Ichien, N., Alfred, K. L., Baia, S., Kraemer, D. J. M., Bunge, S. A., Lu. H., & Holyoak, K. J. (2022). Relation representations in analogical reasoning and recognition memory. In J. Culbertson, A. Perfors, H. Rabagliati, & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Meeting of the Cognitive Science Society.* Cognitive

Relation Representations in Analogical Reasoning and Recognition Memory

Nicholas Ichien[1] ichien@ucla.edu
Katherine L. Alfred[2] katherine.l.alfred.gr@dartmouth.edu
Sophia Baia[3] sbaia@berkeley.edu
David J. M. Kraemer[4] david.j.m.kraemer@dartmouth.edu
Silvia A. Bunge[3, 5] sbunge@berkeley.edu
Hongjing Lu[1, 6] hongjing@ucla.edu
Keith J. Holyoak[1, 7] holyoak@lifesci.ucla.edu

[1] Department of Psychology
[6] Department of Statistics
[7]Brain Research Institute
University of California, Los Angeles
Los Angeles, CA 90095 USA

[2] Department of Psychological and Brain Sciences
[4] Department of Education
Dartmouth College
Hanover, NH 03755 USA

[3] Department of Psychology
[5] Helen Wills Neuroscience Institute
University of California, Berkeley
Berkeley, CA 94720 USA

Abstract

Many computational models of reasoning rely on explicit relation representations to account for human cognitive capacities such as analogical reasoning. *Relational luring*, a phenomenon observed in recognition memory, has been interpreted as evidence that explicit relation representations also impact episodic memory; however, this assumption has not been rigorously assessed by computational modeling. We implemented an established model of recognition memory, the Generalized Context Model (GCM), as a framework for simulating human performance on an old/new recognition task that elicits relational luring. Within this basic theoretical framework, we compared representations based on explicit relations, lexical semantics (i.e., individual word meanings), and a combination of the two. We compared the same alternative representations as predictors of accuracy in solving explicit verbal analogies. In accord with previous work, we found that explicit relation representations are necessary for modeling analogical reasoning. In contrast, preliminary simulations incorporating model parameters optimized to fit human data reproduce relational luring using any of the alternative representations, including one based on non-relational lexical semantics. Further work on model comparisons is needed to examine the contributions of lexical semantics and relations on the luring effect in recognition memory.

Keywords: relational luring, analogy, episodic memory

Introduction

Human reasoning depends on the ability to represent the world not only in terms of individual concepts, such as *beagle* and *dog*, but also in terms of the *relations* between concepts, such as a beagle *is a kind of* dog. Computational models of human analogical reasoning have incorporated explicit representations of relations, so that a relation can link multiple pairs of concepts yet remain distinct from any particular linked concepts (e.g., Falkenhainer, Forbus, & Gentner, 1989; Hummel & Holyoak, 1997). Thus, the relation *is a kind of* can also link *spear* and *weapon*, and an indefinite number of other concept pairs, while maintaining its separate identity.

Relational Luring in Recognition Memory

If relations have explicit representations used in reasoning tasks, then it may be possible to detect their influence in memory tasks that do not directly involve reasoning. Recently, it has been reported that relation similarity can impact episodic memory in recognition tasks, yielding a phenomenon termed *relational luring* (Popov, Hristova, & Anders, 2017). In a typical experiment, participants were shown a sequence of word pairs to commit to memory, and at test were asked to indicate that a given word pair was 'old' if they had seen that exact word pair previously in the sequence, 'recombined' if it was a novel combination of individual words that they had seen before, or 'new' if they had not previously seen either the full word pair or its constituent words. Popov et al. showed that participants were more likely to misclassify 'recombined' word pairs as 'old', and took longer to correctly identify 'recombined' word pairs, when the pair instantiated a relation made familiar by previously-presented pairs as compared to word pairs that did not instantiate the same relation as a prior word pair. Moreover, the degree to which 'recombined' word pairs were misclassified, and correct responses were delayed, increased linearly with the number of instances of that relation a participant had seen previously (see also Challis & Sidhu, 1993; Reder et al., 2000).

On the face of it, relational luring is naturally explained by assuming that an explicit representation of a semantic relation becomes increasingly familiar as it is activated by exposure to specific instances. The accrued familiarity of the relation then serves as a cue that tends to lead to false recognition of recombined word pairs that instantiate the same relation. Thus, relational luring has been interpreted as providing evidence for the role of explicit relations in guiding recognition memory (Popov et al., 2017). However, this assumption has never been formalized in a computational model of recognition memory, nor compared against alternative possibilities. The present paper fills this gap.

Word Embeddings as Predictors of Analogical Reasoning and Word Recognition

Advances in natural language processing (NLP) have generated representations of individual word meanings (e.g., Mikolov et al., 2013; Pennington, Socher, & Manning, 2014; Devlin et al., 2019), referred to as *word embeddings*. These representations are high-dimensional vectors that constitute hidden layers of activation within neural network models trained to predict patterns of text in sequence as they appear in large corpora. Word embeddings have been used to predict human judgments of lexical similarity and probability (for a review see Bhatia & Aka, 2022; for a discussion of and response to critiques of embeddings as psychological models, see Günther et al., 2019.)

Crucially, word embeddings may capture rich aspects of conceptual meaning that go beyond surface features and direct category relations. For example, Utsumi (2020) was able to extract information from embeddings sufficient to predict the values of about 500 words on most of 65 semantic features for which neurobiological correlates have been identified. Such successes suggest that it may be possible to account for relational luring in terms of lexical overlap based solely on embeddings for word pairs, without necessarily involving explicit relation representations. In particular, embeddings might capture information about characteristic relational *roles* that concepts play (Goldwater, Markman, & Stilwell, 2011; Jones & Love, 2007; Markman & Stilwell;, 2001). For example, concatenated embeddings for the word pair *nurse:hospital* might include features that implicitly encode the facts that *nurse* is a human occupation and that *hospital* is a work location, perhaps creating a basis for relational luring.

In the present study we build on recent theoretical developments in which embeddings are used to learn relation representations that can provide a basis for analogical reasoning. A number of alternative methods can be used to define relation similarity, in the sense of similarity between word pairs. In the present study, alternative methods take the same embeddings as inputs, extracted using Word2vec (Mikolov et al., 2013), and all compute relation similarity based on cosine similarity (a measure well-suited for high-dimensional spaces). Critically, relation representations can either be based on explicit re-representations within a new relational space, or implicit in the raw word embeddings (Lu, Chen, & Holyoak, 2012; Lu, Wu, & Holyoak, 2019; Lu, Ichien, & Holyoak, 2022).

We first report an experiment designed to elicit relational luring. Rather than studying word pairs in the context solely of a memory task (Popov et al., 2017), we compared two encoding contexts that were more incidental in nature. One encoding task, involving relatedness judgments, required participants to decide whether or not the two words in a pair were related. Because relatedness judgments do not require identification of any specific relation, they can potentially be made using an implicit relation representation. The second encoding task, verbal analogical reasoning, required participants to decide whether or not an analogy in *A:B :: C:D* format was valid. Evaluating analogies requires attention to the specific relation linking the *A:B* and the *C:D* word pairs, and hence is likely to depend on explicit relation representations (consistent with previous computational modeling; Lu et al., 2019). Each task was followed by a test of recognition memory, with conditions designed to elicit relational luring.

Critically, both the analogy task and the subsequent recognition memory task can be modeled using the same alternative measures of word-pair similarity. Specifically, we compare a measure of *lexical* similarity between individual word meanings, *relational* similarity between explicit relation representations, and a joint measure that combines lexical and relational similarity. Based on previous findings, we predicted that the measure based on relational similarity would prove most effective for the analogy task. The key question is whether recognition memory will be predicted by the same measure of word-pair similarity, or whether a dissociation will be observed between the analogical reasoning and recognition memory tasks.

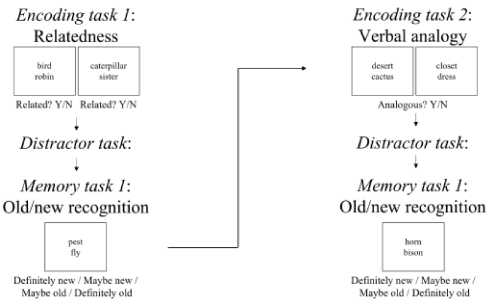

Figure 1. Task structure. Participants completed six tasks, divided into two blocks (columns) of three tasks each. Task order was fixed. The two blocks of tasks were the same except for the encoding task, with assignment of specific word pairs counterbalanced across the two sets.

Experiment

Procedures and analyses were pre-registered on AsPredicted (#66576).

Method

Participants. Participants were 111 undergraduates (M_{age} = 20.12, SD_{age} = 1.94) at either UCLA (n = 93) or at Dartmouth (n = 18; 81 female, 20 male, 1 nonbinary, 9 gender not reported) who completed our tasks online to obtain partial course credit in psychology classes. The study was approved by the Institutional Review Boards at UCLA and at Dartmouth. Participants were self-assessed proficient English speakers, and 82% were native English speakers. We excluded 17 participants whose median correct response time, number of omitted responses, and/or *d'* were 3 standard deviations away from the sample mean on any task (final sample size: 94).

Procedure. All participants completed two blocks, each of which included three tasks. The first task in each block was an incidental encoding task: either relatedness judgments

(first block) or analogical reasoning (second block). The second task in each block was a demanding distractor task involving visuospatial reasoning (a short form of Raven's Progressive Matrices). The third task in each block was a recognition memory task. The assignment of word pairs to each block was counterbalanced across participants. Participants were first shown a list of all the tasks that they would be completing during the experimental session (and thus made aware before starting the experiment that they would be completing memory tasks). The entire test session lasted approximately one hour. Figure 1 presents the sequence of tasks that each participant completed during an experimental session.

Materials and Encoding Tasks. Both encoding tasks involved word pairs that instantiated one of three abstract semantic relations: *category:exemplar* (e.g., *bird:robin*), *part:whole* (e.g., *toe:foot*), and *place:thing* (e.g., *store:groceries*), or else were not semantically related (e.g., *mascara:spoon*). To create the tasks, a total of 200 word pairs were constructed out of 400 unique words. These word pairs were evenly distributed across two 100 word-pair lists. Within each list, 10 unrelated pairs consisted of words with no discernible semantic relation between them. The remaining 90 pairs were evenly distributed across the three abstract semantic relations. Participants saw one list during the relatedness task and the other list during the verbal analogy task; which list was presented during each task was counterbalanced across participants.

Each encoding task consisted of two halves, and each word pair within a given list was presented once during each half. Thus, each half of the relatedness task consisted of 100 trials (with one word pair shown per trial), yielding 200 trials in total. Each half of the verbal analogy task consisted of 50 trials (with two word pairs shown per trial), yielding 100 trials in total. Thus, participants saw each word pair twice across the two halves of each encoding task.

In the relatedness task, participants were presented with a sequence of word pairs and asked to judge whether each pair was comprised of words that were semantically related; this was the case 90% of the time. In the verbal analogy task, participants were sequentially presented with two word pairs on each trial, and were asked to judge whether or not each set constituted a valid analogy; this was the case 54% of the time. Prior to beginning the relatedness task, participants were shown examples of related and unrelated word pairs and then completed seven practice trials. Prior to beginning the verbal analogy task, participants were shown examples of valid and invalid analogies (e.g., *carpenter:hammer* is analogous to *nurse:syringe*, whereas *bowl:cereal* is not analogous to *poverty:money*), and then completed four practice trials. Neither the individual words in the practice trials, nor the relations instantiated by them, overlapped with the word pairs used in the actual encoding tasks. Unlike the relatedness task, the analogy task was expected to require explicit comparison of relations; hence, this task was always delivered after the relatedness task, so as to avoid priming an explicit strategy of identifying abstract relations in the relatedness task.

Recognition Memory Task. Following each encoding task and the intervening distractor task, participants completed a subsequent old/new recognition task, during which they were presented with a sequence of word pairs. Each word pair was constructed out of individual words that participants had seen during their prior encoding task. Participants were asked to identify whether or not they had seen that exact combination of words in the previous encoding task, as well as to rate how confident they were in their judgment using a four-point scale: "Definitely New", "Maybe New", "Maybe Old", and "Definitely Old". The specific word pairs differed across the memory tasks in the two blocks. Participants were given a brief tutorial on the memory task prior to beginning each such task. None of the individual words nor relations instantiated in this tutorial overlapped with those used in the actual task.

A total of 100 word pairs were used for the memory tasks, with each word pair drawn from one of four types. The first type, *intact*, consisted of word pairs that were shown during the relation identification or analogy task. For intact pairs, responses of either "Maybe Old" or "Definitely Old" were scored as correct. The second, third, and fourth types consisted of word pairs that were *not* used in either encoding task; either "Maybe New" or "Definitely New" were scored as correct responses. These three types of word pairs were all constructed by recombining words that had appeared in the immediately prior encoding task, so that individual words were now paired differently, generating novel word pairs distinct from those used in the encoding task. More specifically, *relationally familiar* word pairs consisted of unseen, recombined word pairs instantiating relations to which participants had been exposed during the encoding tasks (i.e., *part:whole*, *category:exemplar*, and *place:thing*). *Relationally unfamiliar* word pairs consisted of unseen, recombined word pairs instantiating a relation type (*similarity*) to which participants had not been exposed. These word pairs included concepts with overlapping salient attributes (e.g., *bartender:cashier*), and hence were relationally similar to one another, but not with respect to any of the three relations included in the encoding tasks. Finally, *unrelated* word pairs consisted of recombined word pairs that were not semantically related in any discernible way.

Based on prior evidence for relational luring (Popov et al., 2017), we hypothesized that participants would false-alarm more often to relationally familiar word pairs than to either relationally unfamiliar or unrelated word pairs.

Experiment Results

Encoding Tasks. Overall, participants performed well on both of the encoding tasks: relatedness task, $\boldsymbol{M_{Acc} = .94}$, $\boldsymbol{SD_{Acc} = .04}$; verbal analogy task, $\boldsymbol{M_{Acc} = .76}$, $\boldsymbol{SD_{Acc} = .11}$. Note that the false alarm rate for unrelated word pairs on the relatedness task was low ($\boldsymbol{M_{FA} = .19}$, $\boldsymbol{SD_{FA} = .18}$), yielding a high d-prime ($\boldsymbol{M_{D'} = 2.77}$, $\boldsymbol{SD_{D'} = .71}$). Thus, even though 90% of the trials involved semantically related word pairs, participants completed the task as instructed, and

did not achieve their high accuracy by simply classifying all word pairs as related.

Recognition Memory. Participants showed good overall performance in recognizing studied word pairs, $M_{Acc} = .80$, $SD_{Acc} = .12$. They correctly recognized intact word pairs as either "Maybe Old" or "Definitely Old" with high accuracy, exhibiting a high hit rate, $M_{Hit} = .88$, $SD_{Hit} = .10$; however, they also sometimes misrecognized recombined word pairs (familiar, unfamiliar, or unrelated), exhibiting a substantial false-alarm rate, $M_{FA} = .25$, $SD_{FA} = .16$.

To test for a relational luring effect, we performed a within-subjects ANOVA on the false alarm data for new pairs with two factors: encoding task (relatedness or verbal analogy) and pair type (familiar, unfamiliar, unrelated). Pair type reliably influenced false alarm rate, $F(2, 186) = 122.21$, $p = < .001$. Planned comparisons revealed that false alarms were more frequent for familiar (.32) than unfamiliar (.22) pairs, and for unfamiliar than unrelated (.10) pairs (both p's < .001). The higher false alarm rate for familiar than unfamiliar pairs reveals a relational luring effect, qualitatively similar to that observed by Popov et al. (2017). The main effect of encoding task was not significant, $F(1, 93) = 0.16$, $p = .69$; nor was the interaction with pair type, $F(2, 186) = 1.96$, $p = .14$.

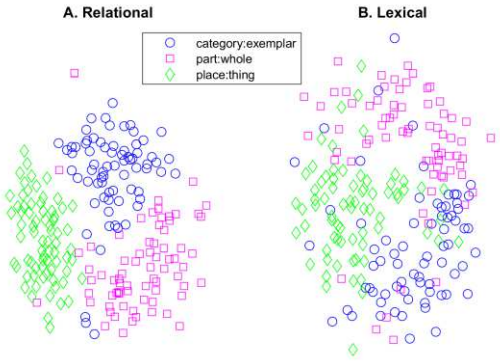

Figure 2. 2-D multidimensional scaling solution of the similarity space derived using relational similarity (Panel A) and lexical similarity (Panel B). Plots show word-pair stimuli instantiating *category:exemplar* (blue circles), *part:whole* (magenta squares), and *place:thing* (green diamonds) relations.

Computational Models

Measures of Word-Pair Similarity

To predict performance on both the analogy task and the recognition memory task, we compared two basic measures of similarity between word pairs: (1) *lexical*: similarity of word pairs computed directly from the similarities of the individual words in each pair; (2) *relational*: similarity of word pairs based on the similarity of the explicit relation between the two words in each individual pair. We also considered the possibility of (3) a *joint* measure that combines both lexical and relational similarity. We implemented specific versions of each of these three possibilities, all rooted in 300-dimensional word embeddings created by Word2vec.

To compute lexical similarity, the meaning of a word pair is represented by a simple aggregate of the semantic vectors of the two individual words. We use f_A to denote the semantic vector for the first word A in a word pair and f_b to denote the semantic vector for the second word B. We compute the distance between word pairs i and j as the mean of the distances between f_{A_i} and f_{A_j} and between f_{B_i} and f_{B_j}:

$$d_{Lex_{ij}} = \frac{cos\left(f_{A_i}, f_{A_j}\right) + cos\left(f_{B_i}, f_{B_j}\right)}{2}. \quad (1)$$

This representation is nonrelational, coding word pairs solely in terms of the meanings of the individual words.

To compute relational similarity, we used relation vectors generated by *Bayesian Analogy with Relational Transformations* (BART; Lu et al., 2012, 2019). BART 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 then uses supervised learning with both positive and negative examples 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. BART uses its pool of 270 learned relations to create a distributed representation of the relation(s) between any two paired words A and B. The posterior probabilities calculated for all learned relations form a 270-dimensional relation vector R_{AB}, in which each dimension codes how likely a word pair instantiates a particular relation. The distance between word pairs i and j is computed as the cosine distance between corresponding relation vectors R_i and R_j :

$$d_{Rel_{ij}} = \cos\left(R_{i,}, R_j\right). \quad (2)$$

Finally, to compute joint similarity, we simply combined lexical and relational representation by taking the unweighted average of the distances generated by each:

$$d_{Joint_{ij}} = \frac{d_{Lex_{ij}} + d_{Rel_{ij}}}{2}. \quad (3)$$

To provide a preliminary sense of how well the two basic measures of word-pair similarity (lexical and relational) capture the categorical distinctions among the three relation types used in the encoding tasks (*category:exemplar*, *part:whole*, and *place:thing*), Figure 2 plots the word pairs used in the experiment on a 2-dimensional projection of the similarity space derived using the two measures. From visual inspection, it is clear that the relational measure (Panel B) separates the three types of pairs into clusters corresponding to semantic categories more clearly than does the lexical measure (Panel A); however, the lexical measure also predicts relation type to some extent.

Modeling Verbal Analogical Reasoning

Performance on the verbal analogy task was modeled directly by the BART model, which in addition to learning relations (as described above), can also be used to predict behavioral (Lu et al., 2019) and neural (Chiang et al., 2021) responses to

analogy problems. In order to predict yes/no decisions about analogy problems, we computed cosine distances between representations of the *A:B* and *C:D* word pairs, and then fit a threshold parameter t such that distances below t indicated a valid analogy and those above t indicated an invalid analogy.

In calculating distance for the purpose of solving analogy problems, we used each of the three similarity metrics described above: lexical, relational, and joint. Based on prior modeling of verbal analogical reasoning (Lu et al., 2019) and of explicit judgments of relation similarity (Ichien, Lu, & Holyoak, 2021), we predicted that the model based on relational similarity would best predict human judgments on the explicit analogy task.

Figure 3 presents the proportion of model and human 'valid' responses broken down by valid analogies (darker bars) and invalid analogies (lighter bars). Overall, BART based on explicit relation similarity achieved the highest accuracy (.75), nearly matching human proportion correct (.76). The alternative model based on lexical (non-relational) similarity performed poorly (.59 correct); this version was overly permissive, detecting valid analogies at a high rate but failing to reject invalid analogies at a similarly high rate. Accuracy for the joint model was intermediate (.65 correct), indicating that incorporating lexical similarity in addition to relational similarity actually impaired model performance on the analogy task.

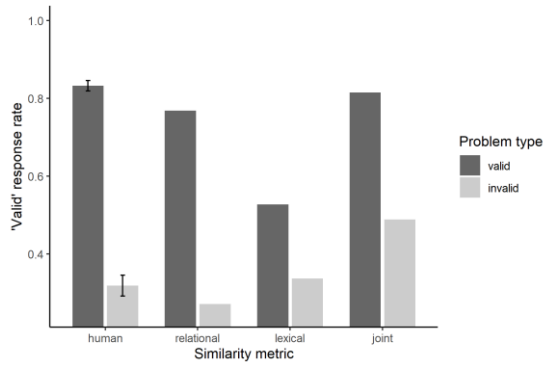

Figure 3. Model and human 'valid' responses on the verbal analogy task. Darker bars represent hits on valid analogies, and lighter bars represent false alarms on invalid analogies. Error bars reflect ±1 standard error of the mean for human responses.

An item-level analysis corroborated these results. We used the *cocor* package in R to test the difference between the extent that each similarity measure correlated with the frequency with which human reasoners judged each analogy as valid (Diedenhofen & Musch, 2015). A Dunn and Clark's (1969) z-test showed that relational similarity was more highly correlated with human responses (r = .47) than were either lexical (r = .21; z = 3.69, p = 2.00 x 10^{-4}) or joint similarity (r = .38; z = 2.04, p = .04). Moreover, because this item-level analysis is based purely on similarity predictions generated with each metric, its results are independent of the decision threshold that was fit to maximize model accuracy in the analogy task. These simulation results thus confirm previous findings indicating that the BART model based on explicit relations outperforms variants based on lexical similarity in tasks involving verbal analogy and explicit judgments of relation similarity (Chiang et al., 2021; Ichien et al., 2021; Lu et al., 2019).

Modeling Recognition Memory

To provide a formal account of relational luring in recognition memory, we adapted an established model of recognition memory, the Generalized Context Model (GCM; Nosofsky, 1988, 1991; Nosofsky & Zaki, 2003). GCM predicts old/new recognition judgments, and is closely related to several other successful cognitive models (e.g., Anderson, 1991; Krushke, 1992; Love, Medin, & Gureckis, 2004). If a version of GCM is able to account for relational luring, we will have demonstrated that this phenomenon is one of many that can be explained within a unified theoretical framework of exemplar-based recognition and categorization.

In the version of GCM implemented here, we assume that recognition of a given word pair on a memory task is based on a comparison of similarities between that word pair and all word pairs presented during a prior encoding task (as described below). The probability with which a participant will classify a word pair i as one they had seen during the encoding task is given by

$$P(old|i) = \frac{F_i}{F_i + k}, \quad (4)$$

where k is a parameter representing a criterion for recognition, and F_i is the familiarity of word pair i which is defined as:

$$F_i = \sum_{j \in J} s_{ij}. \quad (5)$$

Here, J is the set of word pairs shown during the encoding task, and s_{ij} is the similarity between word pair i in the memory task and each word pair j from the encoding task. This similarity follows an exponential decay function (Shepard, 1987) of the psychological distance d_{ij} between word pairs i and j,

$$s_{ij} = e^{-cd_{ij}}, \quad (6)$$

where c is a scaling parameter representing the rate of decline in similarity with psychological distance among word pairs. When GCM is fit to data from individual participants, c is typically interpreted as a measure of a participant's memory sensitivity: i.e., the extent to which they can discriminate between word pairs in memory (Nosofsky, 1988). In the present simulations we fit the model to group-level data, varying the representations for word pairs over which the model operates (details below). In our simulations, c (as it varies across different types of representations) is naturally interpreted as the discriminability between word-pair items within a given representational space. Because our representations are high-dimensional, we adopt cosine distance to compute d_{ij}, rather than the Minkowski power formula typically used in previous work (e.g., Nosofsky, 1988, 1991; Nosofsky & Zaki, 2003).

As the above equations make clear, GCM must be grounded on some measure of similarity between word pairs. We compared the three measures described above (lexical, relational, joint) within the basic GCM framework. Because we found no reliable differences in false alarm rates across the two encoding tasks, we simulated the data obtained by averaging responses across them. Using data for *intact* and *unrelated* word pairs only, we fit the GCM model using each of the three variants of similarity (tuning the criterion and scaling parameters k and c for each) by maximizing the item-wise root mean square deviation (RMSD) between model-generated $P(old|i)$ predictions of the mean frequency with which human participants judged a word pair item to be either "Maybe old" or "Definitely old". Across the three variants, GCM achieved comparable RMSD (where lower RMSD indicates closer fit to human data): lexical: RMSD = .0606; relational: RMSD = .0556; and joint: RMSD = .0584.

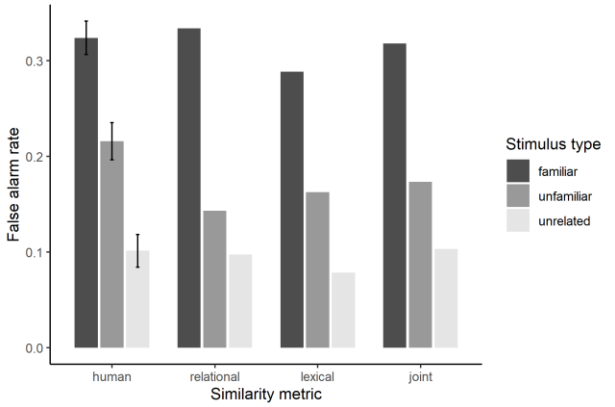

Figure 4. Model and human false-alarm rates on the recognition memory task. Error bars reflect ±1 SEM.

The models were then assessed with respect to their predictions for the critical *relationally familiar* and *relationally unfamiliar* word pairs (not used in parameter estimation). Figure 4 presents false-alarm rates for model-generated $P(old|i)$ predictions and human data, broken down by type of recombined word pairs. Crucially, using each of the alternative similarity calculations, GCM predicts the relational luring effect observed in the human data. We evaluated each variant's ability to account for held-out human data by computing both the Spearman correlation and RMSD between model-generated predictions $P(old|i)$ and the mean frequency of human "old" judgments for *relationally familiar* and *relationally unfamiliar* word pairs. Across the three variants, GCM achieved comparable fits to the human data (where higher ρ indicates closer fit to human data): lexical: RMSD = .1629, ρ = .4623; relational: RMSD = .1588, ρ = .4786; and joint: RMSD = .1535, ρ = .5043.

Given that joint lexical and relational similarity tended to match the human data slightly more accurately (in terms of RMSD) than either lexical or relational similarity alone, we assessed whether each factor may have independently contributed to this overall improvement in model fit. Specifically, we computed semi-partial correlations between the mean frequency of human "old" responses for *familiar* and *unfamiliar* word pairs (thus excluding the *intact* and *unrelated* word pairs used to fit each model), and model-generated $P(old|i)$ predictions based on either lexical or relational similarity, after residualizing the other factor out of the human data. Neither the semi-partial correlation for lexical similarity, $r = .15$, $p = .283$, nor that for relational similarity, $r = .23$, $p = .105$, was reliable. Thus, although we can confidently conclude that the relational luring effect observed in the human data can be fit to a reliable degree using either or both lexical or relational similarity, the evidence from our experiment does not allow us to separate the impact of the two factors.

Discussion

A model based on explicit representations of relations clearly provided the best account of human performance on an analogy task, in accord with previous work (e.g., Chiang et al., 2021; Ichien et al., 2021; Lu et al., 2019). We also replicated the relational luring effect (Popov et al., 2017) in a test of recognition memory, using two alternative encoding tasks. However, computational modeling based on GCM revealed that this luring phenomenon can be predicted using either or both lexical and relational similarity. Relational similarity was more accurate than lexical similarity in clustering word pairs instantiating different categories of semantic relations (see Figure 2); nonetheless, the measure of lexical similarity appears to be crude but "good enough" to reliably predict relational luring. As an instance-based model, GCM effectively computes similarity of any test pair to the entire pool of studied pairs, so even an imperfect measure of word-pair similarity is sensitive to the broad relation types. In contrast, solving a verbal analogy requires fine-grained comparison of one particular word-pair relation (*A:B*) to another (*C:D*), so lexical similarity does not suffice.

Importantly, simulation results reported here are restricted to predictions from models after GCM parameters have been optimized to minimize deviation from human data. Future analyses will examine the extent to which variations in GCM's model parameters impact each similarity metric's ability to reproduce relational luring, thus clarifying *how likely* it is that each of the alternative similarity metrics will reproduce the human phenomenon of relational luring.

In sum, it appears that word embeddings generated by machine learning include implicit information about typical relational roles, so that that in a recognition task, similarity of individual words in pairs can effectively approximate similarity of explicit relations between words. We thus reserve judgment as to whether the phenomenon of relational luring in recognition memory reflects the impact of explicit relational similarity (as previously suggested) and/or lexical similarity.

Acknowledgements

Preparation of this paper was supported by NSF Grants BCS-2022477, 2022357, and 2022369, respectively awarded to S.A.B., D.J.M.K., and K.J.H with H.L.

References

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*, 409-429.

Bhatia, S., & Aka, A. (2022). Cognitive modeling with representations from large-scale digital data. *Current Directions in Psychological Science.* https://doi.org/10.1177/09637214211068113

Challis, B. H., & Sidhu, R. (1993). Dissociative effect of massed repetition on implicit and explicit measures of memory. *Journal of Experiment Psychology: Learning, Memory, & Cognition*, *19*(1), 115-127.

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 of language understanding. In *Proceedings of the 2019 Conference for the North American Chapter of the Association for Computational Linguistics: Human Language Technologies Volume 1*, 4171-4186.

Diedenhofen, B., & Musch, J. (2015). Cocor: A comprehensive solution for the statistical comparison of correlations. *PLoS One*, *10*(4), e0121945.

Dunn, O. J., & Clark, V. A. (1969). Correlation coefficients measured on the same individuals. *Journal of the American Statistical Association*, *64*, 366-377.

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

Goldwater, M. B., Markman, A. B., Stilwell, C. H. (2011). The empirical case for role-governed categories. *Cognition*, *118*, 359-376.

Günther, F, Rinaldi, L, & Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science*, *14*(6), 1006-1033.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, *104*(3), 427-466. https://doi.org/10.1037/0033-295X.104.3.427

Ichien, N., Lu, H., & Holyoak, K. J. (2021). Predicting patterns of similarity among abstract semantic relations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Jones, M., & Love, B. C. (2007). Beyond common features: The role of roles in determining similarity. *Cognitive Psychology*, *55*(3), 196-231.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*, 22-44.

Love, B. C., Medin, D. L., Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*, 309-332.

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

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*(10), 4176-4181.

Markman, A. B., & Stilwell, C. H. (2001). Role-governed categories. *Journal of Experimental & Theoretical Artificial Intelligence*, *13*(4), 329-358.

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

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39-57.

Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 54-65.

Nosofsky, R. M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception and Performance*, *17*(1), 3-27.

Nosofsky, R. M., & Zaki, S. R. (2003). A hybrid-similarity exemplar model for predicting distinctiveness effects in perceptual old-new recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(6), 1194-1209.

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

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.

Reder, L. M., Nhouyvanisvong, A., Schunn, C. D., Ayers, M. S., Angstadt, P., & Hiraki, K. (2000). A mechanistic account of the mirror effect for word frequency: A computational model of remember-know judgments in a continuous recognition paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(2), 294.

Utsumi, A. (2020). Exploring what is encoded in distributional word vectors: A neurobiologically motivated analysis. *Cognitive Science*, *44*, e12844. DOI: 10.1111/cogs.12844