UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Evaluating vector-space models of analogy

Permalink

Journal

Authors

Chen, Dawn
Peterson, Joshua C.
Griffiths, Thomas L.

Publication Date

2017

Evaluating vector-space models of analogy

Dawn Chen (sdawnchen@gmail.com)
Joshua C. Peterson (peterson.c.joshua@gmail.com)
Thomas L. Griffiths (tom_griffiths@berkeley.edu)
Department of Psychology, University of California, Berkeley, CA 94720 USA

Abstract

Vector-space representations provide geometric tools for reasoning about the similarity of a set of objects and their relationships. Recent machine learning methods for deriving vector-space embeddings of words (e.g., `word2vec`) have achieved considerable success in natural language processing. These vector spaces have also been shown to exhibit a surprising capacity to capture verbal analogies, with similar results for natural images, giving new life to a classic model of analogies as parallelograms that was first proposed by cognitive scientists. We evaluate the parallelogram model of analogy as applied to modern word embeddings, providing a detailed analysis of the extent to which this approach captures human relational similarity judgments in a large benchmark dataset. We find that that some semantic relationships are better captured than others. We then provide evidence for deeper limitations of the parallelogram model based on the intrinsic geometric constraints of vector spaces, paralleling classic results for first-order similarity.

Keywords: analogy; `word2vec`; `GloVe`; vector space models

Introduction

Recognizing that two situations have similar patterns of *relationships*, even though they may be superficially dissimilar, is essential for intelligence. This ability allows a reasoner to transfer knowledge from familiar situations to unfamiliar but analogous situations, enabling analogy to become a powerful teaching tool in math, science, and other fields (Richland & Simms, 2015). Computational modeling of analogy has primarily focused on comparing structured representations that contain labeled relationships between entities (Gentner & Forbus, 2011). However, the questions of where these relations come from and how to determine that the relationship between one pair of entities is the same as that between another pair are an unsolved mystery in such models. Some models, such as DORA (Doumas, Hummel, & Sandhofer, 2008) and BART (Lu, Chen, & Holyoak, 2012), try to learn relations from examples, but have only demonstrated success on comparative relations such as *larger*.

Another possibility is that the representations of entities themselves contain the information necessary to infer relationships between entities and that relations do not need to be learned separately. An instantiation of this hypothesis is the parallelogram model of analogy (see Figure 1), first proposed by Rumelhart and Abrahamson (1973) over 40 years ago. In this model, entities are represented as points in a Euclidean space and relations between entities are represented as their difference vectors. Even though two pairs of points may be far apart in the space (i.e., they are featurally dissimilar), they are considered relationally similar as long as their difference vectors are similar. Although Rumelhart and Abrahamson found that this simple model worked well for a small

domain of animal words with vectors obtained using multidimensional scaling, little progress was made on the parallelogram model after the initial proposal, with the exception of a handful of reasonably successful applications (see Ehresman & Wessel, 1978).

However, in the past few years, the parallelogram model was reincarnated in the machine learning literature through popular word embedding methods such as `word2vec` (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) and `GloVe` (Pennington, Socher, & Manning, 2014). These word representations enable verbal analogy problems such as *king* : *queen* :: *man* : ? to be solved through simple vector arithmetic, i.e., $v_{queen} - v_{king} + v_{man}$ results in a vector very close (in terms of cosine distance) to v_{woman}. Word embeddings like `word2vec` and `GloVe` have also been used successfully in a variety of other natural language processing tasks, suggesting that these representations may indeed contain enough information for relations to be inferred from them directly. Recently, researchers in computer vision have been successful in extracting feature spaces that exhibit similar properties in both explicit (supervised) (Radford, Metz, & Chintala, 2015) and implicit (unsupervised) (Reed, Zhang, Zhang, & Lee, 2015) ways, yielding linearized semantic image transformations such as object rotations and high-level human face interpolations. The potential for applying the parallelogram model of analogy to vector space models appears to be domain-agnostic, broadly applicable to both semantic and perceptual domains. This suggests a promising cognitive model and provides the opportunity to evaluate a classic theory in large-scale, ecologically valid contexts.
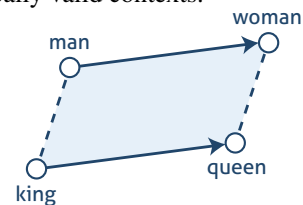

Figure 1: The parallelogram model of analogy completes the analogy *king* : *queen* :: *man* : ? by adding the difference vector between *king* and *queen* to *man*. This forms a parallelogram in the underlying vector space.

In this paper, we evaluate the parallelogram model of analogy as applied to modern vector-space representations of words. Focusing on the predictions that this approach makes about the relational similarity of words, we provide a new dataset of over 5,000 comparisons between word pairs that exemplify 10 different types of semantic relations. We find that the parallelogram model captures human relational sim-

ilarity judgments for some semantic relations, but not others. We then show that human relational similarity judgments for pairs of words violate the geometric constraints of symmetry and triangle inequality, just as Tversky (1977) showed for judgments of first-order similarity between words. This creates a challenge for *any* vector space model that aims to make predictions about relational similarity.

Relational Similarity

One way to evaluate vector space models such as `word2vec` and `GloVe` as accounts of analogy is to compare their assessments of relational similarity – the similarity of the relation between one pair of words to that of another – with human judgments. A good foundation for this task is the SemEval-2012 Task 2 dataset (Jurgens, Turney, Mohammad, & Holyoak, 2012), which contains prototypicality scores based on human data for word pairs that exemplify 79 different semantic relations. These relations were taken from a taxonomy of semantic relations (Bejar, Chaffin, & Embretson, 1991) and are subtypes of 10 general types, such as CLASS-INCLUSION, SIMILAR, and CONTRAST. Participants were given three or four paradigmatic examples of a relation and asked to generate additional examples of the same relation. A total of 3,218 unique word pairs were generated for the 79 relations, with an average of 41 word pairs per relation. A prototypicality score for each participant-generated word pair was calculated based on how often a second group of participants chose the word pair as the best and worst example of the relation among a set of choices. Table 1 shows examples illustrating two representative subtypes of each of the ten general types of relations.

According to the parallelogram model, two pairs of words ($A : B$ and $C : D$) are relationally similar to the extent that their difference vectors ($v_B - v_A$ and $v_D - v_C$) are similar. How appropriate is this geometric relationship for the various semantic relations? As a preliminary investigation of this question, we projected the 300-dimensional `word2vec` vectors into a two-dimensional space using principal components analysis separately for each relational subtype in the SemEval dataset, and visualized the difference vectors for the participant-generated word pairs from each relation. Figure 2 shows the difference vectors for the 20 relational subtypes that are shown in Table 1.

Examining the difference vectors for each relation shows that the parallelogram rule does not appear to capture all relations. CASE RELATIONS *Agent:Instrument* (e.g., *farmer : tractor*) shows a nearly perfect correspondence with what we would expect under the parallelogram model, with all difference vectors aligning. However, many of the relations appear to have no clear geometric pattern. Nevertheless, simply looking at projections of the difference vectors is not sufficient to evaluate the power of geometric models of relational similarity to capture various relations, because information is lost in the projections. What is required is a detailed evaluation on judgments of relational similarity between word pairs

Table 1: Examples of word pairs instantiating each of two representative subtypes from each general relation type in the SemEval-2012 Task 2 dataset

Relation type	Subtype	Example
1. CLASS-INCLUSION	*Taxonomic Class:Individual*	*flower : tulip* *river : Nile*
2. PART-WHOLE	*Object:Component Collection:Member*	*car : engine* *forest : tree*
3. SIMILAR	*Synonymy Dimensional Similarity*	*car : auto* *simmer : boil*
4. CONTRAST	*Contrary Reverse*	*old : young* *buy : sell*
5. ATTRIBUTE	*Item:Attribute Object:State*	*beggar : poor* *coward : fear*
6. NON-ATTRIBUTE	*Item:Nonattribute Object:Nonstate*	*fire : cold* *corpse : life*
7. CASE RELATIONS	*Agent:Instrument Action:Object*	*soldier : gun* *plow : earth*
8. CAUSE-PURPOSE	*Cause:Effect Cause:Compensatory action*	*joke : laughter* *hunger : eat*
9. SPACE-TIME	*Location:Item Time:Associated Item*	*library : book* *winter : snow*
10. REFERENCE	*Sign:Significant Representation*	*siren : danger* *diary : person*

within each relation.

Although the SemEval-2012 dataset contains prototypicality scores for the participant-generated word pairs within each relation, which have been interpreted as the relational similarities between the participant-generated pairs and the paradigmatic pairs, prototypicality is influenced by other factors such as the production frequencies of words (Uyeda & Mandler, 1980). Moreover, because participants were encouraged to focus on the *relation* illustrated by the paradigmatic examples, the prototypicality scores may not have much to do with the particular word pairs chosen as paradigmatic examples. Experiment 1 aims to address these problems.

Experiment 1: Benchmarking Relational Similarity

To overcome the limitations of the SemEval-2012 Task 2 dataset for our purposes, we collected a new large dataset that *directly* measures human judgments of relational similarity between word pairs, focusing on comparisons between word pairs with similar relations.

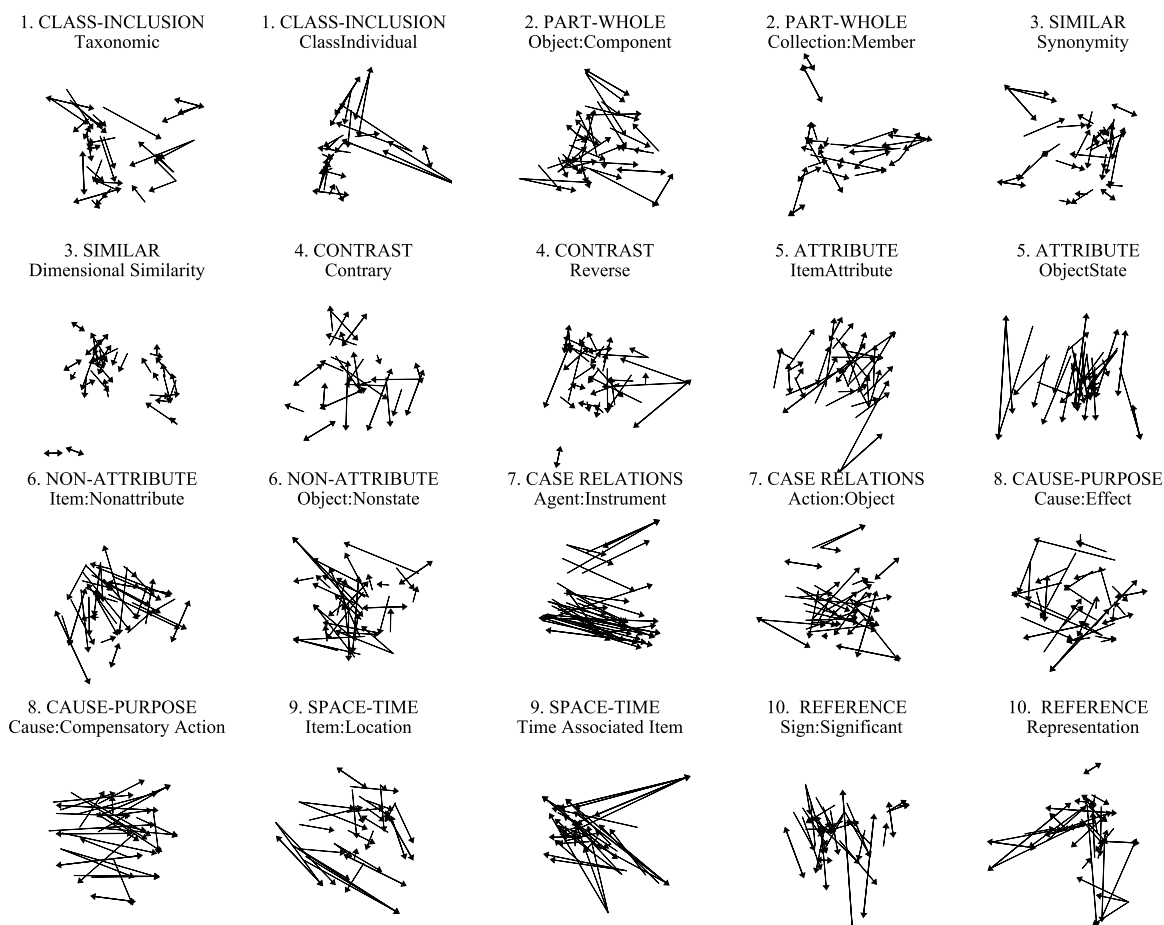**Participants** We recruited 823 participants from Amazon Mechanical Turk. Participants were paid $2.00 for the 20-

Figure 2: Visualizations of difference vectors for 20 relational subtypes using two-dimensional projections of `word2vec` word vectors obtained separately for each relation using principal components analysis.

minute study. We excluded 158 participants from the data analysis because they failed two or more of the attention checks (see below).

Stimuli The stimuli for this study were taken from the SemEval-2012 Task 2 dataset. We were mainly interested in how people rate relational similarities between participant-generated word pairs within each of the 79 relational subtypes. However, because the total number of such "within-subtype" pairwise comparisons is still enormous, we selected the most representative subtype for each relation type out of the two shown in Table 1. The subtype we chose is the first of each pair of subtypes that appears in Table 1. We then randomly chose 30 word pairs out of the entire participant-generated set for each of the 10 subtypes and formed all possible within-subtype comparisons between these word pairs. This created a set of 4,350 within-subtype comparisons. Finally, in order to encourage participants to use the entire rating scale, we added 925 "between-subtype" comparisons, which are comparisons between word pairs from different subtypes within a type (e.g., *Object:Component* and *Collection:Member*, both subtypes of PART-WHOLE), and 925 "between-type" comparisons, which are comparisons between word pairs from the representative subtypes of different relational types (e.g., *Object:Component* and *Taxonomic* CLASS-INCLUSION).

Procedure Participants were given instructions about relational similarity, which included an example of two word pairs that have similar relationships (*kitten : cat* and *chick : chicken*) and an example of word pairs with dissimilar relationships (*chick : chicken* and *hen : rooster*). Participants then viewed two pairs of words side-by-side on each page and were asked to rate the similarity of the relationships shown by the two word pairs on a scale from 1 (extremely different) to 7 (extremely similar). They rated 100 comparisons in a random order, 70 of which were within-subtype, 15 of which were between-subtype, and 15 of which were between-type. The left-right order of the two word pairs on the screen was chosen randomly (but order within pairs was of course maintained). After every 20 trials, there was an attention check question that asked participants to indicate whether two words are the same or different.

Results & Discussion We obtained at least 10 good ratings for each comparison, with an average of 10.74 ratings per comparison. The mean rating across all comparisons was 4.52 ($SD = 2.17$). As expected, we obtained the highest relational similarity ratings for within-subtype comparisons ($M = 5.01$, $SD = 1.98$), mid-level ratings for between-subtype comparisons ($M = 4.02$, $SD = 2.14$) and the lowest ratings for between-type comparisons ($M = 2.70$, $SD = 1.93$).

We calculated relational similarity for each comparison us-

ing `word2vec` and `GloVe` word representations. We used the 300-dimensional `word2vec` vectors trained on the Google News corpus that were provided by Google (Mikolov et al., 2013), and the 300-dimensional `GloVe` vectors trained on a Common Crawl web crawl corpus that were provided by Pennington et al. (2014). We tested two measures of similarity between difference vectors, cosine similarity and Euclidean distance. Specifically, for a given comparison between two word pairs, $A : B$ and $C : D$, letting $\mathbf{r}_1 = v_B - v_A$ and $\mathbf{r}_2 = v_D - v_C$, we calculated the cosine similarity,

$$\frac{\mathbf{r}_1 \cdot \mathbf{r}_2}{\|\mathbf{r}_1\|\|\mathbf{r}_2\|},$$

as well as a similarity measure based on Euclidean distance,

$$1 - \|\mathbf{r}_1 - \mathbf{r}_2\|.$$

Cosine similarity is typically used to measure similarity in vector spaces such as `word2vec` and `GloVe`. However, using Euclidean distance corresponds more closely to the original parallelogram model, in which not only the directions but also the lengths of the difference vectors needed to be similar for two word pairs to be considered relationally similar.

Figure 3 shows Pearson's correlations between predicted relational similarity scores and average human relational similarity ratings on each relation type (including both within-subtype and between-subtype comparisons) for each vector space and similarity measure. There is considerable variation in the performance of `word2vec` and `GloVe` in predicting human relational similarity ratings. As might be expected from examining Figure 2, cosine similarity performs the best on CASE RELATIONS (relation 7). However, cosine similarity completely fails on SIMILAR (relation 3), CONTRAST (relation 4), and NON-ATTRIBUTE (relation 6). Euclidean distance boosts performance on the latter two relations, but still underperforms overall compared to most other relations. Nevertheless, Euclidean distance does perform very well on SPACE-TIME (relation 9).

These results indicate that a single relational comparison strategy cannot capture all semantic relations in the spaces provided. It is unclear if such a result is a reflection of the word embeddings or actual variation in human analogical strategies. Next, we turn to the broader question of the appropriateness of the class of geometric models in general for representing human relational similarity behavior.

Violations of Geometric Constraints

Distance metrics in vector spaces must obey certain geometric constraints, such as symmetry (the distance from x to y is the same as the distance from y to x) and the triangle inequality (if the distance between x and y is small and the distance between y and z is small, then the distance between x and z cannot be very large). Cosine similarity, used to measure similarity between `word2vec` representations, also obeys symmetry and an analogue of the triangle inequality (Griffiths, Steyvers, & Tenenbaum, 2007). However, psychological representations of similarity do not always obey these

constraints (Tversky, 1977). The famous example of this is that people judge North Korea to be more similar to China than the other way around, a violation of symmetry. Griffiths et al. (2007) examined the word representations derived by Latent Semantic Analysis (Landauer & Dumais, 1997), another well-known vector space model, and found that these representations are unable to account for violations of symmetry and the triangle inequality in human word association data. Nevertheless, all prior work has focused on first-order similarity between words, and second-order (relational) similarity between word pairs might be expected to follow a different pattern. In this section, we show that human judgments of relational similarity also do not satisfy the geometric constraints of symmetry and the triangle inequality. Vector space models such as `word2vec` and `GloVe` cannot account for these violations.

Experiment 2: Symmetry

In this experiment, we examined whether there were any pairs of word pairs for which participants' judgments of relational similarity changed when the presentation order was reversed. We might expect such asymmetry to occur when a word pair has multiple relations and shares ones of its less salient relations with another word pair. For example, when presented with *angry : smile – exhausted : run*, one might think, "an angry person doesn't want to smile" and "an exhausted person doesn't want to run," but when presented with *exhausted : run – angry : smile*, one might think, "running makes a person exhausted, but smiling doesn't make a person angry." Thus, participants might give high relational similarity ratings in the first presentation order and low ratings in the second order.

Participants We recruited 1,102 participants from Amazon Mechanical Turk, who gave informed consent and were paid $1.00 for the 10-minute study. We excluded 99 participants from the data analysis because they failed two or more of the attention checks (see below).

Stimuli We randomly selected 220 within-subtype, 220 between-subtype, and 60 between-type comparisons from all possible comparisons formed using the entire SemEval-2012 Task 2 dataset. We created two versions of each comparison, in which the order of the word pairs were switched.

Procedure Participants were given instructions about relational similarity and the two examples used in Experiment 1 illustrating similar and dissimilar relationships. They saw one word pair in each comparison first and were asked to think of the relationship between the words. Then after a 600 ms delay, the other word pair was shown and participants were asked to rate the similarity of the relationships on a 7-point scale. Participants rated 50 comparisons, including 22 within-subtype, 22 between-subtype, and 6 between-type comparisons. Each participant viewed each comparison in only one of its presentation orders. After every 10 trials, there was
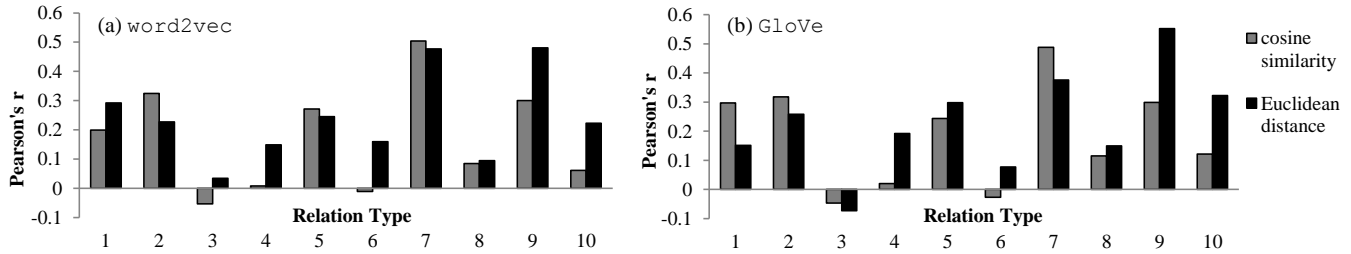
Figure 3: Pearons's *r* between human relational similarity ratings and model predictions on different relation types for (a) `word2vec` and (b) `GloVe`. The name and examples of each numbered relation type are shown in Table 1.

an attention check question that asked participants to indicate whether two words are the same or different.

Results & Discussion We obtained about 50 ratings for each comparison in each presentation order. We conducted a *t*-test for each comparison to see if the two presentation orders resulted in significantly different relational similarity ratings. 77 of these *t*-tests were statistically significant at the .05 level. The number of *t*-tests that we would expect to be significant at the $\alpha = 0.05$ level if presentation order did not matter for any of the comparisons is 25. Assuming that the *t*-tests are independent, a binomial test reveals that this deviation is statistically significant, $p < .001$.

Examining the comparisons for which different presentation orders resulted in significantly different relational similarity ratings confirms our guess as to when people's judgments of relational similarity might not obey symmetry. The previously mentioned example of *angry* : *smile* and *exhausted* : *run* indeed elicited higher ratings in the direction shown here (4.76 mean rating) than in the opposite direction (2.36 mean rating). As another example, people rated *hairdresser* : *comb – pitcher* : *baseball* as more relationally similar (6.10 mean rating) than *pitcher* : *baseball – hairdresser* : *comb* (4.84 mean rating). In the first presentation order, participants might be thinking that "a hairdresser handles a comb" and "a pitcher handles a baseball," whereas in the second presentation order, they might be thinking "a pitcher plays a specific role in baseball," which doesn't fit with *hairdresser* : *comb*.

Experiment 3: Triangle Inequality

For this experiment, we created triads of word pairs for which we expected people's relational similarity judgments to violate the triangle inequality, such as *nurse* : *patient*, *mother* : *baby*, and *frog* : *tadpole*. This triad violates the triangle inequality because *nurse* : *patient* :: *mother* : *baby* is a good analogy (relationally similar), and so is *mother* : *baby* :: *frog* : *tadpole*, but *nurse* : *patient* :: *frog* : *tadpole* is not. In this example, the middle pair has multiple relations and shares one of them with the first pair and a different one with the last pair. We presented the two word pairs in each analogy together and asked participants to rate the quality of the analogy rather than relational similarity, because we wanted to encourage participants to consider the two relations together rather than using one relation as a reference.

Participants We recruited 71 participants from Amazon Mechanical Turk, who gave informed consent and were paid $0.50 for the 5-minute study. This group of participants did not overlap with the participants in Experiment 2. We excluded 11 participants from the data analysis because they failed one of the attention checks (see below).

Stimuli We created twelve triads of word pairs for which analogy quality judgments are likely to violate the triangle inequality. For every triad, the analogy formed between the first and third word pairs was expected to be rated low and the other two analogies were expected to be rated highly.

Procedure Participants were given instructions about verbal analogies and the two examples used in Experiments 1 and 2 as examples of good and bad analogies, respectively. They were then asked to rate the quality of each analogy on a scale from 1 (very bad) to 7 (very good). For each of the twelve triads, each participant viewed one of the three analogies. Each participant received four analogies formed between the first and second word pairs of various triads (analogy type 1-2), four formed between the second and third word pairs (type 2-3), and four formed between the first and third word pairs (type 1-3). Because two thirds of these analogies are expected to be rated highly, participants also viewed four "filler" analogies expected to be given low ratings. Finally, there were two attention check questions that asked to participants to simply choose 1 (or 7) for a bad (or good) analogy.

Results & Discussion We obtained 20 ratings for each analogy. We calculated the mean participant rating for each analogy and conducted a one-way between-subjects ANOVA to test if there was an effect of the analogy type (1-2, 2-3, or 1-3) on the mean analogy quality rating. This revealed a significant effect of analogy type, $F(2, 33) = 45.57$, $p < 0.001$. Post hoc comparisons using the Tukey HSD test indicated that the mean ratings for both type 1-2 analogies ($M = 5.44$, $SD = .99$) and type 2-3 analogies ($M = 5.43$, $SD = .63$) were significantly higher than the mean rating for type 1-3 analogies ($M = 2.99$, $SD = .46$), $p < .001$, whereas the mean ratings for type 1-2 and type 2-3 analogies did not differ significantly from each other. This is consistent with our expectation that types 1-2 and 2-3 analogies would both be rated highly, whereas type 1-3 analogies would be rated lowly. These results indicate that participants' analogy quality ratings violated the triangle inequality.
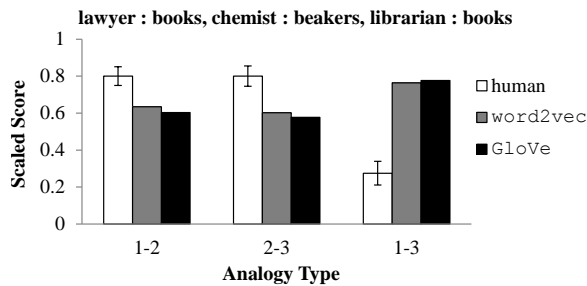
1750

Figure 4: Mean human ratings and predicted relational similarities (scaled to the range 0-1) for the triad *lawyer* : *books*, *chemist* : *beakers*, and *librarian* : *books*. Error bars indicate 1 SEM for scaled human ratings.

We obtained the predicted relational similarity between the word pairs in each analogy by calculating the cosine similarity between difference vectors using `word2vec` and `GloVe` representations. Then we conducted separate ANOVAs for the two representational spaces to test whether there was an effect of analogy type on the predicted relational similarity in each space. Neither ANOVA indicated a significant effect of analogy type. For `word2vec`, $F(2, 33) = 1.20$, $p = .31$. For `GloVe`, $F(2, 33) = .24$, $p = .79$. These results indicate that the predictions of relational similarity made by `word2vec` and `GloVe` do not violate the triangle inequality for our stimuli.

Because each participant contributed ratings to more than one analogy (although only one per triad), the observations are not entirely independent in the first overall ANOVA that we conducted on the participant data. Thus, we conducted a separate between-subjects ANOVA for each of the twelve triads to test if there was an effect of analogy type (1-2, 2-3, or 1-3) on the analogy quality ratings for each triad. All twelve ANOVAs were significant, with every $p < .01$. We then conducted post hoc comparisons using the Tukey HSD test. The pattern we expected was that types 1-2 and 2-3 analogies would have significantly higher ratings than type 1-3 analogies, but would not differ significantly from each other. We observed this pattern for seven of the twelve triads. For every one of the remaining triads, the mean ratings for the type 1-2 and type 2-3 analogies were higher than the mean rating for the type 2-3 analogy, but which differences were statistically significant differed among the triads. Figure 4 shows an example of one of the seven triads with the expected pattern and compares the mean participant ratings and predicted relational similarities for the three analogies.

Conclusions

Our results provide a clearer picture of the utility of vector-space models of analogy. The parallelogram model makes good predictions of human relational similarity judgments for some relations, but is less appropriate for others. For example, consider the word pairs represented as vectors in Figure 2. As one would expect, relation SIMILAR seems to be best represented by a short difference vector rather than the direction of the difference vector. More generally, in more complex analogies with the source and target each consist-

ing of many points in the vector space, one could imagine many ways of describing relationships between the two sets of points.

More challenging are the constraints posed by the geometric axioms. In our datasets, we found considerable violations of two of these axioms, which cannot be overcome through better embedding methods. In light of this, it would be interesting to follow the history of models of first-order similarity in considering the use of featural representations (Tversky, 1977), exploring methods of measuring similarity in vector spaces that are no longer subject to the constraints imposed by the metric axioms (Krumhansl, 1978), or reformulating the problem as probabilistic inference (Griffiths et al., 2007).

References

Bejar, I. I., Chaffin, R., & Embretson, S. (1991). A taxonomy of semantic relations. In *Cognitive and psychometric analysis of analogical problem solving* (pp. 55–91). Springer.

Doumas, L. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological review*, *115*(1), 1.

Ehresman, D., & Wessel, D. L. (1978). *Perception of timbral analogies*. Centre Georges Pompidou.

Gentner, D., & Forbus, K. D. (2011). Computational models of analogy. *Wiley interdisciplinary reviews: cognitive science*, *2*(3), 266–276.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211.

Jurgens, D. A., Turney, P. D., Mohammad, S. M., & Holyoak, K. J. (2012). Semeval-2012 task 2: Measuring degrees of relational similarity. In *Proceedings of the first joint conference on lexical and computational semantics-volume 1* (pp. 356–364).

Krumhansl, C. (1978). Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density. *Psychological Review*, *85*, 450-463.

Landauer, T. K., & Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211.

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

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 *Emnlp* (Vol. 14, pp. 1532–1543).

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Reed, S. E., Zhang, Y., Zhang, Y., & Lee, H. (2015). Deep visual analogy-making. In *Advances in neural information processing systems* (pp. 1252–1260).

Richland, L. E., & Simms, N. (2015). Analogy, higher order thinking, and education. *Wiley Interdisciplinary Reviews: Cognitive Science*, *6*(2), 177–192.

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

Tversky, A. (1977). Features of similarity. *Psychological review*, *84*(4), 327.

Uyeda, K. M., & Mandler, G. (1980). Prototypicality norms for 28 semantic categories. *Behavior Research Methods & Instrumentation*, *12*(6), 587–595.