Linford, B., Ichien, N., Holyoak, K. J., & Lu, H. (2022). Impact of semantic representations on analogical mapping with transitive relations. In J. Culbertson, A. Perfors, H. Rabagliati, & V. Ramenzoni (Eds.), Proceedings of the 44th Annual Meeting of the Cognitive Science Society. Austin, TX: Cognitive Science Society.

Impact of Semantic Representations on Analogical Mapping with Transitive Relations

Bryce Linford (linford@g.ucla.edu)
Nicholas Ichien (ichien@ucla.edu)
Keith J. Holyoak (holyoak@lifesci.ucla.edu)
Hongjing Lu (hongjing@ucla.edu)
Department of Psychology, University of California, Los Angeles
Los Angeles, CA 90095 USA

Abstract

Analogy problems involving multiple ordered relations of the same type create mapping ambiguity, requiring some mechanism for relational integration to achieve mapping accuracy. We address the question of whether the integration of ordered relations depends on their logical form alone, or on semantic representations that differ across relation types. We developed a triplet mapping task that provides a basic paradigm to investigate analogical reasoning with simple relational structures. Experimental results showed that mapping performance differed across orderings based on category, linear order, and causal relations, providing evidence that each transitive relation has its own semantic representation. Hence, human analogical mapping of ordered relations does not depend solely on their formal property of transitivity. Instead, human ability to solve mapping problems by integrating relations relies on the semantics of relation representations. We also compared human performance to the performance of several vector-based computational models of analogy. These models performed above chance but fell short of human performance for some relations, highlighting the need for further model development.

Keywords: analogy; mapping; embeddings; transitive inference

Introduction

The solution of verbal analogy problems (e.g., *tool : hammer :: flower : rose*) is a longstanding focus of work in psychology and educational testing (e.g., Sternberg & Nigro, 1980). More recently, computational models that can solve verbal analogies based on representations of word meanings have been developed both in artificial intelligence (AI) (e.g., Mikolov et al., 2017; Turney, 2013) and cognitive science (Lu, Wu, & Holyoak, 2019). A core problem that these computational models must address is the *eduction of relations* (Spearman, 1923): retrieving or computing the unstated semantic relation between the two words in each pair (e.g., the relation between the source pair *tool* and *hammer*, and that between the target pair *flower* and *rose*). A general solution is to make use of vector representations (*embeddings*) that capture important aspects of the meanings of individual words, generated by machine learning models such as Word2vec (Mikolov et al., 2017), which are trained on large text corpora. The relation between any two words can then be educed either by the generic operation of computing the difference vector between the paired words, or by additional learning mechanisms that enable generation of explicit representations of relations as vectors in a transformed relation space (Lu et al., 2019; Ichien, Lu, & Holyoak, 2022). Once relation vectors have been created, an analogy can be evaluated by assessing the similarity of the relation vectors for the source and target pairs (e.g., by computing cosine similarity).

Solving verbal analogies presented in the form *A:B::C:D* does not require mapping of individual concepts, because the format itself specifies clear correspondences (*A*→*C*, *B*→*D*). In order to extend vector-based computational models of analogy to more complex problems in which each analog involves multiple relations between more than two concepts (necessitating a mapping process), the models must be augmented with some mechanism to integrate multiple relations so as to identify the optimal mappings between concepts in source and target analogs. One approach is to organize vector representations of both concepts and the relations between them into *attributed graphs*, in which concepts correspond to nodes and relations to edges (Lu et al., 2022). Given a pair of attributed graphs, a probabilistic graph matching algorithm can then be applied to identify the optimal mappings between source and target concepts by maximizing graph similarity under a soft isomorphism constraint.

Lu et al. (2022) introduced a paradigm for testing the ability of both humans and computational models to find

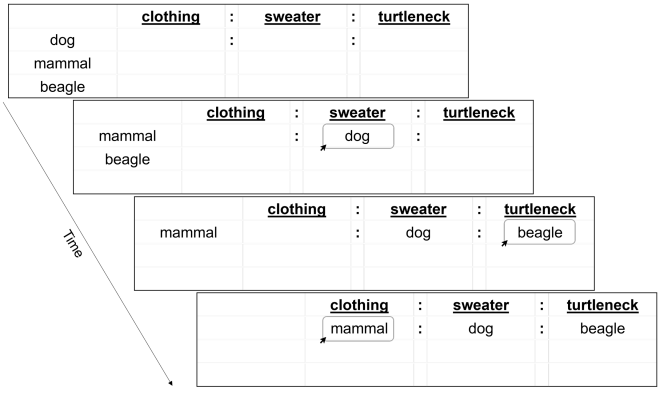

Figure 1: Time-course of an example category triplet problem.

mappings between words in analogy problems minimally more complex than the standard *A:B::C:D* format. Rather than each analog consisting of a single word pair, analogs are *triplets* composed of three words (see Figure 1). One type of problem involved *category* triplets, in which the source was an ordered set of category names (e.g., *clothing : sweater : turtleneck*), and the target consisted of three scrambled words (e.g., *dog*, *animal*, *beagle*) that could also form an ordered set of categories. For each problem participants were asked to create a valid analogy by using their mouse to drag each of the randomly ordered target terms under one of the terms in the ordered source triplet.

The triplet mapping problem provides a basic paradigm for investigating analogical reasoning using simple relational structures. When the source and target analogs involve multiple pairwise relations of the same type, as in category triplets, inherent mapping ambiguities arise. For example, *animal : dog* considered alone could map to either *clothing : sweater* or *sweater* : *turtleneck*, because all of these pairs instantiate the *superordinate-of* relation. Lu et al. (2022) found that people were able to reliably solve such triplet problems; a comparable requirement to integrate multiple relations arises in many other relational reasoning paradigms, such as transitive inference (Andrews & Halford, 1998; Waltz et al., 1999). To resolve ambiguity in local mappings, a reliable analogy model must assess relation similarities and also integrate across relations based on mapping constraints.

Category relations are one of several general types of semantic relations that exhibit the logical property of transitivity (i.e., for relation *r*, *A r B* and *B r C* jointly imply *A r C*). For any transitive relation, it is possible to form triplet mapping problems, the solution of which requires both eduction of relations between pairs of concepts and integration of multiple relations. An important question is whether the solution to mapping problems based on transitive relations depends solely on their logical form, or on the semantic representations of different relations. If the logical form of structures directly determines analogical mapping (as predicted, for example, by structure-mapping theory; Gentner, 1983), we would expect constant mapping performance regardless of semantic relations. In contrast, if mapping performance varies across different transitive relations, this would suggest that the semantics of relations plays an important role in analogical mapping and reasoning.

Here we compare human performance on triplet problems involving three types of transitive relations: *category* (e.g., *bird : parrot : parakeet*), *linear order* (e.g., *pebble : rock : boulder*), and *causal* (e.g., *lightning : fire : smoke*). All of these relations constitute formal structures based on transitive relations. According to a taxonomy of forms proposed by Kemp and Tenenbaum (2008), for categories, the ordering is part of a hierarchy; for linear orders, the relation is itself an ordering; for causal relations, the ordering is a chain within a causal network (Waldmann, 2017).

If mapping of ordered relations depends solely on their formal property of transitivity, then the three relation types would yield mapping problems of approximately the same difficulty. On the other hand, if each type of transitive relation has its own semantic representation (as vector-based models of analogy assume), then mapping difficulty may vary across types. To explore this issue, we performed an experiment to determine how well people are able to solve triplet mapping problems based on the three types of transitive relations. In addition, we also compared human performance with several recent models of mapping based on vector representations of word embeddings and relations.

Experiment: Mapping Triplets Based on Transitive Relations

Method

Participants A total of 561 participants (M_{age} = 40.85, SD_{age} = 12.44, 288 female, 265 male, 6 gender non-binary, 2 gender withheld; located in the United States, United, Kingdom, Ireland, South Africa, New Zealand, Canada, and Australia) were recruited via Amazon Mechanical Turk and received a payment of $1. Of these, 27 participants reported not paying attention while completing the task and were therefore excluded from analyses, resulting in a final sample of 534. The study was approved by the Office of the Human Research Protection Program at the University of California, Los Angeles, and participants provided informed consent. The study was pre-registered online on AsPredicted and can be accessed at: https://aspredicted.org/B2M_28Y.

Materials and Procedure Each participant completed three verbal analogy problems, each based on pairs of triplets (three words) of one of three types. The three triplet types instantiated three classes of semantic relations, each formally transitive: category member, linear order, and cause-effect. The triplets were primarily based on norms of word pairs instantiating the three relations, reported by Jurgens, Mohammed, Turney and Holyoak (2012); some causal word pairs were drawn from stimuli used in a study by Fenker, Waldmann, and Holyoak (2005).

By presenting each participant with just one problem of

Table 1: Examples of Triplets used in Experiment

Relation type	Triplet examples
Category	clothing: sweater: turtleneck weapon: gun: rifle reptile: lizard: iguana
Linear order	second: minute: hour past: present: future penny: nickel: quarter
Causal	exercise: fitness: health nuts: allergy: rash salt: thirst: drink

each type, we minimized any opportunity to learn the general structure of the problems (as our focus was on initial analogical mapping, rather than schema induction). For each problem, an ordered set of three terms (e.g., *clothing : sweater : turtleneck*) appeared in a fixed position on the top of the screen, and a set of three randomly ordered terms (e.g., *dog*, *mammal*, *beagle*) appeared on the left (see Figure 1). Participants were instructed to create a valid analogy by clicking and dragging each of the randomly-ordered terms to a box below the corresponding fixed term. Examples of each type are provided in Table 1. Each problem was formed using two triplets randomly drawn from a pool of eight, and were shown in either order (56 possible pairs for each triplet type). The presentation order of the three triplet problems was counterbalanced across participants.

Before working on the three experimental problems, participants read instructions that explained the triplet analogy task using two examples, each involving different relations than the experimental problems. The triplets in the first example were *barber* : *scissors* : *hair* and *bandage* : *nurse* : *wound*, and the triplets of the second example were *finger* : *hand* : *arm* and *leaf* : *branch* : *tree*. The instructions stated that an analogy is valid if the relations among the terms in the two triplet sets match each other. Participants needed to complete the second example correctly in order to begin the experimental problems.

Results

Human Performance Mapping responses were first coded as correct only if *all three words* were mapped correctly in a problem. As there are six possible orderings of three items, chance-level performance would be 0.17. Mean mapping accuracy of the participants was 0.69 for category triplets, 0.77 for linear order triplets, and .48 for causal triplets. A one-way repeated measures ANOVA, with triplet type (category, linear order, causal) as a within-subjects factor, revealed a significant main effect of semantic relation on mapping accuracy, $F(2,1066) = 68.387$, $p < .001$. Using a Bonferroni correction for multiple comparisons, mapping accuracy was reliably higher for linear order triplets than for category ($p = .003$) or causal triplets ($p < .001$), and accuracy was higher for category triplets than causal triplets ($p < .001$).

We also analyzed mapping accuracy for each of the three individual role positions within each triplet problem. Role-based mapping accuracy was coded as 1 if the correct target word was mapped to its corresponding source word, scored separately for each of the three words in the target triplet. The means are shown in Figure 2. We conducted a two-way ANOVA on mapping accuracy for each role, with triplet type and role position (word 1, 2, and 3) as within-subject factors. Mauchly's test indicated a violation of the sphericity assumption, $\chi^2(9) = 85.949$, $p < .001$. Given a violation of sphericity ($\varepsilon = 9.27$), we report Huyn-Feldt corrected results. This analysis revealed significant main effects of triplet type, $F(1.97, 1051.035) = 70.00$, $p < .001$, and role position, $F(1.94, 1034.16) = 10.40$, $p < .001$, as well as a significant

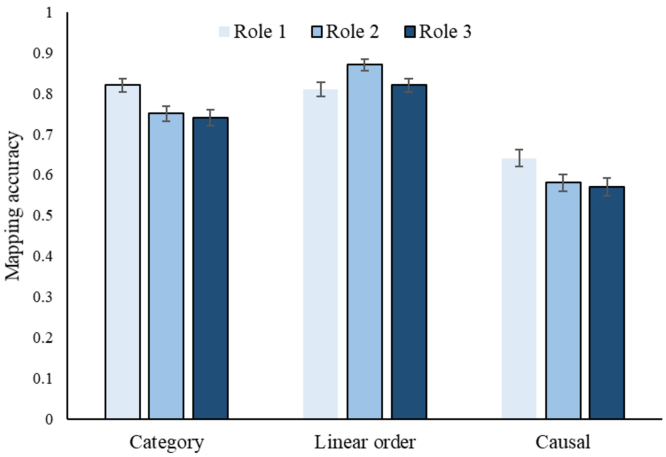

Figure 2: Mean mapping accuracy for words in each of three roles, by triplet type. Error bars reflect ± 1 SEM.

interaction, $F(3.738, 1992.29) = 8.086$, $p < .001$. These results indicate that specific semantic relations affect not only overall mapping accuracy, but also accuracy for individual roles in transitive triplets.

To further examine the impact of semantic relations on mapping accuracy for individual roles, we conducted nine pairwise comparisons between role positions within each triplet type, using a Bonferroni correction for multiple comparisons. For category triplets, accuracy was reliably higher for the first role than the second ($p < .01$) or third ($p < .001$), with no significant difference between the second and third roles. For linear order triplets, accuracy for the second role was reliably higher than for the first ($p = .001$) or third role ($p = .006$), with no reliable difference between the first and third roles. For causal triplets, accuracy was reliably higher for the first role than for the second ($p = .016$) or third ($p = .009$), with no significant difference between the second and third. Thus for category and causal triplets, accuracy was highest for the first word; whereas for linear order triplets, accuracy was highest for the middle word.

Mapping Semantic Relations with Vector-Based Computational Models

We implemented several vector-based models that are capable of computing the semantic relation between any two words, and then integrating multiple relations to identify the optimal mapping between analogs. Each model simulates mapping performance on each of the 56 triplet problems used in the human experiment. For the present simulations, mappings were considered correct only if all three entities in the target were correctly mapped to the source (chance performance = 0.17).

We tested models based on four different methods for creating vector representations of semantic relations. These methods were: two versions based on sentence embeddings generated by a recently-developed model for natural language processing (NLP), *Bidirectional Encoder Representations from Transformers* (BERT) (Devlin et al., 2019); a version based on an earlier NLP method to create

word embeddings, Word2vec (Mikolov et al., 2013; Zhila et al., 2013), and vector representations of word-pair relations generated by a model of relation learning, *Bayesian Analogy with Relational Transformations* (BART) (Lu et al., 2019). Each of these four sets of relation embeddings was used with an exhaustive algorithm for finding the optimal mapping between two triplets. In addition, two of the sets of relation embeddings (based on Word2vec and BART) were also coupled with an algorithm for *Probabilistic Analogical Mapping* (PAM) (Lu et al., 2022), which is more computationally efficient than the exhaustive algorithm. Thus, a total of six computational models were implemented and used to simulate human performance.

In exhaustive mapping, for each problem all alternative mappings are considered between an ordered source triplet (e.g., *tool : ax : hatchet*) and each of the six possible orderings for the entities in a target triplet (e.g., *bird : parakeet : parrot, parrot : bird : parakeet*, etc.). All representations are derived from word embeddings: high-dimensional vector representations of individual word meanings computed from hidden layers of activation in Natural Language Processing (NLP) models (implemented as artificial neural networks) that have been trained to predict word and/or sentence sequences within vast text corpora. For all models based on exhaustive mapping, the predicted correct mapping is obtained by selecting the one of the six possible mappings that minimizes cosine distance.

BERT BERT is an NLP model that takes full sentences as input and is equipped with a transformer block, which enables the model to generate embeddings of individual words in input sentences that are context-dependent: sensitive to both the identity and order of other words used in that sentence (Devlin et al., 2019). Although it represents verbal input as unstructured vectors of activation, BERT embeddings have been used to recover structural properties of sentences that approximate those posited by theoretical linguists (Manning et al., 2020). In the present simulations, we examined the extent that such representations could be used to find correspondences across instances of transitive relations.

We acquired sentence embeddings from BERT through the *Transformer Model for MATLAB* toolbox[1], using the 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). In order to represent each ordering of a given triplet, we used each of two methods. The first employed a *generic sentence* across all three triplet types, in which words representing each entity within a triplet were embedded in the following structure: "A is a related to B, which is related to C." Within this skeletal sentence, we replaced the first word in an ordered triplet with A, the second word with B, and the third word with C (e.g., the ordering *tool : ax : hatchet* yielded "Tool is related to ax, which is related to weapon").

The second method for obtaining embeddings from BERT employed a *specific sentence* for each triplet type, specifying the particular semantic relation instantiated by that triplet: For category triplets: "A is a category of B, which is a category of C;" for linear order triplets: "A goes before B, which goes before C;" and for causal triplets: "A causes B, which causes C."

In order to examine BERT's performance on analogy triplet problems, we adopted two methods for extracting representations of generic and specific sentences, spanning the source analog and the 6 different orders of the target analog for each problem. Using the first method, we computed the mean of the individual word embeddings constituting each input sentence to generate a unified sentence embedding. Using the second method, we simply extracted the embedding for the [CLS] classification token for each input sentence. Because the first method outperformed the second, we report results using the first method.

Word2vec-diff In contrast to context-dependent word embeddings created by BERT, static word embeddings generated from earlier language models like Word2vec (Mikolov et al., 2013) represent individual word meanings using single vectors, regardless of their context of use. In order to compute representations of pairwise relations between words from Word2vec embeddings, we took a generic operation: the vector difference (Word2vec-diff) between words in each pair. This difference-vector approach to representing relations between individual words has been used to solve four-term analogy problems relating similar pairs of concepts (Zhila et al., 2013; but see Peterson, Chen, & Griffiths, 2020, for evidence of limitations). In order to represent the relations instantiated in a triplet *A:B:C*, we concatenated vector differences between vectors representing *A* and *B* as $\boldsymbol{f_A - f_B}$, *B* and *C* as $\boldsymbol{f_B - f_C}$, and *A* and *C* as $\boldsymbol{f_A - f_C}$, for source triplets as $\boldsymbol{S = [f_A - f_B, f_B - f_C, f_A - f_C]}$. Similar operations are used for the target triplet**.**

BART BART uses supervised learning to acquire explicit representations of semantic relations (e.g., *X is a part of Y*) and the individual roles that constitute them (e.g., *part* and *whole*) from unstructured vector representations of individual word meanings (Lu et al., 2019, 2022). For the present simulations, BART was trained using Word2vec word embeddings for word pairs that instantiate a set of relations. The learning model acquires weight distributions over selected feature dimensions of input word vectors. These weight distributions are used to predict the posterior probability that a word pair instantiates a particular relation,

After relation learning, BART has acquired role-based weight distributions that are diagnostic of individual words serving the first role of a given relation (e.g., *part* in the relation *X is a part of Y*), which constitute explicit

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

representations of those relational roles. To do so, BART reapplies Bayesian logistic regression to the element-wise product of prior-learned relation weight distributions and vectors representing the first word of training example word pairs. BART's learning culminates in explicit representations of both full semantic relations and the individual roles that constitute them.

In order to then represent the relation between any pair of words *A:B*, BART applies its learned relation weight distributions to generate a relation vector Rel_{AB} in which each element represents the posterior probability of the word pair instantiating each of learned relations: $Rel_{AB} = \langle P(Rel_1 = 1|f_A, f_B), \ldots P(Rel_k = 1|f_A, f_B)\rangle$.

Ichien et al. (2022) found that applying a power transformation to BART's relation vectors, raising the value along each dimension to a power of 5 (i.e., "winners take most") improves their ability to predict human judgments of relational similarity. We applied that power transformation to relation vectors in the present simulations.

BART uses its learned role weight distributions to generate a role vector $Role_A$ populated by posterior probabilities representing the extent that the first word f_A in a given pair of word vectors f_A and f_B instantiates the corresponding learned role:
$Role_A = \langle P(Role_1 = 1|f_A, f_B), \ldots, P(Role_k = 1|f_A, f_B)\rangle$.
In order to represent the full relational meaning of a given word pair R_{AB}, we concatenated Rel_{AB} and $Role_A$ to form the relation representation $R_{AB} = [Rel_{AB}, Role_A]$.

In the present simulations, we combined two datasets of human-generated word pairs to train BART. The first dataset (Jurgens et al., 2012) consists of at least 20 word pairs (e.g., *engine : car*) instantiating each of 79 semantic relations (e.g., *X is a part of Y*). 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 weight distributions for 135 semantic relations. Since BART's learned relation weights can be expressed as two separate halves (i.e., those associated with the first relational role and those associated with the second relational role), 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 category for Y*, BART can also form a representation of its converse, *Y is a member of category X*, effectively doubling its pool of learned relations from 135 to 270 in total.

Exhaustive Mapping

Each of the four sets of relations embeddings described above was paired with a mapping algorithm that performs an exhaustive search, comparing an ordered source triplet to all six possible orders of a target triplet. This exhaustive mapping algorithm selects mappings based on which ordering of the target $\hat{\boldsymbol{T}}$ maximizes its overall similarity with the ordered source $\boldsymbol{S}$:

$$\hat{T} = \underset{T\in\{T_1,T_2,T_3,T_4,T_5,T_6\}}{\operatorname{argmax}} \; 1 - cos\,(S,T) \qquad (1)$$

Probabilistic Analogical Mapping (PAM)

The second mapping algorithm used in our simulations implements a graph-matching procedure that maximizes the similarity between two *semantic relation networks*, respectively representing the source and target analogs. Formally, semantic relation networks are attributed graphs in which each node $\boldsymbol{N}$ and each edge $\boldsymbol{E}$ is assigned attribute embeddings $\boldsymbol{A}$. Within semantic relation networks, nodes are word embeddings for individual concepts and edges are semantic relation vectors between words. $\boldsymbol{A_{ii}}$ represents the semantic attribute of the $\boldsymbol{i}$th concept, and $\boldsymbol{A_{ij}}$ indicates the relation attribute of the edge between the $\boldsymbol{i}$th concept and $\boldsymbol{j}$th concept. For the present simulations with PAM, we always use Word2vec word embeddings for semantic attribute $\boldsymbol{A_{ii}}$ for the nodes in the attributed graph. In one of two versions, for edge attributes $\boldsymbol{A_{ij}}$ we use Word2vec-diff vectors, $\boldsymbol{f_i - f_j}$; in the other version, we use BART vectors $\boldsymbol{R_{ij}}$.

We represent the source and target analogs as graphs $\boldsymbol{g}$ and $\boldsymbol{g'}$ with concept indices $\boldsymbol{i}$, $\boldsymbol{j}$, and $\boldsymbol{i'}$, $\boldsymbol{j'}$, respectively. $\boldsymbol{M_{ii'} = 1}$ if the $\boldsymbol{i}$th concept node in the source analog maps to the $\boldsymbol{i'}$th concept node in the target analog. The goal of the model is to estimate the probabilistic mapping matrix $\boldsymbol{m}$, which consists of elements denoting the probability that the $\boldsymbol{i}$th node in the source analog maps to the $\boldsymbol{i'}$th node in the target analog, $\boldsymbol{m_{ii'} = P(M_{ii'} = 1)}$. PAM adopts a Bayesian approach to infer a mapping $\boldsymbol{m}$ between concepts in the source and target analogs that maximize its posterior probability:

$$P(m|g, g') \propto P(g, g'|m)P(m),$$

with the constraints

$$\forall_i \textstyle\sum_{i'} m_{ii'} = 1, \forall_{i'} \sum_i m_{ii'} = 1 \qquad (2)$$

The likelihood term $P(g, g'|m)$ uses mapping probabilities as weights to compute likelihood probabilities based on a weighted sum of the semantic similarity between mapped concepts and of the relation similarity between mapped relations. The prior term favors isomorphism, with one-to-one correspondence in graph matching.

To implement the inference in Equation 2, we employ a graduated assignment algorithm (Gold & Rangarajan, 1996) similar to those previously used in matching problems in computer vision (Lu & Yuille, 2005; Menke & Yang, 2020). The algorithm incorporates soft assignments in graph matching, allowing probabilistic mapping values that lie in the continuous range [0,1] rather than requiring deterministic one-to-one mapping values.

Comparisons between Model Predictions and Human Performance

Figure 3 presents mapping accuracy of humans and each of the six computational models for each triplet type. For category triplets, BART with exhaustive search (.75) and with the PAM mapping algorithm (.71) achieved human-level performance (.69). All the other models showed much

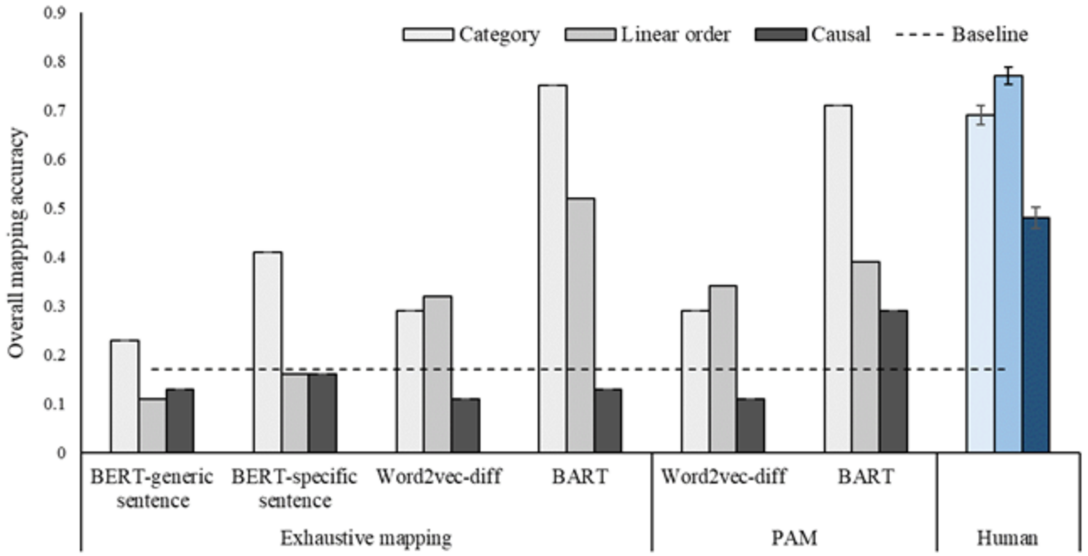

Figure 3: Overall mapping accuracy for models (grey bars) and human reasoners (blue bars) for category (light shade), linear order (middle shade), and causal (dark shade) triplet problems. For models, upper x-axis labels refer to alternative relation representations, and lower x-axis labels refer to alternative mapping algorithms. Dotted line marks chance performance (.17). Errors bars reflect ± 1 SEM.

worse mapping accuracy for category triplet problems (between .23 and .41). For humans, accuracy on linear order triplets was the highest among the three triplet types (.77); however, all models performed poorly on linear order problems. The highest accuracy on linear order triplets was achieved by exhaustive BART (.52) followed by BART coupled with PAM (.39). The Word2vec-diff models reached accuracy around 0.3, and the BERT models showed chance-level performance. For causal triplets, human performance was much lower than for either of the other two types (.48). The models performed even worse, with only BART coupled with PAM achieving above-chance accuracy (.29).

Discussion

Our results show that human performance on mapping problems involving transitive relations differs substantially between different semantic relations: most accurate for linear order relations, followed by category relations, and least accurate for causal relations. These systematic differences among semantic relation types imply that each type of transitive relation has its own semantic representation, and that mapping is influenced by these semantic representations, rather than being based solely on the formal property of transitivity.

One possible explanation for the experimental results is that people have prior schematic knowledge about linear orderings based on magnitude, and such existing schemas are not as easily retrievable for category and causal problems. Future research could explore how people might improve at these problems by learning schemas for the semantic relations (e.g., by completing multiple problems; Gick & Holyoak, 1983).

The differences in mapping performance across relation types also provide insights into how humans represent and map each type of semantic relation in analogical reasoning. In particular, the three types varied in accuracy across the three role positions. For category problems, the first word was mapped most accurately, replicating the pattern reported by Lu et al. (2022). This finding suggests that the most abstract category (superordinate) is the most distinctive of the three. For causal triplets, accuracy was also highest for the first role, consistent with evidence that the root cause in a causal chain is most distinctive (Ahn, Kim, Lassaline, & Dennis, 2000). In contrast, for linear order triplets the middle role was most accurate. This pattern implies that the most common error was a reversal of the order between the source and target (i.e., the first and third roles were reversed, while the middle role was correct because it remains the same regardless of the direction of the ordering).

Vector-based models of relation representations are capable of educing the relation between word pairs; and when coupled with a mapping algorithm, such models can in principle compute mappings that require integration of multiple relations, as is required for our triplet analogies. However, none of the six specific models we implemented proved particularly impressive in capturing the pattern of human performance for all relations examined in the study. It is possible that humans adopt different representation formats for different types of relation representations. For example, a linear ordering could be identified by projecting word vectors onto a magnitude dimension in a semantic space (Grand et al., 2022). Causal relations may be represented using special integration functions (Yuille & Lu, 2007) and learned through interventions. Hence, our comparison of model and human performance highlights the need to develop more sophisticated relation representations (beyond vector-based models) that can support analogical reasoning.

Acknowledgements

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

References

Ahn, W., Kim, N. S., Lassaline, M. E., & Dennis, M. J. (2000). Causal status as a determinant of feature centrality, *Cognitive Psychology*, *41*, 1-55.

Andrews, G., & Halford, G. S. (1998). Children's ability to make transitive inferences: The importance of premise integration and structural complexity. *Cognitive Development, 13*(4), 479–513.

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.

Fenker, D. B., Waldmann, M. R., & Holyoak, K. J. (2005). Accessing causal relations in semantic memory. *Memory & Cognition*, *33*, 1036-1046.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*, 155–170.

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

Gold, S., & Rangarajan, A. (1996). A graduated assignment algorithm for graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *18*(4), 377-388. https://doi.org/10.1109/34.491619

Grand, G., Blank, I. A., Pereira, F., & Fedorenko, E. (2022). Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature Human Behaviou*r. doi: 10.1038/s41562-022-01316-8.

Ichien, N., Lu, H., & Holyoak, K. J. (2022). Predicting patterns of similarity among abstract semantic relations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(1), 108-121.

Jurgens, D. Mohammed, S., Turney, P., & Holyoak, K. J. (2012). SemEval-2012 Task 2: Measuring degrees of relational similarity. *Proceedings of the First Joint Converence on Lexical and Computational Semantics (*SEM)*, 356-364.

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences, USA*, *105*, 10687-10692.

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. (2019). Emergence of analogy from relation learning. *Proceedings of the National Academy of Sciences, USA*, *116*, 4176-4181.

Lu, H., & Yuille, A. (2005). Ideal observers for detecting motion: Correspondence noise. In Y. Weiss, B. Scholkopf, & J. Platt, *Advances in Neural Information Processing Systems*, *18*, 827-834.

Manning, C. D., Clask, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences, USA*, *117*(48), 30046-30054.

Menke, J., & Yang, A. U. (2020). Graduated assignment graph matching for realtime matching of image wireframes. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 5909-5916.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2017). Advances in pre-training distributed word representations. *arXiv preprint: 1712.09405*.

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.

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

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.

Spearman, C. (1923). *The nature of intelligence and the principles of cognition.* London: Macmillan.

Sternberg, R.J., & Nigro, G.N. (1980). Developmental patterns in the solution of verbal analogies. *Child Development, 51*, 27-38.

Turney, P. D. (2013). Distributional semantics beyond words: Supervised learning of analogy and paraphrase. *Transactions of the Association for Computational Linguistics*, *1*, 353–366.

Waldmann, M. R. (2017). Causal reasoning: An introduction. In M. R. Waldmann (Ed.), *Oxford handbook of causal reasoning* (pp. 1–9). New York: Oxford University Press.

Waltz, J. A., Knowlton, B. J., Holyoak, K. J., Boone, K. B., Mishkin, F. S., de Menezes Santos, M., Thomas, C. R., & Miller, B. L. (1999). A system for relational reasoning in human prefrontal cortex. *Psychological Science*, *10*, 119-125.

Yuille, A. L., & Lu, H. (2007). The noisy-logical distribution and its application to causal inference. In *Advances in Neural Information Processing Systems*, *20*.

Zhila, A., Yih, W. -t., 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).

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.