Journal of Experimental Psychology: Learning, Memory, and Cognition

Predicting Patterns of Similarity Among Abstract Semantic Relations

Nicholas Ichien, Hongjing Lu, and Keith J. Holyoak

Online First Publication, July 1, 2021. http://dx.doi.org/10.1037/xlm0001010

CITATION

Ichien, N., Lu, H., & Holyoak, K. J. (2021, July 1). Predicting Patterns of Similarity Among Abstract Semantic Relations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Advance online publication. http://dx.doi.org/10.1037/xlm0001010

Journal of Experimental Psychology:
Learning, Memory, and Cognition

© 2021 American Psychological Association
ISSN: 0278-7393

https://doi.org/10.1037/xlm0001010

Predicting Patterns of Similarity Among Abstract Semantic Relations

Nicholas Ichien[1], Hongjing Lu[1, 2], and Keith J. Holyoak[1, 3]
[1] Department of Psychology, University of California, Los Angeles
[2] Department of Statistics, University of California, Los Angeles
[3] Brain Research Institute, University of California, Los Angeles

Although models of word meanings based on distributional semantics have proved effective in predicting human judgments of similarity among individual concepts, it is less clear whether or how such models might be extended to account for judgments of similarity among relations between concepts. Here we combine an individual-differences approach with computational modeling to predict human judgments of similarity among word pairs instantiating a variety of abstract semantic relations (e.g., contrast, cause–effect, part–whole). A measure of cognitive capacity predicted individual differences in the ability to discriminate among distinct relations. The human pattern of relational similarity judgments, both at the group level and for individual participants, was best predicted by a model that takes representations of word meanings based on distributional semantics as its inputs and uses them to learn an explicit representation of relations. These findings indicate that although the meanings of abstract semantic relations are not directly coded in the meanings of individual words, important aspects of relational similarity can be derived from distributional semantics.

Keywords: relations, similarity, analogy, reasoning, distributional semantics

The ability to consider the relations between entities, rather than solely the features of individual entities, is a central characteristic of human thought. For example, words not only have individual meanings, but also exhibit systematic relations to one another (e.g., rich–poor exemplifies the relation contrast, joke–laughter exemplifies the relation cause–effect). Human intuitions regarding semantic relations exhibit several complexities. Much like object categories (Rosch, 1975); examples of semantic relations form typicality gradients rather than simply being "all or none" (Chaffin, 1992; Chaffin & Herrmann, 1988a; Popov et al. 2020). For example, hot–cold is considered a better example of the contrast relation than is warm–cool (Jurgens et al., 2012). In addition, a single pair of words can instantiate multiple relations to some degree. For example, friend–enemy exemplifies the relation contrast, but also to some degree the relation similar (leading to a potential blended concept, "frenemy"). These graded aspects of human judgments, which suggest that the cognitive and neural representations of semantic relations may be distributed in nature (Chiang et al., 2020), pose problems for models of analogical reasoning that treat relations as atomistic links between symbols (Forbus et al., 2017).

In order to develop models of how relations can be learned and used to make inferences, it is highly desirable to start from inputs that have been created by data-driven methods, rather than simply using the intuitions of researchers. A promising approach for automatically creating representations of word meanings (the natural building blocks for semantic relations) is distributional semantics, the general label for the use of machine-learning models to derive semantic vectors for words by analyzing their statistical distribution in very large text corpora (Bhatia et al., 2019). In the present study we explore models of relational similarity founded on inputs produced by one prominent model of distributional semantics, Word2vec (Mikolov, Sutskever, et al., 2013). For any word, Word2vec generates a vector representing its meaning in a 300-dimensional semantic space. Word2vec and similar models of distributional semantics have been successful in predicting behavioral judgments of lexical similarity or association (Hill et al., 2015; Hofmann et al., 2018; Pereira et al. 2016; Richie & Bhatia, 2020) and neural responses to word meanings (Huth

Nicholas Ichien https://orcid.org/0000-0002-0928-0809
Hongjing Lu https://orcid.org/0000-0003-0660-1176
Keith J. Holyoak https://orcid.org/0000-0001-8010-6267

A preliminary report of Experiment 1 was presented at the 41st Annual Conference of the Cognitive Science Society (July, 2019). The authors declare no competing interests. De-identified data, analysis scripts, and code implementing the computational models are available on the study's OSF website: https://osf.io/stuea/. Conceptualization, methodology and formal analysis were contributed by Nicholas Ichien and Hongjing Lu; investigation was done by Nicholas Ichien; the original draft was written by Nicholas Ichien and Keith J. Holyoak; review and editing was done by all authors; supervision and funding acquisition were done by Hongjing Lu and Keith J. Holyoak. Preparation of this article was supported by NSF Grant BCS-1827374. We thank Mohit Bhatia, Lina Chang, Anvita Diwan, Ali Hepps, Khushi Jain, Juliette LeSaint, and Zhibo Zhang for assistance in data collection.

Correspondence concerning this article should be addressed to Nicholas Ichien, Department of Psychology, University of California, Los Angeles, 405 Hilgard Avenue, Los Angeles, CA 90095-1563, United States. Email: ichien@g.ucla.edu

et al., 2016; Pereira et al. 2018; Zhang et al. 2020); as well as high-level inferences such as assessments of probability (Bhatia, 2017).

The potential for using semantic vectors based on Word2vec to account for relational judgments has also been explored, though with mixed success. One basic approach is to define the generic relation between two concepts as the difference vector between them (e.g., the relation hot–cold is defined by the difference between the semantic vectors for hot and cold). The relation is thus defined only implicitly, without any explicit representation as a concept in its own right (e.g., opposite). Difference vectors have been used to solve semantically close analogy problems (Mikolov, Sutskever, et al., 2013; Mikolov, Yih, et al., 2013; Zhila et al. 2013) and to predict neural patterns associated with some asymmetrical semantic relations between words embedded in continuous text (Zhang et al., 2020). However, models based on difference vectors fail to reliably solve more complex analogy problems (Peterson et al., 2020; Linzen, 2016); or to accurately predict neural responses to semantic relations presented as analogy problems (Chiang et al., 2020). More generally, implicit representations of relations have difficulty accounting for priming effects based on semantic relations (Estes & Jones, 2006; 2009; Jones et al., 2017; Popov et al., 2017).

Judgments of Relational Similarity

A key obstacle to evaluating models of relational similarity is a lack of systematic data on human judgments. Here we report human judgments of similarity between semantic relations (i.e., judgments involving multiple pairs of words), which have been shown to strongly predict human performance on a number of relational reasoning tasks including relation exemplar generation frequency, analogical verification accuracy and response time, and confusability of word pairs in episodic memory judgments (i.e., relational luring in associative memory; Popov et al., 2020). Here we use relational similarity data to evaluate alternative theoretical measures of relational similarity ultimately based on distributional semantics. Rather than assuming that people are uniform in their ability to differentiate semantic relations, we also sought to measure cognitive abilities that may lead to individual differences in relational judgments.

Predicting relational similarity poses methodological as well as theoretical challenges. The effective representation of a relation for any word pair may be context-sensitive to some degree, and therefore may vary depending on the order in which pairs are presented for comparison. In addition, the number of potential pairwise comparisons becomes prohibitive when the total number of pairs grows modestly large. Such problems can be alleviated by using a multiarrangement task, a method for efficiently eliciting similarity judgments for large sets of items (Kriegeskorte & Mur, 2012). The method involves comparisons among a set of items presented together, thus reducing order effects that may arise with pairwise comparisons. The multiarrangement method, which can be viewed as an inverse of standard multidimensional scaling (Shepard, 1962), has previously been applied successfully to judgments of object similarity (Jozwik et al., 2017; Kriegeskorte & Mur, 2012; Mur et al., 2013). In order to assess potential individual differences in the ability to discriminate among relations, we also administered a version of the Ravens Advanced Progressive Matrices (RPM; Arthur et al., 1999); a measure of cognitive capacity, and the Semantic Similarities Test (SST; Stamenković et al., 2019, 2020), a measure of semantic knowledge.

In two experiments, data from human judgments of relational similarity were used to test three alternative computational models, all founded on lexical representations derived by Word2vec. These models instantiate different assumptions about how relations are represented and compared. Further, data estimating individual differences in cognitive capacity were used to examine variability in model predictions of judgments of relational similarity across individual participants. For each model we generated a set of predicted dissimilarities among relation instances that can be compared to dissimilarities derived from human judgments, an approach termed Representational Similarity Analysis, or RSA (Kriegeskorte et al., 2008). To the extent that model-generated dissimilarities approximate human-generated dissimilarities, that model's representation of semantic relations is descriptive of human semantic cognition.

Modeling Relational Dissimilarity

To represent meanings of individual words, each model we tested uses word embeddings produced by Word2vec (Mikolov, Sutskever, et al., 2013). Two of the models we tested derive dissimilarity predictions directly from Word2vec vectors for the individual words in a pair. These two models differ in their assumptions about how (or whether) the relation between the two words is represented. Under Word2vec-concat, the meaning of the words within a pair is a simple aggregate of the semantic vectors of the two individual words. We use f_A to denote the semantic vector for a word A, and use $[f_A \; f_B]$ to denote the concatenated vector that captures the meaning of a word pair consisting of words A and B. The dissimilarity $D_{W2V-concat}$, between two word pairs, A:B versus C:D, is computed by the cosine distance between the two concatenated vectors (top panel in Figure 1):

$$D_{W2V-concat} = cos([f_A f_B], [f_C f_D]). \tag{1}$$

This model is nonrelational, instead capturing semantic dissimilarity across pairs based solely on the meanings of the individual words. Word2vec-concat serves to identify patterns of dissimilarity based on lexical semantics, separate from any representation of the relation between the two words within each pair, and hence can be viewed as a baseline model for comparison to models that actually compute relations.

Under Word2vec-diff, the relation between two words is defined in a generic fashion as the difference between the semantic vectors of each word within a pair, $\boldsymbol{f_A} - \boldsymbol{f_B}$ for the word pair A:B. Dissimilarity of relations, $D_{W2V-diff}$, is assessed by the cosine distance between the difference vectors for two word pairs (middle panel in Figure 1):

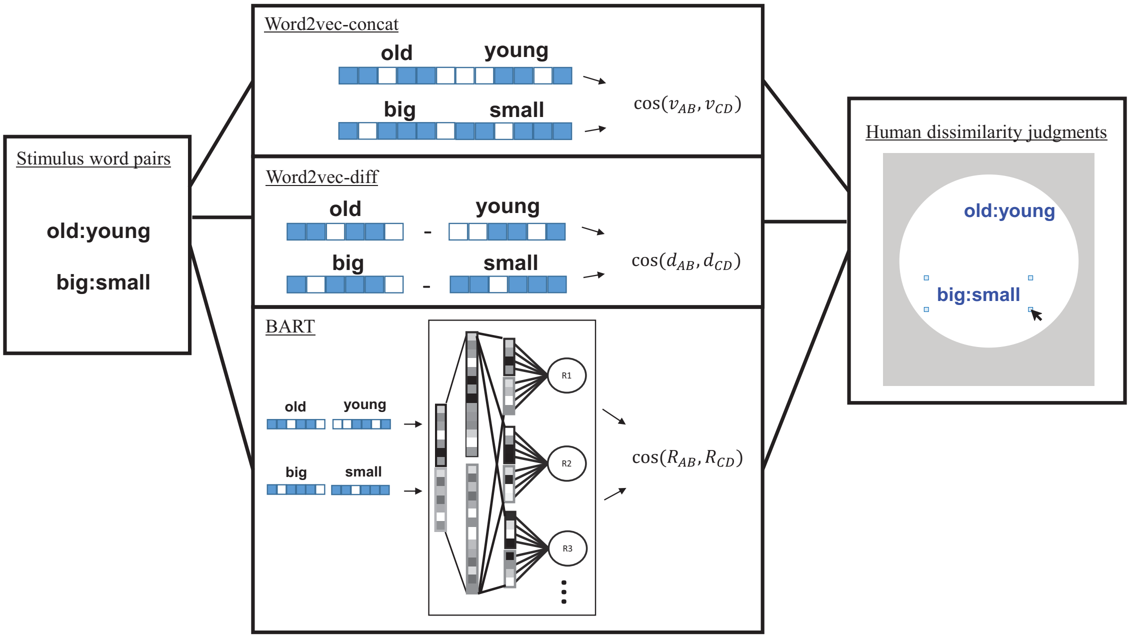
$$D_{W2V-diff} = cos(f_A - f_B, f_C - f_D). \tag{2}$$

This model codes relations only implicitly (i.e., as a difference vector computed from individual words).

The third computational model, Bayesian Analogy with Relational Transformations, or BART (Lu et al., 2012, 2019), assumes that specific semantic relations between words are coded as distributed representations over a set of abstract relations. The BART Model takes concatenated pairs of Word2vec vectors as input, and uses supervised learning with both positive and negative examples

Figure 1

Schematic Illustration of Model-Generated Predictions of Human Dissimilarity Judgments Among Semantic Relations



Note. For any two word pairs (e.g., old:young and big:small), three models are used to predict their dissimilarity based on the cosine distance between vectors representing each individual word pair, using 300-dimensional Word2vec word embeddings as inputs to each model (left). Word2vec-concat (top) concatenates the vectors representing individual words in each pair; Word2vec-diff (middle) defines the relation of each word pair as their difference vector; and BART (bottom) generates a new relation vector for each word pair based on previously learned relations. Human dissimilarity judgments were estimated based on on-screen distances between word pairs arranged in a multiarrangement task (right). See the online article for the color version of this figure.

to acquire representations of individual semantic relations. After learning, BART calculates a relation vector consisting of the posterior probability that a word pair instantiates each of the learned relations.

The BART model uses a three-stage process to learn semantic relations. In its first stage, the model uses difference-ranking operations to partially align relationally important features. The model generates a ranked feature vector based on the same difference values as the raw feature vector, but ordering those values according to their magnitude. Augmenting the raw semantic features with ranked features addresses the issue that across instances different semantic dimensions may be relevant to a relation. This first stage culminates in the generation of a 1,200-dimension augmented feature vector for each word pair, consisting of the concatenation of raw and ranked feature vectors for each word in the pair (second layer from bottom in Figure 2).

In the second stage, BART uses logistic regression with elastic net regularization and the difference vectors as input to select a subset of important features $\boldsymbol{f}_s$ (creating the 3rd layer from bottom in Figure 2) and estimates the associated coefficients β. In the third stage, BART uses Bayesian logistic regression with the selected features of word pairs $\boldsymbol{f}_s$ in training examples to estimate weight distributions w for representing a particular relation R by assuming that selected features and weights are independent. We can apply Bayes rule as:

$$P(\boldsymbol{w}|\boldsymbol{f}_s, R) \propto P(R|\boldsymbol{f}_s, \boldsymbol{w})P(\boldsymbol{w}). \qquad (3)$$

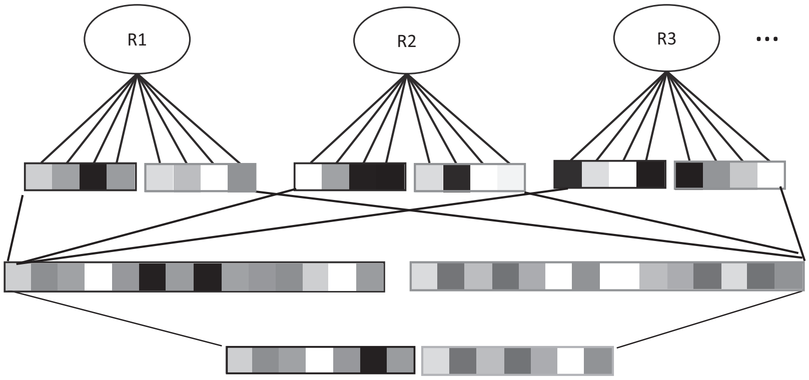
The first likelihood term is defined by a logistic function on the weights $\boldsymbol{w}$ and the features $\boldsymbol{f}_s$ selected in the second stage, $(1 + e^{-\boldsymbol{w}^T \mathbf{f}_s})^{-1}$. The second prior term is the prior distribution of the weights $\boldsymbol{w}$ defined as a multivariate normal distribution, $N(\mu_0, \sigma_0)$, with a mean vector μ_0 consisting of the coefficients β (estimated in the second stage of logistic regression) applied to features of the first entity, and −β (i.e., a contrast prior) applied to features of the second entity.

We trained BART using a dataset of over 3,200 word pairs exemplifying abstract semantic relations (Jurgens et al., 2012). Each word pair in this dataset is an instance of one of 79 specific relations, in turn categorized into one of ten general relation types according to a taxonomy of abstract semantic relations (Bejar et al., 1991). For each of the 79 relations, BART is trained with 20 positive examples of word pairs instantiating the relation to be learned, and 69–74 negative examples including the most prototypical word pairs instantiating relations belonging to a different relation type from the relation to be learned.

For the present simulations, we added a focused training step (similar to that employed in Ponti et al., 2018) to update the representation of two fundamental relation types: similar and contrast. These relation types include variants of synonyms and antonyms, which are typically taught to children in elementary school (Common Core State Standards

Figure 2

Schematic Illustration of BART Model of Relation Learning



Note. The bottom layer represents the concatenated input vector based on the two words in a pair; the top layer indicates the set of learned relations.The ellipses are meant to communicate a continuation of further learned relations beyond a third learned relation, which is denoted by "R3" (i.e., R4, R5, and so on to RN). BART = Bayesian Analogy with Relational Transformations.

Initiative, 2020). In this focused training, BART repeats the third stage of learning using a constrained set of negative examples. To update representations of similar relations, we constrained the set of negative examples to 40 word pairs instantiating contrast relations, and vice versa to update representations of contrast relations.

BART uses its pool of 79 learned relations to create a distributed 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 particular relation. For the purpose of modeling human similarity judgments, we calculated BART's relational representation of a word pair by introducing a nonlinear power transformation with the power parameter α, set at 5. This nonlinear power transformation serves to emphasize the contributions of those relations with higher posterior probabilities. Cosine distances based on these transformed vectors are used to compute the BART-generated relational dissimilarity D_{BART} between word pairs (bottom panel in Figure 1):

$$D_{BART} = cos\left(R^{a}_{AB}, R^{a}_{CD}\right). \tag{4}$$

We used the models described above to calculate predicted dissimilarities between relational word pairs using Word2vec-concat (Eq. 1), Word2vec-diff (Eq. 2), and BART (Eq. 4), respectively. For all three models, the dissimilarity between any two word pairs is computed by the cosine distance between the vectors representing each word pair (see Figure 1). These models are well matched in that each takes the same word embeddings as inputs and uses the same cosine calculation to predict dissimilarities, based respectively on lexical similarity only (Word2vec-concat), generic relation similarity (Word2vec-diff), and learned semantic relations (BART).

Method

We performed two experiments to assess similarity judgments for word pairs instantiating abstract relations. The two experiments used word pairs exemplifying different relations; otherwise the procedures and data analyses were identical. Accordingly, we will report the two experiments together.

Participants

In Experiment 1, 95 participants (M_{age} = 20.15 years, SD_{age} = 4.69; age range = [18–59]; 71 female, 24 male) were recruited from the Department of Psychology subject pool at the University of California, Los Angeles (UCLA). In Experiment 2, 94 different participants (M_{age} = 20.60 years, SD_{age} = 2.85; age range = [18–37]; 71 female, 23 male) were recruited from the same pool. Sample sizes were comparable to those used in previous studies that have assessed individual differences in relational reasoning (Stamenković et al., 2019, 2020; Vendetti et al., 2014). In both experiments, all participants were self-reported fluent English speakers. All participants were 18 years of age or older and provided verbal informed consent and were compensated with course credit. Experimental protocols for both experiments were approved by the UCLA Institutional Review Board.

Stimuli

For both experiments, all stimuli were word pairs taken from a set of norms (Jurgens et al., 2012) based on a taxonomy of abstract semantic relations (Bejar et al., 1991). Word pairs in this dataset express one of 79 specific relations, each falling into one of 10 general types of relations. The multiarrangement task in Experiment 1 used 27 word pairs, with three pairs chosen from each of three specific subrelations of three different general relation types (similar, contrast, and cause–purpose; see Table 1). Word pairs drawn from different subrelations of the same general type (e.g., car:auto instantiates synonymy and rake:fork instantiates attribute similarity, two subrelations of the relation type similar) are differentiated on the basis of relatively subtle relational differences. These stimuli were selected based on prior research in which these word pairs were used in an analogy task, where human participants

Table 1

Full Set of Word Pairs Used in Experiment 1, Organized by Relation Type (Table Headings) and Subrelation (Table Subheadings). Values Next to Each Word Pair Indicate the Mean Number of Times That a Participant Judged That Word Pair in an Experimental Session, and Values in Parentheses Are Standard Deviations

Similar		
Synonymy	Attribute similarity	Change
car:auto − 8.27 (5.04) kid:child − 7.74 (4.58) big:large − 7.40 (3.85)	rake:fork − 7.91 (4.43) sword:knife − 8.09 (5.10) stairs:ladder − 8.04 (4.00)	discount:price − 8.00 (3.75) dim:light − 9.73 (6.69) raise:salary − 8.21 (4.63)
Contrast		
Contrary	Directional	Pseudoantonym
old:young − 10.57 (5.66) big:small − 10.27 (5.40) black:white − 11.06 (5.38)	east:west − 10.27 (5.05) front:back − 9.52 (4.60) north:south − 9.73 (5.05)	right:bad − 16.57 (9.20) good:wrong − 16.64 (9.23) majority:small − 9.99 (4.83)
Cause–purpose		
Cause:effect	Cause:compensatory action	Action/activity:goal
joke:laughter − 7.91 (3.82) injury:pain − 8.69 (4.77) accident:damage − 8.49 (6.02)	hunger:eat − 8.28 (4.15) tiredness:rest − 7.86 (3.49) sadness:cry − 8.06 (3.65)	flee:escape − 7.83 (4.46) study:learn − 7.83 (3.49) work:earn − 8.03 (3.67)

reliably judged word pairs instantiating the same specific subrelation as constituting valid analogies and word pairs instantiating different relation types as constituting invalid analogies (Chiang et al., 2020). Experiment 2 used a different set of 27 word pairs, instantiating different relations than those examined in Experiment 1. Each set of three word pairs instantiated one of nine specific subrelations, consisting of three subrelations within three general types: class inclusion, part–whole, and space–time (see Table 2). Notably, each of the word pairs used in Experiment 2 consisted of noun–noun word pairs so as to control for any effects attributable to syntactic word class. In both experiments word pairs were drawn from among the most prototypical examples in the norms for the relevant subrelation.

Procedure

The basic procedure was identical for the two experiments. We acquired human similarity judgments of semantic relations by asking participants to perform a multiarrangement task. On each trial, participants were presented with a set of word pairs on a computer screen. Participants were asked to first identify the relation between words in each pair silently to themselves, and to then use a mouse to arrange word pairs in a two-dimensional circular space according to the similarity of their relations. Participants were told, “word pairs that involve similar relations should be placed close together,” “word pairs that involve very different relations should be placed far apart,” and “the distance between two word pairs should represent how different their relations are” (see Figure 3). Estimates of similarity were based on the relative on-screen distances between word pairs as arranged by participants on each trial. Estimates of pairwise judgments collected on the first trial were scaled to have a root mean square of 1, and these estimates were used to populate a participant’s Relational Dissimilarity Matrix (RDM). Pairwise judgments collected on subsequent trials were then used to update those estimates. These pairwise judgments were calculated by scaling the on-screen distances between items arranged on the most recent trial so that their root mean square matched the root mean square of the current estimates of the corresponding pairwise judgments in a participant’s RDM. The updated estimates were weighted averages of the current estimates and the rescaled pairwise judgments collected on the most recent trial. Once a participant’s RDM was fully populated with estimates of pairwise judgments between each item in the stimuli set, estimates provided by this RDM were used to predict on-screen distances between items arranged on subsequent trials. These estimates were further updated using deviations between predicted on-screen distances and observed on-screen distances collected on the most recent trial.

On a given trial, participants were presented with a maximum of 20 word pairs. The multiarrangement task adaptively selects stimuli to present on each trial. On the first trial, participants arranged a pseudorandom subset of 20 items from the entire set of 27 items. On subsequent trials, participants arranged a subset of 20 or fewer items selected based on item pairs with the weakest similarity evidence (Kriegeskorte & Mur, 2012). We limited session length to 20 minutes to avoid excessive fatigue.

We investigated whether relation judgments are systematically influenced by individual differences in cognitive capacity (especially working memory and inhibitory control) and/or semantic knowledge. To assess cognitive capacity, we administered a short version of the Ravens Advanced Progressive Matrices test (RPM; Arthur et al., 1999) adapted for computer administration using Matlab software. Participants were presented with a 3 × 3 grid of items with the item in the bottom right corner missing. They were asked to use the pattern instantiated by the presented items to select the most appropriate item to fill that bottom right corner

Table 2

Full Set of Word Pairs Used in Experiment 2, Organized by Relation Type (Table Headings) and Subrelation (Table Subheadings). Values Next to Each Word Pair Indicate the Mean Number of Times That a Participant Judged That Word Pair in an Experimental Session, and Values in Parentheses Are Standard Deviations

Class inclusion		
Taxonomic	Functional	Plural collective
weapon:spear – 8.48 (4.39) tree:oak – 9.34 (5.30) animal:pig – 8.8 (4.52)	tool:hammer – 8.84 (4.67) utensil:spoon – 8.84 (4.72) instrument:violin – 8.35 (3.96)	snacks:chips – 9.66 (5.00) cutlery:forks – 8.73 (3.87) furniture:chairs – 8.31 (3.99)
Part–whole		
Mass:portion	Item:topological part	Object:stuff
hour:seconds – 8.24 (4.42) feet:inches – 8.29 (4.44) week:day – 8.61 (4.34)	hotel:lobby – 8.46 (3.99) hill:top – 8.89 (4.75) airplane:cockpit – 8.44 (4.12)	omelette:eggs – 17.36 (13.02) ocean:water – 17.02 (13.04) wall:bricks – 11.88 (10.04)
Space–time		
Location:process/product	Contiguity	Time:associated item
factory:goods – 8.52 (4.07) mill:flour – 8.96 (4.57) mine:coal – 8.86 (4.91)	bank:river – 8.76 (5.00) shore:lake – 8.86 (5.22) ditch:road – 8.23 (4.4)	childhood:toys – 7.89 (4.36) girlhood:dolls – 7.54 (3.55) infancy:pacifier – 7.67 (3.72)

from a set of eight options. Prior research has shown that superior performance on this test is correlated with performance on tests of analogical reasoning (Gray & Holyoak, 2020; Kubricht et al., 2017; Vendetti et al., 2014). We hypothesized that the RPM measure would be associated with the degree to which people are able to differentiate word pairs that instantiate distinct relations.

In addition to cognitive capacity, the ability to differentiate among semantic relations may vary with knowledge of semantic relations. As a measure of semantic knowledge, we administered the SST. This test was designed to be similar to the similarities subscale of the Weschler Adult Intelligence Scale-III (WAIS), and is correlated with the vocabulary subtest (Stamenković et al., 2019). Participants are presented with 20 pairs of verbal concepts and asked to describe how the concepts in each pair are similar. The concept pairs span a broad range of similarities: some are fairly specific (e.g., bird–airplane, which both fly), some are more general (e.g., tavern–church, which are both public buildings), and some are more metaphorical (e.g., marriage–alloy, which are both bonds between elements). Prior research has shown that scores on the SST correlate positively with metaphor comprehension (Stamenković et al., 2019, 2020).

In Experiment 1, all participants completed the three tasks in the following order: the multiarrangement task, RPM, and SST. In Experiment 2, participants completed the three tasks in one of the following three orders: multiarrangement task, RPM, and SST; RPM, SST, and multiarrangement task; SST, multiarrangement task, and RPM.

Results

Overall Patterns of Relation Dissimilarity

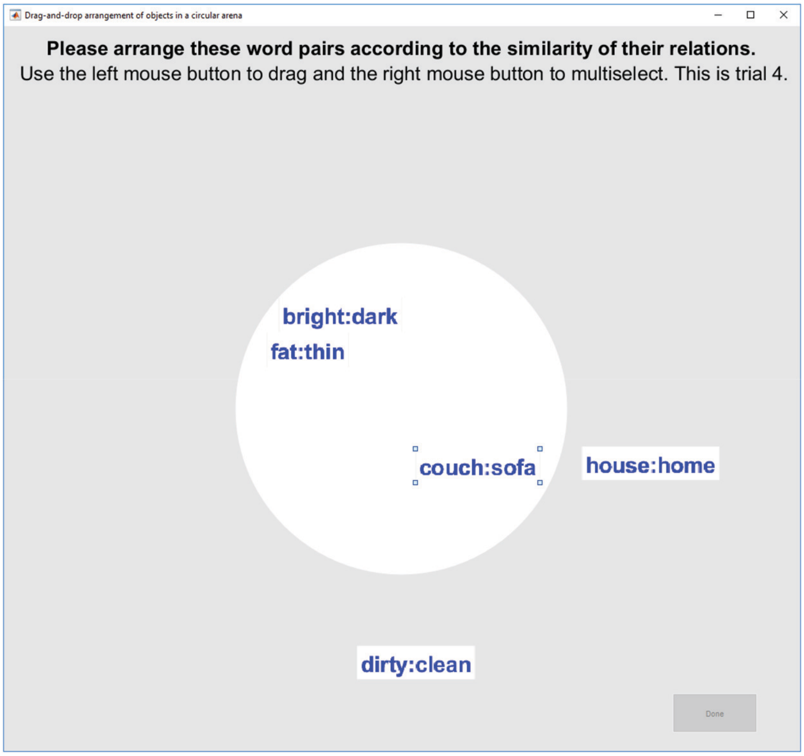
In Experiment 1, 95 participants completed a mean of 19.69 trials (SD = 9.70, range = [2–55]) on the multiarrangement task. Due to program failures, only 88 participants completed the SST, and 90 participants completed the RPM. In Experiment 2, 94 participants completed a mean of 20.59 trials on the multiarrangement task (SD = 12.20, range = [2–64]). Again, due to program failures, only 93 participants completed the SST, and 92 participants completed the RPM. In each experiment all but one participant provided pairwise similarity judgments for all 27 word pairs (351 pairwise comparisons), with the exception providing judgments for 86% of the pairwise combinations.

Figure 4 (top) displays the observed and predicted patterns of dissimilarity judgments for Experiment 1. Human judgments (leftmost panel) are shown as an RDM across the 27 items. This visual display shows that the patterns of similarity judgments were qualitatively different across the three relation types. The human dissimilarity matrix clearly indicates that human judgments differentiate each of the three broad types. The degree of differentiation among subrelations appears to be more subtle.

To assess the reliability of these apparent differences in similarity patterns across relation types, we computed two discrimination indices: (a) within-subrelation distance as the mean dissimilarity between word pairs instantiating the same subrelation, and (b) cross-subrelation/within-type distance as the mean dissimilarity between word pairs instantiating different subrelations but within the same relation type. We then conducted a two-way repeated measures ANOVA using discrimination index (within-subrelation vs. cross-subrelation/within-relation type) and subrelation as within-subject factors. Because we found a significant interaction between discrimination index and subrelation, $F(8, 752) = 25.34$, $p < .001$, we followed up with pairwise comparisons of discrimination indices for each type, using a Bonferroni-adjusted alpha-level of .005. As summarized in the top nine rows of Table 3, the two discrimination indices were significantly different for all three similar subrelations and for all three contrast subrelations. Of the

Figure 3

Example Trial of the Multiarrangement Task Used to Generate a Semantic Space for Relations



Drag-and-drop arrangement of objects in a circular arena

Please arrange these word pairs according to the similarity of their relations.
Use the left mouse button to drag and the right mouse button to multiselect. This is trial 4.

Note. See the online article for the color version of this figure.

three cause-purpose subrelations, cause:effect and cause:compensatory action showed reliable differences, whereas action/activity: goal did not. These analyses indicate that relational similarity judgments were generally sensitive to specific subrelations for the relation types used in Experiment 1.

Figure 4 (bottom) displays the observed and predicted patterns of dissimilarity judgments for Experiment 2. This visual display shows that the patterns of similarity judgments for human judgments (leftmost panel) were qualitatively different across the three relation types. The relation type class inclusion (top) forms a distinct category, but its three subrelations are not clearly differentiated. The relation type part–whole appears to form a weaker category, with clearly differentiated subrelations (diagonal). Finally, space-time (bottom) does not appear to form a unitary cluster as a relation type, though its respective subrelations are well differentiated individually.

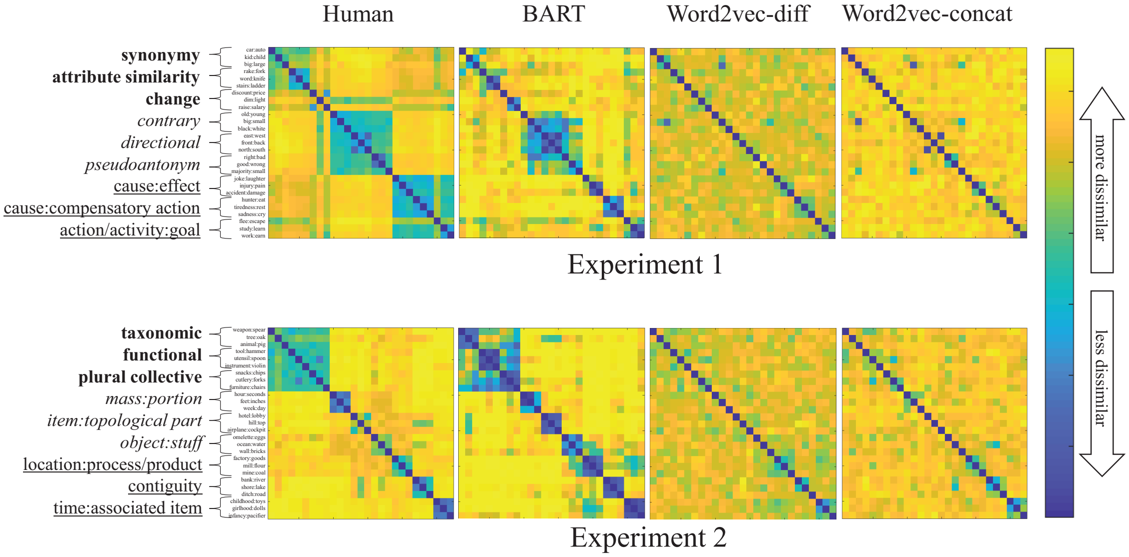
To assess the reliability of these apparent differences in similarity patterns across relation types, we ran similar analyses as for Experiment 1. In order to evaluate the differentiation among subrelations, we conducted a two-way repeated measures ANOVA using discrimination index (within-subrelation vs. cross-subrelation/within-relation type) and subrelation as within-subject factors. Again, we found a significant interaction between discrimination index and subrelation, $F(8, 744) = 137.79$, $p < .001$, so we followed up with pairwise comparisons of discrimination indices for each type, using a Bonferroni-adjusted alpha-level of .005 to correct for multiple comparisons. As summarized in the bottom nine rows of Table 3, for class inclusion, none of the mean cross-subrelation/within type distances were significantly different from the mean within-subrelation distances, indicating lack of differentiation among subrelations. In contrast, the two discrimination indices differed reliably for part–whole and for space–time. These findings indicate that human judgments of relational similarity in Experiment 2 were sensitive to specific subrelations within part–whole and space–time relation types, but not within the class inclusion relation type.

Individual Differences in Relation Discriminability

We also performed analyses to determine whether individual differences in cognitive capacity (as assessed by the RPM) and semantic knowledge (as assessed by the SST) were associated with participants' sensitivity to differences among relations. For each experiment, two independent raters scored the SST based on the criteria summarized in (Stamenković et al., 2019). We assessed the reliability of these raters' scores by testing the average intraclass correlation coefficient across scores using a two-way random model for Experiment 1 (ICC = .866, $F(87, 87) = 9.472$, $p < .001$, 95% CI [.708, .929]) and for Experiment 2 (ICC = .923, $F(93, 93) = 13.396$, $p < .001$, 95% CI [.884, .949]). Given the reliability of these scores across both experiments, for each dataset we used the average SST score across the two raters in the

Figure 4

Dissimilarity Matrices (RDMs) Representing Human Judgments and Model-Generated Predictions



Note. Top: Experiment 1; bottom: Experiment 2. Each row and column represents a word pair, and each cell represents the pairwise dissimilarity between the word pair represented by that cell's row and the word pair represented by that cell's column. Diagonal cells represent the pairwise dissimilarity between a word pair and itself, which is assumed to be 0. Warmer colors indicate greater dissimilarity, while cooler colors indicate less dissimilarity. BART = Bayesian Analogy with Relational Transformations; RDM = Relational Dissimilarity Matrix. See the online article for the color version of this figure.

following analyses. Descriptive statistics for both RPM and SST performance for Experiments 1 and 2 are provided in Table 4.

In order to estimate individual differences in sensitivity to broad distinctions between relation types, we computed a relation type discriminability index for each participant using the following steps. First, we found each participant's cross-type distance by calculating the mean distance for pairwise comparisons between word pairs instantiating different general relation types (e.g., old: young instantiates the relation type contrast, whereas car:auto instantiates the relation type similar). Second, we found each participant's within-type distance by calculating the mean distance for pairwise comparisons between word pairs instantiating the

Table 3

Results of Paired-Sample T-Tests Comparing Within-Subrelation Dissimilarity (M and Standard Deviation Reported Under Mw and SDw, Respectively) With Cross-Subrelation/Within-Relation Type Dissimilarity (M and Standard Deviation Reported Under Mc and SDc, Respectively). Top Nine Rows Show Results from Experiment 1, and Bottom Nine Rows Show Results from Experiment 2

Type	Subrelation	*M*w	*SD*w	*M*c	*SD*c	*df*	*t*	*p*	95% CI	Cohen's *D*
Similar	Synonymy	0.039	0.015	0.048	0.008	94	6.18	<.001	[0.006, 0.011]	0.63
	Attribute similarity	0.028	0.016	0.048	0.008	94	13.34	<.001	[0.017, 0.023]	1.37
	Change	0.042	0.012	0.052	0.008	94	8.40	<.001	[0.008, 0.012]	0.86
Contrast	Contrary	0.029	0.014	0.034	0.011	94	6.12	<.001	[0.004, 0.007]	0.63
	Directional	0.021	0.012	0.036	0.012	94	12.72	<.001	[0.012, 0.017]	1.31
	Pseudo-antonym	0.030	0.013	0.037	0.011	94	7.27	<.001	[0.005, 0.010]	0.75
Cause-purpose	Cause:effect	0.029	0.014	0.035	0.011	94	6.58	<.001	[0.004, 0.008]	0.68
	Cause:compensatory action	0.025	0.016	0.035	0.011	94	7.29	<.001	[0.007, 0.012]	0.75
	Action/activity:goal	0.038	0.013	0.039	0.011	94	0.39	.701	[−0.002, 0.003]	n/a
Class inclusion	Taxonomic	0.036	0.017	0.035	0.015	93	−1.41	.161	[−0.003, 0]	n/a
	Functional	0.031	0.017	0.033	0.014	93	2.02	.046	[0, 0.003]	n/a
	Plural collective	0.032	0.015	0.033	0.015	93	1.04	.300	[0, 0.002]	n/a
Part-whole	Mass: portion	0.017	0.013	0.051	0.010	93	21.47	<.001	[0.031, 0.037]	2.22
	Item:topological part	0.039	0.015	0.051	0.008	93	7.32	<.001	[0.009, 0.015]	0.76
	Object:stuff	0.040	0.017	0.049	0.009	93	5.85	<.001	[0.006, 0.012]	0.58
Space-time	Location: process/product	0.033	0.017	0.053	0.005	93	11.08	<.001	[0.017, 0.025]	1.14
	Contiguity	0.026	0.016	0.055	0.057	93	16.23	<.001	[0.025, 0.032]	1.67
	Time: associated item	0.016	0.013	0.055	0.007	93	25.06	<.001	[0.036, 0.042]	2.58

Table 4
Descriptive Statistics for Individual Difference Measures

Exp.	Measure	*n*	*M*	*SD*	Range
1	RPM	90	0.66	0.24	[0.08–1]
	SST	88	30.35	3.67	[21–36.5]
	Type discriminability index	95	1.53	0.34	[1.00–2.6]
	Subrelation discriminability index	95	1.35	0.29	[0.96–2.74]
2	RPM	92	0.61	0.24	[0.08–1.00]
	SST	93	28.08	4.52	[14.50–37.00]
	Type discriminability index	94	1.36	0.33	[0.95–2.33]
	Subrelation discriminability index	94	1.77	0.72	[0.97–4.54]

Note. RPM = Ravens Advanced Progressive Matrices; SST = Semantic Similarities Test.

same general relation type (e.g., old:young and east:west both instantiate the relation type contrast). Third, we computed each participant's discriminability index by dividing that participant's cross-type distance by their within-type distance. This relation type discriminability index reflects how well a participant discriminated between relation types in their similarity judgments. An index of 1 indicates complete lack of discriminability between word pairs instantiating different relation types and those instantiating the same relation type, whereas higher indices indicate judgments of greater similarity between word pairs instantiating the same relation type than between word pairs instantiating different relation types. Descriptive statistics for this type discriminability index for Experiments 1 and 2 are provided in Table 4.

The patterns of individual differences for this type discriminability index are summarized in Table 5. In Experiment 1, the discriminability indices for relation types were significantly correlated with RPM scores ($r = .364$, $p < .001$) and also with SST scores ($r = .372$, $p < .001$). We then assessed the extent that RPM and SST scores each uniquely explained variation in discriminability indices by computing partial correlations to partition out overlapping variance. These partial correlations were statistically significant for both RPM scores ($r = .236$, $p = .028$) and for SST scores ($r = .236$, $p = .028$). In Experiment 2, discriminability indices for relation types were significantly correlated with RPM scores ($r = .410$, $p < .001$) and also with SST scores ($r = .286$, $p = .005$). The partial correlation between these indices and RPM scores after residualizing out SST scores ($r = .337$, $p = .001$) was statistically significant, but that between these indices and SST scores after residualizing out RPM scores ($r = .096$, $p = .365$) was not. These analyses reveal a consistent association between the discrimination of general relation types with cognitive capacity, and a less consistent link with semantic knowledge.

In order to estimate each participant's sensitivity to more fine-grained distinctions between specific subrelations within general relation types, we also computed a subrelation discriminability index using the following steps. First, we found each participant's cross-subrelation distance by calculating the mean distance for pairwise comparisons between word pairs instantiating different subrelations within the same general relation type (e.g., old:young instantiates the subrelation contrary, and east:west instantiates the subrelation directional, where both instantiate the relation type contrast). Second, we found each participant's within-subrelation distance by calculating the mean distance for pairwise comparisons between word pairs instantiating the same subrelation (e.g., old:young and black:white both instantiate the subrelation contrary). Third, we computed each participant's subrelation discriminability index by dividing each participant's cross-subrelation distance by their within-subrelation distance. This subrelation discriminability index reflects how well a participant was able to discriminate between specific subrelations within a relation Type in their similarity judgments. An index of 1 would indicate a complete lack of discriminability between word pairs instantiating different subrelations and those instantiating the same subrelation, whereas higher indices indicate judgments of greater similarity between word pairs instantiating the same subrelation than between word pairs instantiating different subrelations. Descriptive statistics for this subrelation discriminability index for Experiments 1 and 2 are provided in Table 4.

The patterns of individual differences for this subrelation discriminability index are summarized in Table 5 For reference, Table 5 also includes the intercorrelations between the Type and subrelation discriminability indices for both experiments. In Experiment 1, these fine-grained discriminability indices for subrelations showed a significant correlation with RPM scores ($r = .367$, $p < .001$), and also with SST scores ($r = .251$, $p = .018$). A partial correlation between these indices and RPM scores after residualizing out SST scores was statistically significant ($r = .291$, $p = .006$), but that between these indices and SST scores after residualizing out RPM scores was not ($r = .090$, $p = .408$).

Table 5
Pearson Correlations Between Participant RPM and SST Scores and Type and Subrelation Discriminability Indices Computed from Dissimilarity Judgments in Experiments 1 and 2. Partial Correlations Between RPM Scores and Discriminability Indices Residualize Out SST Score, and Partial Correlations Between SST Scores and Discriminability Indices Residualize Out RPM Score

			Bivariate correlations			Partial correlations		
Exp.	Discriminability index	Measure	*r*	*p*	95% CI	*r*	*p*	95% CI
1	Type	RPM	.364	<.001	[.14, .53]	.236	.028	[.03, .42]
		SST	.372	<.001	[.20, .53]	.236	.028	[.03, .42]
	Subrelation	RPM	.367	<.001	[.18, .50]	.291	.006	[.10, .45]
		SST	.251	.018	[.05, .41]	.090	.408	[−.12, .26]
	Type Index × Subrelation Index		.416	<.001	[.19, .55]			
2	Type	RPM	.410	<.001	[.25, .55]	.337	.001	[.14, .50]
		SST	.286	.005	[.10, .46]	.096	.365	[−.12, .31]
	Subrelation	RPM	.353	<.001	[.17, .51]	.257	.014	[.10, .41]
		SST	.310	.003	[.09, .50]	.163	.123	[−.06, .37]
	Type Index × Subrelation Index		.612	<.001	[.50, .73]			

Note. RPM = Ravens Advanced Progressive Matrices; SST = Semantic Similarities Test.

In Experiment 2, these fine-grained discriminability indices for subrelations showed a significant correlation with RPM scores (r = .353, p = .001), and also with SST scores (r = .310, p = .003). A partial correlation between these indices and RPM scores after residualizing out SST scores was statistically significant (r = .257, p = .014), but that between these indices and SST scores after residualizing out RPM scores was not (r = .163, p = .123). These convergent results indicate that there is a reliable association between the discrimination of specific subrelations within relation types with cognitive capacity, but not with semantic knowledge.

To provide a visualization of the difference between high and low discriminability, Figure 5 presents multidimensional scaling (MDS) solutions (Shepard, 1962) for the distance matrices of a participant in Experiment 1 with both a low relation type and a low subrelation discriminability index (left), and of a participant with both a high relation type and a high subrelation discriminability index (right). The latter solution shows a much greater degree of clustering into distinct relation types as well as into subrelations.

Model Predictions

We assessed each of the three computational models of relation similarity as predictors of mean human relational similarity ratings (see Figure 1). Specifically, we computed the correlation between the pairwise cosine distance between two word pairs predicted by each model with human judgments in the resulting dissimilarity matrix using the RSA approach (Kriegeskorte et al., 2008). Confirming the visual impression (relatively close match of pattern for humans with BART in Figure 4, top), in Experiment 1 BART-generated predictions of relational similarity yielded the highest Pearson correlation with human judgments, followed by Word2vec-concat, and then Word2vec-diff (Table 6, top three rows).

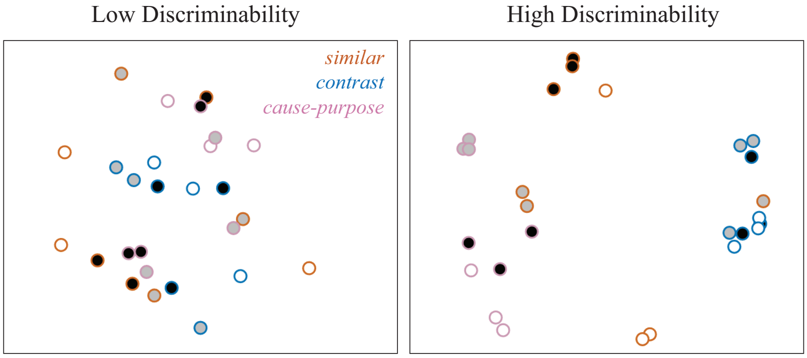
We went on to examine the unique variance in human relational similarity judgments explained by our two relational models, BART and Word2vec-diff, after controlling for the variance explained by Word2vec-concat (which is based solely on lexical similarity). We used semipartial correlations to examine the extent that BART and Word2vec-diff explained the variance in human similarity judgments, after residualizing out the variance in judgments explained by Word2vec-concat and hence attributable solely to lexical similarity. For Experiment 1, the resulting semipartial correlation was highly significant for BART (r = .439, p < .001), but yielded a small but significant negative correlation for Word2vec-diff (r = −.157, p = .003). BART was also the strongest predictor of human judgments in Experiment 2 (Table 6, bottom three rows), for which the semipartial correlation was again highly significant for BART (r = .571, p < .001), but not for Word2vec-diff (r = −.002, p = .964). These analyses confirm that for both data sets, BART provided a superior account of human relational similarity judgments than did the competing computational models.

It could be argued that BART is favored over the competing models because the test items used in the present study were a subset of those used to train BART. However, BART's training task is quite different from that used in the present study. BART's training task consists simply of learning individual subtypes based on labeled examples. In contrast, similarity judgments for word pairs require generating vectors for individual word pairs across all 79 subtypes. Thus, BART was never trained to make similarity judgments for any particular word pairs. Nonetheless, we ran a cross-validation simulation in which we trained BART after removing all the pairs used in the present study (see Tables 1 and 2). When BART is trained on this reduced dataset, it still provides a superior account of human judgments of relational similarity relative to the competing relation model, Word2vec-diff. After controlling for the variability in human similarity ratings accounted for by raw semantic similarity (based on Word2vec-concat), the version of BART with a reduced training set achieves a semipartial correlation of .237 (p < .001) with human similarity ratings in Experiment 1 and .400 (p < .001) with human similarity ratings in Experiment 2.

In order to assess the relative contributions of BART's major components in predicting human judgments of relational similarity, we tested four additional control models. Each control model is a variant of BART from which one individual component has been removed,

Figure 5

Visualization of Relation Similarities From Two Representative Participants in Experiment 1



Note. Left: MDS solution for a participant with low discriminability indices (relation type discriminability index = 1.02; subrelation discriminability index = .98). Right: solution for a participant with high discriminability indices (2.08 and 2.74, respectively). Each marker indicates a single word pair. Marker outline color indicates word pair relation type, and marker shading indicates subrelation within relation type. See the online article for the color version of this figure.

Table 6

Pearson Correlations Between Each Set of Model-Generated Dissimilarity Predictions and Human Judgments in Experiments 1 and 2. Semipartial Correlations Between Human Judgments and BART and Word2vec-Diff Predictions Control for Baseline Word2vec-Concat Predictions

		Bivariate correlations			Semipartial correlations		
Exp.	Model	*r*	*p*	95% CI	*r*	*p*	95% CI
1	BART	.574	<.001	[.482, .659]	.439	<.001	[.333, .545]
	Word2vec-diff	.037	.498	[−.098, .183]	−.157	.003	[−.330, .027]
	Word2vec-concat	.387	<.001	[.236, .520]			
2	BART	.786	<.001	[.725, .832]	.571	<.001	[.484, .643]
	Word2vec-diff	.346	<.001	[.182, .489]	−.002	.964	[−.132, .120]
	Word2vec-concat	.602	<.001	[.486, .696]			

Note. BART = Bayesian Analogy with Relational Transformations.

while holding all other components constant. In Control 1, the power transformation that contributed to BART's final estimation of relational similarity was removed. Without it, BART's correlation with human judgments drops from .57 to .35 in Experiment 1 and from .79 to .41 in Experiment 2. In Control 2, the focused training step (similar vs. contrast relations) was removed. Without it, BART's performance drops from .57 to .30 in Experiment 1 and from .79 to .77 in Experiment 2. The particularly severe drop in performance in Experiment 1 is to be expected, given that the stimuli used in that experiment exemplified similar and contrast relations. Control 3 removed Bayesian logistic regression, the third step of BART's learning algorithm. (Because focused training depends on Bayesian logistic regression, Control 3 necessarily excluded that learning phase as well.) After removing these components, BART's performance drops from .57 to .12 in Experiment 1 and from .79 to .51 in Experiment 2. Control 4 removed BART's use of ranked features. Without these features, BART's performance drops from .57 to .14 in Experiment 1 and from .79 to .64 in Experiment 2.

We also assessed the effectiveness of the three major models in predicting individual differences among participants in their judgments of relation dissimilarity. First, we computed correlations between human judgments of relation dissimilarity and predictions of relation dissimilarity generated by each of the three models (BART, Word2vec-diff, and Word2vec-concat). Recall that in analyzing model predictions of overall relation dissimilarity, we computed the correlation between model predictions and mean human judgments. In contrast, the present analysis involved computing correlations between model predictions and individual human judgments, which resulted in a set of correlation coefficients across individual participants. As in our analyses of overall relation dissimilarity, we computed bivariate correlations between participant judgments and predictions from each of the three models, and then computed semipartial correlations between participant judgments and predictions from BART and Word2vec-diff, after controlling for predictions from Word2vec-concat. This analysis yielded five correlation coefficients for each participant, representing the degree to which each model predicted an individual's judgments of relation dissimilarity. Because RPM performance (i.e., the measure of cognitive capacity) emerged as the stronger predictor of relation discriminability in both Experiments 1 and 2, we examined the relationship between participant RPM scores and the degree of correspondence between participant judgments and model predictions.

Merging the data sets from Experiments 1 and 2, we computed bivariate correlations between RPM scores and each of the five correlation coefficients described above, representing the degree that each of the three models predicted individual participants' dissimilarity judgments. The results of these analyses are summarized in Table 7. RPM scores were correlated with bivariate correlation coefficients between participant and BART dissimilarity ratings (r = .249, p = .001), as well as with semipartial correlation coefficients between participant and BART dissimilarity ratings after controlling for Word2vec-concat dissimilarity ratings (r = .291, $p < .001$). In contrast, no reliable relationships were obtained in corresponding analyses correlating RPM scores with Word2vec-concat (r = −.037, p = .617) or Word2vec-diff bivariate correlation coefficients (r = −.042, p = .569), or with Word2vec-diff semipartial correlation coefficients controlling for Word2vec-concat (r = −.008, p = .910).

Table 7

Pearson Correlations Between RPM Scores and (a) Bivariate Correlation Coefficients Indicating the Degree of Correspondence Between Individual Participant RDMs and Model Predictions (Left Columns), and (b) Semipartial Correlation Coefficients Indicating the Degree of Correspondence Between Individual Participant RDMs and Model Predictions, Controlling for Baseline Word2vec-Concat Predictions (Right Columns)

	RPM × Bivariate Coefficients			RPM × Semipartial Coefficients		
Model	*r*	*p*	95% CI	*r*	*p*	95% CI
BART	.249	<.001	[.082, .395]	.291	<.001	[.153, .424]
Word2vec-diff	−.042	.569	[−.175, .109]	−.008	.910	[−.151, .134]
Word2vec-concat	−.037	.617	[−.190, .114]			

Note. BART = Bayesian Analogy with Relational Transformations; RPM = Ravens Advanced Progressive Matrices; SST = Semantic Similarities Test; RDM= Relational Dissimilarity Matrix.

These results converge to indicate that participants with higher scores on the RPM also generated relational dissimilarity ratings that were better predicted by BART, but not by Word2vec-diff or Word2vec-concat. That is, participants with superior cognitive capacity produced more BART-like judgments of similarity between semantic relations.

Discussion

By testing alternative computational models of relation dissimilarity, all founded on semantic representations of words derived using the same model of distributional semantics (Word2vec), we were able to distinguish between rival accounts of how humans code semantic relations. Across two distinct data sets using six general relation types, human judgments of relational similarity were best predicted by BART, a model that assumes semantic relations are coded by distributed representations across a pool of learned relations. This model made reliable predictions even after statistically removing the predictive power of a baseline model (Word2vec-concat) sensitive only to lexical similarity. Further, the accuracy of BART's predictions of relational similarity judgments increased with a measure of cognitive capacity at the level of individual participants. An alternative model (Word2vec-diff) that assumes relations are coded solely by a generic function (the difference between the vectors for two words in a pair) was unable to robustly predict human judgments. These results converge with prior research that has revealed the limitations of efforts to explain human judgments about abstract relations using an untransformed semantic space (Linzen, 2016; Peterson et al., 2020). At the same time, the success of BART in predicting human judgments of relation dissimilarity provides evidence that abstract semantic relations can indeed be learned from nonrelational inputs (semantic vectors for individual words) created using the approach of distributional semantics. These findings have the potential to advance models of analogical reasoning by providing an objective basis for learning relations from nonrelational inputs without any hand coding.

The present results are particularly striking considering the diverse similarity patterns observed across the three data sets we examined. The data generated in Experiment 1 for three major relation types (similar, contrast, cause–purpose) revealed a different similarity structure than did the data from Experiment 2 for three other relation types (class inclusion, part–whole, space–time). In Experiment 2, human judgments did not distinguish between subrelations within the class inclusion relation type, and did not show space–time to be a distinct and coherent relation category. This apparent diversity in human judgments across different relation types deserves further exploration. But despite this variability in human similarity judgments across relation types, for each dataset BART yielded the best fit among the computational models that were tested.

The present findings address a longstanding question concerning the mental and neural representation of semantic relations: whether relations are simply implicit (e.g., defined by the vector distance between the representations of the entities being related, as assumed by Word2vec-diff), or have explicit representations based on the meanings of specific relations (Chaffin, 1992; Chaffin & Hermann, 1988b; Popov et al., 2017, 2020). Several lines of evidence now converge on the conclusion that semantic relations are coded explicitly by a distributed code, as postulated by the BART model. A distributed representation created by learning from examples is able to account for human behavioral data both for within-relation structure—the typicality gradient observed for semantic relations (Chaffin & Herrmann, 1988a; Jurgens et al., 2012; Lu et al., 2019; Popov et al., 2020)—and also for between-relation structure—the pattern of similarity judgments across a diverse set of relations, as shown here. In addition, the same basic model can predict patterns of neural similarity at the item level within a frontoparietal network for relations computed during a verbal analogy task (Chiang et al., 2020). Finally, a model based on a distributed code for relations can explain the emergence of analogical reasoning from a basic process of relational comparison (Lu et al., 2019).

The present behavioral findings also converge with neural evidence that relational reasoning is heavily dependent on circuitry (primarily in prefrontal cortex) that supports aspects of human cognitive capacity, particularly working memory and inhibitory control (Bunge et al., 2005; Cho et al., 2010; Knowlton et al., 2012; for a review see Holyoak & Monti, 2020). The RPM, a basic measure of cognitive capacity, is a general predictor of relational reasoning (Gray & Holyoak, 2020; Kubricht et al., 2017; Vendetti et al., 2014). Here we found that RPM scores predicted both the degree to which people draw clear distinctions among semantic relations, and also the degree to which the BART Model predicts the similarity patterns observed for individual participants.

References

Arthur, P. L., Tubre, T. C., Paul, D. S., & Sanchez-Ku, M. L. (1999). College-sample psychometric and normative data on a short form of the Raven Advanced Progressive Matrices Test. *Journal of Psychoeducational Assessment*, *17*(4), 354–361. 10.1177/0734282999 01700405

Bejar, I. I., Chaffin, R., & Embretson, S. E. (1991). *Cognitive and psychometric analysis of analogical problem solving*. Springer-Verlag.

Bhatia, S. (2017). Associative judgment and vector space semantics. *Psychological Review*, *124*(1), 1–20. 10.1037/rev0000047

Bhatia, S., Richie, R., & Zou, W. (2019). Distributed semantic representations for modeling human judgment. *Current Opinion in Behavioral Sciences*, *29*, 31–36. 10.1016/j.cobeha.2019.01.020

Bunge, S. A., Wendelken, C., Badre, D., & Wagner, A. D. (2005). Analogical reasoning and prefrontal cortex: Evidence for separable retrieval and integration mechanisms. *Cerebral Cortex*, *15*(3), 239–249. 10.1093/cercor/bhh126

Chaffin, R. (1992). The concept of a semantic relation. In A. Lehrer & E. Kittay (Eds.), *Frames, fields, and contrasts: New essays in semantic and lexical organization* (pp. 253–288). Erlbaum.

Chaffin, R., & Herrmann, D. J. (1988a). Effects of relation similarity on part-whole decisions. *The Journal of General Psychology*, *115*(2), 131–139. 10.1080/00221309.1988.9711096

Chaffin, R., & Herrmann, D. J. (1988b). The nature of semantic relations: A comparison of two approaches. In M. Evens (Ed.), *Relational models of the lexicon: Representing knowledge in semantic networks* (pp. 289–334). Cambridge University Press.

Chiang, J. N., Peng, Y., Lu, H., Holyoak, K. J., & Monti, M. M. (2020). Distributed code for semantic relations predicts neural activity during analogical reasoning. *Journal of Cognitive Neuroscience*. Advance online publication. 10.1162/jocn_a_01620

Cho, S., Moody, T. D., Fernandino, L., Mumford, J. A., Poldrack, R. A., Cannon, T. D., Knowlton, B. J., & Holyoak, K. J. (2010). Common and dissociable prefrontal loci associated with component mechanisms of analogical reasoning. *Cerebral Cortex*, *20*(3), 524–533. 10.1093/cercor/bhp121

Common Core State Standards Initiative. (2020). *English language arts standards (Language, grade 4)*. http://www.core standards.org/ELA-Literacy/L/4/5/c/

Estes, Z., & Jones, L. L. (2006). Priming via relational similarity: A COPPER HORSE is faster when seen through a GLASS EYE. *Journal of Memory and Language*, *55*(1), 89–101. 10.1016/j.jml.2006.01.004

Estes, Z., & Jones, L. L. (2009). Integrative priming occurs rapidly and uncontrollably during lexical processing. *Journal of Experimental Psychology: General*, *138*(1), 112–130. 10.1037/a0014677

Forbus, K. D., Ferguson, R. W., Lovett, A., & Gentner, D. (2017). Extending SME to handle large-scale cognitive modeling. *Cognitive Science*, *41*(5), 1152–1201. 10.1111/cogs.12377

Gray, M. E., & Holyoak, K. J. (2020). Individual differences in relational reasoning. *Memory & Cognition*, *48*(1), 96–110. 10.3758/s13421-019-00964-y

Hill, F., Reichart, R., & Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, *41*(4), 665–695. 10.1162/COLI_a_00237

Hofmann, M. J., Biemann, C., Westbury, C., Murusidze, M., Conrad, M., & Jacobs, A. M. (2018). Simple co-occurrence statistics reproducibly predict association ratings. *Cognitive Science*, *42*(7), 2287–2312. 10.1111/cogs.12662

Holyoak, K. J., & Monti, M. M. (2020). Relational integration in the human brain: A review and synthesis. *Journal of Cognitive Neuroscience*. Advance online publication. 10.1162/jocn_a_01619

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453–458. 10.1038/nature17637

Jones, L. R., Wurm, L. H., Calcaterra, R. D., & Ofen, N. (2017). Integrative priming of compositional and locative relations. *Frontiers in Psychology*, *8*, 359. 10.3389/fpsyg.2017.00359

Jozwik, K. M., Kriegeskorte, N., Storrs, K. R., & Mur, M. (2017). Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. *Frontiers in Psychology*, *8*, 1726. 10.3389/fpsyg.2017.01726

Jurgens, D. A., Mohammad, S. M., Turney, P. D., & Holyoak, K. J. (2012). SemEval-2012 Task 2: Measuring degrees of relational similarity. *Proceedings of the First Joint Conference on Lexical and Computational Semantics (SEM)* (pp. 356–364). Association for Computational Linguistics.

Knowlton, B. J., Morrison, R. G., Hummel, J. E., & Holyoak, K. J. (2012). A neurocomputational system for relational reasoning. *Trends in Cognitive Sciences*, *16*(7), 373–381. 10.1016/j.tics.2012.06.002

Kriegeskorte, N., & Mur, M. (2012). Inverse MDS: Inferring dissimilarity structure from multiple item arrangements. *Frontiers in Psychology*, *3*, 245. 10.3389/fpsyg.2012.00245

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis—Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4. 10.3389/neuro.06.004.2008

Kubricht, J. R., Lu, H., & Holyoak, K. J. (2017). Individual differences in spontaneous analogical transfer. *Memory & Cognition*, *45*(4), 576–588. 10.3758/s13421-016-0687-7

Linzen, T. (2016). Issues in evaluating semantic spaces using word analogies. *Proceedings of the First Workshop on Evaluating Vector-Space Representations for NLP* (pp. 13–18). Association for Computational Linguistics.

Lu, H., Chen, D., & Holyoak, K. J. (2012). Bayesian analogy with relational transformations. *Psychological Review*, *119*(3), 617–648. 10.1037/a0028719

Lu, H., Wu, Y. N., & Holyoak, K. (2019). Emergence of analogy from relation learning. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(10), 4176–4181. 10.1073/pnas.1814779116

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, *26*, 3111–3119.

Mikolov, T., Yih, S. W., & Zweig, G. (2013). Linguistic regularities in continuous space word representations *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-2013)* (pp. 746–751). Association for Computational Linguistics.

Mur, M., Meys, M., Bodurka, J., Goebel, R., Bandettini, P. A., & Kriegeskorte, N. (2013). Human object-similarity judgments reflect and transcend the primate-IT object representation. *Frontiers in Psychology*, *4*, 128. 10.3389/fpsyg.2013.00128

Pereira, F., Gershman, S., Ritter, S., & Botvinick, M. (2016). A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. *Cognitive Neuropsychology*, *33*(3–4), 175–190. 10.1080/02643294.2016.1176907

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S. J., Kanwisher, N., Botvinick, M., & Fedorenko, E. (2018). Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications*, *9*(1), 963. 10.1038/s41467-018-03068-4

Peterson, J. C., Chen, D., & Griffiths, T. L. (2020). Parallelograms revisited: Exploring the limitations of vector space models for simple analogies. *Cognition*, *205*, 104440. 10.1016/j.cognition.2020.104440

Ponti, E. M., Vulić, I., Glavaš, G., Mrkšić, N., & Korhonen, A. (2018). Adversarial propagation and zero-shot cross-lingual transfer of word vector specialization. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 282–293). Association for Computational Linguistics.

Popov, V., Hristova, P., & Anders, R. (2017). The relational luring effect: Retrieval of relational information during associative recognition. *Journal of Experimental Psychology: General*, *146*(5), 722–745. 10.1037/xge0000305

Popov, V., Pavlova, M. & Hristova, P. (2020). *The internal structure of semantic relations: Effects of relational similarity and typicality*. https://psyarxiv.com/fqd4b/

Richie, R., & Bhatia, S. (2020). *Similarity judgment within and across categories: A comprehensive model comparison*. https://psyarxiv.com/5pa9r/

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, *104*(3), 192–233. 10.1037/0096-3445.104.3.192

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika*, *27*(2), 125–140. 10.1007/BF02289630

Stamenković, D., Ichien, N., & Holyoak, K. J. (2019). Metaphor comprehension: An individual-differences approach. *Journal of Memory and Language*, *105*, 108–118. 10.1016/j.jml.2018.12.003

Stamenković, D., Ichien, N., & Holyoak, K. J. (2020). Individual differences in comprehension of contextualized metaphors. *Metaphor and Symbol*, *35*(4), 285–301. 10.1080/10926488.2020.1821203

Vendetti, M., Wu, A., & Holyoak, K. J. (2014). Far-out thinking: Generating solutions to distant analogies promotes relational thinking. *Psychological Science*, *25*(4), 928–933. 10.1177/0956797613518079

Zhang, Y., Han, K., Worth, R., & Liu, Z. (2020). Connecting concepts in the brain by mapping cortical representations of semantic relations. *Nature Communications*, *11*(1), Article 1877. 10.1038/s41467-020-15804-w

Zhila, A., Yih, W., Meek, C., Zweig, G., & Mikolov, T. (2013). *Combining heterogenous models for measuring relational similarity*. Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 1000–1009). Association for Computational Linguistics.

Received July 21, 2020
Revision received December 8, 2020
Accepted December 10, 2020 ■