Emergence of analogy from relation learning

Hongjing Lu[a,b,1], Ying Nian Wu[b], and Keith J. Holyoak[a]

[a]Department of Psychology, University of California, Los Angeles, CA 90095; and [b]Department of Statistics, University of California, Los Angeles, CA 90095

Edited by James L. McClelland, Stanford University, Stanford, CA, and approved January 9, 2019 (received for review August 28, 2018)

By middle childhood, humans are able to learn abstract semantic relations (e.g., antonym, synonym, category membership) and use them to reason by analogy. A deep theoretical challenge is to show how such abstract relations can arise from nonrelational inputs, thereby providing key elements of a protosymbolic representation system. We have developed a computational model that exploits the potential synergy between deep learning from "big data" (to create semantic features for individual words) and supervised learning from "small data" (to create representations of semantic relations between words). Given as inputs labeled pairs of lexical representations extracted by deep learning, the model creates augmented representations by remapping features according to the rank of differences between values for the two words in each pair. These augmented representations aid in coping with the feature alignment problem (e.g., matching those features that make "love-hate" an antonym with the different features that make "rich-poor" an antonym). The model extracts weight distributions that are used to estimate the probabilities that new word pairs instantiate each relation, capturing the pattern of human typicality judgments for a broad range of abstract semantic relations. A measure of relational similarity can be derived and used to solve simple verbal analogies with human-level accuracy. Because each acquired relation has a modular representation, basic symbolic operations are enabled (notably, the converse of any learned relation can be formed without additional training). Abstract semantic relations can be induced by bootstrapping from nonrelational inputs, thereby enabling relational generalization and analogical reasoning.

semantic relations | analogy | word embeddings | learning | generalization

Human intelligence depends on the capacity to think about the relations between things, rather than simply about individual entities. This ability makes it possible to understand an indefinite number of instantiations of the same abstract relation. The capacity to reason using abstract relations is much more developed in humans than in any other species, perhaps constituting a qualitative difference in intelligence (1). We can grasp, for example, that "love" and "hate" are related to one another in much the same way as "rich" and "poor," and that "blindness" and "sight" are related in the same way as "poverty" and "money" are. It is known that the ability to reason about abstract semantic relations emerges during early childhood (2), with children being taught the concepts of antonym and synonym in elementary school (3); however, how abstract relations might be learned remains unclear.

This question is tightly linked to the question of how humans actually represent relations. Within psychology and cognitive science, the dominant view has been that people acquire explicit representations of relations that are stored in semantic memory. For example, most computational models of analogical reasoning assume that a binary relation such as "lack of" has the structure of a two-place predicate, allowing an indefinite number of instantiations (e.g., blindness is the lack of sight, ignorance is the lack of knowledge) (4–7). However, there is no agreement about how (or even if) structured relations are acquired. One approach to addressing this issue has been to generate structural representations of relatively formal relations in a top-down manner by assuming an innate grammar of relations (8); however, such models have not addressed the acquisition of the less-clearly defined semantic relations that arise in natural languages. Other models have addressed learning of semantic relations within a variety of neural network architectures. Some of these models have aimed to create structured relations in which constituent roles can be distinguished (e.g., "lack of" consists of a role expressing a need, such as "sight," and a role expressing its absence, such as "blindness") (9, 10), whereas other models represent relations without explicit roles (11, 12). Most models of relation learning have been initially applied to small hand-coded inputs, with fewer being tested on realistic inputs not specifically chosen to enable relation learning.

In contrast to models that explicitly aim to learn relations, recent deep learning models, such as Word2vec (13, 14) and GloVe (15), have raised the possibility that basic relational reasoning can be achieved without explicit representations of relations. These models take a large text corpus as input, extract distributional statistics that allow each word to predict neighboring words in sentences (local context), and output a vector representation for each individual word, termed "word embedding." Vectors for similar words are located close together in a high-dimensional semantic space. Relational words (e.g., verbs, prepositions) are represented in the same space as nouns and adjectives, without any structured roles. These models have achieved some degree of success in solving verbal analogies based on difference vectors for pairs of words (as anticipated in ref. 16). Difference vectors exhibit a parallel relation (typically defined by cosine distance) between some analogous word pairs (e.g., "king:queen::man:woman"). In this example, a semantic relation that a human might interpret as an antonym is represented only implicitly in the model by parallel difference vectors (but for important caveats, see ref. 17). Nonetheless, despite their suggestive successes, a profound gap continues to separate relational processing in AI models from human capacities. Learning in current deep learning models relies on massive data (e.g., ref. 18); however, humans (even young children) can learn

Significance

The ability to learn and make inferences based on relations is central to intelligence, underlying the distinctively human ability to reason by analogy across dissimilar situations. We have developed a computational model demonstrating that abstract relations, such as synonymy and antonymy, can be learned efficiently from semantic feature vectors for individual words and can be used to solve simple verbal analogy problems with close to human-level accuracy. The approach illustrates the potential synergy between deep learning from "big data" and supervised learning from "small data." Core properties of high-level intelligence can emerge from relatively simple computations coupled with rich semantics. The model illustrates how operations on nonrelational inputs can give rise to protosymbolic relational representations.

Author contributions: H.L. and K.J.H. designed research; H.L., Y.N.W., and K.J.H. performed research; H.L. and K.J.H. analyzed data; and H.L., Y.N.W., and K.J.H. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: hongjing@ucla.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1814779116/-/DCSupplemental.

Published online February 15, 2019.

new relational concepts rapidly from small numbers of examples, primarily by transferring learned knowledge to facilitate the acquisition of new relational concepts (19).

Here we propose an approach to relation learning based on operations that progressively re-represent implicit relational knowledge in increasingly explicit forms. Relational knowledge develops as the result of initial biases that guide bottom-up statistical learning, coupled with a series of bootstrapping operations that transform unstructured representations (semantic vectors for individual words) into protosymbolic representations that can support relational inferences in high-level cognition. For example, after acquiring a representation of the relation "category:instance" (e.g., "tree:oak"), this relation representation can be transformed into its corresponding converse relation "instance:category" ("oak:tree"), without any additional training with examples.

The model of relation learning described here aims to integrate the computational approaches developed in cognitive science with those underlying current AI models. Specifically, we use word embeddings as a starting point for the induction of structured relations that can support more complex analogical reasoning. Our general goal is to combine the type of learning that produces semantic vectors from "big data" with supervised learning that acquires explicit relations from "small data." Arguably, it is the combination of these different mechanisms for learning that enables human relational reasoning. The text corpora used by deep learning models can be viewed as a proxy for the massive linguistic input that a human normally encounters over many years. A gradual learning process operating over this large input yields rich semantic representations of individual concepts, coded as modular feature vectors. Because these vectors are derived from predictions about co-occurring words, some features are likely to be correlated with semantic relations between words. These vectors in turn support bootstrapping to extract modular relation representations based on supervised learning applied to a modest number of training examples consisting of related pairs of entities.

The model incorporates a heuristic solution to the feature alignment problem, identifying common patterns across examples based on distinct sets of features. The model is thereby able to learn that, for example, "love:hate" and "rich:poor" exemplify the same relation, even though the basis for the relation differs radically across the two examples (with the former pair contrasting on a dimension of emotional attitude and the latter on economic status). The model also creates a distributed representation of relations between any pair of lexicalized concepts, allowing the model not only to assess whether the pair instantiates a given relation (e.g., "love" and "hate" are antonyms), but also to represent the likelihood of the two entities instantiating other abstract semantic relations (e.g., while "love" and "hate" contrast as antonyms, they also have important similarities, such as being types of emotions). These acquired relational representations support analogical reasoning and enable basic symbolic operations, notably the ability to create the converse of a learned relation by a rule-like transformation, without any additional training.

Model of Relation Learning

The model described here is Bayesian Analogy with Relational Transformations (BART). (MATLAB code for the BART model is available at cvl.psych.ucla.edu/BART2code.zip.) An earlier version of this model had the limited function of learning comparative relations (e.g., "larger," "smarter") (10, 20, 21). BART takes as inputs feature vectors for pairs of words that constitute positive or negative examples of a semantic relation. For example, a vector formed by concatenating the individual vectors for "love" and "hate" would be a positive example of the antonymy relation, but a negative example of the category membership relation. The feature vectors of lexicalized concepts used in the present work were taken from word embeddings with 300 feature dimensions produced by the Word2vec model trained on a corpus of articles published in Google News (13, 14). The SemEval-2012 Task-2 dataset (22) was used to teach BART the representations for 79 abstract semantic relations. This dataset is based on a taxonomy of semantic relations (23) and includes 10 general types (e.g., class inclusion, similar, contrast, cause-purpose). The dataset includes 3,215 word pairs, with 35~48 pairs for each of the 79 relations (*SI Appendix*).

From a psychological perspective, the key property of the SemEval-2012 Task-2 dataset is to provide human ratings of the prototypicality for each relation. Just as instances of basic object categories, such as fruit or furniture (24), follow a typicality ordering (e.g., an orange is a more typical fruit than a watermelon), people are sensitive to differences in the typicality of instances of abstract relations (e.g., for the reverse relation, the pair "fail:succeed" is considered a better instantiation than "eat:starve"). Accordingly, a basic test of a model of relation learning is to predict human typicality gradients for the normed relations. The mean split-half reliability of human typicality ratings across all 79 relations (calculated as described in ref. 25) was 0.83, a value that provides an approximate upper bound on the success any model could attain in predicting the human judgments. More than a dozen machine-learning algorithms have been applied to this task, and the highest mean rank-order correlation reported so far is 0.41 (26).

Training Inputs. BART was trained on each of the 79 abstract relations included in the SemEval-2012 Task-2 dataset. Training was conducted separately for each relation. Just as children generally learn object categories from typical examples (24), it is highly plausible that people's first encounters with relations will involve a small number of typical examples. Accordingly, a small number of the best example pairs served as positive instances for learning each relation. We used 20 positive training examples for most of the simulations reported below. In addition, a fixed set of 64–74 negative instances was selected, using the top example for each relation from general types other than that of the target relation.

Relation Learning and Inference. The BART model consists of a three-stage process for learning a broad range of abstract semantic relations (Fig. 1). Each relation representation is modular; that is, each relation is represented with a distinct distribution of weights. In its first stage, BART exploits the heuristic that features playing similar functional roles will tend to occupy similar ranks in an ordering of differences between paired words. Specifically, the model computes the difference vector between two paired words and sorts them to derive the rank order of the differences. Features are then dynamically remapped to align them by position in the ranking, rather than by identity. The difference-ranked feature vector highlights semantic features that tend to be aligned with respect to functionally relevant differences between the two words in a pair. For example, the various features that provide a distributed representation of emotional attitude for "love-hate" will tend to show greater differences, resulting in higher ranks; however, for "rich-poor," higher-ranked features will likely be those relevant to financial status, reflecting larger differences in these semantic dimensions. Thus, the first stage of the model yields augmented feature vectors $(\mathbf{f'}_1, \mathbf{f'}_2)$, including the raw vector for the word pair $(\mathbf{f}_1, \mathbf{f}_2)$ (bottom layer in Fig. 1) and the vector sorted according to ranked differences (second layer, in which the second and fourth feature sets are reorderings of the first and third sets, respectively). Such a sorting algorithm can be implemented by a three-layer feedforward neural net of a size polynomial in n (number of items to be sorted) (27).

Note that unlike in previous computational models of analogical mapping, this postulated alignment process is presymbolic, in that it is performed before any structured relations have been acquired. Features are functionally aligned without the support of preexisting relations or role bindings.

In the second stage, the model selects a subset of important features using logistic regression on the augmented difference vector (both raw and ranked; 600 dimensions total). The regression includes sparsity regularization to produce the third layer in Fig. 1. The selected subset of dimensions will include those highly relevant in discriminating the target relation from alternative semantic relations. Across the 79 relations, the mean

PSYCHOLOGICAL AND COGNITIVE SCIENCES

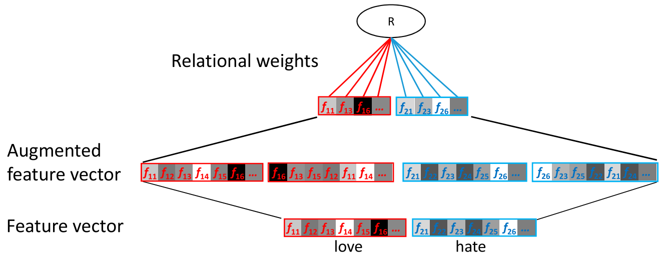

Fig. 1. Illustration of the BART model for learning a semantic relation, *R*, from feature vectors for word pairs. Colors indicate features and weights associated, respectively, with the first word (red) and the second word (blue). Note that semantic roles (parts of networks based on first word and second word, respectively) are distinctly separated in the distribution of relation weights. The feature alignment problem is addressed by augmenting the raw feature vectors (layer 1 at bottom) by concatenating features ranked by difference (layer 2), followed by feature selection (layer 3), and learning of relational weight distributions using the variational Bayesian method with a contrast-based prior.

number of selected dimensions was 127 (range, 75–216), and the mean proportion of selected dimensions based on ranked features was 0.16 (range, 0–0.45). Ranked features were selected in largest number for relations of the general types "contrast" (typically dimensions with extreme difference values; i.e., large positive or negative feature differences) and "similar" (typically dimensions with small feature differences).

In the third stage, BART uses the selected features of word pairs, $\mathbf{f}_s$, in training examples to estimate weights distributions, $\mathbf{w}$, for representing a particular relation, R, by applying the Bayes rule as

$$P(\mathbf{w}|\mathbf{f}_s, R) \propto P(R|\mathbf{f}_s, \mathbf{w})P(\mathbf{w}), \qquad [1]$$

in which the prior on mean weights, $P(\mathbf{w})$, is derived from a contrast-based empirical prior. Specifically, the means for weights associated with the features of the first word are set as the coefficients estimated in stage 2 for the corresponding feature dimension; to form a contrast, the means for the weights associated with the features of the second word are generated by reversing the sign on the corresponding weights for the first role. The likelihood term is defined by a logistic function on the relational weights and selected semantic features, $(1+e^{-\mathbf{w}^T f})^{-1}$. The inference is implemented using the variational Bayesian method to approximate integrals involved in probabilistic models (28). A formal statement of the model and pseudocode are provided in *SI Appendix*.

After learning the distribution of weights associated with the subset of semantic features $\mathbf{f}'_L, R_L$, the model can estimate the posterior probability that a new word pair, $\mathbf{g}'$, instantiates the learned relation, R_i, by marginalizing the weight distribution for this relation:

$$P(R_i|\mathbf{g}') = \int P(R_i|\mathbf{g}', \mathbf{w})P(\mathbf{w}|\mathbf{f}'_L, R_L). \qquad [2]$$

Results

Predicting Typicality for Semantic Relations. During training, the BART model was provided only with binary labels for examples of each relation (positive or negative examples); thus, the model was not provided with any information about the typicality ordering of the positive examples. To assess the model's ability to predict typicality for semantic relations, the model calculated the posterior probability of instantiating the relation for each word pair listed in ref. 22 for that relation (i.e., the 20 training positive examples and the remaining 15~28 word pairs not included in the training phase). Spearman rank-order correlation coefficients were calculated between the model-generated posterior probabilities and human typicality judgments for each relation.

Fig. 2 depicts the correlation values averaged across all relations within each of the 10 types. Predictions derived from BART were compared with those derived from the Word2vec model, based on the cosine distance between the difference vector for a word pair and that for a paradigmatic example for that relation (14). For all 10 relation types, BART achieved high rank-order correlations between human typicality ratings and predicted probabilities derived from the model. Across all 79 individual relations, the model's mean Spearman correlation with the human ordering was 0.81 (range, 0.65–0.91). The performance of BART considerably exceeded the mean correlation of 0.34 achieved using Word2vec as a baseline.

Note that the mean split-half reliability of the orderings across all 79 relations was 0.83, a value that provides an approximate upper bound on the success in predicting the human typicality gradients that any model could attain. The fact that BART's correlation with human judgments approached this upper bound demonstrates that supervised learning with small training samples (20 positive examples) can significantly enhance the representational sensitivity of a computational model of semantic relations. Furthermore, the consistent improvement obtained across diverse semantic relations indicates that learning mechanisms originally developed for learning comparatives generalize very well to learning a broad range of abstract semantic relations.

Solving Verbal Analogy Problems Based on Learned Relations. Solving verbal analogy problems requires specifying the full relationship between each pair of concepts, which may often have nontrivial posterior probabilities of instantiating multiple relations. In BART, the pool of learned relations affords a natural mechanism to create a more refined representation of the relation(s) between two paired words. The posterior probabilities calculated for all known relations form a relation vector, with each element indicating how likely a word pair instantiates a relation. Thus, the result of this operation is to create a distributed representation of the relation(s) between two words, with the original semantic features being projected into a transformed space that can be used to assess relation probabilities. The use of a distributed representation across learned relations allows the model to capture both the fact that word pairs vary in the degree to which they instantiate any given relation (i.e., the relation typicality effect, discussed above) and the fact that they often instantiate multiple relations, enabling the solution of simple verbal analogy problems.

Fig. 3 displays the vector of posterior probabilities for four word pairs that instantiate the specific relation of "contradictory" (within the general type of "contrast"). These are ordered from a highly typical example of contradictory ("hot-cold") to a less typical (but still positive) example ("abundant-scarce"). Each of these four examples was given to BART as a positive example for the learned relation of contradictory. These pairs yield high to moderate posterior probabilities for multiple specific relations

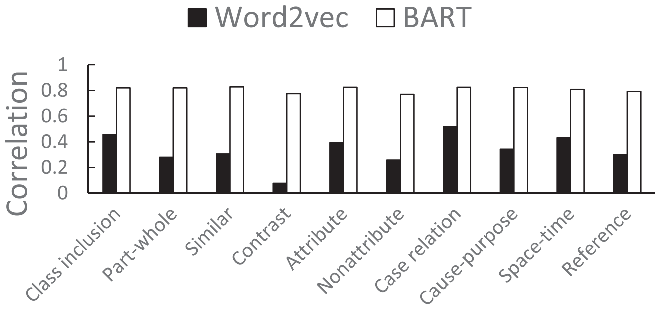

Fig. 2. Predictions of relation typicality. Correlations between human typicality ratings and model predictions for 10 types of abstract relations after training with 20 positive examples for each relation by BART and for the baseline Word2vec model.

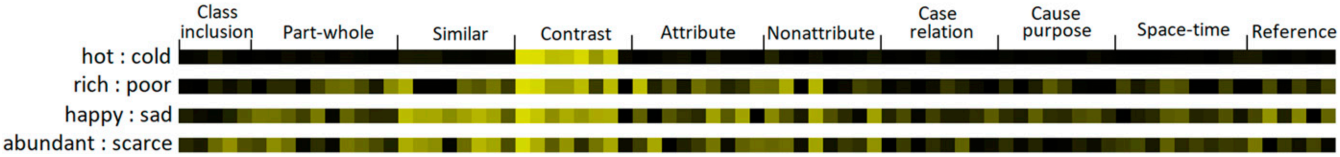

Fig. 3. Illustration of relation vectors, in which each element indicates how likely a word pair is to instantiate a relation. High probabilities are represented in yellow; low probabilities, in black. Relation vectors are shown for four positive examples of the specific contradictory relation, ordered by typicality with respect to that relation. As the typicality of the word pairs with respect to the contradictory relation declines, the relation vector becomes increasingly distributed across a broader range of relation types.

within the general type of contrast. Notably, as the typicality of the word pairs with respect to contradictory declines, the relation vector becomes increasingly distributed across a broader range of relation types (particularly for the general type of similar).

To solve a verbal analogy problem in the form "A:B :: C:D," the basic requirement is to identify the relation or relations linking "A" to "B" and "C" to "D," and to then assess whether the two relational representations satisfy some criterion for matching one another. In the absence of any explicit relational representations, models such as Word2vec simply code the generic relation between any two words as the difference vector between the two word embeddings, and then compute the cosine distance between the resulting difference vectors (where an ideal analogy will yield parallel difference vectors, hence cosine distance of 0) (14). In BART, the relational similarity between two word pairs can be readily estimated by computing the cosine distance between the two relation vectors (Fig. 3). Whereas the cosine distance computed by Word2vec depends solely on the word semantic features, the distance computed by BART operates on the set of explicit relations that the model has previously learned, providing greater sensitivity to the specific relations linking a given pair.

To assess the performance of BART and the baseline Word2vec model in solving verbal analogies, we developed the UCLA Verbal Analogy Test (VAT; *SI Appendix*, Table S2). This test consists of 80 analogy problems, with 20 items based on each of four types of relations, which we term categorical, function, antonym, and synonym (loosely based on the types "class inclusion," "case relation," "contrast," and "similar," respectively, in ref. 22). An example (for the categorical relation) is "insect: bee::fish:halibut" vs. "fish:water." Note that the D′ foil ("water") is chosen to be a highly associated word to the C term ("fish"). The correct answer and the foil are always based on the same C term, and the same word classes (in this case, noun-noun). All 80 problems have a similar structure. Here 94% of the word pairs used to construct VAT items are new pairs that did not appear in the set of pairs (22) used to train BART, ensuring a strong test of generalization of acquired relation knowledge to novel instantiations presented as analogy problems.

Fig. 4 displays the proportion correct for the VAT problems achieved by Word2vec, BART, and humans. Both models performed better than chance level (0.50) for problems based on each of the four relation types. The mean accuracy for BART was 0.84 (range, 0.70–0.95 across the four types), consistently higher than that of 0.69 (range, 0.55–0.80) achieved by Word2vec. For comparison, the VAT was administered to 57 human participants (minimum education level of high-school graduation, located in the United States) using Amazon Mechanical Turk (approved, including informed consent procedures, by the UCLA Office of the Human Research Protection Program). This group achieved a mean solution rate of 0.84 (range, 0.81–0.86 across relation types). BART achieved a performance level comparable to that of humans for the relations of categorical, function, and synonym, but performed less well for the antonym relation. A possible reason for this finding is that the BART model was trained on each relation in isolation, with negative examples distributed across all the other nontarget relation types. Some concepts that form VAT antonyms (e.g., "friend:enemy") are quite similar in some respects (e.g., both are social roles). The concepts used in the VAT foils (CD′ pairs) also tend to be similar to one another, which may have made analogies based on the relation of contrast (the type most related to antonym in VAT items) especially confusable for BART. Unlike BART, children are generally taught in school that antonym (or opposite) is mutually exclusive with synonym; this learning experience may enhance discriminability between the two relations for humans.

To assess the developmental course of BART's relation learning and analogical reasoning, we tested the model as the number of positive training examples was varied over the range of 1–20 (*SI Appendix*, Fig. S1). Four positive training examples sufficed for BART to clearly exceed Word2vec's performance in predicting typicality orderings. For the analogy task, an advantage for BART relative to Word2vec emerged after BART had been trained with approximately eight positive examples.

Comparisons of BART Model with Control Simulations. A number of alternative models and control variants were implemented and tested (see *SI Appendix*, with summary in *SI Appendix*, Fig. S2). A feedforward neural network with a hidden layer (10 units) yielded chance level performance for the VAT. We tested five variants of BART to isolate the impact of individual components of the model. Each variant was designed to assess the importance of a specific component of the BART model by altering or removing a single component while keeping the rest of the model identical to the full BART. After removal of Bayesian learning in stage 3 (control 1) or removal of the empirical priors (control 2), performance on the analogy test dropped by roughly 20%. When only ranked features (control 3) or only raw features (control 4) were used, performance on the analogy test was reduced by 24% and 9%, respectively. When a transformed semantic space for word features was used (control 5), performance dropped by 10%.

Generating Converse Relations. The relation representations learned by BART are modular (each corresponding to a separate and identifiable weight distribution) and inherently structured, in

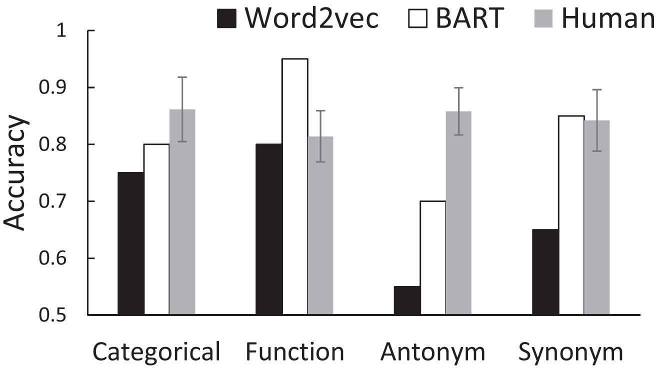

Fig. 4. Proportion of correct solutions to UCLA VAT problems achieved by BART and by Word2vec for each of four relation types. An accuracy level of 0.50 represents chance performance. Error bars for human performance indicate 95% confidence intervals.

that the learned weight distribution for a binary relation can be readily decomposed into weights on its two constituent roles, that is, on features of the first and second word, respectively, in a pair instantiating the relation. This concept is illustrated in Fig. 1, in which a color code distinguishes features and weights defining the two roles.

These properties of BART's relation representations (i.e., modularity and role structure) enable the performance of basic symbolic computations on them. In particular, BART can capture a general relation between relations: any relation R(a, b) can be used to define a converse relation, R'(b, a), by reversing the role assignments; for example, "instance of (a, b)" has the converse relation "category of (b, a)". Because BART's relation representations are modular, a converse can simply be added to the model's set of known relations without altering the relation used to generate it. The converse constitutes a distinct relation only when the roles are asymmetrical, so that $R' \neq R$. A criticism of previous neural network models of analogy is that they do not learn relations structured in terms of roles and thus cannot solve analogies based on relations that are converses of those on which the model was trained without explicitly retraining the model on examples of the converse relations (29).

For many relations in the taxonomy used to train BART (22), the converse is also included, so the model is quite robust when word pairs are reversed for the asymmetrical relations used to form VAT items. To provide an initial demonstration of the potential for using converse creation to further increase BART's capacity to solve analogy problems, we examined a more semantically-distant analogy (30). When presented in the order "blindness:sight:: poverty:money" vs. "finance:money," BART correctly picks "poverty:money" as the better analogical alternative. Note that none of the word pairs in this example was used in the learning phase of the BART model. However, when the word pairs are reversed ("sight:blindness:: money:poverty" vs. "money:finance"), the model incorrectly chooses the foil "money:finance." This sensitivity to pair order is due to the fact that the A:B pair "blindness:sight" and C:D pair "poverty: money" both have moderately high posterior probabilities of instantiating multiple asymmetrical relations included in the taxonomy on which BART was trained, whereas the reversed B:A and D:C pairs do not.

We then allowed BART to create converses of its learned relations by simply reversing the covariance matrix, swapping weights (and their associated variances and covariances) across the two roles, and again attempt the reversed analogy problem. Rather than requiring retraining to learn the converse of a learned relation, BART forms the converse with a rule-like operation by reordering the weights without any additional learning. When the converse of each of BART's nonsymmetrical learned relations is computed and added to the distributed representation used to solve the analogy, BART correctly selects "money:poverty" as the analogical answer to the reversed problem.

As another challenging example in which enriching relational knowledge by introducing converse relations can enhance analogical reasoning, we considered a verbal analogy problem discussed in ref. (12): "pig:boar:: dog:wolf" vs. "dog:cat." This problem is difficult in part because the words in the foil ("dog:cat") are more closely associated with each other than are those in the analogical option ("dog:wolf"), creating strong competition between the choice based on relation similarity and the foil based on simple association (31). Using its basic vectors of posterior probabilities over the 79 relations on which it was trained, BART incorrectly selects the associative foil. However, if converses are created in the manner described above and used to expand its set of relation vectors, BART chooses the analogical option.

Discussion

The present study provides an example of the synergy made possible by combining learning of semantic feature vectors from big data with more focused learning that enables induction of relational representations from small data. Arguably, it is the combination of these different types of learning that supports human relational reasoning. Because relations are central to analogy, an account of relation learning must provide a necessary building block for a computational theory of analogical reasoning. To assess whether one word pair (e.g., "hot:cold") is analogous to another (e.g., "rich:poor"), the reasoner must first determine what (unstated) relation(s) each word pair instantiates and then determine whether the two pairs share a set of relations that yields sufficiently high relational similarity to make them analogous. However, most previous analogy models (4–7) can neither learn relations from nonrelational inputs nor generate the relations required to solve simple verbal analogies. In this paper, we propose computational mechanisms to enable learning of relations and to use these relations to solve verbal analogy problems.

BART's use of weight distributions (rather than pools of units) to represent relations within a neural network allows the representation of each relation to be modular (similar to the approach in ref. 11). In agreement with other neural network models that aim to integrate relation learning with analogical learning (9, 12), BART views the learning and comparison of relations as core mechanisms that drive analogical reasoning rather than treating the latter as an entirely separable process. The ability to solve analogy problems increases in a graded fashion with increased training on the relevant relations (*SI Appendix*, Fig. S1), consistent with both developmental evidence (32) and previous computational models (12, 33).

At the same time, BART has a number of properties that distinguish it from related neural-network models. BART is able to directly learn two-place relations (e.g., synonym) that cannot be constructed in any apparent way from one-place predicates (a limitation of the model described of ref. 9). BART's relation representations, which have identifiable constituent roles, enable the model to systematically transform any learned relation into its converse (without retraining), thereby extending the power of analogical reasoning. Similarly, for the special case of comparative relations (21), the model is able to make role-based transitive inferences with arbitrary role fillers (e.g., given "A larger than B" and "B larger than C," the model can infer "A larger than C").

Many hurdles remain to be overcome to achieve the goal of providing a model of human relation learning that can support the full range of relational reasoning. Important psychological questions remain to be addressed concerning the triggers for forming converse relations. BART's procedure for feature remapping provides a partial solution to the feature alignment problem, but the model's limited success in solving analogy problems based on antonyms suggests that its account of feature alignment needs to be improved. In addition, the BART model is based on supervised learning with labeled examples. Children in early elementary school typically receive explicit instruction that includes labeled examples for at least some core semantic relations (e.g., antonym and synonym); however, it has been argued that human concepts may be introduced by direct instruction with a few salient examples, followed by a process of semisupervised learning based in part on unlabeled examples (34). How to effectively integrate supervised and unsupervised learning in relation learning remains an open question for computational models.

So far, the BART model has been tested only on simple verbal analogy problems based on the types of general relations on which it was trained. However, verbal analogies can be constructed using highly specific relations (e.g., "one:five:: soloist: quintet"). The model would need to be trained to recognize a much broader pool of relations to solve a more comprehensive set of analogy problems. Moreover, a full model of analogical reasoning requires the capacity to identify correspondences between elements organized into more complex propositional structures (e.g., analogies between stories). In addition, humans clearly learn relations from visuospatial as well as linguistic inputs and can solve analogy problems posed as pictures of either individual objects or as scenes depicting multiple interacting objects (35, 36). AI models of machine vision (e.g., ref. 37) are likely to contribute to future advances in modeling human relation learning and analogical reasoning.

Another direction for future work is to relate computational models more closely to neuroscientific evidence. For example, forms of brain damage that disrupt semantic knowledge (e.g., damage to the anterior temporal cortex) are known to impair reasoning with verbal analogies (31). Such damage could be modeled by adding noise to BART's feature vectors and/or weight distributions (cf. refs. 12, 31, and 38). Frontal damage, which increases the tendency to err by choosing a semantically associated foil rather than the analogical completion, could be modeled by impairment in mechanisms required to form, maintain and compare relation vectors for A:B and C:D pairs (cf. ref. 12). For healthy individuals performing reasoning tasks, BART's quantitative account of relation similarity can potentially be used to predict patterns of neural activity associated with specific word pairs, using the methods of representational similarity analysis (39). A full understanding of relation learning and analogical reasoning will require greater integration of computational models with neural mechanisms.

ACKNOWLEDGMENTS. Preparation of this paper was supported by a Google Faculty Research Award and National Science Foundation Grant BCS-1827374 (to H.L. and K.J.H.).

1. Penn DC, Holyoak KJ, Povinelli DJ (2008) Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behav Brain Sci* 31:109–130, discussion 130–178.
2. Glass AL, Holyoak KJ, Kossan NE (1977) Children's ability to detect semantic contradictions. *Child Dev* 48:279–283.
3. Common Core State Standards Initiative (2017) English language arts standards (language, grade 4). Available at www.corestandards.org/ELA-Literacy/L/4/5/c/. Accessed August 11, 2017.
4. Falkenhainer B, Forbus KD, Gentner D (1989) The structure mapping engine: Algorithm and examples. *Artif Intell* 41:1–63.
5. Halford GS, Wilson WH, Phillips S (1998) Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behav Brain Sci* 21:803–831, discussion 831–864.
6. Hummel JE, Holyoak KJ (2003) A symbolic-connectionist theory of relational inference and generalization. *Psychol Rev* 110:220–264.
7. Petrov AA (2013) *Associative Memory-Based Reasoning: A Computational Model of Analogy-Making in a Decentralized Multi-Agent Cognitive Architecture* (Lambert Academic, Saarbrücken, Germany).
8. Tenenbaum JB, Kemp C, Griffiths TL, Goodman ND (2011) How to grow a mind: Statistics, structure, and abstraction. *Science* 331:1279–1285.
9. Doumas LAA, Hummel JE, Sandhofer CM (2008) A theory of the discovery and predication of relational concepts. *Psychol Rev* 115:1–43.
10. Lu H, Chen D, Holyoak KJ (2012) Bayesian analogy with relational transformations. *Psychol Rev* 119:617–648.
11. Paccanaro A, Hinton GE (2001) Learning distributed representations of concepts using linear relational embedding. *IEEE Trans Knowl Data Eng* 13:232–244.
12. Kollias P, McClelland JL (2013) Context, cortex, and associations: A connectionist developmental approach to verbal analogies. *Front Psychol* 4:857.
13. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. *Adv Neural Inf Process Syst* 26: 3111–3119.
14. Le Q, Mikolov T (2014) Distributed representations of sentences and documents. *Proceedings of the 31st International Conference on Machine Learning, PMLR* 32: 1188–1196.
15. Pennington J, Socher R, Manning CD (2014) GloVe: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp 1532–1543. Available at aclweb.org/anthology/D14-1162. Accessed January 15, 2019.
16. Rumelhart DE, Abrahamson AA (1973) A model for analogical reasoning. *Cognit Psychol* 5:l–28.
17. Chen D, Peterson JC, Griffiths TL (2017) Evaluating vector-space models of analogy. *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*, pp 1746–1751. Available at mindmodeling.org/cogsci2017/. Accessed January 15, 2019.
18. Santoro A, et al. (2017) A simple neural network module for relational reasoning. *Proceedings of Neural Information Processing Systems 2017*, pp 4967–4976. Available at https://papers.nips.cc/paper/7082-a-simple-neural-network-module-for-relational-reasoning. Accessed January 15, 2019.
19. Kotovsky L, Gentner D (1996) Comparison and categorization in the development of relational similarity. *Child Dev* 67:2797–2822.
20. Chen D, Lu H, Holyoak KJ (2014) The discovery and comparison of symbolic magnitudes. *Cognit Psychol* 71:27–54.
21. Chen D, Lu H, Holyoak KJ (2017) Generative inferences based on learned relations. *Cogn Sci* 41:1062–1092.
22. Jurgens DA, Mohammad SM, Turney PD, Holyoak KJ (2012) SemEval-2012 Task 2: Measuring degrees of relational similarity. *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, pp 356–364. Available at dl.acm.org/citation.cfm?id=2387693. Accessed January 15, 2019.
23. Bejar II, Chaffin R, Embretson SE (1991) *Cognitive and Psychometric Analysis of Analogical Problem Solving* (Springer, New York).
24. Rosch E (1975) Cognitive representations of semantic categories. *J Exp Psychol Gen* 104:192–233.
25. Mohammad SM, Kiritchenko S (2016) Capturing reliable fine-grained sentiment associations by crowdsourcing and best-worst scaling. *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp 811–817. Available at www.aclweb.org/anthology/N16-1095. Accessed January 15, 2019.
26. Turney PD (2013) Distributional semantics beyond words: Supervised learning of analogy and paraphrase. *Trans Assoc Comput Linguist* 1:353–366.
27. Siu K-Y, Bruck J, Kailath T, Hofmeister T (1993) Depth efficient neural networks for division and related problems. *IEEE Trans Inf Theory* 39:946–956.
28. Jaakkola TS, Jordan MI (2000) Bayesian parameter estimation via variational methods. *Stat Comput* 10:25–37.
29. Holyoak KJ, Hummel JE (2008) No way to start a space program: Associationism as a launch pad for analogical reasoning. *Behav Brain Sci* 31:388–389.
30. Green AE, Kraemer DJ, Fugelsang JA, Gray JR, Dunbar KN (2010) Connecting long distance: Semantic distance in analogical reasoning modulates frontopolar cortex activity. *Cereb Cortex* 20:70–76.
31. Morrison RG, et al. (2004) A neurocomputational model of analogical reasoning and its breakdown in frontotemporal lobar degeneration. *J Cogn Neurosci* 16:260–271.
32. Goswami U (1991) Analogical reasoning: What develops? A review of research and theory. *Child Dev* 62:1–22.
33. Doumas LAA, Morrison RG, Richland LE (2018) Individual differences in relational learning and analogical reasoning: A computational model of longitudinal change. *Front Psychol* 9:1235.
34. Xu F, Tenenbaum JB (2007) Sensitivity to sampling in Bayesian word learning. *Dev Sci* 10:288–297.
35. Krawczyk DC, et al. (2008) Distraction during relational reasoning: The role of prefrontal cortex in interference control. *Neuropsychologia* 46:2020–2032.
36. Richland LE, Morrison RG, Holyoak KJ (2006) Children's development of analogical reasoning: Insights from scene analogy problems. *J Exp Child Psychol* 94:249–273.
37. Mao J, et al. (2015) Learning like a child: Fast novel visual concept learning from sentence descriptions of images. *Proceedings of the IEEE International Conference on Computer Vision*, pp 2533–2541.Available at https://ieeexplore.ieee.org/document/7410648. Accessed January 15, 2019.
38. Hoffman P, McClelland JL, Lambon Ralph MA (2018) Concepts, control, and context: A connectionist account of normal and disordered semantic cognition. *Psychol Rev* 125: 293–328.
39. Kriegeskorte N, Mur M, Bandettini P (2008) Representational similarity analysis – connecting the branches of systems neuroscience. *Front Syst Neurosci* 2:4.