Ichien, N., Kan, A., Holyoak, K. J., & Lu, H. (2022). Generative inferences in relational and analogical reasoning: A comparison of computational models. In J. Culbertson, A. Perfors, H. Rabagliati, & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Meeting of the Cognitive Science Society.* Cognitive Science Society.

Generative Inferences in Relational and Analogical Reasoning: A Comparison of Computational Models

Nicholas Ichien[1] ichien@ucla.edu
Angela Kan[1] angelakan2002@gmail.com
Keith J. Holyoak[1] holyoak@lifesci.ucla.edu
Hongjing Lu[1,2] hongjing@ucla.edu

[1] Department of Psychology
[2] Department of Statistics
University of California, Los Angeles
Los Angeles, CA 90095 USA

Abstract

A key property of human cognition is its ability to generate novel predictions about unfamiliar situations by completing a partially-specified relation or an analogy. Here, we present a computational model capable of producing generative inferences from relations and analogs. This model, BART-Gen, operates on explicit representations of relations learned by BART (*Bayesian Analogy with Relational Transformations*), to achieve two related forms of generative inference: reasoning from a single relation, and reasoning from an analog. In the first form, a reasoner completes a partially-specified instance of a stated relation (e.g., *robin is a type of* ____). In the second, a reasoner completes a target analog based on a stated source analog (e.g., *sedan:car :: robin:*____). We compare the performance of BART-Gen with that of BERT, a popular model for Natural Language Processing (NLP) that is trained on sentence completion tasks and that does not rely on explicit representations of relations. Across simulations and human experiments, we show that BART-Gen produces more human-like responses for generative inferences from relations and analogs than does the NLP model. These results demonstrate the essential role of explicit relation representations in human generative reasoning.

Keywords: relational reasoning, analogy, cognitive modeling, embeddings

Introduction

Human reasoners are remarkably sensitive to structural similarities. For example, despite the superficial differences between generational wealth accumulation and blood clotting, a brief elaboration of each reveals a clear analogy. In the first case, initial financial success allows a family to pass on wealth to the subsequent generation, which then grants that new generation access to social resources enabling its own financial success, affording further wealth to pass onto future generations. In the second case, an initial injury attracts blood platelets to cling to the injured site. Upon recognizing even this hint of a shared relational structure across these two processes, a reasoner can more easily map entities playing corresponding roles, such as wealth and blood platelets. Crucially, the reasoner could also generate the inference that the presence of blood platelets would then attract yet more blood platelets to the injured site.

Computational models of such relational reasoning have been developed both in cognitive science (e.g., Falkenhainer, Forbus, & Gentner, 1989; Hummel & Holyoak, 1997; Lu, Ichien, & Holyoak, 2022) and in artificial intelligence (e.g., Battaglia et al., 2018; Santoro et al., 2017; Shanahan et al., 2020). Models of analogical reasoning developed in cognitive science typically include explicit representations of relations, such that a relation is distinct from, but bound to, the entities it relates. This property supports the recognition of structural similarity by enabling a direct comparison of the relations constituting each analog. Crucially, explicit relation representations can also prompt the *generation* of predictions about a target analog based on the source. Indeed, the generative capacity afforded by relation representations is the core of analogical inference, which human reasoners can exploit in everyday problem solving (Gick & Holyoak, 1980; 1983), technological innovation (Kittur et al., 2019), and scientific discovery (Gentner, 2002; Holyoak & Thagard, 1995; Nersessian, 1992).

Here we introduce a new computational model of generative relational and analogical inference. We then present the results from three simulations, in which we examine the model's ability to capture the human capacity to reason from a relation (Simulations 1a and 1b) and from an analog (Simulation 2). In addition, we compare the performance of the model to that of a leading model of Natural Language Processing (NLP).

The model presented here, BART-Gen, operates on explicit relation representations generated by BART (*Bayesian Analogy with Relational Transformations*) (Lu, Chen, & Holyoak, 2012; Lu, Wu, & Holyoak, 2019; see also Chen, Lu, & Holyoak, 2017), a model of relation learning that acquires representations of relations from unstructured vector representations of individual word meanings. Many previous analogy models have relied on representations that are hand-coded by the modeler, and thus bypass the problem of relation acquisition altogether (see Chalmers, French, & Hofstadter, 1992, for an early critique of such models). In contrast, BART deals directly with the problem of learning relations from non-relational inputs, taking as inputs embeddings for individual words produced by machine-learning algorithms. BART's relation representations have been used to predict human judgments of relational similarity among word pairs (Ichien, Lu, & Holyoak, 2021), to support human-like analogical reasoning on simple four-term verbal problems (e.g., *artificial : natural :: friend : enemy*) (Lu et al., 2019), and to predict patterns of similarity in neural responses to relations during analogical reasoning (Chiang, Peng, Lu, Holyoak, & Monti, 2021). BART also can support analogical

mapping in problems requiring finding correspondences between multiple entities across complex relational systems (e.g., mapping the solar system to atomic structure) (Lu et al., 2022).

We first provide an overview of BART's relation learning algorithm, and then detail how BART-Gen uses the representations learned by BART to perform generative relational and analogical inference.

Relation representation in BART

BART[1] learns explicit representations of the semantic relations between word pairs from unstructured vector representations of individual word meanings (Lu et al., 2012; 2019). In the present simulations, BART's input consists of concatenated pairs of word vectors from Word2vec[2] (Mikolov et al., 2013) and uses supervised learning with positive and negative examples to acquire each relation representation individually. For example, a vector formed by concatenating the individual vectors for *old* and *young* would constitute a positive example for the relation *X is the opposite of Y* and might also serve as a negative example of the relation *X is a synonym of Y*. After learning, BART computes 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 a broad range of semantic relations. In its first stage, BART 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 1200-dimension augmented feature vector for each word pair, consisting of the concatenation of raw and ranked feature vectors for each word in the pair.

In the second stage, BART uses logistic regression with elastic net regularization to select a subset of important feature dimensions across word pairs f_s. In the third stage, BART uses Bayesian logistic regression with f_s to estimate weight distributions w for representing a particular relation R by applying Bayes rule as:

$$P(w|f_s, R) \propto P(R|f_s, w)P(w). \quad (1)$$

The first term is the likelihood defined by a logistic function on w and f_s (selected in the second stage), $\frac{1}{1+e^{-w^T f_s}}$. The second term is the prior distribution of w, defined as a multivariate normal distribution, $N(\mu_0, \Sigma_0)$, with a mean vector $\mu_0 = (\beta, -\beta)$, consisting of the β values of weights estimated in the second stage of logistic regression.

We trained BART by combining two datasets of human-generated word pairs, each chosen as an example of a specific semantic relation. The first dataset (Jurgens, Mohammed, Turney, & Holyoak, 2012) consists of at least 20 word pairs (e.g., *bird:robin*) instantiating each of 79 semantic relations (e.g., *X is a type of Y*) taken from a taxonomy proposed by Bejar, Chaffin, and Embretson (1991), which includes 10 major relation categories (e.g., *class inclusion*). The second dataset consists of at least 10 word pairs instantiating each of 56 additional semantic relations (Popov, Hristova, Anders, 2017). Across both datasets, BART acquired 135 semantic relations via supervised learning. Since BART's learned weights w can be expressed as two separate halves (i.e., those associated with the first relational role, w_1, and those associated with the second relational role, w_2), BART can automatically generate representations of the converse of each learned relation by swapping the relation weights associated with each individual relational role. Thus, upon learning a representation of *X is a type of Y*, BART can also learn a representation of its converse, *Y is a superordinate of X*, the same relation but with the roles flipped. This operation effectively doubles BART's pool of learned relations from 135 to 270 in total.

After learning weight distributions associated with selected feature dimensions across word pairs in its training set f_L,R_L, BART can estimate how likely any novel pair of words A and B instantiates a learned relation R_i, $P(R_i|f_A, f_B)$ by marginalizing the weight distribution for that relation:

$$P(R_i|f_A, f_B) = \int P(R_i|f_A, f_B, w)P(w|f_L, R_L)dw. \quad (2)$$

Hence, given any pair of words A: B, BART can perform this operation for each of its learned relations and then generate a relation vector R_{AB}, in which the value of each element is a posterior probability reflecting how good an example A and B are of that particular relation. For example, given that *old* and *young* constitute a good example of the relation *X is the opposite of Y* but a poor example of the relation *X causes Y*, $R_{old:young}$ would have a high value for the dimension corresponding to the first relation, but a low value for the dimension corresponding to the second dimension. Ichien et al. (2021) added a power transformation to these relation vectors, raising each relation dimension to a power of 5, and found that adding this transformation ("winners take most") improves the model's ability to capture human judgments of relational similarity. Accordingly, we incorporated the same power transformation in the present simulations.

Generative inference in BART-Gen

BART-Gen uses the relation representations acquired by BART to perform generative relational and analogical inference. We first detail its algorithm for reasoning from a relation, and then describe the extended algorithm for generative reasoning via analogy.

Reasoning from a relation in BART-Gen. Recall that the second stage of BART's learning algorithm uses logistic regression with elastic net regularization to select a subset of informative feature dimensions of a word pair, $\boldsymbol{f_s}$. Given the individual words combined within that word pair, these selected feature dimensions can be separated into those

[1] https://cvl.psych.ucla.edu/wp-content/uploads/sites/162/2021/04/BART2code.zip

[2] https://code.google.com/archive/p/word2vec/

corresponding to one word $\boldsymbol{C}$, $\boldsymbol{f_{s_C}}$, and those corresponding to the other word $\boldsymbol{D}$, $\boldsymbol{f_{s_D}}$. Given $\boldsymbol{C}$, and the hypothesis that a relation $\boldsymbol{R}$ holds between $\boldsymbol{C}$ and some predicted $\boldsymbol{D}$, BART-Gen generates a probability distribution of $\boldsymbol{f_{s_D}}$, using the following inference:

$$P(f_{s_D}|R, f_{s_C}) \propto P(R = 1|f_{s_C}, f_{s_D})P(f_{s_D}|f_{s_C}). \quad (3)$$

The likelihood term, $P(R = 1|f_{s_C}, f_{s_D})$, is the probability that R holds for the predicted f_{s_D} and the known f_{s_C}. As with Equation 1, the likelihood term $P(R = 1|f_{s_C}, f_{s_D})$ is defined using a logistic function:

$$P(R = 1|f_{s_C}, f_{s_D}, w) = \frac{1}{1+e^{-w_C^T f_{s_C} - w_D^T f_{s_D}}}. \quad (4)$$

In Equation 4, learned weights w are written as two separate halves: those associated with C's relational role, w_C, and those associated with D's relational role, w_D. Correspondingly, the selected feature dimensions of a given word pair f_s are rewritten as those corresponding to C, f_{s_C}, and D, f_{s_D}.

The prior term, $P(f_{s_D}|f_{s_C})$, follows a multivariate normal distribution conditional on f_{s_C}, which is defined as:

$$P(f_{s_D}|f_{s_C}) = N(f_{s_C}, \sigma^2 I). \quad (5)$$

BART-Gen uses the semantic embedding of word C as a starting point for generating D, in that the means of the prior $P(f_{s_D}|f_{s_C})$ are the feature values of C, reflecting the assumption that D is semantically associated with C. The prior term also assumes equal variance σ^2 for semantic features of word D. σ^2 is a free parameter that controls the degree to which the predicted D is semantically associated with C in the prior. Larger values of σ^2 correspond to a weaker degree of semantic association in the prior. The BART-Gen inference balances the likelihood guided by relation representation and the prior guided by semantic similarity to the query word, so as to generate maximum a posteriori (MAP) estimates of feature values for D words on selected dimensions, $\hat{f}_{s_D}$. Based on initial test simulations we set the variance parameter at 50 for all simulations reported below.

Note that f_{s_D} is only a subset of all feature dimensions along which D is represented, f_D. In order to generate semantic embedding for D along the feature dimensions that were *not* selected by BART's learning algorithm, BART-Gen simply copies over the corresponding feature values for C, f_{ns_C}. Hence, by combining the generated feature values for selected dimensions and copying values for unselected feature dimensions, BART-Gen specifies a complete prediction for f_D for a specific query word C and a relation:

$$f_D = \{f_{ns_C}, \hat{f}_{s_D}\}. \quad (6)$$

Reasoning from an analog in BART-Gen. Solving a generative analogy problem, *A:B :: C:?,* requires generating a D word such that the word pair formed by C and generated D instantiate the same relations as the source word pair consisting of A and B. To solve this task, BART-Gen needs to perform relation identification on the word pair *A:B*, and then use the inferred relations and word C to generate the missing D word. The model generates the D word by marginalizing all possible relations:

$$P(f_D|\, f_C, f_A, f_B) = \sum_r P(f_D|r, f_C)P(r|f_A, f_B). \quad (7)$$

BART-Gen relies on a *distributed* vector representation of the relation holding between a pair of concepts $\boldsymbol{A}$ and $\boldsymbol{B}$, $\boldsymbol{R_{AB}}$, which consists of a set of posterior probability each corresponding to a distinct relation learned by BART (see Equation 2). BART-Gen iterates through each of these relations, using the algorithm described in the previous section to compute a specific prediction of word embedding for $\boldsymbol{D}$ from the learned relation corresponding to that dimension. That is, BART-Gen repeats its algorithm for reasoning from a relation, using each relation for which BART has learned an explicit representation. Given 270 learned relations, BART-Gen generates 270 distinct predictions of word embeddings for $\boldsymbol{D}$. Then according to Equation 7, BART-Gen computes a weighted average of the set of generated D embeddings, scaled by the normalized relation vector. Thus, predictions from the particular relations for which $\boldsymbol{A}$ and $\boldsymbol{B}$ constitute a good example contribute much more to the final prediction of $\boldsymbol{D}$ than those relations for which $\boldsymbol{A}$ and $\boldsymbol{B}$ constitute a poor example.

Baseline model: BERT for generative inference without explicit relations

For comparison with BART-Gen, we also derived generative inferences from a major NLP model, *Bidirectional Encoder Representations from Transformers* (BERT; Devlin et al., 2019), developed in artificial intelligence (AI) research. BERT (no relation to BART!) is a prominent example of a transformer architecture. Like other similar NLP models, BERT is trained to predict words in sequence within a huge text corpus. Given an incomplete sentence such as "A robin is a type of ____," BERT is trained to predict words that would complete that sentence with the highest probability. Importantly, BERT and similar models routinely solve generation tasks without any explicit relation representations, instead relying solely on the statistics of word usage in their training corpora.

Recent evidence supports the possibility that BERT captures important aspects of human conceptual knowledge. Bhatia and Richie (in press) have shown that a version of BERT fine-tuned to complete sentences related to human-generated semantic feature norms (e.g., "Cat is a four-footed animal") can model several phenomena characteristic of human semantic cognition: predicting semantic verification times, typicality judgments, feature distribution judgments, and semantic similarity judgments.

BERT thus provides an impressive model of human verbal behavior that (unlike BART) does not rely on explicit relation representations. Moreover, the basic training regime for BERT is based on massive experience with sentence generation tasks, which make the model a natural candidate to predict human performance in generative inference tasks with relations and analogies. In the simulations reported here, we use BERT as a non-relational model to predict the pattern of human generative inferences from relations and analogs.

We used *Transformer Models for MATLAB* toolbox[3], a default bert-base model pre-trained on the BooksCorpus (800M words) (Zhu et al., 2015) and the English Wikipedia corpus (2,500M words) (Devlin et al., 2019).

Reasoning from a Single Relation

In the first simulations, we test BART-Gen's ability to reason from a single relation. We operationalize this capacity as generating a word *D* (e.g., *bird*) that best instantiates a known relation *R* (e.g., *is a type of*) with a query word C (e.g., *robin*); i.e., completing relational sentences such as "A robin is a type of ____." We restrict our analyses to those relations for which BART has learned an explicit representation, comparing the performance of BART-Gen with that of BERT.

Simulation 1a: SemEval 2012 Task 2 Dataset

We compared model performance using a series of problems derived from statements consisting of three components: a word pair and a semantic relation that that word pair was generated to instantiate, in the form *word1-relation-word2* (e.g., *robin is a type of bird*). To construct these problems, we used the dataset of human-generated word pairs used to train BART (Jurgens et al., 2012), thus ensuring that BART-Gen had an explicit representation of each relation instantiated in these statements. We generated statements using the 20 most typical word pairs for each of the 79 semantic relations from Jurgens et al. (2012), yielding 1,580 statements in total.

Relation completion problems (BART-Gen). Each statement yielded two relation completion problems, which omitted either the first word in its word pair (e.g., *bird*) or the second word (e.g., *robin*), yielding 3,160 of these problems with which to evaluate BART-Gen. Solution to each of these problems involved generating the omitted word.

Relation sentence completion problems (BERT). To construct corresponding sentence completion problems with which to test BERT, we used relation descriptions (e.g., *Y is a type of X*) provided by Bejar et al. (1991). We embedded either one of the words in each word pair (e.g., *bird:robin*) into its relation description to generate a problem either omitting the first word of the word pair (e.g., "Robin is a type of ____") or the second word (e.g., "____ is a type of bird"). As with the relation completion problems, each statement yielded two sentence completion problems, yielding a total of 3,160 problems to evaluate BERT.

Results and discussion. Across all problems, each model generated a set of words ranked according to the model's confidence in the corresponding prediction (i.e., first-ranked word among a model's set of predictions represented the word for which the model was most confident). In order to evaluate models, we took each model's ranking of the most typical answer provided in the Jurgens et al. (2012) dataset, which was defined as the correct answer. In computing rankings, we excluded any strings containing non-letter symbols (e.g., #, !, /) (sometimes generated by BERT). A lower ranking for the correct answer on a particular problem indicates more accurate model performance. Because each model's predicted words were generated from that model's dictionary, these rankings were sensitive to the overall size of each model's dictionary, such that smaller dictionaries may systematically yield lower (i.e., better) rankings. Given that BERT's dictionary was considerably smaller (30,522 words) than the Word2vec dictionary used by BART-Gen (929,022 words), our analyses favored BERT due to the smaller size of its dictionary.

Despite this difference in dictionary size, BART-Gen outperformed BERT, consistently generating a lower rank for the correct answer across relations. Figure 1 shows the median ranks of correct answers, broken down according to the 10 relation categories defined by Bejar et al. (1991). These results demonstrate superior performance of BART-Gen relative to BERT as a model of generative relational inference. BERT constitutes a demonstration that explicit relation representations are not necessary for generating predictions on this relation completion task; however, BART-Gen, which is guided by such representations, proved much more successful in generating human-like completions preferred by humans.

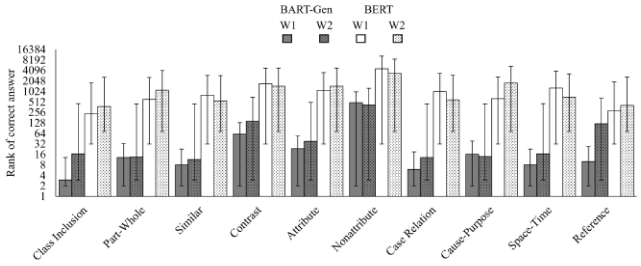

Figure 1. Results from Simulation 1a with generative relation problems (e.g., *robin is a type of ?*) for 10 relation categories (lower ranks indicate better performance). W1 represents problems for which models were tasked with generating the first word of each word pair, given the second word and relation; W2 represents problems for which models were tasked with generating the second word, given the first word and relation. Error bars indicate interquartile range in this paper. The *y*-axis is represented on a log (base 2) scale.

Note that BART-Gen was exposed to all of the word pairs used to construct each of the relation completion problems during explicit relation learning. To ensure that BART-Gen's superior performance is not due to its exposure to word pairs during relation learning in BART, we further tested a version of BERT that was similarly exposed to these word pairs: For each relation completion problem (e.g., "Robin is a type of ____"), we provided BERT with 19 complete relational phrases based on all the other word pairs that were used to instantiate the same relation (e.g., "Spear is a type of weapon and oak is a type of tree and pig is a type of animal…robin is a type of ____"). Providing this input improved BERT's performance; however, across all 10 relational categories, BART-Gen's median rank for the correct answer (median

[3] https://github.com/matlab-deep-learning/transformer-models

rank = 21) was still considerably lower than that achieved by this input-rich BERT (median rank = 181).

Notably, the problems used in Simulation 1a were constructed using a dataset for which human reasoners provided intact word pairs as examples of various semantic relations (Jurgens et al., 2012). Thus, although these word pairs were indeed human-generated, the task within which human reasoners provided these word pairs differed slightly from the task that models reproduced in the simulation. In particular, we defined the "correct" response as that which people rated as most typical of a relation, rather than a response that people directly generated. In order to better evaluate model performance, in Simulation 1b we directly measured human responses in the generative inference task.

Simulation 1b: Human experiment

We collected human responses on a selection of sentence completion problems used in Simulation 1a. These problems were generated from 16 statements, each consisting of a different relation and a word pair that was highly typical of the relation. These relations were evenly divided among four relation categories from Bejar et al. (1991): *class inclusion*, *part-whole*, *case relation*, and *cause-purpose*. Since each statement was used to generate two problems (differing in which word was omitted), we acquired responses to 32 problems in total.

We separated these problems into two 16-problem lists, counterbalanced and presented in randomized orders across participants. Each list consisted of a single problem generated from each statement. Procedure and analyses were pre-registered on AsPredicted (#84748).

Participants. Participants were 100 MTurk workers (M_{age} = 39.06, SD_{age} = 9.19; 45 female, 55 male) who completed our tasks online for payment of $2. The study was approved by the Institutional Review Board at UCLA. Participants had a minimum education level of a U.S. high school graduate, and were sampled from the following English-speaking countries: Australia, Canada, Ireland, New Zealand, South Africa, the United Kingdom, and the United States. We excluded data from 2 participants who reported having trouble paying attention while completing the study, as well as 2 other participants who provided nonsensical responses. Since each participant completed 16 out of the total 32 problems, roughly 50 participants provided responses for each problem.

Results and discussion. Across problems, participants generated a variety of responses, which were largely sensible. Figure 2 shows the proportions of human-generated responses for two sentence completion problems constructed out of the same statement. The most frequent human responses matched the 'correct' response included in the Jurgens et al. (2012) norms for 24 out of the 32 problems. When human responses yielded asymmetries between the two problems generated from the same statement (i.e., easier to perform the completion task for one query word than the other), BART-Gen's predictions were consistent with human responses (67% of the cases) more often than were BERT's (33% of the cases). For the present simulations, we evaluated model performance by finding the rank of the most frequent human-generated response among *all* human-generated responses, aggregated across all problems. As shown in Figure 3, BART-Gen outperformed BERT for all relation categories, consistently ranking the most frequent human response in top ten, lower than BERT did.

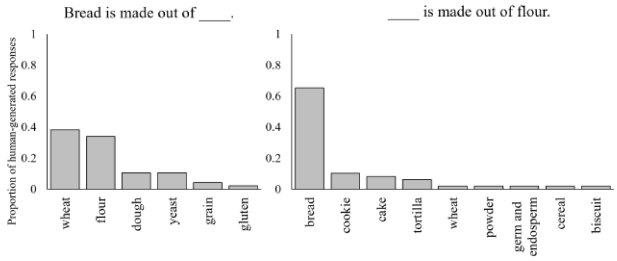

Figure 2. Proportion of human-generated responses to two sentence completion problems, constructed from the same statement. These statements are based on the word pair *bread:flour* and the relation *X is made out of Y*.

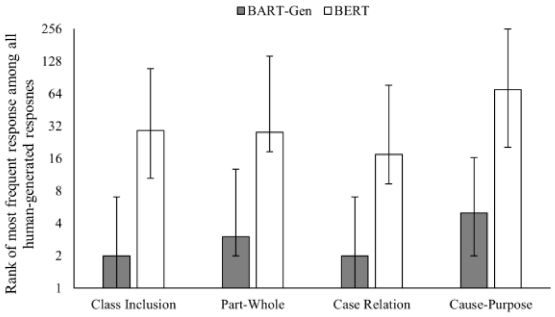

Figure 3. Results from Simulation 1b with generative relation problems (e.g., *robin is a type of ?*), showing median ranks for the most frequent human-generated response, among all human-generated responses across the task (lower ranks indicate better performance).

The results of Simulations 1a and 1b indicate that BART-Gen shows considerable promise as a model of generative relational inference. In outperforming BERT across simulations, BART-Gen generated more human-like predictions than did BERT for the type of sentence completion task on which BERT had been originally trained. The present results are consistent with other evidence supporting the importance of explicit relation representations in accounting for human-like relational reasoning (e.g., Ichien et al., 2021; Lu et al., 2019).

Reasoning from an Analog

In the final simulation, we shift focus from inference based on a single relation to solving analogy problems based on untrained relations for BART. We operationalize the capacity to reason from analogs as the ability to generate a word *D* (e.g., *bird*) that, when linked to a given word *C* (e.g., *robin*), is most analogous to another pair of words *A* (e.g., *sedan*) and *B* (e.g., *car*). We compared the performance of BART-Gen with that of BERT on the task of completing analogical

sentences such as, “Sedan is related to car, just as robin is related to ____.”

Simulation 2: Peterson et al. (2020) Exp. 1a

For Simulation 2 we used a set of 80 four-term analogy problems developed by Green et al. (2010) and adapted for generative analogical inference by Green et al. (2012). Half of these problems consist of *near* analogies, in which the *A* and *B* terms are semantically associated with the *C* and *D* terms (e.g., *answer:riddle :: solution:problem*). The other half consists of *far* analogies in which the corresponding terms are semantically distant (e.g., *answer:riddle :: key:lock*). In general, human reasoners have greater difficulty solving far than near problems (Green et al., 2010; 2012). Importantly, this set of problems is based on very specific relations that BART had not acquired during training; hence this dataset constitutes a strong test of generalization for BART’s relation representations, as well as a natural basis for evaluating BART-Gen’s algorithm for generating relational inferences from any analog.

To create generation problems, the fourth term of each analogy problem was removed (e.g., *answer:riddle :: key:lock* becomes *answer:riddle :: key:?*). We compared the performance of BART-Gen with that of BERT, which completed matched analogical sentences, such as "Answer is related to riddle, just as key is related to ____.” In order to evaluate both models, we compared their responses to human-generated responses collected by Peterson et al. (2020, Experiment 1a).

Results and discussion. As in Simulation 1b, we evaluated model performance by finding the rank of the most frequent human-generated response to each problem among all human-generated responses across all problems. As shown in Figure 4, BART-Gen outperformed BERT, generating lower ranks for the most frequent human responses across problems. These results reveal that BART-Gen can produce human-like responses on a generative analogy task. Moreover, BART-Gen (but not BERT) proved robust to variations in the semantic distance of analogies in terms of accounting for human judgments in generative analogical inference, emphasizing the importance of explicit relation representation for human-like analogical generalization.

General Discussion

We introduce BART-Gen, a new model capable of two related forms of generative inference: reasoning from a single relation, and reasoning from an analog. In the first form, a reasoner completes a partially-specified instance of a stated relation (e.g., *robin is a type of* ____). In the second, a reasoner completes a target analog based on a stated source analog (e.g., *sedan:car :: robin:*____). BART-Gen operates on explicit representations of relations learned from non-relational inputs (word embeddings produced by Word2vec).

We compared BART-Gen to a widely used NLP model, BERT (Devlin et al., 2019). BERT lacks explicit relation representations, but nevertheless appears to produce human-like behavior on several verbal reasoning tasks after fine-tuning on human-generated feature norms (Bhatia & Richie, in press). Across simulations, BART-Gen approximated inferences produced by humans more closely than did completions generated by BERT. This advantage for BART-Gen was obtained even though the tasks we simulated are formally equivalent to the basic sentence-completion task on which BERT was originally trained, or the tasks involving relations that were not trained for BART. Our results thus support the importance of explicit relation representations in human reasoning (Ichien et al., 2021; Lu et al., 2019, 2022).

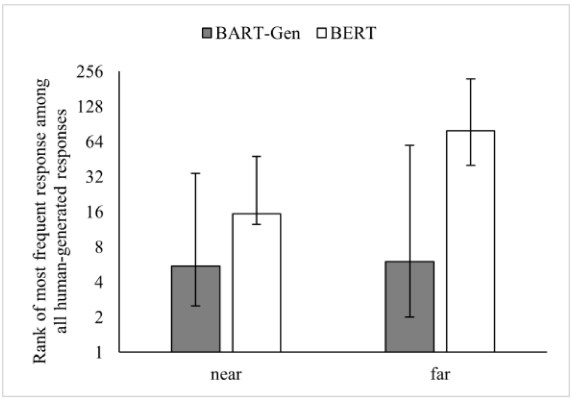

Figure 4. Results from Simulation 2 with generative analogy problems (e.g., *answer:riddle :: key:?*) across semantically *near* and *far* analogies. Lower ranks indicate better performance.

Although BART-Gen tended to rank the “best” analogical completion relatively low in the comparison set of potential responses, the model’s choice usually was not ranked first by humans. One way to potentially improve BART-Gen’s performance on analogy problems would be to employ a “generate-test” strategy: given a limited number of lower-ranked choices produced by BART-Gen, the BART model itself could be used to evaluate the similarity of the *A:B* and *C:D* in analogy. The model’s final choice of the best *D* term would be whichever lower-ranked option maximizes relational similarity.

The present work has focused exclusively on completion of semantic relations, presented either alone (Simulations 1a and 1b) or as part of a four-term analogy problem (Simulation 2). In more general analogical reasoning and problem solving (e.g., Gick & Holyoak, 1980, 1983), inferences are generated on the basis of more complex systems of relations, each involving more than two entities and higher-order causal relations (e.g., Yuille & Lu, 2007). An important future direction will be to extend a model of analogical mapping based on vector representations (e.g., Lu et al., 2022) to include mechanisms for generative inferences, as well as the induction of more general relational schemas.

Acknowledgements

Preparation of this paper was supported by NSF Grant BCS-1827374 awarded to K.J.H and IIS-1956441 H.L.

References

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V.,…Pascanu, R. (2018). Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*.

Bejar, I. I., Chaffin, R., & Embretson, S. (1991). *Cognitive and psychometric analyses of analogical problem solving*. New York: Springer-Verlag.

Bhatia, S. (2017). Associative judgment and vector space semantics. *Psychological Review*, *124*(1), 1-20.

Bhatia, S., & Richie, R. (in press). Transformer networks of human conceptual knowledge. *Psychological Review*.

Chalmers, D. J., French, R. M., & Hofstadter, D. R. (1992). High-level perception, representation, and analogy: A critique of artificial intelligence methodology. *Journal of Experimental and Theoretical Artificial Intelligence*, *4*(3), 185-211.

Chen, D., Lu, H., & Holyoak, K. J. (2017). Generative inferences based on learned relations. *Cognitive Science*, *41*, 1062-1092.

Chiang, J. N., Peng, Y., Lu, H., Holyoak, K. J., & Monti, M. M. (2021). Distributed code for semantic relations predicts neural similarity during analogical reasoning. *Journal of Cognitive Neuroscience*, *33*(3), 377-389.

Devlin, J., Chang, M-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT 2019*, 4171-4186.

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

Gentner, D. (2002). Analogy in scientific discovery: The case of Johannes Kepler. In L. Magnani & N. J. Nersessian (Eds.), *Model-based reasoning: Science, technology, values* (pp. 21-39). New York: Springer Science+Business Media.

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology*, *12*, 306-355.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, *15*, 1-38.

Green, A. E., Kraemer, D. J., Fugelsang, J. A., Gray, J. R., & Dunbar, K. N. (2010). Connecting long distance: Semantic distance in analogical reasoning modulates frontopolar cortex activity. *Cerebral Cortex*, *20*(1), 70-76.

Green, A. E., Kraemer, D. J. M., Fugelsang, J. A., Gray, J. R., & Dunbar, K.N. (2012). Neural correlates of creativity in analogical reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 264–272.

Holyoak, K. J., & Thagard, P. (1995). *Mental leaps: Analogy in creative thought*. Cambridge, MA: MIT Press.

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

Kittur, A., Yu, L., Hope, T., Chan, J., Lifshitz-Assaf, H., Gilon, K., Ng, F., Kraut, R. E., & Shahaf, D. (2019). Scaling up analogical innovation with crowds and AI. *Proceedings of the National Academy of Sciences, USA*, *116*(*6*), 1870-1877.

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

Lu, H., Ichien, N., & Holyoak, K. J. (2022). Probabilistic analogical mapping with semantic relation networks. *Psychological Review.* https://doi.org/10.1037/rev0000358

Lu, H., Wu, Y. N., & Holyoak, K. J. (2019). Emergence of analogy from relation learning. *Proceedings of the National Academy of Sciences, USA*, *116*, 4176-4181.

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*, 311-3119.

Nersessian, N. (1992). How do scientists think? Capturing the dynamics of conceptual change in science. In R. Giere (Ed.), *Cognitive models of science* (pp. 3-44). Minneapolis: University of Minnesota Press.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. *Prioceedings of the 2014 Conference on Empirical Methods in Natural Language Processing,* 1532-1543.

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.

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

Rumelhart, D. E., & Abrahamson, A. A. (1973). A model for analogical reasoning. *Cognitive Psychology*, *5*(1), 1-28.

Santoro, A., Raposo, D., Barrett, D. G., Malinowski, Pascanu, R., Battaglia, P., & Lillicrap, T. (2017). A simple neural network module for relational reasoning. *Advances in Neural Information Processing Systems,* 4967-4976.

Shanahan, M., Nikiforou, K., Creswell, A., Kaplanis, C., Barrett, D., & Garnelo, M. (2020). An explicitly relational neural network architecture. *Proceedings of the 37th International Conference on Machine Learning*, *119*, 8593-9603.

Yuille, A. L., & Lu, H. (2007). The noisy-logical distribution and its application to causal inference. *Advances in neural information processing systems*, *20*.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *Proceedings of the IEEE International Conference on Computer Vision*, 19-27.