Modelling metric violations in (geometric) conceptual spaces

Karthikeya Kaushik (karthikeya.kaushik@berkeley.edu)
Department of Psychology, UC Berkeley

Bill Thompson (wdt@berkeley.edu)
Department of Psychology, UC Berkeley

Abstract

Understanding how people represent similarity relations between concepts is one of the most fundamental problems in cognitive science, with implications for many theories of learning and reasoning. Human judgments of similarity violate basic metric assumptions, leading to effects such as judgment asymmetry and the triangle inequality. These effects have been difficult to capture with modern geometric representations of conceptual structure such as vector embeddings. Here we introduce a similarity function related to a feature-based view of concepts. We show how this function can be applied to geometric representations and that the resulting algorithm can account for classic judgment effects. Using representations extracted from a Large Language Model, we computed the predictions of this approach to similarity relations among a set of everyday concepts (world countries), and evaluated these predictions against human judgments of similarity in a behavioral experiment. The model's predictions correlate with human judgments. These results offer insight into human judgments of similarity relations and the design of algorithms that align with human reasoning.

Keywords: Vector space representations; Similarity; Concepts

Introduction

Similarity is one of the most fundamental theoretical constructs in cognitive science (Medin, Goldstone, & Gentner, 1993). judgments of similarity between objects or situations play a role in theories of information processing in many domains of cognition, including categorization and category learning, decision-making, memory, and core functions of perception such as object recognition and depth perception. In addition to being a core computation in many cognitive algorithms, similarity plays an important role in conceptual structure, being central to our ability to learn and deploy the semantic categories that underpin natural language, and the relational principles that facilitate abstract reasoning (Block, 2016).

Despite its centrality and seeming simplicity, there are still key gaps in our understanding of similarity. The computation of similarity is deeply related to the representational format in which concepts are expressed (Roads & Love, 2024; Richie & Bhatia, 2021). Modern views of conceptual structure typically represent concepts using the notion of a geometric space. Each concept is constructed as a point in a high dimensional space, and similarity and difference can then be expressed using a distance function between these points. This modeling approach is popular because of its natural relation to the high dimensional properties of lexical concepts and their graded relationship to each other.

However, geometric representations of conceptual structure have some well-known difficulties. In particular, geometric representations of conceptual structure are difficult to reconcile with aspects of human similarity judgments that violate the principles of a metric. For example, one classic finding in human judgments of similarity is deviation from symmetry (Tversky & Gati, 1982). That is, given two concepts x and y, a participant may judge that "x is similar to y", but disagree that "y is similar to x". This violates one of the three assumptions of the mathematical structure underlying a metric function – that of symmetry. Another difficulty relates to the triangle inequality. The triangle inequality suggests that given a triplet of concepts – x,y, and z, the perceived distance from x to y added to the distance from y to z should be greater than or equal to the perceived distance from z to x. However, this property is routinely violated in human similarity judgments - dogs are similar to wolves, and wolves are similar to bears, but dogs are not similar to bears. (Tversky & Gati, 1978, 1982; Tversky, 1977).

In a series of classic experiments, (Tversky & Gati, 1982) showed yet another seemingly paradoxical effect relating judgments of similarity and dissimilarity. They found that when participants were presented with a prominent pair of countries to judge (e.g USA and Canada), and a non-prominent pair (e.g Bulgaria and Albania), they often chose the prominent pair as being both more similar than the non-prominent pair *and* more dissimilar relative to the non-prominent pair! Tversky and Gati explained this effect through the suggestion that people weigh similar features more in tasks involving similarity, and different features more in tasks involving dissimilarity. A prominent pair of countries is known to participants to share more common features than the non-prominent pair, but is also known to participants to have more features that are distinct.

Deviations from simple metric assumptions highlight the need for models of similarity that can accommodate these effects. Classic approaches relied primarily on a set-based representation of concepts, with set-like properties attached to any similarity function. For example, in the contrast model (Tversky & Gati, 1982), similarity is a weighted sum of the common and distinctive binary valued features between two concepts. Others, like the distance-density model (Krumhansl, 1978) use the notion of density to loosen met-

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

ric assumptions. Connecting these models with powerful modern representations of conceptual structure (such as distributed vector representations) is an active area of research (Griffiths, Steyvers, & Tenenbaum, 2007; Nematzadeh, Meylan, & Griffiths, 2017; Jones, Gruenenfelder, & Recchia, 2018).

Recent advances in natural language processing offer new ways to re-examine these longstanding questions, by integrating distributed representations of semantics with auxiliary algorithms for the computation of similarity that go beyond simple metrics such as cosine distance. One example of such an approach is a recent analysis by (Nematzadeh et al., 2017)(see also (Nachshon, Cohen, Ben-Artzi, & Maril, 2022)). In this model, concepts are represented as *high dimensional vector summaries*. The similarity between two concept vectors, $v_x, v_y \in \mathbf{R}^D$ can be computed by first taking its dot product - $v_x.v_y$, which results in a value between $[-1,1]$, and then scaling down this value by the sum of all other vector similarities between the source and other concepts in a model's repository. That is,

$$p(x|y) = \frac{exp(v_x.v_y)}{\Sigma_{v_j} exp(v_j.v_x)} \quad (1)$$

where $p(x|y)$ is the conditional probability for a given pair of words, seen as the probability of eliciting word x, given y (See also (Jones et al., 2018) for a process model using Luce's choice algorithm on top of geometric models). However, this approach does not immediately generate asymmetry in isolation, since it relies on computations over not just a specific pair of concepts, but a whole vocabulary of other possible pairs too. While this is an important advance over previous approaches, it is at odds with our capacity to flexibly make judgments about pairs of concepts taken at one pair at a time. In other words, this approach does not exploit the internal structure of concepts when calculating similarity.

In this paper, we evaluate an approach to similarity that makes use of concepts' internal representational structure and can be applied to modern semantic representations. In line with classic theories of concepts (Rosch, 1978; Medin & Schaffer, 1978) our approach identifies a concept by a statistical distribution over features. However, unlike binary valued features, we use high dimensional vector representations of words and phrases relevant to a concept (e.g. "A dog has a fluffy coat", "A table has four legs"). These representations can be readily extracted from language models for arbitrary concepts. While the semantic representations within modern language model models may be opaque (e.g none of their dimensions maps straightforwardly into human interpretable features), they are internally coherent (Piantadosi & Hill, 2022) and therefore suitable for similarity computations. Leveraging this internal coherence, we generate sets of concept-specific descriptive phrases, and extract High Dimensional (HD) vector representations of these phrases, essentially sampling from a HD subspace that is representative of a concept. With such a representative subspace, we then construct a distance function that acts over sets of vectors to capture conceptual similarity. First, we show how the approach can produce some well known deviations from metric assumptions. Then, we use this approach to predict similarity judgments for a large set of broadly known concepts – world countries. Finally, we evaluate the predictions of the model through a behavioral experiment in which participants judged the similarity and dissimilarity between pairs of countries.

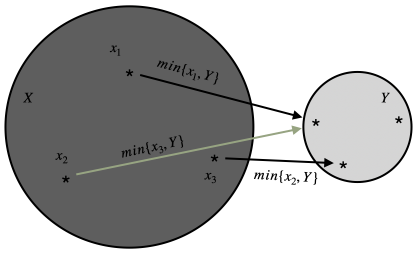

Figure 1: Computing Hausdorff distance : For each point in the left circle, the shortest line is found, and the longest of those is called the Hausdorff distance, *d_H(left,right)* (In green)

Model

The Hausdorff metric, also called Pompeiu–Hausdorff distance, measures how far apart two sets are from each other. For an intuitive explanation of how this is computed, consider the distance between the USA and Spain. Both the US and Spain are not single points in a 3D space, but consist of large areas of land. Therefore, one convenient way to answer the question of "distance between countries" is to first find the distance from every city with an international airport in the US, to its nearest destination airport in Spain. Next, we find the maximum of these distances, and call that the distance from the US *to* Spain. But we still don't have the distance from Spain *to* the US. We perform the same computations - identify airports in Spain, and from every Spanish airport, find the nearest American airport, and then take the maximum of that. Then, as shown in Eq. (2), we have two maximum distances - the maximum of this pair is the Hausdorff distance. More formally, if we consider a two set of points $x \in X, y \in Y$, the symmetric Hausdorff distance between them is given by:

$$d_H(X,Y) = max\{max_{x \in X} d(x,Y), max_{y \in Y} d(X,y)\} \quad (2)$$

However, in this paper we will use a modified *directional* version of the Hausdorff distance, given by eq. (3):

$$d_H(X,Y) = max_{x \in X} d(x,Y) = max_{x \in X}\{min_{y \in Y} d(x,y)\} \quad (3)$$

.

Where $max_{x \in X}$ denotes the maximum over distances between points $x \in X$ to all points in Y. Furthermore, the distance $d(x,Y)$ is given by $d(x,Y) = min_{y \in Y}(d(x,y))$. We use

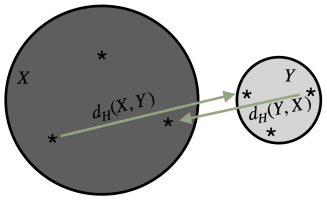

Figure 2: Asymmetry in the Hausdorff metric - $d_H(X,Y) > d_H(Y,X)$

Minkowski distance to generalize $d(x,y)$ to the setting of two vector arguments $v,u \in \mathbf{R}^D$:

$$d_{minkowski}(v,u) = (\Sigma_i^D(|v_i - u_i|)^\rho))^{\frac{1}{\rho}} \tag{4}$$

Setting $\rho = 1$ recovers Manhattan distance, and $\rho = 2$ Euclidean distance.

Deriving judgment Asymmetry

Figure 2 illustrates how asymmetry arises from the Hausdorff metric in Eq. (3). For each of the points in the left circle, compute the minimum distance to any point on the right circle; the resulting values vary depending upon where the origin point on the left circle was located. However, for points in the smaller right circle, the nearest point in the left circle does not vary substantially in distance from one origin point to another. Therefore, since

$$\begin{aligned} &max_{p_l \in Circle_{left}}\{min_{p_r \in Circle_{right}} d(p_l,p_r)\} > \\ &max_{p_r \in Circle_{right}}\{min_{p_l \in Circle_{left}} d(p_r,p_l)\}, \\ &d_H(left,right) > d_H(right,left) \end{aligned} \tag{5}$$

Put simply, under the Hausdorff metric the right circle is farther from the left circle than the left is from the right. This formulation is relevant to conceptual similarity because when comparing two concepts (two sets of features), people are potentially accounting for the shape of concepts (their extent in a semantic space). The more skewed the distance between the two sets, the more we observe the effect of direction on this similarity computation. For example, consider a containment relationship: a dog is an animal, but an animal need not be a dog, because the span of features representing an animal contains the span of features representing a dog. In this case, we would find $d_H(dog,animal) > 0$, but $d_H(animal,dog) = 0$.

Recovering the Triangle Inequality

We approach the triangle inequality by drawing on recent research from (Yearsley, Barque-Duran, Scerrati, Hampton, & Pothos, 2017), which proposes a reformulation of this effect using similarity terms instead of distance terms. The key intuition, attributed to (Shepard, 1987), is that similarity is related by an inverse exponential relationship to psychological distance, $sim(x,y) \propto exp(-D(x,y))$. Yearsley et al. rewrite the triangle inequality as:

$$\begin{aligned} D(x,y) + D(y,z) &\geq D(x,z) \\ exp(-D(x,y) - D(y,z)) &\leq exp(-D(x,z)) \\ exp(-D(x,y)) * exp(-D(y,z)) &\leq exp(-D(x,z)) \\ sim(x,y) * sim(y,z) &\leq sim(x,z) \end{aligned} \tag{6}$$

The motivation for this reformulation is that similarity is not usually additive, and the multiplicative term sets a lower bound on the inequality.

Recovering the similarity-dissimilarity contradiction

We have described a metric in which $d(X,X) = 0$. That is, from any point $x \in X$, the smallest distance to any other point in X is 0 since $x \in X$. However, consider a reformulation $\hat{d}_H$ in which we first find the maximum distance, and then the minimum over distances between two sets:

$$\hat{d}_H(X,Y) = min_{x \in X}\{max_{y \in Y} d(x,y)\} \tag{7}$$

When X contains more than one element, we see that :

$$\begin{aligned} \hat{d}_H(X,X) &= min_{x \in X}\{max_{x_i \in X} d(x,x_i)\} \\ &= min_{x \in X}\{max(d(x,x_1), d(x,x_2), d(x,x_3)..)\} \end{aligned} \tag{8}$$

Here, in the second term, since we have at least one $x_i \neq x$, we also have a $d(x,x_i) > 0$, therefore $\{max(d(x,x_1), d(x,x_2), d(x,x_3)..)\} > 0$. Substituting this back into the equation, we find $\hat{d}_H(X,X) > 0$. This formulation provides a way to characterize the "overlap" between sets, which the previous measure could not. In figure 3, we see this property in action. Since Y is a subset of X, the original metric would tell us that $d_H(Y,X) = 0$ since Y is contained in X. However, with the reformulation in 7, we find $\hat{d}_H(X,Y) > 0, \hat{d}_H(Y,X) > 0$. Consider $\hat{d}_H(Y,X)$: we first find the *maximum* distance from each point in Y to any point in X: all of the maximum distances point from Y to the top left corner. The minimum of these goes from the top left of Y to the top left of X. Similarly, $\hat{d}_H(X,Y) > 0$. Therefore, it is intuitive to associate "distances" with the regular Hausdorff metric, and to associate a form of "overlap" with this reformulated metric. Combining the two formulations will allow us to recover the surprising relationship between similarity and dissimilarity documented by (Tversky & Gati, 1978).

Methods

Behavioral Experiment

Participants We recruited 200 participants (2 excluded due to platform error) on Prolific recruitment platform. Participants performed a simple similarity judgment task in which they judged the similarity of multiple pairs of countries. Participants were US-based and were paid 1$ for participation, which took at most five minutes.

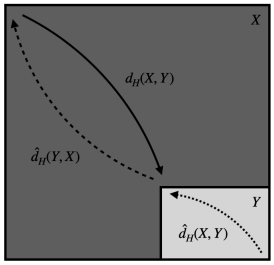

Figure 3: Comparing d_H and $\hat{d}_H$ when Y is a subset of X

Procedure The experiment consisted of two blocks. In the first block, on each trial participants were asked to rate how similar a country X is to another country Y. On a previous instruction trial, they were told not to worry if they didn't know too much about a country, but instead to focus on how similar the country X is *to* the country Y. Participants selected a position on the slider, which recorded their response from 0-100, with 100 being labelled "Identical", and 0 being "No similarity". In the second block, participants were asked to rate how different a country X is *from* another country Y, with ratings from 0-100, where 100 was labelled "Completely different", and 0 labelled "Identical".

Stimuli In each block of the experiment, participants rated 19 pairs of countries, chosen in random order, from a list of 334 pairs of countries. These 344 pairs had 167 unique combinations, with each combination appearing in two orders in the stimuli set. Since collecting all permutations of 195 countries was prohibitive, we constructed this list so that pairs of countries would include high and low similarities, and vary in geographic location and other features. We chose countries because it was a convenient domain in which to sample concepts that span well known to lesser known, and which people were likely to have basic intuitions about.

Dataset & Data Processing In total, we collected 3762 similarity, and 3762 dissimilarity ratings. Each pair of countries received on average 11 ratings. Next, we normalised the scores to make them comparable across participants, by first z-scoring within participant, within condition (similarity vs dissimilarity) data. This way, we obtained zscores for similarities between pairs of countries for participants that had rated those pairs, which was then averaged to obtain a composite zscore rating per country pair. These scores were then converted to probabilities by consulting the cumulative ztable, giving us probabilities in the range $[0,1]$ corresponding to composite zscores. These constitute the data for our subsequent analyses. Sample country pairs, and their similarity and dissimilarity scores are shown in table .

Generating Model Predictions

To apply the Hausdorff metric over sets of features in a high dimensional space, we first needed to construct the feature space. Therefore, we extracted feature descriptions for each country from OpenAI's ChatGPT 3.5 model using different prompts. For example, "Give me 50 phrases describing Luxembourg's culture, geography, and language. Be balanced, it is okay to use both positive and negative language in your description." Our motivation for using ChatGPT in this work is purely instrumental - it gives us access to rich feature descriptions without the processing steps needed to generate a similar description from traditional text sources like wikipedia.

Table 1: Sample similarities and dissimilarities

Country left	Country right	Similarity	Dissimilarity
Poland	China	0.067	0.959
USA	Germany	0.238	0.731
Algeria	Jordan	0.807	0.374

For each prompt, we obtained 50 phrases for each country. We then used the BERT (Bidirectional Encoder Representations from Transformers) model pretrained on the BookCorpus, a dataset consisting of 11,038 books and English Wikipedia, to generate contextual embeddings for these phrases. Each phrase or sentence was first tokenized using the BERT tokenizer - a token usually corresponds to a meaningful sub-word unit, and a sequence of tokens acts as the input to the model. BERT then builds its internal state by pooling together information from different parts of the sequence of tokens, and generates a contextual embedding for each token. For example, the sentence "Luxembourg is a landlocked country in Western Europe." is tokenized as : ['[CLS]', 'luxembourg', 'is', 'a', 'land', '##lock', '##ed', 'country', 'in', 'western', 'europe', '.', '[SEP]']", where 'CLS' and 'SEP' are reserved tokens indicating start and end of sequence respectively. The hidden state built by BERT is given by the 13x768 dimensional tensor, with 768 dimensions per token. We took the mean over all tokens to obtain a 1x768 dimensional vector for each phrase, and subsequently, a 50*768 dimensional matrix for all 50 phrases. This is denoted $V \in \mathbf{R}^{N\times D}$, where $D = 768$, and $N = 50$.

For every pair of countries, we now have two such matrices belonging to the country $c1, c2$, denoted V_{c_1}, V_{c_2}. The Hausdorff metric is then obtained by applying the distance function over the two sets of feature vectors :

$$d_H(V_{c_1}, V_{c_2}) = max_{n_1 \in N_{c_1}}(min_{n_2 \in N_{c_2}}(d(v_{n_1,c_1}, v_{n_2,c_2}))), \text{with}$$
$$d(v_{n_1,c_1}, v_{n_2,c_2}) = (\Sigma_i(|v^i_{n_1,c_1} - v^i_{n_2,c_2}|^2))^{\frac{1}{2}} \quad (9)$$

For further analyses, we converted these raw distances into probabilities. To do this, we first converted distances into zscores ($z = (d-\mu)/\sigma$). Next, we converted zscores into probabilities by taking the area of the unit normal curve less than a given zscore ($P(x < z)$). These formed our distance measures from phrasal embeddings. Further, we constructed the probabilistic model proposed by (Nematzadeh et al., 2017). Under this model, we have a set of countries $c \in C$, and D di-

mensional vector representations of those countries $v_c \in \mathbf{R^D}$. Here, v_c is simply the representation offered by BERT for the input "India", "China" etc. To recap, the metric here is a probability value given by :

$$p(c_1|c_2) = \frac{exp(v_{c_1}.v_{c_2})}{\Sigma_{c_j} exp(v_{c_j}.v_{c_1})} \tag{10}$$

Results

We first assessed whether the particular prompt and number of features influenced the result of the Hausdorff metric. To test this, we used two different prompts eliciting phrases listing (i) the country's culture, geography, and language (General) (ii) everything great about the country (Everything great). We found that the prompt eliciting general overall descriptions led to better alignment (measured by the spearman correlation coefficient) with human similarity and dissimilarity data (4). In all figures, the gray dotted line shows the competing probabilistic model from (Nematzadeh et al., 2017)

To see the effect of varying number of features, we performed 100 simulation runs in which we randomly sampled subsets of features from the 50 phrases, varying the number of features from 5 to 45. Figure 4 shows how set sized influenced correlation with human judgments – error bars show standard error. Note that the directions of the correlations to similarity have been reversed to make interpretation easier.

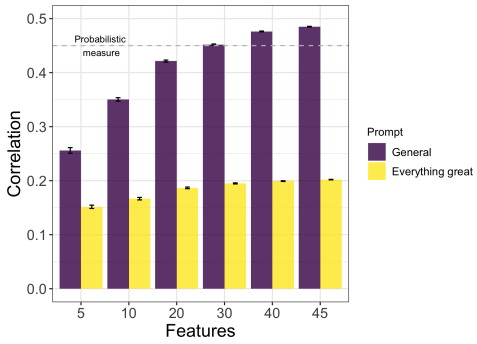

Figure 4: Number of features versus correlation to similarity data.

Judgment Asymmetry

We compared our model's predictions to the judgment asymmetries evidenced in participant responses. As a reminder, judgment asymmetries refer to the idea that e.g. people judge the concept X to be more similar to concept Y than concept Y is to concept X. For all 334 pairs of countries, we computed the ratio of asymmetry among participant responses as:

$$\sigma_{human} = \frac{sim(c_1, c_2)}{sim(c_2, c_1)} \tag{11}$$

We computed analogous ratios of asymmetry using our model and the probabilistic baseline model as:

$$\sigma_{d_H} = \frac{d_H(c_2, c_1)}{d_H(c_1, c_2)}$$
$$\sigma_{pr} = \frac{p(c_1|c_2)}{p(c_2|c_1)} \tag{12}$$

Figure 5 shows the correlation of asymmetry coefficients in human judgments and asymmetry coefficients computed using 12. Specifically, this plot shows a correlation of ratios. Perfect positive correlation implies that the inverted ratio of distances matches the ratio of similarities - a model using one single vector for each country would always have an asymmetry ratio=1, and as a result, cannot produce asymmetry. These effects were largest for the prompt listing only positive things about a country. As seen in 5, our model does significantly better than the probabilistic model. Further, while using the "general" prompt provides a better fit to human similarity judgements overall, it performs worse than the "everything great" when considering judgement asymmetry. We intend to explore the reason for this sensitivity to prompting and feature content in future work.

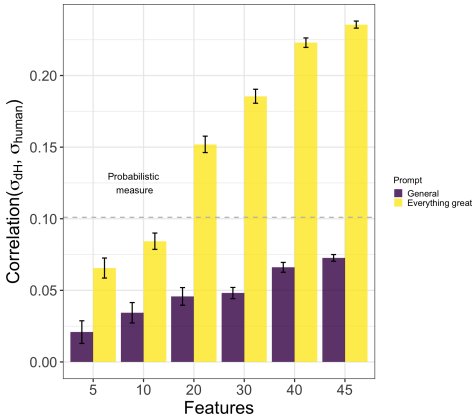

Figure 5: Correlation of asymmetry coefficients

Similarity-dissimilarity contradiction

We computed the average similarity and dissimilarity in participant data for all pairs of countries using the mean participant judgment of directional similarity and the mean participant judgment of directional dissimilarity. Note the difference in interpretation - a score of 0.9 on similarity indicates a high degree of similarity, while 0.9 on dissimilarity indicates a highly distinct pair. Under conditions of complementarity, similarity and dissimilarity ratings for a given pair of countries should sum to approximately 1. We calculated this sum in our participant's responses for specific country pairs. Figure 6 shows the distribution of these sums. Substantial deviations from 1 indicate differences between similarity and dissimilarity ratings within a pair. A value less than 1 indicates that participants judge the pair to be both minimally similar *and* minimally dissimilar. A value greater than 1 indicates

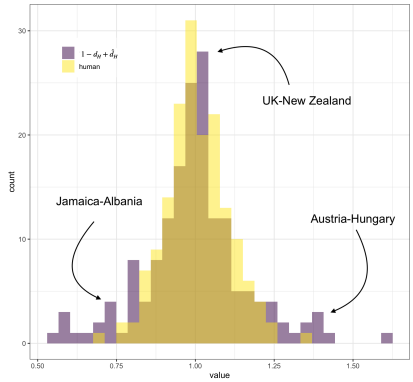

Figure 6: The histogram of summed values of similarity and dissimilarity

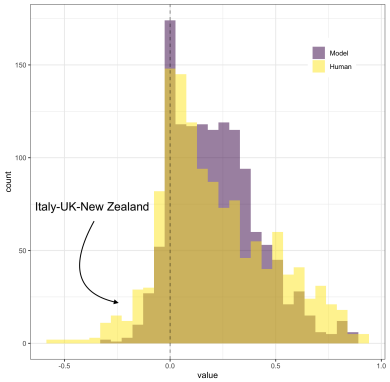

Figure 7: The histogram of human and model predictions on the triangle inequality - items to the left of the 0 line violate the inequality

that participants judged the pair to be both highly similar and highly dissimilar.

To assess whether our model can account for these effects, we make use of d_H and $\hat{d}_H$ (Eq. 7). Following our reasoning above, the mean of $\hat{d}_H$ measures average overlap, and the mean of d_H measures average distance. Therefore, we propose a push-pull relationship of the form : $1 - d_H + \hat{d}_H$. The interpretation here is - a high overlap and low distance gives a prominent pair, while high distance and low overlap gives a non-prominent pair. We then modelled a linear regression considering this new composite variable, and found a positive relationship between our composite variable, and human responses (Coefficient = 0.98, $R^2 = 0.965$, $p < 0.001$).

Triangle inequality

Figure 7 shows a histogram of values from the multiplicative triangle inequality. As described in the previous section, when the condition $sim(x,y).sim(y,z) - sim(x,z) \leq 0$ is met for a triplet of concepts x, y, z, the triangle equality is violated. The triplets we analyzed are fully connected in the experimental data (all pairwise similarity judgments are attested among participant responses). We first obtained the value of $sim(x,y) * sim(y,z) - sim(x,z)$ for human responses, and then obtained the model's prediction of the same via cosine similarity (since we are operating over unit normal vectors). We compared our model's predictions to the predictions from the probabilistic model and another basic word embedding corresponding to the country eg. getting BERT's embedding for "India". We find that our model performs better than both the probabilistic and basic word embedding models ($\rho_{d_H} = 0.446$, $\rho_{prob} = 0.433$, $\rho_{basic} = 0.351$).

Conclusion

Understanding how people judge similarity is a longstanding research program in cognitive science with important implications for theories of cognition in many domains. Current models have difficulty explaining violations of metric structure that have been evidenced in human judgments. We introduced a set-based metric that can be applied to geometric representational structures and compared its predictions to human judgments of similarity among a set of everyday concepts – world countries. Our application of the Hausdorff metric can account for traditional failures of the metric axioms in geometric spaces, as shown by our application of this approach to vector-based feature sets elicited from a large language model. In future work, we would like to explore the connections between the Hausdorff metric and classic models like the contrast model further. Additionally, the specific prompts used to elicit concept features have an impact on model predictions, which we have not yet fully explored. Understanding how prompt optimization can be leveraged to improve alignment with human judgements is an exciting avenue for future research.

Acknowledgements

We thank Steve Piantadosi for his help in locating relevant research articles, and anonymous reviewers for their helpful suggestions.

References

Block, N. (2016). Semantics, conceptual role. In *Routledge Encyclopedia of Philosophy* (1st ed.). London: Routledge.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211–244.

Jones, M. N., Gruenenfelder, T. M., & Recchia, G. (2018, August). In defense of spatial models of semantic representation. *New Ideas in Psychology*, *50*, 54–60.

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

Medin, D., Goldstone, R., & Gentner, D. (1993, April). Respects for Similarity. *Psychological Review*, *100*, 254–278.

Medin, D., & Schaffer, M. (1978). Context Theory of Classification Learning. *Psychological Review*, 207–238.

Nachshon, Y., Cohen, H., Ben-Artzi, M., & Maril, A. (2022, August). *A Model of Similarity: Metric In a Patch* (preprint). PsyArXiv. doi: 10.31234/osf.io/fyd4q

Nematzadeh, A., Meylan, S. C., & Griffiths, T. L. (2017). Evaluating Vector-Space Models of Word Representation, or, The Unreasonable Effectiveness of Counting Words Near Other Words. In *CogSci.*

Piantadosi, S. T., & Hill, F. (2022, August). *Meaning without reference in large language models.* arXiv.

Richie, R., & Bhatia, S. (2021). Similarity Judgment Within and Across Categories: A Comprehensive Model Comparison. *Cognitive Science*, *45*(8), e13030. (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cogs.13030) doi: 10.1111/cogs.13030

Roads, B. D., & Love, B. C. (2024). Modeling Similarity and Psychological Space. *Annual Review of Psychology*, *75*(1).

Rosch, E. (1978). Principles of Categorization. In A. Collins & E. E. Smith (Eds.), *Readings in Cognitive Science, a Perspective From Psychology and Artificial Intelligence.*

Shepard, R. N. (1987, September). Toward a Universal Law of Generalization for Psychological Science. *Science*, *237*(4820), 1317–1323.

Tversky, A. (1977, July). Features of similarity. *Psychological Review*, *84*(4), 327–352. doi: 10.1037/0033-295X.84.4.327

Tversky, A., & Gati, I. (1978). Studies of Similarity. In E. Rosch & B. Lloyd (Eds.), *Cognition and Categorization.*

Tversky, A., & Gati, I. (1982). Similarity, Separability, and the Triangle Inequality. *Psychological Review*, *89*(2), 32.

Yearsley, J. M., Barque-Duran, A., Scerrati, E., Hampton, J. A., & Pothos, E. M. (2017, November). The triangle inequality constraint in similarity judgments. *Progress in Biophysics and Molecular Biology*, *130*, 26–32.