Distributional semantic representations predict high-level human judgment in seven diverse behavioral domains

Russell Richie, Wanling Zou
Department of Psychology
Sudeep Bhatia
Department of Psychology, Wharton Marketing
University of Pennsylvania
{drrichie,wanlingz,bhatiasu}@sas.upenn.edu

Abstract

The complex judgments we make about the innumerable objects in the world are made on the basis of our representation of those objects. Thus a model of judgment should specify (a) our representation of the many objects in the world, and (b) how we use this knowledge for making judgments. Here we show that word embeddings, vector representations for words derived from statistics of word use in corpora, proxy this knowledge, and that accurate models of judgment can be trained by regressing human judgment ratings (e.g., femininity of traits) directly on word embeddings. This method achieves higher out-of-sample accuracy than a vector similarity-based baseline and compares favorably to human inter-rater reliability. Word embeddings can also identify the concepts most associated with observed judgments, and can thus shed light on the psychological substrates of judgment. Overall, we provide new methods and insights for predicting and understanding high-level human judgment.

Keywords: judgment; semantic memory; machine learning; word embeddings

Introduction

People are constantly perceiving, judging and evaluating entities in the world, on the qualities that these entities possess. They may consider, for example, whether a food item is nutritious, whether a political candidate is competent, whether a consumer brand is exciting, or whether the work of an occupation is significant. Such judgments influence every sphere of life, determining the social, professional, consumer, and health outcomes of individuals, as well as the political and economic makeup of our societies. It is thus of critical importance to cognitive and behavioral scientists to develop predictive and explanatory models of human judgment. To have good empirical coverage and practical utility, such models must apply to naturalistic objects and concepts, i.e., the vast range of entities people encounter every day and have rich knowledge about. They should be able to quantify what people know about these entities, and specify how people map this knowledge onto the diverse array of complex judgments they make on a day-to-day basis.

To date, building such models has been elusive, as it has been difficult to represent the detailed knowledge people have about the millions of entities in the world that they judge. Traditional psychometric methods of formally specifying object knowledge – multidimensional scaling or simply asking people to rate objects on dimensions theorized to be core to a domain – are costly and typically yield sparse representations. Thus, a technique is needed which cheaply delivers rich, high-dimensional knowledge representations for a large number of objects and concepts, which can then be used to model judgments. Fortunately, such a technique can be found in word embeddings, real-valued vector representations of word meaning derived from the statistics of word use in language corpora, such that words that occur in similar linguistic contexts yield similar vectors (see Lenci (2018) for a review). Word embeddings are a useful tool for many practical natural language processing and artificial intelligence applications. However, they also mimic aspects of human semantic cognition: they can be used to predict judgments of word similarity and relatedness, patterns of free word association, strength of semantic priming, and semantic search (Hill, Reichart, & Korhonen, 2015; Hofmann et al., 2018; Hills, Jones, & Todd, 2012; Jones, Kintsch, & Mewhort, 2006). Most relevant, researchers have also found that word embeddings predict certain association-based probability judgments, social judgments, and consumer judgments (Bhatia, 2017, 2018; Caliskan, Bryson, & Narayanan, 2017)

In this paper we show that the structure of knowledge captured by word embeddings can be used to model a very wide range of complex human judgments, including judgments that are not easily captured by association-based measures of vector similarity. More specifically, we find that with some training data in the form of human judgments about a set of words or phrases, it is possible to learn a mapping from these entities word embeddings to the judgment dimension in consideration, and subsequently make accurate predictions for nearly any entity in that domain. In other words, we use word embeddings as feature vectors for supervised machine learning models and predict out-of-sample judgment ratings with high accuracy. We also show that these learnt mappings can be used to identify the concepts that are most related to each judgment, and thus understand the most important psychological factors underlying judgments.

Method

To illustrate the broad applicability of our method, we use study fourteen types of judgment across seven different domains of mental and behavioral life: masculinity and femininity of traits (Bem, 1974), dread and unknowability of potential risk sources (Slovic, 1987), warmth and competence of people (Rosenberg, Nelson, & Vivekananthan, 1968; Cuddy, Fiske, Glick, & Xu, 2002), taste and nutrition of

foods (Raghunathan, Naylor, & Hoyer, 2006), significance and autonomy of occupations (Hackman & Oldham, 1976), sincerity and excitement of consumer brands (Aaker, 1997), and hedonic and utilitarian value of consumer goods (Batra & Ahtola, 1990). The judgment dimensions, items, participant instructions, and various implementation details for this study and for the resulting analysis, have been pre-registered on OSF here and here.

Experimental Details

We recruited 354 participants (mean age = 31.89 years, 46.19% female) through Prolific Academic. We limited our data collection to participants who were from the U.S. and had an approval rate above 80%. Participants were only allowed to participate once, and they were paid $4.40 each. Using a between-subjects design, we randomly assigned each participant to one of the seven judgment domains brands (N = 54), consumer goods (N = 51), traits (N = 46), foods (N = 55), occupations (N = 49), risk sources (N = 49), people (N = 51). These domains were chosen to span a diverse range of cognitive and behavioral sciences. Additional details about the generation of these items and other methodological details can be found on this project's OSF page and especially supplemental information here. After being randomly assigned to one judgment domain, participants were instructed to rate 200 items (e.g., occupations) on two dimensions from -100 (e.g. not at all significant) to 100 (e.g. extremely significant), one item at a time.

Word Embeddings

For our primary analyses, we used a pre-trained word embedding model, word2vec, obtained using the skip-gram technique (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013)[1], applied to a very large dataset of Google News articles. This space has vectors for 3 million words and short phrases, with each vector being defined on 300 dimensions. Although there are other training methods as well as other pre-trained semantic spaces, we base our analysis on the Google News space because of its rich vocabulary, which includes all of the naturalistic entities used in our study (including multiword entities, such as famous people and various consumer brands, which are often absent from other spaces). This pre-trained space has also been shown to accurately capture human ratings on linguistic and semantic judgment tasks (Pereira, Gershman, Ritter, & Botvinick, 2016).

[1]This technique relies on a multilayer feedforward neural network that slides over windows of text in a large corpus, and attempts to predict the words in the periphery of the window, given the word in the center of the window. By learning to predict context words in this way, the weight matrix of the network gradually learns to encode information about the relationships between words, such that semantically related words have similar (weight) vectors. The rows of the weight matrix from the input layer to the hidden layer are precisely the word embeddings we use.

Results

Predictive Accuracy of Mapping Approach

We first evaluated the predictive accuracy of our mapping method for average participant judgments (i.e. averages of the ratings made on each the fourteen judgment dimensions). We tested the ability of a variety of (regularized) regression techniques (ridge and lasso regressions, k-nearest neighbor regression, and support vector regressions with radial basis function, linear, and polynomial kernels), across a range of hyperparameters, to map our word embeddings to judgments in a pre-registered cross-validation exercise (see pre-registration form for more details). A range of models performed well, but we focus here on our best-performing model, a ridge regression with regularization hyperparameter λ set to 10, which achieved an average r-squared of .54 and an average RMSE of 21. Figure 1 shows, for each judgment dimension, scatterplots of actual judgments and predicted judgments, along with Pearson correlation coefficients, for this method. Each predicted judgment in the scatterplot was obtained by leave-one-out cross-validation (LOOCV): we trained our ridge regression model on the vectors for all but one judgment target, and then used the trained model to predict the rating for the left-out judgment target based on the target's vector. As can be seen in Figure 1, our approach was able to predict participant judgments with a high degree of accuracy, with an average correlation rate of .77 across the fourteen judgment dimensions, and all fourteen judgments yielding statistically significant positive correlations (all $p < 10^{-20}$). Our approach can also be applied to individual-level judgments, thereby accommodating participant heterogeneity. We obtain average correlations of .52 for predicted vs. observed judgments, for the individual participants in each of our fourteen tests. These accuracy rates are lower than those obtained on the aggregate level, likely due to the fact that averaging participant ratings reduces variability in data.

Comparison to Model and Human Baselines

We then compared the vector mapping approach with a simpler, baseline approach that relies only on the relative similarity of a judgment target to words denoting high vs. low ends of a particular judgment dimension (Grand, Blank, Pereira, & Fedorenko, 2018). This method works as follows: First, we select words reflective of high and low ends of some judgment dimension. For example, the occupation significance dimension was represented by the words significant, meaningful, important and insignificant, meaningless, unimportant, pointless. Where possible, we chose words used in previous literature to define the dimensions. Then, for each judgment dimension, the average pairwise vector difference between each possible pair of high and low words is computed to obtain a single vector d representing that dimension. Last, to obtain a score for a judgment target entity on that dimension, we compute the dot product between the target entity's embedding x_i and the dimension embedding, $d * x_i$. This method essentially

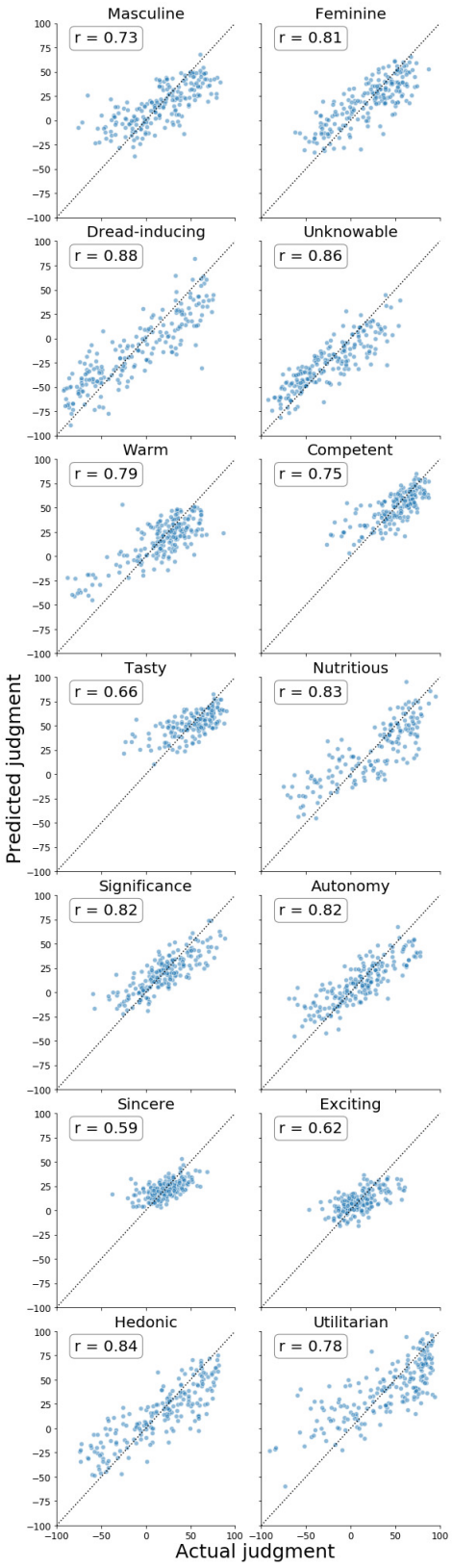

Figure 1: Scatterplots of actual judgments and predicted judgments using leave-one-out cross-validation for each judgment dimension.

computes the similarity of a judgment target (e.g., surgeon) to words high (significant, meaningful, important) relative to words low (insignificant, meaningless, unimportant, pointless) along the dimension of interest. Last, to transform these relative similarities to the range of our human judgment data, we trained OLS models predicting the human judgments from the measures of relative vector similarity, in a leave-one-out cross validation procedure. We found that the average correlation using this method was .30, which is much lower than that obtained using the vector mapping method. Additionally, the similarity method yields significant ($p < .05$) correlations for only eleven out of the fourteen tests. The baseline approach also performs worse on individual-level judgments, for which it generates average correlations of .21. As the baseline approach uses the same distances on the semantic space, for all participants, it cannot substantively accommodate participant heterogeneity (though this approach does allow for different participants to map vector similarities onto responses in different ways).[2]

We also compared the predictive accuracy of our mapping method with human inter-rater reliability, as human inter-rater reliability is often thought to place an upper bound on machine performance (Hill et al., 2015; Grand et al., 2018). To asses models predicting average models, we computed reliability two ways. First, we computed the inter-subject correlation (IS-r, (Grand et al., 2018)), which is the average correlation between one participants ratings and the average of the rest (Hill et al., 2015). This is a commonly used metric in assessing word embeddings' ability to model semantic judgments (e.g., Grand et al., 2018) and is sometimes taken to place an upper bound on machine performance (Pilehvar & Camacho-Collados, 2018). This correlation came out to 0.60, whereas our main model surpassed this with an average correlation of 0.77 across judgments. However, given that our main model is predicting an average judgment rating with word embeddings that more or less constitute the average of human knowledge reflected in word use, it may be more sensible to compare our models' performance to split-half reliability, or the correlation between the average of half the participants with the average of the other half of the participants. Thus, for each judgment dimension, we split participants into two sets, averaged judgment ratings within each set, computed the correlation between the averages, and repeated this process 100 times. The resulting split-half reliability in our judgments averaged across all judgment dimensions is .88, ranging from .69 for taste judgments to .97 for dread-inducing judgments. To assess the individual-level models relative to inter-rater reliability, we again computed reliability two ways. First, we computed the average pairwise correlation between raters (Hill et al., 2015). This correlation

[2]It is perhaps unsurprising that our baseline approach, an unsupervised method, is not as accurate as the mapping method, which is supervised. However, we maintain that this approach is the appropriate baseline to the extent that most previous applications of word embeddings in cognitive science rely on simple relative similarities like our baseline approach does.

came out to 0.34, whereas our individual-level model predictions correlated with actual judgments at an average correlation of 0.53. We can also compare individual-level model accuracy with IS-r rates, since IS-r reflects the ability to predict an individual judgment from the mean of other judgments. As stated above, mean IS-r was .60, somewhat above our average individual-level model accuracy of .53. Overall, for both average- and individual-level judgments, our model performs favorably in comparison to human inter-rater reliability, either exceeding inter-rater reliability or approaching it, depending on choice of inter-rater reliability metric.

Amount of Information Required for Prediction

A natural question for the present work is how much information in the 300-dimensional embeddings is actually required to represent our judgment targets, and hence predict our participants judgments. To this end, we measured predictive accuracy through leave-one-out cross-validation with our primary ridge model ($\lambda = 10$) after reducing the embedding spaces with principal components analysis. Specifically, for each domain, we fit a PCA on the training data design matrix (approximately 199 items, by 300 word2vec dimensions), applied the learned transformation to both the training and held-out data, discarded all but a certain number of initial principal components, and then tested how our ridge model trained on these dimension-reduced matrices predicted the held-out judgment. We emphasize that this approach obtains a *different* reduced space for every domain (cf. retraining word2vec models *for the entire vocabulary* at lower dimensional hidden layers). Figure 2 has predicted vs. actual Pearson correlations for every judgment dimension and number of retained principal components we tested. As can be seen, the 300-dimensional word embeddings can be compressed drastically to $< 10\%$ of their initial dimensionality while preserving predictive performance, with only, on average, a 3-point drop in correlation strength when retaining only the first 25 PCs, and a 7-point drop when retaining only the first 10 PCs. This suggests that, within a domain, the representational space needed to predict the present kinds of judgments is much sparser than the space provided by word2vec. Theoretically, this shows that people may only be evaluating a relative handful of (latent) dimensions when making the kinds of judgments studied here. At the same time, that much of the information relevant to making these judgments is present in the initial principal components further validates previous claims that these 14 dimensions are core dimensions along which we represent objects in these seven domains (Bem, 1974; Slovic, 1987; Rosenberg et al., 1968; Cuddy et al., 2002; Raghunathan et al., 2006; Hackman & Oldham, 1976; Aaker, 1997; Batra & Ahtola, 1990). Practically, these results indicate that future applications of the tested method need not utilize all 300 dimensions, and that successful predictions can be obtained using standard, non-regularized regression methods in the behavioral sciences applied to 10- or 25-dimensional target spaces. What kinds of information the individual principal components represent is an important question for future research, but we believe these dimension-reduced spaces are a step towards more interpretable yet highly predictive models of judgment, as a modeler now has far fewer dimensions (10 to 25, vs 300) to examine or relate to interpretable psychological quantities (by, for example, extracting the words that project onto high and low ends of the principal component's).

Psychological Substrates of Judgment

The ridge regression approach used in most of the above tests involves learning a (regularized) linear mapping from the semantic space to the judgment dimension. The best-fit weights for this mapping have the same dimensionality as the semantic space, and can thus be seen as representing a vector in this space. Judgment items whose vectors project strongly onto the weight vector (typically judgment items whose vectors are highly similar to the weight vector) will be predicted to have the highest judgment ratings. Given this interpretation, we can ask what other objects and concepts (that may not necessarily be judgment targets themselves) project strongly onto the weight vector. Intuitively, these would be the objects and concepts that are most related to the judgment, and may correspond to the judgment-relevant qualities that people evaluate when generating their responses. Thus, we took the 5000 most frequent words in the Corpus of Contemporary American English that were not also judgment targets, and fed their word2vec embeddings through our trained ridge regressions to determine their association with our 14 judgment dimensions. We then computed the difference between a words predicted association with one dimension (e.g., masculinity) and its predicted association with the complementary dimension (e.g., femininity), to find the words most strongly associated with one dimension relative to the other. Figure 3 has word clouds of these words, sized according to the strength of their association with one dimension relative to the other. These word clouds conform with expectations of the bases of these judgments. For example, traits seem to be masculine to the extent they suggest aggression, and feminine to the extent they suggest pro-sociality. A degree of artistry in a job may contribute to perceptions of autonomy, while directly guiding or helping others especially in a medical setting makes for perceptions of significance. Perceived brand sincerity may depend on brand proximity to food, family, and home; perceived brand excitement may depend on brand proximity to science, technology, and the arts.

Discussion

Despite the ubiquity of human judgment, until now we have had limited ability to predict arbitrary human judgments of objects and concepts, as capturing the rich knowledge used to make predictions has been difficult or impossible. Here we demonstrated in a pre-registered study that word embeddings, vector representations for words and concepts based on statistics of language use, proxy this knowledge and can predict 14 diverse judgments across the behavioral sciences with a high degree of accuracy. Our approach to judgment pre-

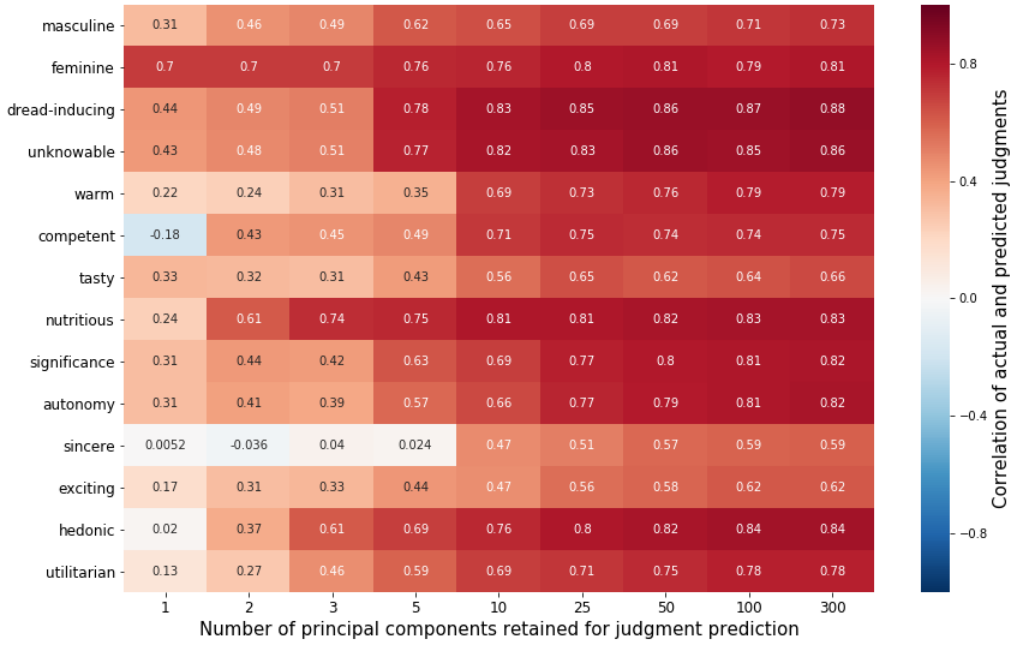

Figure 2: Pearson correlations between predicted and actual judgments for every judgment dimension and varying numbers of retained principal components. Judgment domains (brands, goods, traits, etc.) can be compressed to 5 to 25 principal components while preserving judgment prediction accuracy.

diction learning a (linear) mapping directly from word embeddings to judgment ratings surpassed a similarity-based baseline and compared favorably to human inter-rater reliability. We also showed that, despite our word embedding space (word2vec) being very rich (300 dimensions), predictive accuracy barely dropped when reducing this space to 25, 10, or even fewer dimensions, suggesting that people may only be evaluating a relative handful of pieces of information when making the present kinds of judgments. Finally, we showed that the learned mapping from word embeddings to judgments can also be used to explore the conceptual underpinnings of judgments, by mapping non-judgment target entities onto the judgment dimension.

We view the present approach as a modern extension to classical psychometric approaches used to uncover the underlying representations used for making judgments (Shepard, 1980; Slovic, 1987). However, the present approach offers several advantages over classical techniques. First, the only human data that our approach requires is a (relatively) small number of judgment ratings to train a predictive model. Once a satisfactory model has been trained, no new human psychometric data is required to predict judgments for new entities. Second, word embeddings capture more knowledge about judgment targets than can realistically be collected from human participants, especially when the relevant knowledge used to make a particular judgment is not already theoretically well-understood and thus surveyed from human participants. Capturing a great degree of knowledge leads to the high predictive accuracy we have achieved here, which we suggest may be high enough for applications in downstream behavioral sciences and technologies. For example, marketers could use predicted hedonic and utilitarian values for consumer goods to optimally advertise each of their hundreds or thousands of products, while health policy designers could use predicted risk and food perceptions to guide risk education or nutrition intervention campaigns tailored to individual perceptions.

The present research can be extended in many directions. Besides simply modeling new judgment dimensions for additional domains and entities, one promising avenue is to attempt to model different subpopulations judgments. One way to do this is simply training different regression models for different subpopulations of participants (e.g., Democrats and Republicans), but another is training word embeddings on different corpora more reflective of one population than another (e.g., MSNBC vs. Fox News articles). Under this approach, words and concepts that have somewhat different meanings and associations for different subpopulations, like the word immigrant may for Democrats and Republicans, will be located in different parts of the word embedding spaces for the corresponding representative corpora. Thus, differences in judgments about, say, the warmth and competence of immigrants, elicited from Democrats and Republicans could be predicted from their different word embeddings.

Despite the strength of our approach, it is not without limitations. Cognitive scientists, who are accustomed to inter-

Figure 3: Non-judgment target words with strong association with one judgment relative to its within-domain complement. These suggest potential conceptual underpinnings of judgments.

pretable models, may be most concerned that the dimensions of the most common word embedding techniques including word2vec, which we use here are not themselves interpretable. We attempted to mitigate this problem by using our learnt mappings to predict judgment associations for non-judgment targets, and we suggested that our PCA results were a step towards interpretable models, insofar as they reduced the number of dimensions a modeler would need to examine and relate to psychologically meaningful quantities. Another approach is to train models that predict interpretable psychological qualities that are theorized to subserve different judgments. For example, the unknowability of a potential risk source is theorized to be a composition of its observability, knowledge to the exposed, the delay of their effects, and other specific factors. Thus, one could train a model to predict these quantities from word embeddings, and then train a model predict unknowability from these predicted quantities. It is also worth pointing out that classic psychometric techniques do not always avoid this problem; multi-dimensional scaling is not guaranteed to uncover dimensions corresponding to meaningful psychological qualities. Thus, word embeddings are not always a step down in interpretability relative to other empirical methods of quantifying conceptual knowledge. Finally, cognitive scientists have traditionally focused on interpretable, explanatory models, at the expense of models that make accurate out-of-sample predictions (Yarkoni & Westfall, 2017). Of course, this is undesirable to the extent that we think a good model requires external validity; having statistically significant, interpretable model coefficients is ultimately of limited use if a model cant predict new behavior with any accuracy. Thus, our work can be seen as part of the trend to rebalance the concerns of prediction and explanation in cognitive science.

References

Aaker, J. L. (1997). Dimensions of brand personality. *Journal of Marketing Research*, 347–356.

Batra, R., & Ahtola, O. (1990). Sources of the hedonic and utilitarian measuring attitudes consumer. *Consumer Attitudes*, *2*(2), 159–170.

Bem, S. L. (1974). The measurement of psychological androgyny. *Journal of Consulting and Clinical Psychology*, *42*(2), 155–162.

Bhatia, S. (2017). Associative judgment and vector space semantics. *Psychological Review*, *124*(1), 1–20.

Bhatia, S. (2018). Semantic processes in preferential decision making. *Journal of Experimental Psychology. Learning, Memory, and Cognition*.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186.

Cuddy, A. J., Fiske, S. T., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and

competition. *Journal of Personality and Social Psychology*, *82*(6), 878–902.

Grand, G., Blank, I. A., Pereira, F., & Fedorenko, E. (2018). Semantic projection: recovering human knowledge of multiple, distinct object features from word embeddings. *arXiv preprint arXiv:1802.01241*.

Hackman, J. R., & Oldham, G. R. (1976). Motivation through the design of work: Test of a theory. *Organizational Behavior and Human Performance*, *16*(2), 250–279.

Hill, F., Reichart, R., & Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, *41*(4), 665–695.

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological Review*, *119*(2), 431–440.

Hofmann, M. J., Biemann, C., Westbury, C., Murusidze, M., Conrad, M., & Jacobs, A. M. (2018). Simple co-occurrence statistics reproducibly predict association ratings. *Cognitive Science*, *42*(7), 2287–2312.

Jones, M. N., Kintsch, W., & Mewhort, D. J. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language*, *55*(4), 534–552.

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).

Pereira, F., Gershman, S., Ritter, S., & Botvinick, M. (2016). A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. *Cognitive Neuropsychology*, *33*(3-4), 175–190.

Pilehvar, M. T., & Camacho-Collados, J. (2018). Wic: 10, 000 example pairs for evaluating context-sensitive representations. *CoRR*, *abs/1808.09121*. Retrieved from http://arxiv.org/abs/1808.09121

Raghunathan, R., Naylor, R. W., & Hoyer, W. D. (2006). The unhealthy= tasty intuition and its effects on taste inferences, enjoyment, and choice of food products. *Journal of Marketing*, *70*(4), 170–184.

Rosenberg, S., Nelson, C., & Vivekananthan, P. (1968). A multidimensional approach to the structure of personality impressions. *Journal of Personality and Social Psychology*, *9*(4), 283–294.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, *210*(4468), 390–398.

Slovic, P. (1987). Perception of risk. *Science*, *236*(4799), 280–285.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122.