Snefjella, B., Ichien, N., Holyoak, K. J., & Lu, H. (2022). Predicting human judgments of relational similarity: comparison of computational models based on vector representations of meaning. 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.

Predicting Human Judgments of Relational Similarity: A Comparison of Computational Models Based on Vector Representations of Meaning

Bryor Snefjella (bsnefjella@psych.ucla.edu)

Nicholas Ichien (ichien@g.ucla.edu)

Keith J. Holyoak (holyoak@psych.ucla.edu)

Hongjing Lu (hongjing@ucla.edu)
Department of Psychology, University of California Los Angeles

Abstract

Computational models of verbal analogy and relational similarity judgments can employ different types of vector representations of word meanings (embeddings) generated by machine-learning algorithms. An important question is whether human-like relational processing depends on explicit representations of relations (i.e., representations separable from those of the concepts being related), or whether implicit relation representations suffice. Earlier machine-learning models produced static embeddings for individual words, identical across all contexts. However, more recent Large Language Models (LLMs), which use transformer architectures applied to much larger training corpora, are able to produce contextualized embeddings that have the potential to capture implicit knowledge of semantic relations. Here we compare multiple models based on different types of embeddings to human data concerning judgments of relational similarity and solutions of verbal analogy problems. For two datasets, a model that learns explicit representations of relations, Bayesian Analogy with Relational Transformations (BART), captured human performance more successfully than either a model using static embeddings (Word2vec) or models using contextualized embeddings created by LLMs (BERT, RoBERTa, and GPT-2). These findings support the proposal that human thinking depends on representations that separate relations from the concepts they relate.

Keywords: analogy; transformers; embeddings; relation learning

Introduction

A core property of human thinking is that the mental representation of a relation is separable from the representations of the concepts it relates (Hummel & Holyoak, 1997). That is, for humans a relation representation is at some cognitive level *explicit*, so that it can itself serve as an input to mental processes. In particular, judgments of similarity show dissociations between the contributions of entity-based and relational similarity. Relational similarity tends to be more potent when overall relational similarity across analogs is relatively high (Goldstone, Medin, & Gentner, 1991), when the objects in visual analogs are sparse rather than rich (Markman & Gentner, 1993), and for older as compared to younger children (Gentner & Rattermann, 1991). When reasoning by analogy, human adults can sometimes identify correspondences between situations based primarily on similar relations, even when the entities involved are very dissimilar (Gick & Holyoak, 1980). Explicit relation representations provide the basis for flexible generalization to novel instantiations of relational patterns (e.g., Doumas, Puebla, Martin, & Hummel, 2022).

An important goal for cognitive science is to characterize the representation of relations, and in particular to show how explicit relation representations could be acquired from non-relational inputs, while avoiding hand-coding of the inputs (Lu, Chen, & Holyoak, 2012). In recent years, advances in machine learning have enabled the generation of high-dimensional vectors of continuous-valued features, termed *embeddings*, which can be interpreted as representations of word meanings (for a general overview see Günther, Rinaldi, & Marelli, 2019). Embeddings correspond to activation states in the hidden layer of a neural network that has been trained to predict patterns of words that co-occur in large text corpora. Notable early embedding models include Word2vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) and GloVe (Pennington, Socher, & Manning, 2014), both of which have been used to model human judgments based on similarity between words (Bhatia, Richie, & Zou, 2019). Embedding models such as Word2vec have also been used to solve four-term verbal analogies by computing the cosine distance between difference vectors for *A:B* and *C:D* pairs (Zhila, Yih, Meek, Zweig, & Mikolov, 2013). A difference vector provides an implicit representation of the specific relation between two particular concepts. But although Word2vec achieved some success for analogies based on semantically-close concepts, it fails to reliably solve problems based on more dissimilar concepts (Linzen, 2016; Peterson, Chen, & Griffiths, 2020).

An alternative approach, developed in a model termed *Bayesian Analogy with Relational Transformations* (BART) (Lu et al., 2012; Lu, Wu, & Holyoak, 2019), is to use embeddings of individual words as inputs to a learning mechanism that yields explicit representations of relations. BART operates on word embeddings, taking Word2vec embeddings for pairs of individual words as inputs. From Word2vec embeddings, BART learns dimensions of disentangled relation vectors in a transformed space. As illustrated in Figure 1, BART effectively re-represents the relation between two specific concepts as a vector in a new semantic space (for a re-

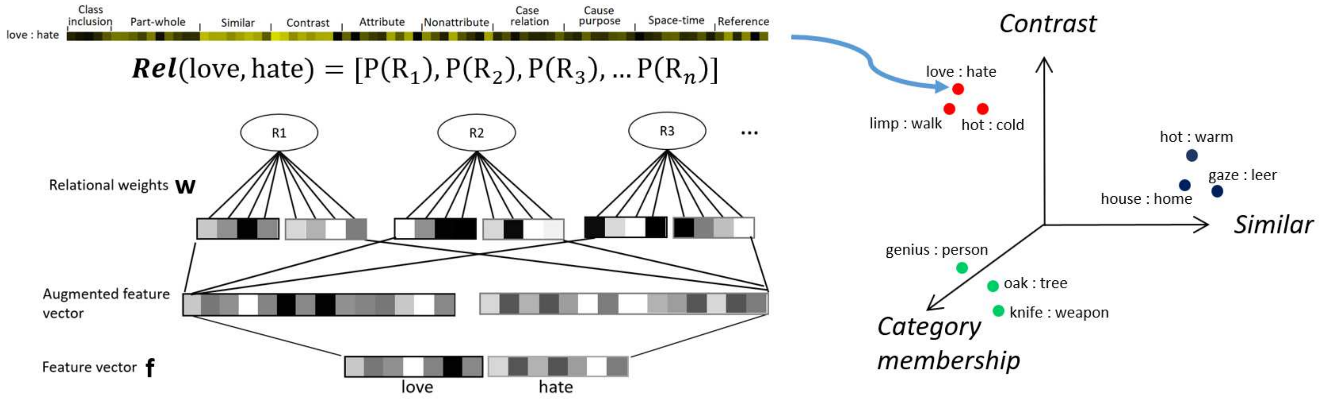

Figure 1: Learning Explicit Relation Vectors from Embeddings
Left: Schematic illustration of BART model architecture for relation representation. The bottom layer of the BART model is a concatenated input vector based on the two words in a pair; the top layer indicates the set of learned relations (ellipses indicate additional relations beyond the three illustrated here). After learning, the semantic relation between any two words is represented as a vector of the posterior probabilities of each learned relation; the relation vector (Rel) linking love and hate is shown on the top as an illustration. Right: A schematic illustration of semantic relations formed by BART to generate a transformed (and disentangled) space in which pairs instantiating similar sets of relations tend to show similar patterns in relation vectors, and hence are located close to one another in the relation space.

lated approach see Roads & Love, 2020). The dimensions in BART's relation vectors are meaningful semantic relations that have been identified in classic psychometric and psycholinguistic research (Bejar, Chaffin, & Embretson, 1991; Chaffin, 1989), including major classes such as *class inclusion* (*tree : oak*), *part-whole* (*hand : finger*), *similar* (*road : highway*), *contrast* (*hot : cold*), *case relation* (*read : book*), and *cause-purpose* (*joke : laughter*). Each element in BART's relation vector corresponds to the posterior probability that a particular meaningful relation holds between the concepts. These distributed (but disentangled) representations enable the model to generalize to new word pairs that may be linked by relations on which the model had not been specifically trained. By comparing the similarity between relation vectors (assessed by cosine distance), semantic relation representations derived by BART have been used to solve verbal analogies in *A:B :: C:D* format (Lu et al., 2019), to predict human judgments of relation typicality and similarity (Ichien, Lu, & Holyoak, 2021), and to predict patterns of similarity in neural responses to relations during analogical reasoning (Chiang, Peng, Lu, Holyoak, & Monti, 2021). In each case BART's performance exceeds that of the baseline Word2vec model, supporting the importance of incorporating explicit relation representations into models of human reasoning.

However, machine-learning models developed in the field of Natural Language Processing (NLP) provide new potential alternatives as predictors of human judgments of relational similarity. *Large Language Models*[1] (LLMs) equipped with self-attention mechanisms (Vaswani et al., 2017) have driven large increases in performance in many natural language processing tasks (for a review see Kalyan, Rajasekharan, & Sangeetha, 2021). The innovation behind the success of LLMs in natural language processing is known as the *transformer block*, which employs a multi-head attention mechanism (for a visual introduction to transformers see Alammar, 2018, and for visualization of the self-attention mechanism see Vig & Belinkov, 2019). For each token in an input sequence, a transformer block creates a representation for that token through multiple weighted combinations of the representations of all the tokens in the sequence. The multihead attention mechanism determines the weights given to each representation in the sequence. These reweighted representations are then fed through a fully connected neural network layer. The full LLM is a deep stack (typically 12 or more) of these transformer blocks with one or more task-specific final output layers. LLMs are typically trained using either masked language modeling (where some input tokens are corrupted and the model attempts to predict the masked words) or autoregressively on next-word prediction tasks. This training is done over multi-billion word corpora of text.

In addition to improving performance on applications to natural language processing, LLMs have been shown

[1] We use the term LLM rather than transformer because the transformer block itself is a deep neural network module that is not specific to text input.

to better predict brain activity and behavior during language processing than static word embeddings, including during naturalistic story comprehension (Schrimpf et al., 2020). Accuracy of an LLM on next word prediction is related to measures of processing difficulty of words during reading (Wilcox, Gauthier, Hu, Qian, & Levy, 2020). Probing tasks have found that LLMs learn structural representations of sentences similar to those posited by theoretical linguists (Manning, Clark, Hewitt, Khandelwal, & Levy, 2020). Although devised as an "engineering" solution for pretrained representations for NLP, there is potential for these models to inform the study of brain and behavior.

An LLM uses transformer blocks to yield a context-specific representation for each token in the input sequence that is conditioned on itself and the other tokens in the sequence. As little as 5% of the variance in higher, contextualized layers of LLMs can be accounted for by the initial embedding layer, the LLM's analogue to static word embeddings (Ethayarajh, 2019). These context-specific representations distinguish LLMs from static word embedding techniques such as Word2vec and GloVe. Static word embeddings are insensitive to local word contexts and ordering, and instead (either in an explicit or implicit fashion) perform factorization of the global co-occurrence matrix of all words in the training corpus (Levy, Goldberg, & Dagan, 2015). The success of LLMs can be attributed in part to their ability to learn from complex interactions between words. For example, an ambiguous word such as *bank* can be disambiguated by local context (e.g. *I swam by the river* **bank** versus *I dropped my deposit in the* **bank**). An LLM will yield a separate representation for *bank* in each sentential context, whereas static word embeddings will yield only a single representation for *bank*, encoding ambiguity only implicitly by clustering semantic neighbours of its senses in semantic space (Günther et al., 2019).

Despite these important differences, representations derived from LLMs, like static word embeddings, contain no explicit relational component. The adequacy of LLMs for capturing human abstract generalization remains open to question (Balasubramanian, Jain, Jindal, Awasthi, & Sarawagi, 2020; Ettinger, 2020). With regards to analogy, the performance of LLMs is often worse than that of static word embeddings (Ushio, Espinosa-Anke, Schockaert, & Camacho-Collados, 2021). For a review of work on unpacking the success of LLMs and the nature of their internal representations, see Rogers, Kovaleva, and Rumshisky (2020).

In the present paper we compare contextual word embeddings derived from LLMs, static word embeddings, and BART, as predictors of human judgments in two studies of relational similarity. The first study we consider (Peterson et al., 2020) derived crowd-sourced human judgments of relational similarity on a subset of relation pairs taken from a set of norms created by Jurgens, Mohammad, Turney, and Holyoak (2012). The second study we consider (Lu et al., 2019) measured human ability to solve a set of verbal analogy problems in *A:B :: C:D* format.

Methods

Materials

Crowdsourced Relational Similarity Judgments Peterson et al. (2020) collected crowdsourced human similarity judgments for 6191 relation pairs. Each pair was drawn from one of the ten major relation types in the taxonomy proposed by Bejar et al. (1991), which provided the basis for the normative data collected by Jurgens et al. (2012). On each trial, participants were presented with two pairs of words, and were instructed to rate on a 1-7 scale the degree to which the two word pairs instantiate the same semantic relation. We compare computational models as predictors of mean human judgments of relational similarity for individual pair combinations.

Static Word Embeddings To obtain static word embeddings, we used pretrained Word2vec word embeddings [2]. To directly predict human judgments of relational similarity, the relation between any given pair is represented by the difference vector between its constituent words. The degree of (dis)similarity between pairs is predicted by the cosine distance between the two difference vectors (a measure termed *Word2vec-diff*).The alternative models (see below) all also derive their predictions based on cosine distance.

BART Model of Relational Similarity BART represents the relation between a given word pair as a vector, in which each dimension corresponds to a different learned relation. The value along each dimension is based on posterior probability that the word pair instantiates a particular relation. For instance, *finger : hand* is a good example of the relation *X is a part of Y* but a poor example of the relation *X causes Y*, so its vector would have a high value for the dimension corresponding to the former relation but a low value for the dimension corresponding to the latter. BART is trained using Word2vec embeddings of pairs of individual words. We used a version of BART that generates vectors based on 79 dimensions, where each dimension represents one of 79 relations learned from word pairs provided by Jurgens et al. (2012). In predicting human similarity judgments, previous work has shown that fits are improved by raising the value on each dimension to the power of 5, thereby increasing the relative weight of the most probable relations (Ichien et al., 2021). This transformation also proved helpful for BART in the simulations reported here.

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

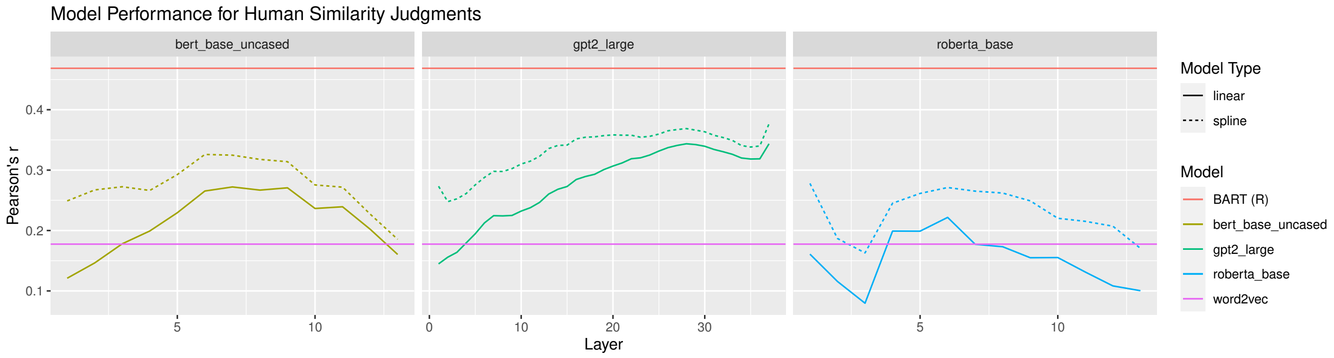

Figure 2: Model Comparisons for Human Judgments of Relational Similarity
Predictions of each model (and each layer of LLMs) for human relational similarity judgments obtained by Peterson et al. (2020). BART (red line, all plots) outperforms Word2vec (purple line, all plots), which is based on static embeddings, as well as all layers of all three LLMs, which are based on contextual embeddings. Among the LLMs, the highest layers of gpt2-large yield the predictions with the greatest correlation with human judgments. Solid lines of the LLMs indicate goodness of fit from a linear effect of distance and dashed lines from a spline fit to distance.

LLM Relation Representations To examine performance on relational similarity judgments for transformer models that yield contextual embeddings of words, we selected three of the most used LLMs[3]: BERT (bert-base-uncased) Devlin, Chang, Lee, and Toutanova (2018), RoBERTa (roberta-base) (Liu et al., 2019), and GPT-2 (gpt2-large) (Radford et al., 2019). We extracted representations using the transformers library (Wolf et al., 2020). We use capitalized names to refer to specific LLM architectures, and lowercase names with hyphens to indicate a specific set of pretrained weights available in the transformers library using that architecture. BERT uses bidirectional attention and is trained on a masked language-modeling task. RoBERTa shares its architecture with BERT, but receives longer training with larger batch sizes and with corrupted words randomized during each epoch of training, leading to improved empirical performance over BERT on many NLP tasks. GPT-2 is trained with a next-word prediction task and differs in the attention mechanism in its transformer blocks: the representation for a token at each point can only attend to its own representation and the representation of the tokens that precede it. This constraint differs from BERT and RoBERTa, in which attention mechanisms allow tokens to attend to all other tokens in the sequence at all points. As well, gpt2-large is a much larger model than bert-base-uncased and roberta-base, containing some 750 million parameters and 36 layers, compared to 110 million and 12 layers for bert-base and 125 million and 12 layers for roberta-base.

For each of the LLMs, we extracted semantic distances between relation pairs. In doing so, an important concern is that previous work has shown that raw hidden states from LLMs tend to have a small number of hidden units with large "outlier" activations, leading to poor performance on tasks involving the calculation of distances between states of the network (Sajjad, Alam, Dalvi, & Durrani, 2021; Timkey & van Schijndel, 2021). To avoid this problem, we normalized activations obtained from each network by inputting 10,000 random sentences from the Blog section of COCA (Davies, 2008), recording the mean and standard deviation of activations for each hidden unit, and then applying a z-score transformation to activations prior to any further transformation.

To obtain embeddings for comparison to human similarity judgments, we placed the two words in each pair into a sentence in the form "A is related to B". This sentence form matches the input to human similarity judgments in the study conducted by Peterson et al. (2020), where no explicit semantic relation was given to participants, and word pairs were only described as "related." From each layer we extracted representations for only the word tokens A and B (averaged across the token dimension if the model's tokenizer splits the word into subword units). These values became a 768-1280 dimensional vector representation for words A and B. As for Word2vec, distance between relation pairs was predicted using the cosine distance between difference vectors. Thus each LLM generated a set of predictions using each of its layers, allowing us to assess which layer(s) provided the best match to human judgments of relational similarity.

Results of Model Comparisons: Relational Similarity

Figure 2 presents the results of regressing relation pair distances derived from BART, Word2vec (static embeddings), and three LLMs (contextual embeddings) on

[3] According to the most-used online API for these models, `https://huggingface.co/models`, bert-base-uncased has been downloaded 12.6 million times, roberta-base 5.89 million times, and gpt-2 14.9 million times.

human judgments of relational similarity obtained by Peterson et al. (2020). Among all the models tested, BART distances between relation pairs yielded the most accurate predictions of human relation similarity judgments ($r = 0.46$). The highest performing LLM layers were gpt2-large layer 28 ($r = 0.34$), roberta-base layer 6 ($r = 0.22$), and bert-base-uncased layer 7 ($r = 0.27$). Among the three LLMs, higher layers of gpt2-large outperformed Word2vec. Some middle layers of bert-base-uncased also exceeded the performance of Word2vec, while roberta-base was the worst performing LLM, exceeding the accuracy of Word2vec only for a few middle layers. It is noteworthy that none of the LLMs exceed Word2vec performance using their first layer (which like Word2vec is based on static embeddings).

BART distances underwent a power transformation; accordingly, to ensure that BART's superior performance was not purely the result of this transformation, we tested non-linear transformations of the LLM distances. For distances from each layer and model, we fit a thin plate regression spline with a maximum of 10 degrees of freedom with R package mgcv (Wood, 2017). Effective degrees of freedom exceeded 3 for all models, indicating the effect of distance on relation similarity is nonlinear. However, while nonlinear models improved goodness of fit across the board, the improvements never increased LLM performance to the level of BART (see dashed lines, Figure 2).

Model Comparisons for Verbal Analogies

Judgments of relational similarity are closely related to the solution of verbal analogies. In a second set of model comparisons, we assessed the performance of the same alternative computational models on a set of verbal analogy problems for which data on human performance is available.

Materials

UCLA Verbal Analogy Test The UCLA Verbal Analogy Test (VAT) (Lu et al., 2019) consists of 80 analogy problems in the form *A:B :: C:D* versus *C:D'*, with 20 items representing each of four general relations: *category member*, *function*, *antonym*, and *synonym*. The task requires selection of one of two forced-choice alternatives (*D*) as the better analogical completion, where the incorrect option (*D'*) is also closely related to the *C* term; e.g., *artificial : natural :: friend : enemy* (correct) versus *friend : relative* (incorrect).

Deriving Model Predictions For all models, representations were derived in the same manner as in the simulations of relational similarity judgments reported above (except that the power transformation was omitted for the BART simulation to be consistent with Lu et al., 2019). A model was considered to select the correct answer to a problem if it yields a cosine distance between *A:B* and *C:D* less than that between *A:B* and *C:D'*.

Results of Model Comparisons: Verbal Analogies

Figure 3 presents performance of each type of model for each of the four relation types in the UCLA VAT. BART has the highest overall performance at .84 correct, which matches the human mean of .84 correct reported by Lu et al. (2019). The next highest accuracy is achieved by gpt2-large (across layers: min=0.575, max=.80, mean=0.70), followed by Word2vec (0.69 correct), followed by BERT (across layers: min=0.6, max=.725, mean=0.66), and roberta-base (across layers: min=0.59, max=.74, mean=0.66). The only LLM variants that numerically exceed BART's accuracy are a single layer of BERT on the *antonym* relation and a few layers of gpt2-large on the *antonym* and *synonym* relations. However, for the LLMs the best performing layers are not consistent across relations, and the best performing single layer does not exceed BART's overall accuracy.

Discussion

We compared multiple models of relational comparisons, each based on a different source of vector representations of meanings, to human data concerning judgments of relational similarity and solution of verbal analogy problems. The basic decision criterion for determining relation similarity (cosine distance between vector representations) was identical across all models. For both human judgments of relational similarity between word pairs (Peterson et al., 2020) and human solution of verbal analogy problems (Lu et al., 2019), the best match to human performance was obtained using BART, a model that learns explicit representations of relations, coded as vectors in a transformed similarity space. BART captured human performance more successfully than either a model using static embeddings (Word2vec) or models using contextualized embeddings created by LLMs (BERT, RoBERTa, and GPT-2).

However, considerably more work will be required to assess the potential of LLMs to support human-like relational reasoning. LLMs have great flexibility in how they can be applied to solve verbal analogies and similar tasks. Not only can embeddings be taken from different layers in the network (as explored here), but the contextualized embeddings can be produced by indefinitely many different text contexts. Here we used a very simple context, a sentence in the form "A is related to B", which is arguably a natural expression for a similarity statement. Other choices for formatting the inputs to LLMs (e.g., sentences with linking phrases more specific than "is related") can be explored in future work.

In the present study we derived predictions using contextualized embeddings directly produced by pretrained LLMs. Another avenue is to use forms of "fine tuning":

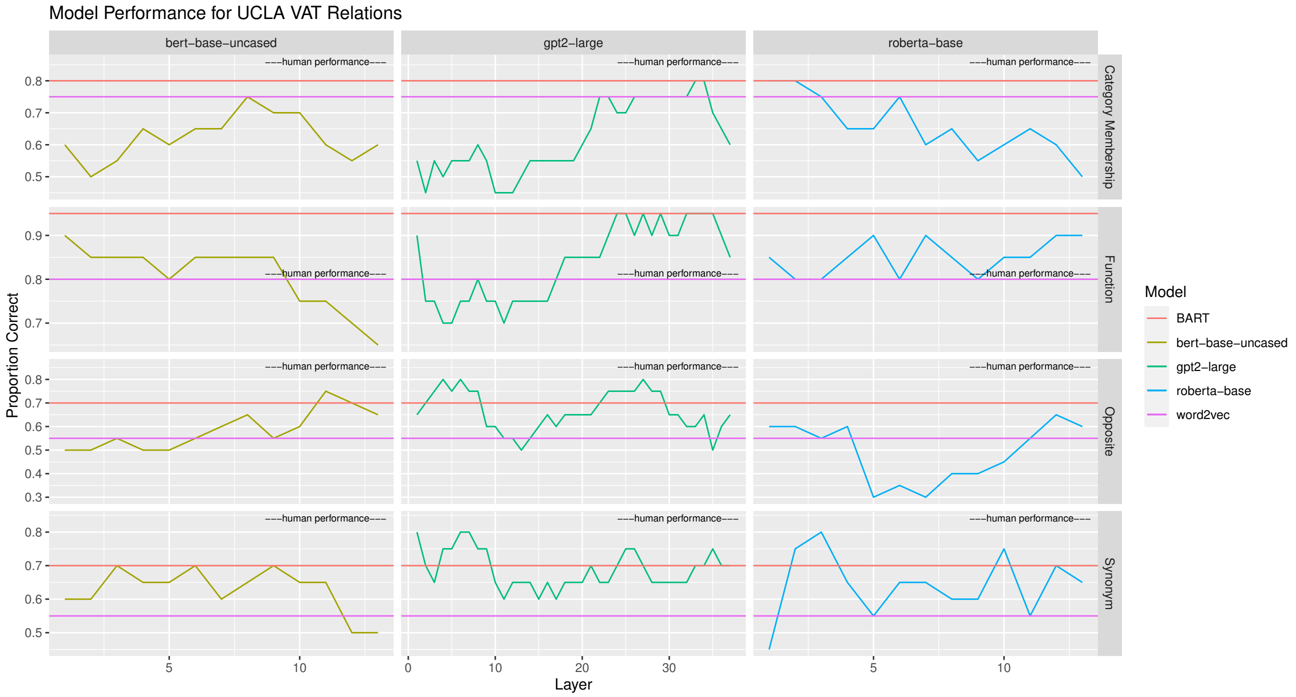

Figure 3: Model Comparisons with Human Accuracy in Solving Verbal Analogy Problems
Accuracy of each model (and each layer of LLMs) for problems in the UCLA VAT based on each of four relation types. Mean human accuracy from Lu et al. (2019) is indicated at right side in each panel.

copying weights from an LLM trained on masked or autoregressive language modeling and retraining the model on a new task, such as document classification. An LLM can be fine-tuned on specific cognitive tasks (Bhatia & Richie, in press), potentially including verbal analogy tasks. However, the cognitive interpretation of transfer learning via a fine-tuning step is not clear. Insofar as the aim is to model human cognition, it is necessary to explain how people accomplish analogical reasoning without direct training on analogy as a task.

Our approach to using LLMs in the present study was to directly extract hidden states of a network and then use difference vectors for these extracted states to predict human judgments. Language models can also function in a generative fashion: a verbal prompt can be fed to the network, based on which the network generates a completion. Indeed, it has been shown that providing LLMs with a few examples of the desired output can improve performance (Brown et al., 2020). An important direction for future research will be to evaluate the use of LLMs to solve analogy problems in the generative mode.

A further limitation of the present study is that we were unable to test recent state-of-the-art LLMs such as GPT-3 (Brown et al., 2020). These models, with hundreds of billions of parameters, exceed the capabilities of consumer hardware and/or are not publicly available for use in extracting representations. There is a concern as to whether the sheer scale of the data used to train the latest LLMs so far exceeds human capacity as to render these models implausible as the basis for cognitive models. Moreover, the representativeness and balance of LLM training corpora compared to the natural language use to which an average speaker is exposed is questionable.

Finally, an important future direction is to explore whether contextualized word embeddings can be used as inputs to a model such as BART, which aims to learn explicit representations of relations in a transformed space. BART has been trained on a variety of vector representations of word meanings (Lu et al., 2012). BART in the present paper was trained using static Word2vec embeddings. There is some evidence that explicit training on relations can yield LLMs representations that in some cases outperform static word embeddings (Bouraoui, Camacho-Collados, & Schockaert, 2020). By taking contextualized embeddings as inputs to learn vector representations of relations, it may be possible to better capture the human ability to solve complex analogies. More complex analogies that require mapping of more than two concepts in each analog can be performed using vector representations organized into attributed graphs (Lu, Ichien, & Holyoak, 2022) and using higher-order relations such as causal relations (Yuille & Lu, 2007). More generally, advances in machine learning can continue to create opportunities for synergistic advances in cognitive modeling.

Acknowledgments

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

References

Alammar, J. (2018). *The illustrated transformer.* Retrieved from https://jalammar.github.io/illustrated-transformer/

Balasubramanian, S., Jain, N., Jindal, G., Awasthi, A., & Sarawagi, S. (2020). What's in a name? are BERT named entity representations just as good for any other name? *arXiv preprint arXiv:2007.06897*.

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

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

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

Bouraoui, Z., Camacho-Collados, J., & Schockaert, S. (2020). Inducing relational knowledge from BERT. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 34, pp. 7456–7463).

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... others (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Chaffin, R. (1989). The nature of semantic relations: a comparison of two approaches. In *Relational models of the lexicon* (pp. 289–334).

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.

Davies, M. (2008). *The corpus of contemporary american english (coca).* Retrieved from https://www.english-corpora.org/coca/

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Doumas, L. A. A., Puebla, G., Martin, A. E., & Hummel, J. E. (2022). A theory of relation learning and cross-domain generalization. *Psychological Review.*. doi: https://doi.org/10.1037/rev0000346

Ethayarajh, K. (2019). How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. *arXiv preprint arXiv:1909.00512*.

Ettinger, A. (2020). What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, *8*, 34–48.

Gentner, D., & Rattermann, M. (1991). Language and the career of similarity. In S. Gelman & J. Brynes (Eds.), *Perspectives on language and thought: Interrelations in development* (pp. 225–277). Cambridge University Press.

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive psychology*, *12*(3), 306–355.

Goldstone, R. L., Medin, D. L., & Gentner, D. (1991). Relational similarity and the nonindependence of features in similarity judgments. *Cognitive Psychology*, *23*(2), 222–262.

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.

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

Jurgens, D., Mohammad, S., Turney, P., & Holyoak, K. (2012). Semeval-2012 task 2: Measuring degrees of relational similarity. In ** sem 2012: The first joint conference on lexical and computational semantics–volume 1: Proceedings of the main conference and the shared task, and volume 2: Proceedings of the sixth international workshop on semantic evaluation (semeval 2012)* (pp. 356–364).

Kalyan, K. S., Rajasekharan, A., & Sangeetha, S. (2021). Ammus: A survey of transformer-based pretrained models in natural language processing. *arXiv preprint arXiv:2108.05542*.

Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the association for computational linguistics*, *3*, 211–225.

Linzen, T. (2016). Issues in evaluating semantic spaces using word analogies. In *Proceedings of the 1st workshop on evaluating vector-space representations for nlp* (pp. 13–18).

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Stoyanov, V. (2019). RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lu, H., Chen, D., & Holyoak, K. J. (2012). Bayesian analogy with relational transformations. *Psychological review*, *119*(3), 617.

Lu, H., Ichien, N., & Holyoak, K. J. (2022). Probabilistic analogical mapping with semantic relation networks. *Psychological Review*. doi: 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*, *116*(10), 4176–4181.

Manning, C. D., Clark, 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*, *117*(48), 30046–30054.

Markman, A. B., & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive Psychology*, *25*(4), 431–467.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 1532–1543).

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

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8), 9.

Roads, B. D., & Love, B. C. (2020). Learning as the unsupervised alignment of conceptual systems. *Nature Machine Intelligence*, *2*(1), 76–82.

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in bertology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, *8*, 842–866.

Sajjad, H., Alam, F., Dalvi, F., & Durrani, N. (2021). Effect of post-processing on contextualized word representations. *arXiv preprint arXiv:2104.07456*.

Schrimpf, M., Blank, I., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., ... Fedorenko, E. (2020). The neural architecture of language: Integrative reverse-engineering converges on a model for predictive processing. *BioRxiv*.

Timkey, W., & van Schijndel, M. (2021). All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. *arXiv preprint arXiv:2109.04404*.

Ushio, A., Espinosa-Anke, L., Schockaert, S., & Camacho-Collados, J. (2021). BERT is to nlp what alexnet is to cv: Can pre-trained language models identify analogies? *arXiv preprint arXiv:2105.04949*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008).

Vig, J., & Belinkov, Y. (2019). Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 acl workshop blackboxnlp: Analyzing and interpreting neural networks for nlp* (pp. 63–76).

Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., & Levy, R. (2020). On the predictive power of neural language models for human real-time comprehension behavior. *arXiv preprint arXiv:2006.01912*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... Rush, A. M. (2020, October). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations* (pp. 38–45). Online: Association for Computational Linguistics. Retrieved from `https://www.aclweb.org/anthology/2020.emnlp-demos.6`

Wood, S. (2017). *Generalized additive models: An introduction with r* (2nd ed.). Chapman and Hall/CRC.

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

Zhila, A., Yih, W.-t., Meek, C., Zweig, G., & Mikolov, T. (2013). Combining heterogeneous models for measuring relational similarity. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 1000–1009).