Learning and Generalizing Cross-Category Relations Using Hierarchical Distributed Representations

Dawn Chen[1] (sdchen@ucla.edu)

Hongjing Lu[1, 2] (hongjing@ucla.edu)

Keith J. Holyoak[1] (holyoak@lifesci.ucla.edu)
Departments of Psychology[1] and Statistics[2]
University of California, Los Angeles
Los Angeles, CA 90095 USA

Abstract

Recent work has begun to investigate how structured relations can be learned from non-relational and distributed input representations. A difficult challenge is to capture the human ability to evaluate relations between items drawn from distinct categories (e.g., deciding whether a truck is larger than a horse), given that different features may be relevant to assessing the relation for different categories. We describe an extension of *Bayesian Analogy with Relational Transformations* (BART; Lu, Chen & Holyoak, 2012) that can learn cross-category comparative relations from autonomously-generated and distributed input representations. BART first learns separate representations of a relation for different categories and creates second-order features based on these category-specific representations. BART then learns weights on these second-order features, resulting in a category-general representation of the relation. This hierarchical learning model successfully generalizes the relation to novel pairs of items (including items from different categories), outperforming a flat version of the learning model.

Keywords: relation learning; generalization; distributed representations; Bayesian models

Introduction

Learning Relations from Non-relational Inputs

A hallmark of human intelligence is the ability to learn and make inferences based on *relations* between entities, rather than solely on features of individual entities (for a review see Holyoak, 2012). A challenge for cognitive science is to explain how relations can be acquired. Some approaches to relation learning postulate some sort of grammar that generates possible relations, tacitly assuming that the origin of relational concepts is top-down (e.g., Tenenbaum, Kemp, Griffiths & Goodman, 2011). Doubtless some relations are constructed in a top-down fashion, but there is strong evidence that at least some relations are formed by bottom-up processes (Mandler, 1992). For example, children seem to acquire comparative relations such as *larger than* in stages, first learning features of individual objects, then extracting specific attributes of individual objects (e.g., a size value), and eventually linking attributes of paired objects to form a binary relation (Smith, 1989). Thus a basic problem for cognitive science is: How can relations be acquired from *non-relational* inputs?

A few models based on neural-network architectures (Doumas, Hummel & Sandhofer, 2008; Smith, Gasser & Sandhofer, 1997) have had some success in modeling bottom-up relation learning. However, it is difficult to fully evaluate the adequacy of proposed models of relation learning without first controlling the nature of the elementary inputs on which learning is based. A well-known limitation of models of analogy (for which relational knowledge is central) is that modelers typically create their own "toy" input representations, which may be tailored (perhaps inadvertently) so as to reduce task difficulty (Chalmers, French & Hofstadter, 1992). In modeling basic relation learning, it is critical to ensure that the non-relational inputs on which learning operates are autonomously created (rather than hand-coded by the modeler), and are of realistic complexity. When a model of relation learning is forced to operate on realistic inputs, theoretical issues that might have gone unnoticed with simpler inputs are brought to the fore.

Here we address one key issue that arises in learning relations from non-relational and realistically complex inputs: How can a learned relation be generalized to novel items, which have representations dissimilar to the items used to train the system? We first describe the basic model that served as our starting point, and then demonstrate how it could be extended to overcome apparent limits on its capacity to generalize.

Bayesian Model of Relation Learning

Recently, discriminative Bayesian models have been used to learn relations in a bottom-up fashion. A key idea is that an *n*-ary relation can be represented as a function that takes an ordered set of *n* objects as its input and outputs the probability that these objects instantiate the relation. The model learns a representation of the relation from labeled examples, and then applies the learned representation to determine whether the relation holds for novel examples. A second key idea is that relation learning can be facilitated by incorporating *empirical priors*, which are derived using some simpler learning task that can serve as a precursor to the relation learning task (Silva, Heller & Ghahramani, 2007).

These ideas were incorporated into *Bayesian Analogy with Relational Transformations* (BART), a discriminative

model that can learn comparative relations from non-relational inputs (Lu, Chen & Holyoak, 2012). Given independently-generated feature vectors representing pairs of animals that exemplify a relation, the model acquires representations of first-order comparative relations (e.g., *larger*, *faster*) as weight distributions over the features. The richest and most complex feature representations we have used are derived by applying the topic model (Griffiths, Steyvers, & Tenenbaum, 2007) to the English Wikipedia corpus. The output of the topic model is used to create a real-valued feature vector for each word. The simulations presented here are based on topic vectors.

BART represents a relation using a joint distribution of weights, **w**, over object features. A relation is learned by estimating the probability distribution $P(\mathbf{w} \mid \mathbf{X}_S, \mathbf{R}_S)$, where $\mathbf{X}_S$ represents the feature vectors for object pairs in the training set, the subscript **S** indicates the set of training examples, and $\mathbf{R}_S$ is a set of binary indicators, each of which (denoted by *R*) indicates whether a particular object (or pair of objects) instantiates the relation or not. The multivariate distribution of weights, **w**, constitutes the learned relational representation, which can be interpreted as quantifying the influence of the corresponding feature dimensions in **X** on judging whether the relation applies. The weight distribution can be updated based on examples of ordered pairs that instantiate the relation. Formally, the posterior distribution of weights can be computed by applying Bayes' rule using the likelihood of the training data and the prior distribution of **w** (which we assume to be independent of the object-pair features in the training set, $\mathbf{X}_S$):

$$P(\mathbf{w} \mid \mathbf{X}_S, \mathbf{R}_S) = \frac{P(\mathbf{R}_S \mid \mathbf{w}, \mathbf{X}_S) P(\mathbf{w})}{\int_{\mathbf{w}} P(\mathbf{R}_S \mid \mathbf{w}, \mathbf{X}_S) P(\mathbf{w})}. \quad (1)$$

The likelihood is defined as a logistic function for computing the probability that a pair of objects instantiates the relation, given the weights and feature vector:

$$P(R = 1 \mid \mathbf{w}, \mathbf{x}) = \frac{1}{1 + e^{-\mathbf{w}^T \mathbf{x}}}. \quad (2)$$

The prior, $P(\mathbf{w})$, is a Gaussian distribution and is constructed using a bottom-up approach in which initial learning of simple concepts provides empirical priors that guide subsequent learning of more complex concepts. Specifically, BART extracts empirical priors from weight distributions for one-place predicates such as *large* to guide the acquisition of two-place relations such as *larger*. Lu et al. (2012) trained BART on the eight one-place predicates (e.g., *large*, *small*, *fierce*, *meek*) that can be formed using the extreme animals at each end of the four relevant continua (size, speed, ferocity, and intelligence). When learning a two-place relation, BART automatically chooses the most relevant one-place predicate based on the training pairs for the relation, from which the empirical prior weight distribution is derived. For additional details on the operation of the model, see Lu et al. (2012).

BART's learned relations support generalization to new animal pairs. After receiving 100 training pairs represented using topic feature vectors, the model discriminates between novel pairs that instantiate a relation and those that do not with about 70-80% accuracy. The model yields the classic symbolic distance effect (Moyer & Bayer, 1976), as discrimination accuracy increases monotonically with the magnitude difference between items in a pair. Moreover, BART's learned weight distributions can be systematically transformed to solve analogies based on higher-order relations between the learned first-order relations (e.g., *opposite*). A simpler version of the model can predict magnitude values (based on human ratings) for individual objects (Chen et al., 2014), and an extension can use its learned relational representations to generate novel items instantiating the relation (Chen, Lu & Holyoak, 2013).

Relations Linking Distinct Categories

BART has thus demonstrated the promise of using a bottom-up approach to bootstrap relation learning. However, after being trained on animal pairs, the initial model failed on tests requiring generalization to inanimate objects (e.g., deciding if a toaster is larger than a shoe), performing only slightly above chance (about 57% accuracy).

This failure illustrates the importance of using realistic, independently-generated inputs. It would have been easy to hand-code a local feature representing a discriminative dimension (e.g., a size value) into the representations of all physical entities, in which case the model would readily generalize a relation learned from pairs within a specific domain (e.g., animal pairs) to all pairs of entities. But topic inputs do not provide such discriminative features. As a consequence, each relation acquired by BART is represented by a highly distributed pattern of weights across many feature dimensions. The pattern acquired using one subset of entities (animals) may not match the pattern needed to make relational discriminations for a different subset (inanimate objects). It is easy to imagine features that would impact a relational discrimination very differently for different classes of entities. For example, the feature "is from Africa" might predict that an animal is relatively large, but that a building is relatively small. Similarly, a topic feature of "found in nature" might be associated with certain large objects (e.g., mountain, ocean), but have weak predictive power for the sizes of most animals.

Semantic hierarchies are built out of disjunctions of different categories (Hampton, 1988), and more general categories are typically more difficult to learn than specific ones (e.g., Horton & Markman, 1980). Ecologically, it seems very likely that comparative relations most commonly are learned using pairs of entities drawn from a relatively specific category (e.g., a dog is larger than a cat; a bowl is larger than a glass). Nonetheless, adults are quite accurate in judging relative sizes of dissimilar entities they may never have previously considered together (e.g., a toaster is larger than a sparrow; Holyoak, Dumais & Moyer,

1979). Thus, comparative relations can be evaluated not only for items drawn from a single specific category, but also for items drawn from different categories. Here we describe a hierarchical extension of the BART model that addresses generalization across different categories, and report tests comparing its performance with a "flat" (non-hierarchical) version of the same model.

Hierarchical Model of Relation Learning

Overview

The computational goal is to learn comparative relations, such as *larger*, that span multiple categories (animals and inanimate objects), from mostly within-category examples (animal-animal pairs and object-object pairs) and a small number of cross-category examples (animal-object and object-animal pairs). We have developed a two-layer model for this task, illustrated in Figure 1. The bottom layer contains the raw input features, which we term *first-order* features. Based on within-category examples described by first-order features, the model first learns a separate, specialized representation of the relation for each category. From these initial category-specific relational representations, the model derives a small number of more abstract features (*second-order* features), which comprise the model's second layer of features. These second-order features have similar interpretations across different categories, abstracting away differences among the categories in how their first-order features influence relational judgment. The model then learns a second layer of weights that operate on second-order features to predict whether a pair of entities (possibly from different categories) instantiates the comparative relation. These second-order weights can be learned from cross-category as well as within-category examples. We use a small number of cross-category examples in most of our simulations, but we also experiment with using only within-category examples.

Domain and Inputs

Although in principle our model could learn any comparative relation that encompasses multiple categories, here we focus on the *larger* and *smaller* relations between animals and inanimate objects. To establish the "ground truth" of whether various pairs of entities instantiate these relations, we used a set of human ratings of size for a mixed set of animals and objects (Holyoak et al., 1979). After ambiguous words (e.g., "match") were removed, there were 32 animals and 111 inanimate objects for which topic representations were available.

To obtain topic feature vectors, we ran the topic model on Wikipedia corpus to obtain 300 topics. Note that deriving a higher number of topics would take a very long time on such a large corpus. The algorithm was used to generate a Markov chain. The first sample was taken after 1,000 iterations, and sampling was repeated once every 100 iterations until eight samples were produced. Each sample yielded a matrix in which the (i, j)th entry is the number of times that word i has been assigned to topic j. From this matrix, we derived a vector for each word based on the conditional probability of each topic given that word. We averaged the word vectors created from different samples of the Markov chain because they contained very similar topics (determined by examining the most probable words for each topic).

Finally, we reduced the dimensionality of the topic vectors by automatically choosing 50 *effective features* (those for which learning is enabled) for animals and objects separately. These were simply the features most associated with the items in each category (i.e., those with the highest values summed across the items). Thus, animals and objects are represented by different (but possibly overlapping) sets of effective features. As we will see, the hierarchical model allows entities from different categories to be represented by different sets of effective features. Figure 2 provides a visualization of the topic vectors (reduced to 10 features) for 10 animals and 10 objects. Note that the topic vector for each word is not constrained to be a probability distribution over topics.

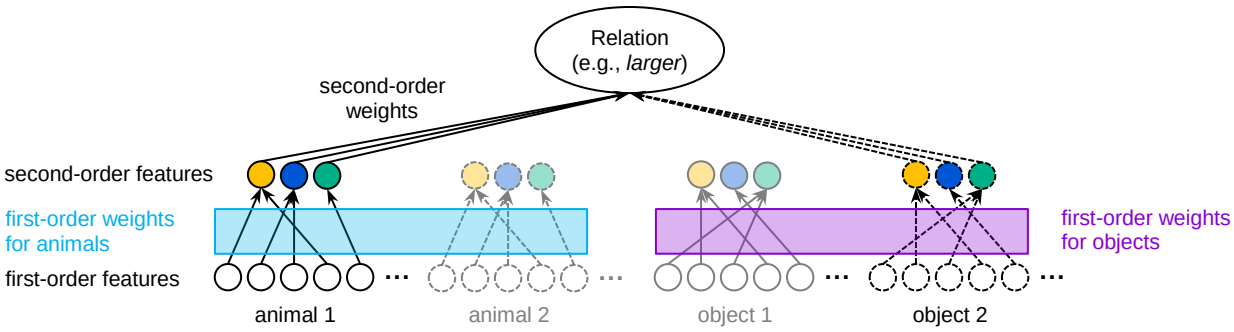

Figure 1: Illustration of the hierarchical model of relation learning with an example in which an animal occupies the first role and an object occupies the second role (e.g., *whale-ocean*). The first-order relational weights for animals are highlighted in light blue and the weights for objects are highlighted in purple. Features and weights for the second relational role are indicated with dashed lines. Each second-order feature is distinguished by a different color.

Table 1: The top 10 words associated with some example topics found by the topic model using the Wikipedia corpus.

Topic	Top 10 words
1	fish marine fishing sea species water waters ocean shark whale
2	food rice meat made milk foods served cuisine cooking eating
3	wear worn wearing hair dress wore made fashion clothing black
4	disease virus infection cases infected HIV diseases spread human AIDS
5	animals animal harry potter wild bear hunting lion horn sheep
6	forest species plant tree plants trees wood forests native found
7	land agricultural farm farmers food farming agriculture crops rural production
8	park hill creek mount parks area mountain trail located rock

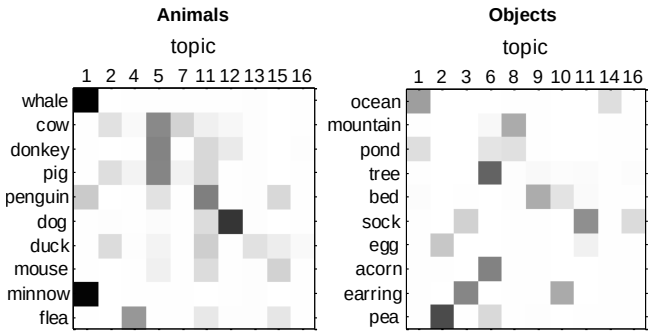

Figure 2: Illustration of topic vectors (reduced to 10 features to conserve space) for some example animals and objects, which are sorted by their sizes. The cell intensities represent feature values (dark indicates high values and light indicates low values).

Operation of the Model

Learning First-Order Weights In order to learn the initial category-specific relational representations, the original BART model is trained separately on 40 animal-animal pairs and 40 object-object pairs that instantiate *larger* (or *smaller*; we will use the *larger* relation to illustrate the operation of the model in the rest of this paper). This phase of learning yields two weight distributions: one that represents the *larger* relation for animals and one for objects.

Deriving Second-Order Features For each category, the model uses *k*-means clustering to separate the input features into three clusters, based on the pattern of learned first-order weights across the two relational roles (i.e., *larger-object* and *smaller-object*). We chose $k = 3$ because we hypothesized that the weights for comparative relations would generally follow three distinct patterns: (1) positive for the first role and negative for the second role, associated with features that predict largeness for the particular category, animals or objects (the *pro* cluster); (2) negative for the first role and positive for the second role, associated with features that predict smallness (the *con* cluster); and (3) around zero for both roles (with a few weights that are positive or negative for both roles), corresponding to features that are uncorrelated with size (the *neutral* cluster). These patterns were indeed the three clusters that the *k*-means algorithm typically found for the learned weights, as illustrated by the example in Figure 3. (The choice of $k = 3$ is further justified by a plot of the sum of all within-cluster point-to-centroid distances as a function of *k*, in which an "elbow" occurs at $k = 3$.)

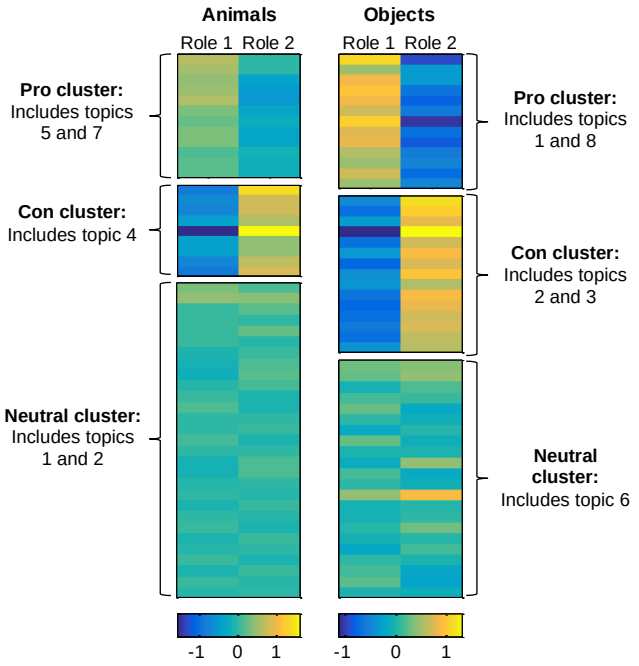

Figure 3: Illustration of the first-order weights that BART learned for each category and the three clusters found by the *k*-means algorithm for one run of the model. Each row within each set of weights represents a single topic dimension. See Table 1 for interpretation of some of the topics in each cluster. Note that corresponding clusters for the two categories include different features.

A single second-order feature is derived from each cluster. For a given pair of entities, each second-order feature is computed by taking the dot product between the first-order feature values in the corresponding cluster and the learned first-order weight means on those features. As illustrated in Figure 1, the weights for the appropriate item category and relational role are used. For example, given the pair *whale-ocean*, the role-1 weights for animals are used to calculate the second-order features for *whale*, and the role-2 weights for inanimate objects are used for *ocean*. Since there are three clusters for each role, a total of six second-order features are calculated for each pair of entities. Critically, the second-order features are indifferent to the identity and number of the first-order features included in each cluster.

Learning Second-Order Weights We use the original BART model with an uninformative prior (zero means and identity covariance matrix) to learn the weights on the second-order features from a small number (10 or fewer) of cross-category pairs that instantiate *larger*. In our simulations, we experiment with using different numbers of such training pairs.

Baseline Model

We compare the hierarchical model to a "flat" model that does not include second-order features and weights. This is simply the original BART model, given 40 within-category examples of each type and 10 cross-category examples. For this model, entities from different categories must be represented by the same set of effective features. We use the union of the two sets of 50 features selected for animals and objects, resulting in a set of 79 features.

Simulation Results

We evaluated the models by running them on ten different sets of training pairs and novel test pairs. These pairs were randomly chosen from the set of all possible pairs that instantiate a specific relation. Each model calculates the probability of instantiating the relation for each test pair (e.g., *penguin-flower*) and its reverse (*flower-penguin*), so the test set contains an equal number of positive and negative examples of the relation. The model is considered to be correct on a test pair if the pair instantiates the relation and its predicted probability is greater than .5, or if the pair does not instantiate the relation and its predicted probability is less than .5. Results are averaged over the ten runs. Here we report the results of several tests of the generalization ability of each model.

Testing on Pairs of Each Type

In the first test, we trained each model on 40 animal-animal pairs, 40 object-object pairs, and 10 cross-category pairs. We then tested each model on 100 novel animal-animal, object-object, or cross-category pairs. Figure 4 shows the mean accuracy on the *larger* relation across ten runs for the two models on each type of test pair (the results for *smaller* were similar). The hierarchical model outperformed the baseline model by about 19 percentage points across the three types of test pairs. In subsequent tests, we focus on cross-category test pairs, which are the most interesting type because the models must consider items from different categories together (the type that occurs least frequently in the training set).

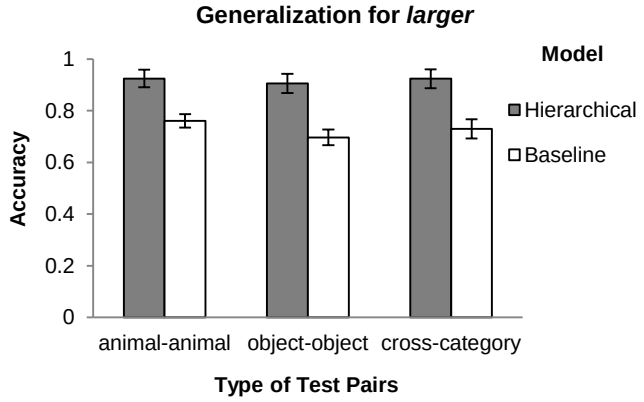

Figure 4: The models' mean accuracy on a generalization test for *larger* across ten runs for different types of test pairs. Error bars indicate 1 standard deviation.

Number of Cross-Category Training Examples

We varied the number of cross-category training examples while keeping constant the number of within-category examples of each type (40), and tested the models on 100 novel cross-category pairs. Figure 5 shows the mean accuracy on *larger* across ten runs for the two models as a function of the number of cross-category training pairs (ranging from 0 to 10). (Once again, the results for *smaller* were similar.) The hierarchical model performed at chance level when no cross-category examples were provided, because its second-order weights had not been learned and were simply the prior (zero means and identity covariance matrix). The baseline model was insensitive to the number of cross-category training pairs, whereas accuracy for the hierarchical model increased with the number of cross-category pairs, besting the baseline model after just four examples.

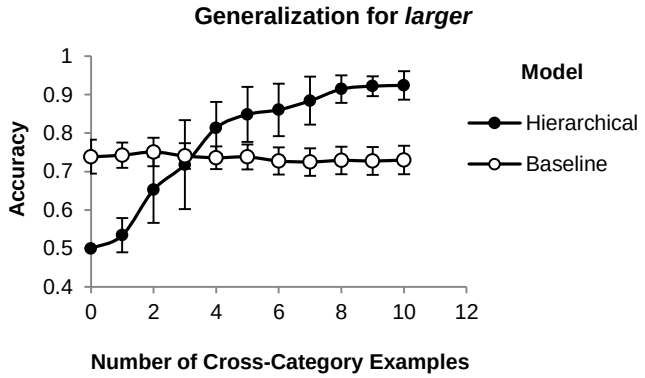

Figure 5: Learning curves for the two models: mean accuracy of models on a generalization test for *larger* across 10 runs as a function of number of cross-category training pairs. Error bars indicate 1 standard deviation.

Type of Examples for Second-Order Weights

The hierarchical model's second-order weights apply equally to pairs drawn from the same or different categories because they operate on the same set of second-order features for each category. Thus, the hierarchical model improves performance on all pair types (see Figure 4). It follows that the model should perform well on cross-category test pairs even when its second-order weights are learned from within-category examples only. We again trained the first-order weights using 40 animal-animal and 40 object-object pairs. We then trained the model's second-

order weights on 10 additional pairs: either 10 animal-animal pairs, 10 object-object pairs, 5 within-category pairs of each type, or 10 cross-category pairs. Finally, we tested the model on 100 novel cross-category pairs. The model achieved accuracies of 91%, 92%, and 91% on *larger* when trained respectively on 10 animal-animal pairs, 10 object-object pairs, and 5 within-category pairs of each type. In comparison, accuracy was 93% when the model's second-order weights were learned from 10 cross-category examples. Thus, the hierarchical model can learn to make relational judgments for cross-category pairs without ever encountering a single example of such pairs.

General Discussion

We have demonstrated that a hierarchical extension of the BART model can learn and generalize comparative relations across dissimilar categories, using non-relational topic vectors as inputs. The key to the model's performance is its creation of higher-order features based on patterns of category-specific first-order weights applied to primitive features representing individual entities.

Insight into the superior performance of the hierarchical model can be provided by examining the topic dimensions from which the model created each second-order feature. One topic that appears in the feature representations of both animals and objects involves fish and the sea (topic 1; see Table 1). For objects, this topic was a part of the pro feature for *larger* (objects related to the sea tend to be large, such as *ocean*, *pond*, and *boat*), whereas for animals it was a part of the neutral feature (marine animals span the full range of sizes). Another feature shared by animals and objects is a topic related to food (topic 2 in Table 1). This feature was a part of the con feature for objects (food items and objects related to cooking tend to be small), but a part of the neutral feature for animals (animals of various sizes are eaten).

As we expected, first-order features impacted relational judgments differently for different categories of entities. For each of the topic dimensions mentioned above, the baseline model is forced to learn a single weight that applies to both animals and objects, and therefore cannot capture these differences between categories. In contrast, the hierarchical model accommodates these differences by assigning each topic dimension to the cluster corresponding to the most appropriate second-order feature (pro, con, or neutral), which may differ for each category. These second-order features have similar interpretations across different categories, and hence simplify the complex and distributed input representations from which relations can be acquired.

Acknowledgments

Preparation of this paper was supported by a grant from the National Science Foundation (BCS-135331).

References

Chalmers, D. J., French, R. M., & Hofstadter, D. R. (1992). High-level perception, representation, and analogy: A critique of artificial intelligence methodology. *Journal of Experimental and Theoretical Artificial Intelligence*, *4*, 185–211.

Chen, D., Lu, H., & Holyoak, K. J. (2013). Generative inferences based on a discriminative Bayesian model of relation learning. In M. Knauf, M. Pauven, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 2028-2033). Austin, TX: Cognitive Science Society.

Chen, D., Lu, H., & Holyoak, K. J. (2014). The discovery and comparison of symbolic magnitudes. *Cognitive Psychology*, *71*, 27-54.

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

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*, 211-244.

Hampton, J. A. (1988). Disjunction of natural concepts. *Memory & Cognition*, *16*, 579-591.

Holyoak, K. J. (2012). Analogy and relational reasoning. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 234-259). New York: Oxford University Press.

Holyoak, K. J., Dumais, S. T., & Moyer, R. S. (1979). Semantic association effects in a mental comparison task. *Memory & Cognition*, *7*, 303-313.

Horton, M. S., & Markman, E.M. (1980). Developmental differences in the acquisition of basic and superordinate categories. *Child Development*, *51*, *708-719.*

Lu, H., Chen, D., & Holyoak, K. J. (2012). Bayesian analogy with relational transformations. *Psychological Review*, *119*, 617-648.

Mandler, J. M. (1992). How to build a baby: II. Conceptual primitives. *Psychological Review*, *99*(4), 587.

Moyer, R. S., & Bayer, R. H. (1976). Mental comparison and the symbolic distance effect. *Cognitive Psychology*, *8*, 228–246.

Silva, R., Heller, K., & Ghahramani, Z. (2007). Analogical reasoning with relational Bayesian sets. In M. Mella & X. Shen (Eds.), *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*.

Smith, L. B. (1989). From global similarities to kinds of similarities: The construction of dimensions in development. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 147-177). Cambridge, UK: Cambridge University Press.

Smith, L. B., Gasser, M., & Sandhofer, C. M. (1997). Learning to talk about the properties of objects: A network model of the development of dimensions. In R. L. Goldstone, D. L. Medin & P. G. Schyns (Eds.), *Advances in the psychology of learning and motivation*, *Vol. 36: Perceptual learning* (pp. 219-255). San Diego, CA: Academic Press.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*, 1279-1285.