Thibodeau, P., Flusberg, S. J., Sternberg, D. A. & Glick, J. J. (2011). On the emergence of analogical inference. Proceedings of the 33rd annual meeting of the Cognitive Science Society, Boston.

On the Emergence of Analogical Inference

Paul H. Thibodeau (pthibod1@stanford.edu)
Stephen J. Flusberg (sflus@stanford.edu)
Jeremy J. Glick (jjglick@stanford.edu)
Daniel A. Sternberg (sternberg@stanford.edu)
Department of Psychology, 450 Serra Mall, Bldg. 420
Stanford, CA 94305 USA

Abstract

What processes and mechanisms underlie analogical reasoning? In recent years, several computational models of analogy have been implemented to explore this question. One feature of many of these models is the assumption that humans possess dedicated analogy-specific cognitive machinery – for instance, a mapping or binding engine. In this paper, we question whether it is necessary to assume the existence of such machinery. We find that at least for some types of analogy, it is not. Instead, some forms of analogical processing emerge naturally and spontaneously from relatively simple, low-level learning mechanisms. We argue that this perspective is consistent with empirical findings from the developmental literature and with recent advances in cognitive neuroscience.

Keywords: Analogy; metaphor; relational reasoning; development; connectionism; computational model.

Introduction

In the past three decades, there has been a growing appreciation for the possibility that analogy lies at the core of human cognition (Gentner, 1983; Hofstadter, 2001; Holyoak, Gentner, & Kokinov, 2001; Penn, Holyoak, & Povinelli, 2008). On this view, it is our ability to understand, produce, and reason with analogies that allows us to create the wonderfully rich and sophisticated intellectual and cultural worlds we inhabit.

In an attempt to illuminate the cognitive mechanisms that underlie analogical processing, several detailed computational models have been developed that capture key components of the analogical reasoning process (see French, 2002 for a review). Among the most influential of these models are the Structure Mapping Engine (SME: Falkenhainer, Forbus, & Gentner, 1989), and Learning and Inference with Schemas and Analogies (LISA: Hummel & Holyoak, 1997). These models vary drastically in many ways; however, they share a fundamental commitment to explicitly structured symbolic or hybrid representations (e.g. of objects and relations), together with the existence of a dedicated analogical mapping or binding mechanism that operates over these representations. Indeed, proponents of these approaches argue that analogical inference is beyond the reach of models that lack these properties, including fully distributed connectionist models (e.g. Gentner & Markman, 1993; Holyoak & Hummel, 2000).

While the structured approach has successfully captured adult behavior in numerous analogical reasoning tasks (e.g. Markman & Gentner, 1997; Hummel & Holyoak, 1997), it is unclear how this analogy-specific machinery comes to exist in the brain over the course of development. Even developmentally-oriented models such as DORA (Doumas, Hummel, & Sandhofer, 2008), which attempts to learn the structure used by LISA, assume a great deal of analogy-specific cognitive machinery without specifying how this machinery comes to exist in the first place.

Here, we address this issue by proposing that some forms of analogical processing may emerge gradually over the course of development through the operation of low-level domain general learning mechanisms (Flusberg, Thibodeau, Sternberg, & Glick, 2010; Leech, Mareschal, & Cooper, 2008). In support of this view we describe a set of simulations carried out using the Rumelhart network (Rumelhart, 1990), a neurally inspired model that has succeeded in capturing many results from the literature on semantic development in children (e.g. Rogers & McClelland, 2004) and whose variants have been used to understand the deterioration of conceptual knowledge in semantic dementia (e.g. Dilkina, McClelland, & Plaut, 2008).

Simulations

Our learning task is inspired by Hinton's (1986) family tree model, one of the first attempts to address relational learning in a connectionist network. The task of the model is to learn "statements" that are true about the various members of a family, including identity information, perceptual features, and relations between family members.

Input to the model consists of activating a Subject unit, corresponding to a particular family member, and a Relation unit. The Relation units correspond to the different kinds of relationships that can hold between subjects and objects (e.g. "is_named", "parent_of"). The network is wired up in a strictly feed-forward fashion, as shown in Figure 1, such that the input propagates forward through the internal layers, resulting in a set of predictions over the Object layer.

Over the course of training, the network's weights change (via backpropagation of the cross-entropy error on the output units) in order to better predict which Object outputs correspond to each combination of Subject and Relation inputs. As the model also contains intervening layers of units between the input and output layers, it is forced to re-represent the inputs as a distributed pattern of activation over these internal layers.

The underlying model parameters were identical in all of the simulations that we present. In all cases, the learning rate was .005 and the network was trained for 10,000 epochs. Results were averaged over 10 runs of each network in order to provide statistical tests. The hidden layers were identical in each case: 6 Subject Representation units and 16 Integration units. In all presented simulations, error on the training patterns was very low by the end of training (average cross-entropy error < .35).

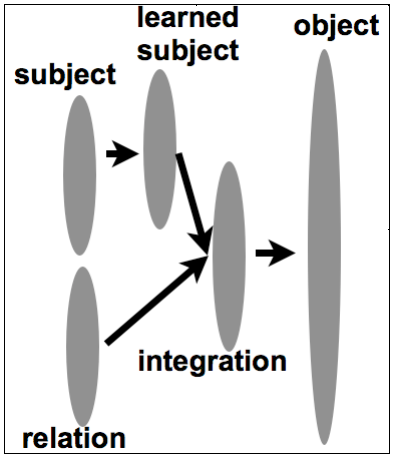

Figure 1: The network architecture.

The Basic Model

In the first simulation, the network learns about the Stripes and the Solids – two families with isomorphic relational structure (detailed in Table 1 and pictured in Figure 2) – with a single fact omitted about the Solid family. While the network knows that the daughter of the Solids owns their dog, it receives no information about who walks their dog. This network does a good job of learning the facts on which it is trained, but the question of interest is whether it can extend its knowledge to answer a question on which it received no training: who walks the Solids' dog?

SUBJECT	*RELATION*	OBJECT
$\text{Daughter}_{\text{Stripe}}$	*is_named*	Stripe, $\text{Daughter}_{\text{Stripe}}$
$\text{Daughter}_{\text{Stripe}}$	*is_a*	Stripe, human, child, daughter
$\text{Daughter}_{\text{Stripe}}$	*has*	blond hair, blue eyes
$\text{Daughter}_{\text{Stripe}}$	*daughter_of*	$\text{Mom}_{\text{Stripe}}$, $\text{Dad}_{\text{Stripe}}$
$\text{Daughter}_{\text{Stripe}}$	*sister_of*	$\text{Son}_{\text{Stripe}}$
$\text{Daughter}_{\text{Stripe}}$	*owner_of*	$\text{Dog}_{\text{Stripe}}$
$\text{Dog}_{\text{Stripe}}$	*walked_by*	$\text{Daughter}_{\text{Stripe}}$
$\text{Daughter}_{\text{Solid}}$	*is_named*	Solid, $\text{Daughter}_{\text{Solid}}$
$\text{Daughter}_{\text{Solid}}$	*is_a*	Solid, human, child, daughter
$\text{Daughter}_{\text{Solid}}$	*has*	brown hair, ponytail, green eyes
$\text{Daughter}_{\text{Solid}}$	*daughter_of*	$\text{Mom}_{\text{Solid}}$, $\text{Dad}_{\text{Solid}}$
$\text{Daughter}_{\text{Solid}}$	*sister_of*	$\text{Son}_{\text{Solid}}$
$\text{Daughter}_{\text{Solid}}$	*owner_of*	$\text{Dog}_{\text{Solid}}$
$\text{Dog}_{\text{Solid}}$	*walked_by*	??? ($\text{Daughter}_{\text{Solid}}$)

Table 1: A subset of the information that the network learns about each family member.

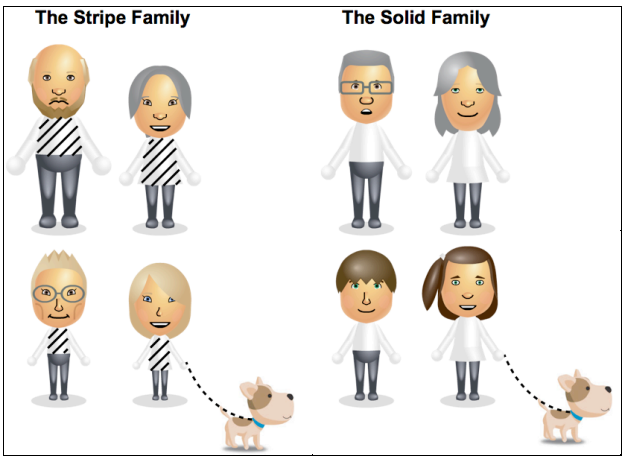

Figure 2: An illustration of the Stripe and Solid families, which served as the source and target domain.

We can contrast two major predictions. Naively, one might think that the network runs on raw association. As the Solids' dog is most similar to the Stripes' dog, the network would therefore conclude that the Stripes' daughter walks the Solids' dog! Alternatively, we might expect that the network will encode the relational structure between the two families, and so will correctly conclude that the person in the appropriate position within the Solid family -- namely, the daughter -- will be the one who walks their dog. In fact, the latter is the case: the network decides that within the Solid family, the daughter walks the dog. A paired t-test contrasting the activation levels of the Stripes' daughter with the Solids' daughter was highly significant, $t[9] = 7.75$, $p < .001$ (see Figure 3).

To ensure that the network used the relational similarity between the two families in making this inference, we ran a second simulation, in which the model was trained only on the Solid family, with no information about the Stripe family. In this network, the model does not conclude that the daughter walks the dog. Instead, the network decides that the dog walks itself! A paired t-test contrasting the activation levels of the Solids' dog with the Solids' daughter was highly significant, $t[9] = 5.61$, $p < .001$ (see Figure 3).

Simulations 1 and 2 do not, however, distinguish another set of predictions. It is possible that the network has learned to align the two families, either with respect to their relational structure or shared perceptual features, but only in an exact way. On this account, the model may have placed both mothers, both daughters, and both dogs in correspondence.

On the other hand, perhaps the network has learned the details of the family relations within each family as well as across families. In this case, it could learn a regularity like "whoever owns the dog, walks the dog," which is driven neither by perfect, global structural alignment nor by associations between surface features. This kind of relational binding is closely related to those tasks that previous researchers have argued can only be done using a distinct mapping mechanism operating over explicit symbols (e.g. Gentner & Markman, 1993; Holyoak &

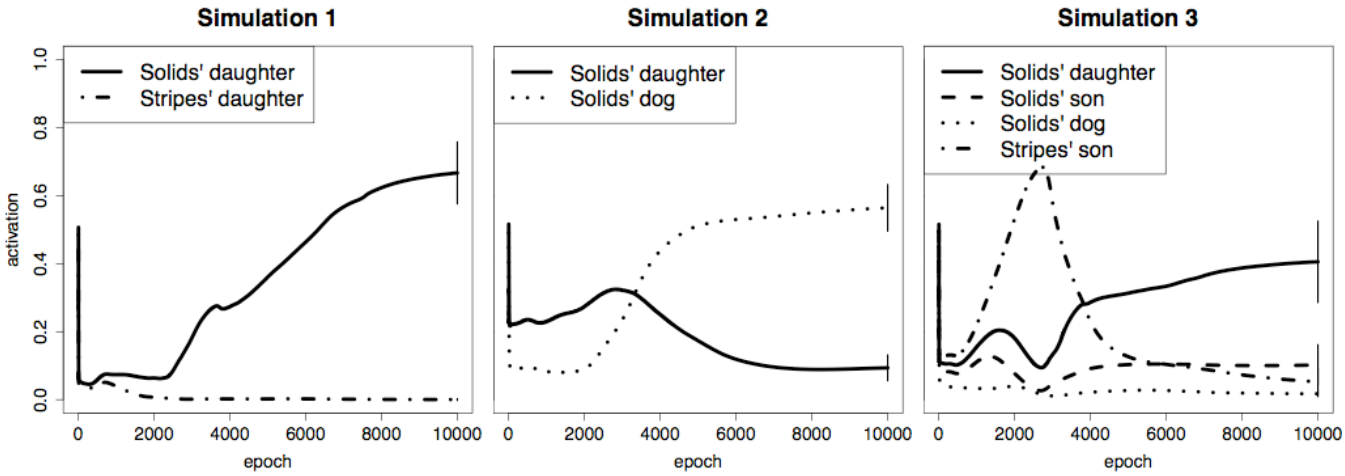

Figure 3: Activation levels for the target units in the first three simulations.

Hummel, 2000). Therefore, it would be a surprising and exciting finding if this network were able to succeed in such an abstract relational mapping task.

In order to distinguish between these hypotheses, we ran a third simulation, very similar to the first, except that in the Stripe family, the *son*, not the daughter, both owns and walks the dog. In this case, the network can only succeed in inferring that the Solids' daughter walks the dog if it learns the details of the relational structure, and in particular the regularity between owning a dog and walking it. This is precisely what occurs. Separate tests contrasting the activation level of the Solids' daughter with the activation level of the Stripes' son, $t[9] = 2.58$, $p < .05$, the Solids' son, $t[9] = 2.95$, $p < .05$, and the dog, $t[18] = 3.35$, $p < .01$, are all significant (see Figure 3).

This demonstrates that raw co-occurrence, or other simple associative processes which are often believed to underlie the performance of error-driven learning models (e.g., Hummel, 2010 in reply to Ramscar, Yarlett, Dye, Denny, & Thorpe, 2010), is not the key to learning in this model. It is, however, interesting to notice that the Stripes' son is the model's choice early in training, suggesting that the network first tends to make judgments predominately based on surface similarity, but over time shifts towards judgments based on relational similarity. This "relational shift" has been widely observed in the literature on the development of analogical reasoning abilities (e.g. Goswami, 1992). Intriguingly, this pattern is observed throughout the various simulations presented in this paper.

Extending The Model

We have shown a basic set of simulations that succeed in performing analogical inference from a family that is fully described to one that is less fully described. In the simulations below, we will extend the basic model in several directions, addressing possible objections to our claim that it is in fact succeeding at analogical inference. Each of these models will extend the third simulation, in which the son of one family owns and walks the dog, and the task of the model is to infer that the daughter of the other family, who owns the dog, also walks it.

Inexact Match – Can the model align non-isomorphic structures? We can investigate the extent to which the network relies on perfectly overlaying the two families by making the family structures only approximately match. In the fourth simulation, the Stripes have three children, two sons and a daughter, and one of the sons again owns (and walks) their dog. The Solid family still has two children, one son and one daughter, and their daughter owns the dog. Despite these changes, the model continues to make the inference that she probably walks the dog as well. A paired t-test contrasting the activation levels of the Solids' daughter with the Stripes' son was highly significant, $t[9] = 4.28$, $p < .01$. This demonstrates that the network can learn to draw inferences over structures, like many of those in previous work (such as Falkenhainer et al., 1989), which are only partially alignable.

Distributed Inputs – Does the model rely on implementing symbols? We have claimed that the success of this network depends on its development of distributed, subsymbolic representations, with which it can integrate the perceptual and the relational information about the family members within a high-dimensional representational space. Others might argue instead that the network is simply implementing symbols, and succeeds by performing some syntax-like transformation on those symbols. Such an argument may point to the localist input units representing the family members. We argue that the localist inputs are a useful simplification, but that focusing on them is a distraction, as the network can never directly exploit these localist units. Instead, it is required to re-represent each item as a pattern of activation over a hidden layer, as described above.

To make this point more clearly, we ran a fifth simulation that used distributed input representations for the family members. Following a model by Rogers and McClelland (2004), these were simply chosen to be each

family member's corresponding perceptual features. This should not assist the network in acquiring the relational structure; if anything, it should bias the network towards the surface-level perceptual features for generalization. Nevertheless, the network still infers that the owner of the dog walks it, transferring from the Stripes' son to the Solids' daughter. A paired t-test contrasting the activation levels of the Solids' daughter with the Stripes' son was highly significant, $t[9] = 6.05$, $p < .001$.

Non-overlapping Outputs – Does the model require perceptual overlap? On the other hand, one might argue that the architecture is biased in the opposite direction: the more direct overlap between the two families at the feature level (that is, at the Output layer), the less work the model needs to do to align their structures. What if only the relational similarity is available, as might be the case when constructing analogical mappings across very different domains of knowledge? This kind of analogy may be critical for explaining how analogy can subserve cognition and reasoning more generally.

To test this, we constructed a sixth simulation that had completely non-overlapping output units. The network essentially had two copies of each output property, so that each family's target representations were totally distinct. To succeed in generalizing the relation between the two families, the network would need to align the structures even in the absence of any surface-level similarity between the two families. And this is precisely what it did. Again, when the network is told that, in the Stripe family, the son owns and walks the dog, it concludes that for the Solids, the owner of the dog -- the daughter -- must also walk it. A paired t-test contrasting the activation levels of the Solids' daughter with the Stripes' son was significant, $t[9] = 3.58$, $p < .01$.

Scaling up – Can the model make inferences when given more than two families? Finally, it remains to be shown that the ability of the model to make analogies does not depend on it living in a world with only two different structures. Is it able to extend its learning to multiple families?

In this final simulation, the network learned about four rather than two distinct families (adding the Dash family and the Dot family). In this training set, a different person walks the dog in each family. Additionally, two of the families have slightly different structures: one has only a son, another has two sons and a daughter. Despite this added complexity, the network infers that in the target family, the daughter must also walk the dog. A within-subjects ANOVA using a planned contrast comparing the activation values of the Solids' daughter with the Stripes' son, the Dashes' mother, and the Dots' father (each a dog walker in their respective family) was significant, $F[1,36] = 42.40$, $p < 0.01$. Paired t-tests contrasting the activation levels of the Solids' daughter with the dog walkers in each of the other families including the Solids' son (t [9]=7.39, $p < .001$), the Dashes' mother ($t[9]=7.37$, $p < .01$), and the Dots' father ($t[9]=7.31$, $p < .001$) were also significant.

Discussion

To summarize the results of the above simulations, we have demonstrated that analogical reasoning can emerge from a general, neurally inspired connectionist model of semantic learning and reasoning. Critically, this analogical inference: (1) is driven by generalization from a source domain to a target domain; (2) relies on abstract relational structure, not surface-level similarities or direct featural associations or co-occurrences; (3) parallels important features of the development of analogy in children; (4) can operate over structures which only approximately match, or which are only partially alignable; (5) exploits structural similarity even in the absence of explicit overlap, allowing the possibility of cross-domain analogical inference in guiding learning; and (6) scales up to more complex training sets.

How is it that a connectionist model can succeed at this kind of analogical inference task? As we have demonstrated in several variations of the model, it is not due to any direct co-occurrence of feature, nor is it due to any kind of surface-level similarity between the items. Instead, we argue that part of the answer involves the progressive differentiation of its representations over the course of development. Initially, all the weights are set to very small random values, so the network essentially treats every family member, and every relation, as being the same. Over the course of training, the model learns to "pull apart" those representations that must be differentiated in order to produce the right answers. However, it only does so in response to erroneous predictions. This biases the network to reuse as much representational structure as it can get away with.

In this particular network, the families share a great deal of structure. As a result, the network's representations of the families become aligned over the course of training – since this allows the network to learn more efficiently (i.e., to reduce error more quickly). The side effect of this representational overlap is that when the network learns a fact about one family (e.g. one dog's owner walks it), the representations of the members of the other family (e.g. between that dog and its owner) get to come along for the ride. This is not to say that the model is stuck with its first guess about the structure of the world. As we indicated in the description of Simulation 3, and as is visible in other simulations, the model undergoes a developmental shift from predominantly perceptual to predominantly relational inference, when the environment warrants such a shift.

We can observe the process of progressive differentiation in this network by looking at a clustering diagram of activation patterns along the Subject Representation layer at different points in time for simulation 3 (see Figure 4). Early in training, the network groups items essentially at random, since the weights were initialized to very small values. Later in training, the network's representations capture both the surface similarities and the relational similarities between items. Progressive differentiation in semantic networks has been explored more extensively in previous work (Flusberg et al., 2010; Rogers & McClelland, 2004).

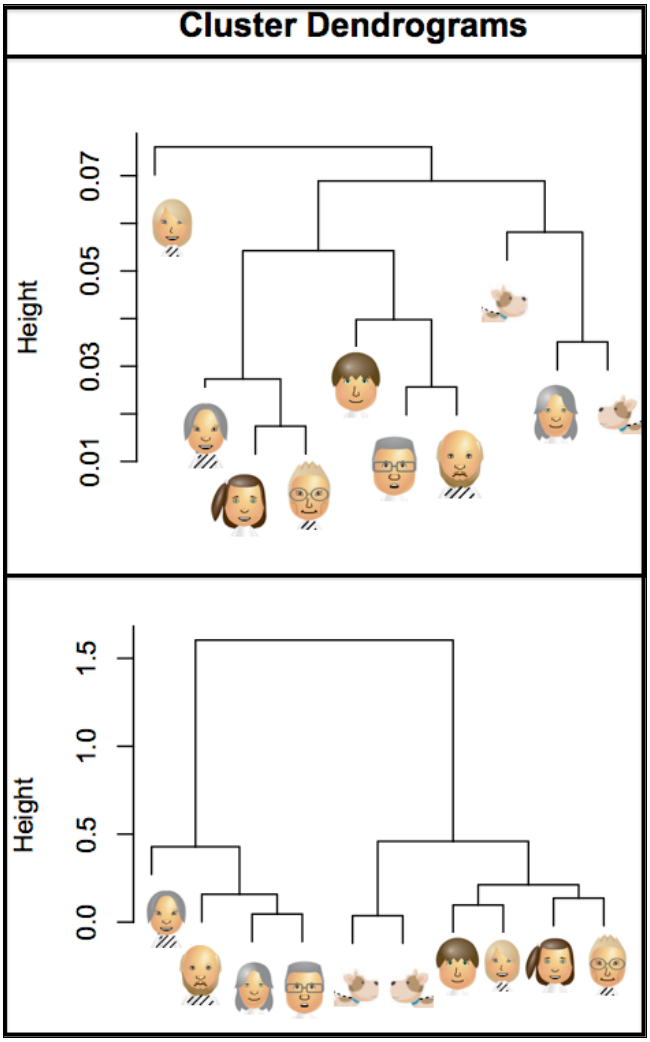

Figure 4: The hierarchical clusters above illustrate the similarity structure of the learned Subject representations in Simulation 3. Early in training (the upper panel), the network does not group individuals by family or relation. Later in training, at 1,300 epochs (the lower panel), the network has aligned the families according to their relational similarity.

It is also important to clarify what aspects of the environment we believe are encoded in our training patterns. Many of these patterns, such as those representing the visual features of the family members, might be thought of as arising from perception. However, others, particularly those representing familial relations such as "mother_of" and "owner_of", are much more likely to be encoded linguistically than visually. That is, part of our story is that learners hear language describing the people and things around them at the same time as they experience them directly, and these different sources of information are integrated whenever (as we think is almost always the case) there is some coherent covariation of information between the several sources (Rogers & McClelland, 2004). This is consistent with a great deal of empirical work demonstrating that relational language facilitates analogical inference and drives the relational shift in analogical development (Gentner, Simms, & Flusberg, 2009; Loewenstein & Gentner, 2005). Therefore, this approach views relational labels as another set of environmental regularities, serving the function of augmenting the statistical structure of the environment in ways that facilitate learning analogical representations (rather than as explicitly symbolic representations in the brain).

In one sense, then, our model supports the view that analogy is a special component of cognition, because it allows us to draw inferences about things we haven't directly experienced. In another sense, however, analogy is not special, in that we do not posit a separate set of cognitive machinery in order to accomplish analogical inference. Instead, these inferences emerge as a byproduct of learning to predict outcomes in an environment that contains relevant relational structure.

Previous work has highlighted the difficulties of pursuing subsymbolic accounts of analogy (e.g. Gentner & Markman, 1993; Holyoak & Hummel, 2000). In part because of the lack of progress in this direction, some researchers have gone so far as to claim that analogical reasoning requires at least some explicitly symbolic representations, or even that a subsymbolic account is impossible in principle. Our model is, of course, not the first to counter these claims (see, e.g., Leech et al. 2008). On the other hand, it may be the first to demonstrate that a model equipped with subsymbolic representations can make novel analogical inferences. Leech and colleagues (2008) pointed to a possible reframing similar to our own (and to the principle of coherent covariation described in Rogers & McClelland, 2004), suggesting that, "analogical inferences might best be understood as novel generalizations governed by the distributional information about which input features and relations co-vary across the base and target domains" (p. 403).

This is not to say that this kind of semantic network can account for all of human cognition. Far from it! We do not believe that these models can even explain all of human analogy. Many of the analogy tasks used in previous work, which models like SME and LISA can capture so well, rely on cognitive processes which we do not even attempt to model (e.g. Markman & Gentner, 1997; Morrison et al., 2004). In particular, we would agree that some of these tasks may rely on strong working memory and cognitive control processes, one-shot learning and episodic memory, and much richer linguistic abilities than we implement in this model. In our model, we treat relational language as a simple environmental cue, encoding a certain kind of statistical structure that is then used to shape semantic representations. While this is one important role of language in analogical reasoning, it is not the only one; the ability to verbally re-describe a situation to oneself, for example, is an important tool in many higher-level reasoning tasks (Williams & Lombrozo, 2010).

Therefore, we would like to suggest that one major unsolved problem is the integration of the kind of slow-learning semantic cognition model described in this paper with the online, structurally explicit models already in place. The extensive and valuable work on models such as

SME and LISA over the past twenty years, no less than the connectionist models we have implemented, must be used to guide future research into analogical processing across development, in behavior, and in the brain.

Acknowledgments

The authors would like to thank Jay McClelland and Lera Boroditsky for inspiration and helpful, critical discussion of the content of this paper and the issues it addresses. This material is based on work supported under a National Science Foundation Graduate Research Fellowship and a Stanford Graduate Fellowship.

References

Bowdle, B. and Gentner, D. (1997). Informativity and asymmetry in comparisons. *Cognitive Psychology*, 34(3):244–286.

Dilkina, K., McClelland, J., and Plaut, D. (2008). A single-system account of semantic and lexical deficits in five semantic dementia patients. *Cognitive Neuropsychology*, 25(2):136–164.

Doumas, L., Hummel, J., and Sandhofer, C. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review*, 115(1):1–43.

Falkenhainer, B., K.D., F., and Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. *Artificial Intelligence*, 41(1):1–63.

Flusberg, S. J., Thibodeau, P. H., Sternberg, D. A., and Glick, J. J. (2010). A connectionist approach to embodied conceptual metaphor. *Frontiers in Psychology*, 1(0):12.

French, R. (2002). The computational modeling of analogy-making. *Trends in Cognitive Sciences*, 6(5):200–205.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2):155–170.

Gentner, D. (2010). Bootstrapping the mind: Analogical processes and symbol systems. *Cognitive Science*, 34(5):752–775.

Gentner, D. and Markman, A. B. (1993). Analogy - watershed or waterloo? structural alignment and the development of connectionist models of cognition. In *Advances in Neural Information Processing Systems 5, [NIPS Conference]*, pages 855–862, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Gentner, D., Simms, N., and Flusberg, S. (2009). Relational language helps children reason analogically. In Taatgen, N. A. and van Rijn, H., editors, *Proceedings of the 31th Annual Conference of the Cognitive Science Society*.

Goswami, U. (1992). *Analogical reasoning in children*. Psychology Press.

Hinton, G. (1986). Learning distributed representations of concepts. In *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*, volume 1, page 12. Amherst, MA.

Holyoak, K., Gentner, D., and Kokinov, B. (2001). The place of analogy in cognition. In Gentner, D., Holyoak, K., and Kokinov, B., editors, *The analogical mind: Perspectives from cognitive science*, pages 1–19. MIT Press, Cambridge, MA.

Holyoak, K. and Hummel, J. (2000). The proper treatment of symbols in a connectionist architecture. In Dietrich, E. and Markman, A., editors, *Cognitive dynamics: Conceptual and representational change in humans and machines*, pages 229–264. Lawrence Erlbaum Associates, Hillsdale, NJ.

Hummel, J. and Holyoak, K. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, 104(3):427–466.

Hummel, J. E. (2010). Symbolic versus associative learning. *Cognitive Science*, 34(6):958–965.

Leech, R., Mareschal, D., and Cooper, R. (2008). Analogy as relational priming: A developmental and computational perspective on the origins of a complex cognitive skill. *Behavioral and Brain Sciences*, 31(04):357–378.

Loewenstein, J. and Gentner, D. (2005). Relational language and the development of relational mapping. *Cognitive Psychology, 50*(4):315–353.

Markman, A. and Gentner, D. (1997). The effects of alignability on memory. *Psychological Science*, 8(5):363–367.

Morrison, R., Krawczyk, D., Holyoak, K., Hummel, J., Chow, T., Miller, B., and Knowlton, B. (2004). A neurocomputational model of analogical reasoning and its breakdown in frontotemporal lobar degeneration. *Journal of Cognitive Neuroscience*, 16(2):260–271.

Penn, D., Holyoak, K., and Povinelli, D. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences*, 31(02):109–130.

Ramscar, M., Yarlett, D., Dye, M., Denny, K., and Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science*, 34(6):909–957.

Rogers, T. T. and McClelland, J. L. (2004). *Semantic Cognition*. MIT Press, Cambridge, MA.

Rumelhart, D. (1990). Brain style computation: Learning and generalization. In Zornetzer, S., Davis, J. L., and Lau, C., editors, *An introduction to neural and electronic networks*, pages 405–420. Academic Press.

Williams, J. and Lombrozo, T. (2010). The role of explanation in discovery and generalization: evidence from category learning. *Cognitive Science*, 34(5):776–806.