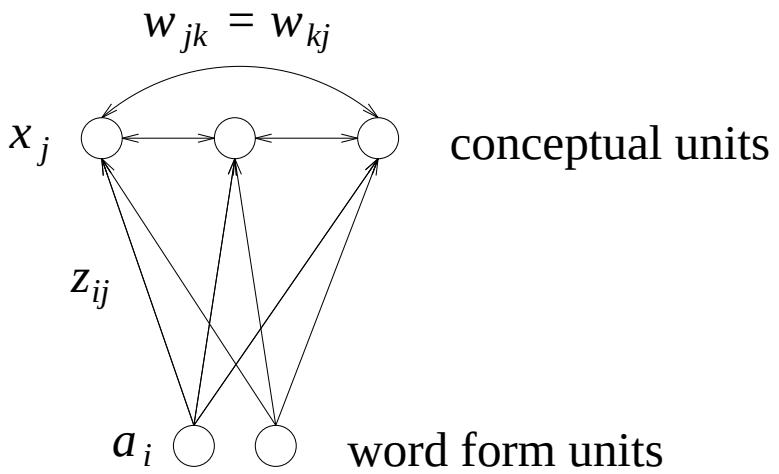

Figure 1: Schematic of the model's architecture. The implemented model contained 379 word form units and 646 conceptual units.

the words listed above) onto dissimilar outputs (e.g., their meanings). This observation led them to suggest that the computation of word meaning may consist of two overlapping stages. First, roughly, a computation from word form puts the system into the appropriate region of semantic space (i.e., into the correct basin of attraction). Next, encoded correlations among semantic properties are used to drive the system to a stable state.

In the next section, we describe a modified Hopfield network in which correlations among properties are naturally encoded and are used to construct basins of attraction that are critical to its ability to compute the meaning of words.

The Model

Figure 1 shows the model's architecture. The input was a distributed representation of word form. Specifically, each of the 379 input units represented a triple of letters (including leading and trailing blanks) that occurred in at least one of the included concept words. This scheme provided a sparse distributed representation of word spelling that encoded letter order and roughly preserved item similarity. The critical theoretical aspect of the input was that no systematic relationship existed between the input patterns and conceptual representations. We assume that, when all words are represented, because of the lack of regularity in the mapping from orthography or phonology to meaning, concepts are not directly computed. Our model possessed this characteristic.

Output conceptual representations were based on semantic property production norms that are described in McRae et al. (1993). There was an output unit for each property that occurred as part of one or more of the 84 concepts on which the model was trained (46 artifacts and 38 biological kinds). Output units were fully interconnected. Each input unit was unidirectionally connected to each of the 646 output units, but input units were not interconnected.

The weights between the output units were calculated using the modified Hopfield rule of Tsodyks and Feigelman (1988). That is

$$w_{ij} = \frac{1}{n_1} \sum_p [(x_{ip} - \mu_p)(x_{jp} - \mu_p)]$$

where w_{ij} represented the weight of the connection between unit i and unit j, n_1 was the number of output units (646), μ_p was the number of properties possessed by concept p expressed as a proportion of the total number of properties, and x_{ip} was 1 if the p^{th} pattern possessed property i and 0 otherwise.

This modified rule allows a network to store a greater number of sparse patterns (Tsodyks & Feigelman, 1988). It is also intuitively sensible in that it captures the idea that in a sparse conceptual space, the absence of a property carries little meaning.

The connections from the input to the output units were calculated using a similar Hebbian (1949) style learning rule. Specifically

$$z_{ij} = \frac{1}{n_2} \sum_p [(a_{ip})(x_{jp} - \mu_p)]$$

where z_{ij} represented the weight of the connection between input unit i and conceptual unit j, n_2 was the number of input units (379), μ_p and x_{ip} were as above, and a_{ip} was the normalized activation of the i^{th} input unit in pattern p. Patterns were normalized to remove the effect

of word-length according to

$$a_{ip} = \frac{2u_{ip}}{\|u_p\|} \qquad \forall p$$

where u_{ip} is 1 if pattern p contains the triple of letters represented by word form unit i and 0 otherwise; and $\|u_p\|$ is the Euclidean norm of the pattern vector of components u_{ip}.

To simulate the computation of word meaning, the appropriate normalized word form pattern was clamped. To simulate an arbitrary starting state, 60 randomly chosen semantic units were initialized to 0.25 activation (all others began at 0). Activation was then propagated according to:

$$x_i(t+1) = g(c_1 \sum_j (z_{ji} a_j) + c_2 \sum_j (w_{ji} x_j(t)) - \theta)$$

where $g(\cdot)$ was a sigmoidal function defined by

$$g(x) = 0.5 tanh(c_3 x) + 0.5$$

x_i was the activation of the i^{th} semantic unit, and a_i was the activation of the i^{th} input unit as above.

For the simulations presented below, we used the following values for the constants $c_1 = 0.85$, $c_2 = 0.33$, $c_3 = 400$, $\theta = 0.0105$, although similar results were obtained with slight variations.

The model can be understood as a Hopfield network with the effect of the connections between the word form and conceptual units as thresholds that are variable over units and patterns. Thus the conceptual interconnections determine the general topology of the energy function (which combinations of properties were stable). The connections from the word form units lower the energy of states involving semantic properties that occur with words that activate those units. They also provide an initial state to the network.

The network learned to compute an appropriate conceptual representation for 80 of the 84 trained patterns. Because the model's behavior was primarily driven by encoded property co-occurrence information, intercorrelational density of a concept's properties influenced convergence. The network made an error of omission on the concept YAM, because it consisted of few and sparsely intercorrelated properties. The network made a number of other interesting "errors". When computing a concept, properties were sometimes activated that were not part of the concept according to how the model was trained, but were strongly intercorrelated with its properties. In the vast majority of these cases, although fewer than 5 of 30 subjects had listed the activated property for the concept in the norms, it was nonetheless true of the concept. The model activated: *has wings* for BUDGIE, *is an animal* for BUZZARD, CANARY, and EAGLE; *is large* for CANARY, CHICKEN, and DUCK; *is dangerous* for CANNON; *is edible* for CARROT, RADISH, and ZUCCHINI; *has leaves* for CARROT; *eats* for HAWK; *has four legs* for MOUSE; *is loud* for CANARY and MISSILE; and *worn by women* for TROUSERS. With the exception of *is large* for CANARY, as well as bird-like properties that were activated for CAT and JET (not listed above), these properties are reasonably true of their respective concepts.

Simulating Semantic Priming

We simulated Experiment 2 of McRae (1992). In that study, subjects made semantic decisions (e.g., is it an object?) to a target (e.g., CHANDELIER) when it was preceded by either a similar (e.g., LAMP) or a dissimilar prime (e.g., GOOSE). On each trial, the prime was presented for 200 ms, followed by a mask for 50 ms, and then the target. The priming effect is the facilitation in decision latency to a target when it is preceded by a similar as compared to a dissimilar prime. In McRae, the magnitude of priming effects was predicted from prime-target similarity in terms of individual and correlated properties. Concept similarity in terms of individual properties predicted priming effects for artifacts, but similarity in terms of correlated properties predicted priming effects for biological kinds.

Short interval priming can be understood in terms of inter-concept distance in semantic space (for similar proposals, see Masson, 1992; Sharkey, 1989). Computing the meaning of a word can be understood as driving the semantic system from its present state to the desired end state, which corresponds to the meaning of the word that is being read or heard. In a short interval semantic priming task, because the prime determines the start state for the computation of the target, prime-target similarity determines amount of facilitation. A simulated trial began by clamping the prime's word form pattern and computing its conceptual representation. With the prime's meaning active, the word form representation of the target was clamped and convergence latency was recorded for the target concept.

The basic idea of this simulation was to determine if the factors that influence priming in humans also influence priming in the model. Specifically, it was important to show the interaction between type of representation (individual versus correlated properties) and type of entity (artifact versus biological kind). To accomplish this, we conducted regression analyses analogous to those of Experiment 2 of McRae (1992). Predictions were again derived from individual and correlated properties representations.

In the individual properties representation, each concept was represented as a vector in which the value for

element *i* was the number of subjects who listed property *i* for that concept in the norms. In the correlated properties representation, a concept was represented as a vector in which each element corresponded to a property pair that was significantly correlated as determined in an analysis of the norms. If a concept possessed both members of the pair, the element was given a positive value; if it possessed neither, it was given a value of 0; if it possessed one property but not the other (i.e., it violated the correlation), the element was given a negative value. Concept similarity was computed as the cosine between the vectors representing the prime and the target. Linear regression analyses were used to predict priming effects (in terms of convergence rates) for artifact and biological kind prime-target pairs from concept similarity in terms of individual and correlated properties. Three measures of convergence were used: the number of iterations for error (defined as the sum of squared error between the target and computed vectors over the conceptual units) to drop below 1.0, for error to be within 0.1 of its value when the concept stabilized, and for error to be within 0.01 of the stabilization point. Only 40 of the 42 similar priming pairs were used because, as described above, the model did not learn to compute the meanings of CAT and JET. Convergence rate for a target preceded by a dissimilar prime was estimated by averaging across three primes that shared no properties with the target. It was assumed that convergence latency for a concept in the model was monotonically related to the time required for people to compute a word's meaning, and was therefore monotonically related to how quickly a person could answer a question concerning its meaning.

In Experiment 2 of McRae (1992), individual properties predicted priming effects for artifacts (r^2 = .15), but not for biological kinds (r^2 = .04). In contrast, correlated property pairs predicted priming effects for biological kinds (r^2 = .21), but not for artifacts (r^2 = .00). The model produced the same pattern of results. Concept similarity as measured by overlap of individual properties significantly predicted priming effects for artifacts for each of the three convergence measures: less than 1: r^2 = .27, $F(1, 19) = 7.065$, $p < .02$; within 0.1: r^2 = .59, $F(1, 19) = 27.074$, $p < .001$; within 0.01: r^2 = .58, $F(1, 19) = 26.166$, $p < .001$. Individual properties did not predict biological kind priming effects: less than 1: r^2 = .07, $F(1, 15) = 1.138$, $p > .3$; within 0.1: r^2 = .04, $F < 1$; within 0.01: r^2 = .01, $F < 1$. In contrast, after similarity in terms of individual properties had been entered into the equation, similarity in terms of correlated property pairs predicted priming effects for biological kinds: less than 1: r^2 = .25, $F(1, 14) = 4.609$, $p < .05$; within 0.1: r^2 = .44, $F(1, 14) = 10.925$, $p < .006$; within 0.01: r^2 = .41, $F(1, 14) = 9.582$, $p < .009$. Correlated properties did not predict artifact priming: less than 1: r^2 = .03, $F < 1$; within 0.1: r^2 = .17 in the wrong direction; within 0.01: r^2 = .01, $F < 1$. More detailed analyses of the human study and simulation data can be found in McRae et al. (1993).

In summary, this simulation demonstrated that the same factors influence short interval semantic priming effects in both humans and the model. Specifically, similarity in terms of individual properties predicted priming effects for artifacts, but similarity in terms of correlated property pairs predicted priming effects for biological kinds. Note that this interaction was found with a model in which artifacts and biological kinds were treated identically. These results concur with Keil (1989) who has claimed that biological kinds cohere around clusters of intercorrelated properties but artifacts cohere around the intended function of the creator.

Simulating Property Verification

In this section, the model is used to guide an explanation of the influence of property intercorrelations on verification. In Experiment 1 of McRae (1992), a concept (e.g., DEER) was presented for 400 ms, followed by a target property (e.g., *hunted by people*). Subjects indicated as quickly as possible whether the target property was reasonably true of the entity to which the concept name referred. With normed production frequency equated, target properties that were strongly intercorrelated with other properties within a concept (e.g., DEER–*hunted by people*) were verified more quickly than matched targets that were weakly intercorrelated (e.g., DUCK–*hunted by people*).

We base our explanation on the assumption that production frequency is less sensitive than verification latency as a measure of a property's activation by a concept name. Production frequency measures the probability that a property is one of the first few generated by a subject when reading a concept name. For example, if subjects list an average of eight properties for a concept, production frequency for any property roughly indicates the frequency (or probability) that it was generated as one of the top eight. In contrast, in the property verification task, a subject was asked to indicate, as quickly as possible, whether or not a specific property was reasonably true of an entity. Thus, verification latency provided a more direct, precise, and sensitive measure of the relationship between a concept and a property. It might be assumed that subtle differences existed in a target property's activation that were too small to influence production frequency, but large enough to be detected in the property verification task. It might further be assumed that these differences resulted from encoded knowledge of property co-occurrences. That is, properties that were strongly intercorrelated with a number of other properties within a concept received a boost in activation when that concept

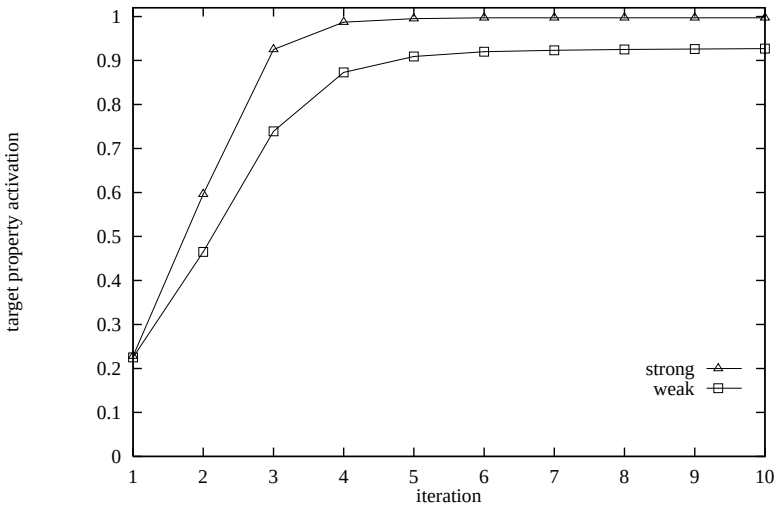

Figure 2: Mean activation of target properties for the strongly and weakly intercorrelated groups. Properties that are more strongly intercorrelated with others of a concept receive a boost in activation.

was computed.[1]

The model contained 14 items from the property verification experiment of McRae (1992) that differed in terms of the strength with which the target property was correlated with other properties of the concept. The groups were also roughly equated on three variables: production frequency, concept familiarity, and number of individual properties listed per concept. For this subset of 14 items, human verification latency was significantly faster for the strongly intercorrelated group (788 ms, 34 ms) than for the matched weakly intercorrelated group (998 ms, 51 ms), $t(13) = 7.595$, $p < .001$ by items.

To test the hypothesis that properties that were strongly intercorrelated with a number of other properties within a concept received a boost in activation when that concept was computed, the word form representation of each of the fourteen concepts was clamped and the network was allowed to iterate 10 times. Activation of the target property was recorded after each iteration. Five runs were used, each with a different random starting configuration. Figure 2 shows the mean activation of the fourteen target properties as the concepts converged. A one-way repeated measures analysis of variance showed that the activation of the target properties was significantly higher when a concept from the strongly intercorrelated group was being computed, $F(1, 13) = 10.277$, $p < .008$. An interaction between group and time step also obtained, $F(9, 117) = 5.057$, $p < .0001$. Simple main effects revealed that target properties from concepts from the strongly intercorrelated group were not significantly more activated after time step 1, but were after time steps 2 through 5 ($p < .05$). The advantage was marginally significant throughout time steps 6 to 10 ($p < .07$).

In summary, the model provides for a reasonable explanation of the influence of correlated properties on property verification. When a property is highly intercorrelated with other properties of a concept, it becomes more accessible. This increased accessibility is too subtle to be picked up in a production task but facilitates verification latency.

Discussion

The modeling presented here illustrates that a network whose computational dynamics are dependent on attractor basins formed from encoding property intercorrelations captures some basic characteristics of human performance on semantic tasks. The semantic priming and property verification results were naturally modeled using a connectionist framework, specifically, a modified Hopfield (1982, 1984) network. Furthermore, although it was not presented in this article, in McRae et al. (1993), we empiricially verified a primary characteristic of the model, that is, that intercorrelational density of a concept's properties influences its convergence latency.

The most theoretically important aspect of the model was the way in which basins of attraction resulting from encoding the relationships among semantic properties were central to the computation of conceptual representations from words. This characteristic is shared with other

[1] Note that this explanation is different from the one in McRae (1992). Both explanations are simulated in McRae et al. (1993).

recent connectionist models of word recognition (Hinton & Shallice, 1991; Masson, 1992; Plaut 1991; Sharkey, 1989)[2]. The concept of attractor basins may be useful in understanding a number of primary word recognition phenomena. For example, the lexicon is typically conceptualized as a set of nodes, each of which corresponds to a lexical item (e.g., Morton, 1969). Alternatively, the lexicon may be conceptualized as a set of attractor basins. In this view, a lexical item corresponds to a stable state in representational space. Hinton and Shallice (1991) and Plaut (1991) have used this approach to show that attractor basins are critical to understanding phenomena that characterize deep dyslexia.

This idea may also shed light on a recent issue of great interest to word recognition researchers, mediated priming (McNamara, 1992). It has been shown that, although two concepts are not directly associated according to free association norms (e.g., LION and *stripes*), they may prime one another (Balota & Lorch, 1986; McNamara & Altarriba, 1988). McNamara interprets these results within a semantic network theory such as Collins and Loftus (1975). He claims that LION primes *stripes* because activation of the LION node spreads to TIGER and then to *stripes*. Although the notion of association is not a primitive in the conceptual system that our model represents, it provides a plausible explanation for mediated priming effects. LION shares properties with TIGER. These properties are intercorrelated. Therefore, when LION is computed, *stripes* will be slightly activated because properties of LION are correlated with *stripes*. A small priming effect will result, as has typically been found in these studies. This explanation is currently being explored.

References

Balota, D.A., & Lorch, R.F. (1986). Depth of automatic spreading activation: Mediated priming effects in pronunciation but not lexical decision. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **12**, 336–345.

Collins, A.M., & Loftus, E.F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, **82**, 407–428.

Hebb, D.O. (1949). *The Organization of Behavior*. New York: Wiley.

Hinton, G.E., & Shallice, T. (1991). Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review*, **98**, 74–95.

Keil, F.C. (1989). *Concepts, Kinds, and Cognitive Development*. Cambridge, MA: MIT Press.

Masson, M.E.J. (1992). A distributed memory model of semantic priming. manuscript submitted for publication.

McNamara, T.P. (1992. Priming and constraints it places on theories of memory and retrieval. *Psychological Review*, **99**, 650–662.

McNamara, T.P., & Altarriba, J. (1988). Depth of spreading activation revisited: Semantic mediated priming occurs in lexical decisions. *Journal of Memory and Language*, **27**, 545–549.

McRae, K. (1992). Correlated Properties in Artifact and Natural Kind Concepts. In *Proceedings of the 14th Annual Meeting of the Cognitive Science Society*, Bloomington, IN. July 30–August 1, 1992. (pp. 349–354). Hillsdale NJ: Erlbaum.

McRae, K., de Sa, V.R., & Seidenberg, M.S. (1993). The Role of Correlated Properties in Accessing Conceptual Memory. manuscript submitted for publication.

Medin, D.L., Altom, M.W., Edelson, S.M., & Freko, D. (1982). Correlated Symptoms and Simulated Medical Classification. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **8**, 37–50.

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review*, **76**, 165–178.

Plaut, D.C. (1991). *Connectionist neuropsychology: The breakdown and recovery of behavior in lesioned attractor networks*. Unpublished doctoral dissertation, Carnegie Mellon University, Pittsburgh.

Sharkey, N.E. (1989). The lexical distance model and word priming. In *Proceedings of the 11th Annual Conference of the Cognitive Science Society*, Ann Arbor, Michigan. August 16–19, 1989. (pp. 860–867). Hillsdale NJ: Erlbaum.

Tsodyks, M.V., & Feigelman, M.V. (1988). The Enhanced Storage Capacity in Neural Networks with Low Activity Level. *Europhysics Letters*, **6**, 101–105.

Younger, B.A, & Cohen, L.B. (1983). Infant Perception of Correlations among Attributes. *Child Development*, **54**, 858–867.

[2] A major advantage of our model is that our conceptual representations were based on production norms, whereas theirs were either created by the researchers themselves or were random patterns of activation.