Measures of Semantic Distance from Word Embeddings Predict Neural Responses During Inferences about People and Objects

Dilara Berkay (dberkay@sas.upenn.edu)
Adrianna C. Jenkins (acjenk@upenn.edu)
University of Pennsylvania, 3720 Walnut St. Philadelphia, PA, United States

Abstract

Recent advances in Natural Language Processing (NLP) make it possible to quantify relationships among different words extracted from large-scale human text corpora. Using a word embeddings model, we quantified the semantic distance between pairs of adjectives that could describe people or objects (e.g., smart, friendly; round, wooden) and scanned participants using fMRI while they had the opportunity to generalize from one known attribute to an unknown attribute across parametrically varying degrees of semantic distance (e.g., given that this person is smart, how likely are they to be friendly?; given that this furniture is round, how likely is it to be wooden?). Across categories, we observed a positive parametric effect of semantic distance on activation in the dorsomedial prefrontal cortex (DMPFC). Results connect to this region's role in abstraction and inference under reducible uncertainty, with implications for understanding how people generalize beyond what they know to make inferences about novel individuals, items, or experiences.

Keywords: word embeddings; NLP; semantic distance; fMRI; DMPFC; social cognition; generalization

Introduction

When making decisions, people rarely have complete information. Instead, people often need to fill in gaps in knowledge in light of the information to which they do have access. For instance, when choosing a meal, shopping online, or forming an impression of a person they meet, individuals can generalize what they know about the foods, articles of clothing, or people they have encountered in past situations to inform predictions about unknown properties of current ones (Addis et al., 2009; Xia, Solomon, Thompson-Schill & Jenkins, 2023).

This ability to generalize from past experiences to make predictions about related, but not identical, current or future ones may be especially useful for making predictions about other people. Unlike toasters, kites, or bicycles, no two people are exactly the same. Moreover, their behavior is much more variable, and it is only through indirect cues that observers can discern the drivers of their behavior (e.g., their beliefs, feelings, or intentions) (Berkay & Jenkins, 2023; Plate, Ham, & Jenkins, 2022). However, it is possible that generalization in social and nonsocial contexts may still rely on shared cognitive processes.

A feature of the inference space that may be relevant to generalization across both social and nonsocial inferences is diagnosticity. That is, the more related an old item or experience is to the new one, the more possible it should be to apply knowledge directly from one to the other. For example, if you know that a person is friendly, you might have a good sense of whether this person is also helpful, provided that you think friendliness is diagnostic of helpfulness. The less related the items, the more it may be necessary to consider abstract relationships between them and/or to rely on a combination of a larger number of past experiences to inform the current one (Jenkins & Mitchell, 2010). For instance, knowing that a person is friendly might not help you as much to predict whether they are also smart (unless you think being friendly is diagnostic of being smart).

Recent advances in machine learning techniques applied to natural language data make it possible to quantify the distance between any two items in a massively multidimensional semantic space derived from large-sale human text corpora. Word embeddings analysis relies on statistics of co-occurence of words in language across contexts to uncover the structure of semantic relations between them (Lenci, 2018; Mikolov et al., 2013). The more semantically similar two words are, the more similar their vectors will be and the closer they will be situated to each other in this semantic space. Word embeddings analysis accordingly provides a quantifiable measure of semantic relatedness between words, which has been shown to predict human performance in semantic judgment tasks, as well as probabilistic judgment and social judgment tasks (Bhatia, Richie, & Zhou, 2019) . This puts us in a position to ask if word embeddings can provide means to quantify diagnosticity in an objective manner and study its effects on inference processes across social and nonsocial contexts.

Here, we specifically ask (i) to what extent measures of semantic distance between pairs of attributes from aggregate natural language data are associated with people's perceptions of diagnosticity of one attribute for another and (ii) whether and how these measures of semantic distance between pairs of attributes are related to activation in particular brain regions when people try to generalize from one attribute to another.

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

Semantic distance in the brain

Although this is, to our knowledge, the first study to use word embeddings to predict participants' brain activation during inference, a number of previous findings support predictions about a particular region in which an effect of semantic distance might be observed.

Previous fMRI studies show that the dorsomedial prefrontal cortex (DMPFC) is indicated in processes that may be especially relevant when making inferences under uncertainty, including imagination and mental simulation (Addis et al., 2009; Andrews-Hanna, 2012; Buckner & Carroll, 2007; Hassabis & Maguire, 2009; Jenkins & Mitchell, 2010; Spreng et al., 2009). These mental processes can aid in using the semantic representations one has to infer characteristics of objects and people based on what is known about them. The DMPFC has also been implicated in inference when known information incompletely constrains people's inferences about others' mental states (Jenkins & Mitchell, 2010) and may play a particular role when people make judgments and decisions under reducible, but not irreducible uncertainty (Berkay & Jenkins, 2023). Together, these findings point at the possibility that the DMPFC is important for generating inferences based on available information, which makes it a good candidate for tracking semantic distance.

If the DMPFC is indeed important for making inferences under uncertainty, we would expect activation in this brain region to correlate with the semantic distance to be traversed so as to arrive at a judgment. The shorter the semantic distance between two concepts, and the more diagnostic one concept is for the other, the less semantic space one would need to traverse in order to make this connection. On the other hand, the farther away two concepts are in the semantic space, and the less diagnostic one concept is for the other, the more one would need to rely on cognitive processes important for bridging this gap. Intriguingly, DMPFC activation has previously been found to track with the semantic distance traversed in analogical reasoning (Green, 2006).

Our aim was to examine whether and how semantic distance, as measured by word embeddings, relates to brain activation during inference. In an fMRI study, participants made semantic inferences about people and objects based on one piece of information given about them, which varied in diagnosticity for a second piece of information. On each trial, participants were presented with one characteristic that describes i) the personality of another person, ii) the physical appearance of another person, or iii) the physical appearance of a piece of furniture, and were asked to make an inference about the likelihood of a second characteristic describing the same person or object. We calculated the cosine similarity values reflecting the semantic relatedness between the given characteristic and the characteristic to be inferred using word embeddings and examined the brain regions that tracked this measure. This analysis allowed us to examine how the semantic relatedness as calculated based on the cooccurrence of different concepts in natural language data is represented in the individual brain.

Method

Participants

Fifty individuals with no reported history of neurological conditions participated in the fMRI experiment in exchange for payment. All participants were right-handed, had normal or corrected-to-normal vision, and were fluent English speakers. All participants gave written informed consent before the experiment. Four participants were excluded from the sample prior to fMRI analysis due to excessive head motion, not engaging with the task, or technical issues. The final sample size consisted of 46 individuals (33 females, 13 males; age range: 18-44, mean age: 24). The experiment was approved by the Institutional Review Board at the University of Pennsylvania.

Procedure

Stimulus Development Stimuli were taken from a previous pilot study where we collected informativeness ratings from 145 individuals for a superset of stimulus pairs. On each trial, participants were presented with a given characteristic (e.g., "ambitious"; "freckled"; "wooden"), that described a person or piece of furniture, and were asked to evaluate how informative this first characteristic is to make a judgment about whether or not a second characteristic (e.g., "motivated?"; "red-headed?"; "smooth?") also describes the same person or piece of furniture on a scale from 0 (not informative at all) to 100 (very informative). This stimulus set gave us 72 pairs of characteristics describing people's physical appearances, 72 pairs of characteristics describing people's personality traits, and 72 pairs of characteristics describing furniture. For the current report, we focused on the semantic relatedness within each stimulus pair as our primary measure of interest. In order to get at semantic relatedness, we measured the cosine similarity between each pair of characteristics using word2vec (Mikolov et al., 2013). This measure reflected the semantic distance between these characteristics and therefore the space that participants needed to traverse in order to go from the information that is provided to the information to be inferred. The semantic distance scores did not show a significant difference across categories ($F(2,193) = 2.46$, $p > .05$; see Figure 1B).

Behavioral Procedure Participants were screened for contraindications for MRI prior to the experimental session. Upon arrival, participants gave informed consent, were given instructions, and completed practice trials outside the scanner. In the scanner, participants made a total of 216 inferences (72 per category) across six runs (6.93 mins each). On each trial, a cue to the category was displayed (i.e., "Personality", "Physical", or "Furniture") along with a characteristic that described a member of that category (e.g., "smart"). After a variable interval (2-4 s), a question about a second characteristic appeared (e.g., "sincere?"). After a

fixed delay of 2.5 s, a response scale ranging from 1 (very unlikely) to 4 (very likely) was presented and remained on screen for 2 s. Participants were instructed to report how likely the second characteristic (i.e., the inferred characteristic) was to describe the same member of that category, given the first characteristic (i.e., the given characteristic) using a 4-button response box held in their right hand. Trials were separated by a jittered intertrial interval ranging between 1-16 s during which a central fixation cross was presented on the screen (Figure 1A). Trial order was pseudo-randomized such that each run consisted of 36 trials, with the number of trials per condition and general uncertainty level (categorized for this purpose only as low, medium, and high) was constant across runs. The task was programmed using PsychoPy47 (v3).

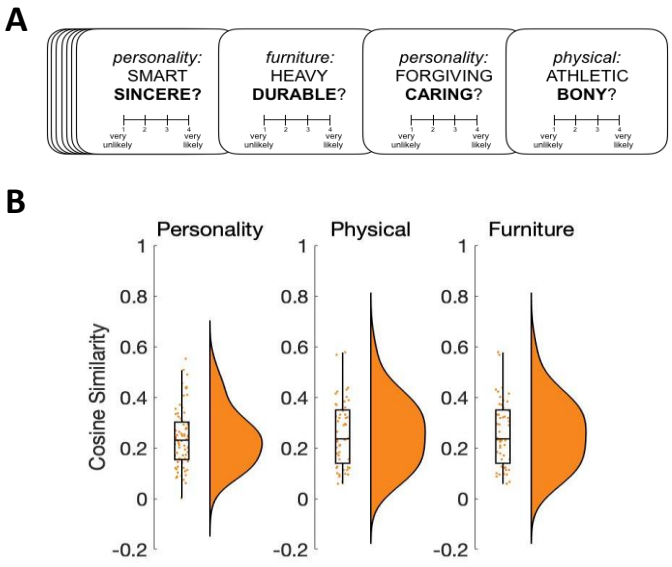

Figure 1: Stimulus examples and cosine similarity values for stimulus pairs across categories. (A) Participants made 216 inferences across three conditions in a pseudorandomized, interleaved fashion. Participants first viewed a characteristic of a person or piece of furniture (the given characteristic), then reported how likely it was that a second characteristic (the inferred characteristic) described the same person or piece of furniture. (B) We calculated the cosine similarity for each stimulus pair. There was no significant difference across cosine similarity values of word pairs in personality, physical, and furniture conditions ($F(2,193) = 2.46$, $p > .05$).

MRI data acquisition and analysis MRI data were acquired using a Siemens MAGNETOM Prisma 3T MRI scanner with a 32-channel head coil at the Center for Advanced Magnetic Resonance Imaging and Spectroscopy (CAMRIS) at the University of Pennsylvania. High-resolution T1-weighted structural images were acquired using a magnetization-prepared rapid-acquisition gradient-echo (MPRAGE) pulse sequence (voxel size = 0.9 x 0.9 x 1 mm, 160 axial slices, TR = 1850 ms, TE = 3.91 ms, flip angle = 8°, TA = 3.30 mins). T2*-weighted functional images were acquired using a multiband echo-planar imaging (EPI) sequence (multiband acceleration factor = 2, 3-mm isotropic voxels, 62 interleaved axial slices, TR = 2 s, TE = 30 ms, flip angle = 70°, TA = 6.93 mins, FOV = 200 x 200 mm). To minimize frontal signal dropout, we used a tilted acquisition angle of 30° to the anterior commissure-posterior commissure line. We acquired functional data across 6 runs, comprising 214 volumes each. There were 4 additional dummy volumes acquired at the beginning of each run which were automatically discarded.
MRI data were preprocessed and analyzed using SPM12 (Wellcome Department of Cognitive Neurology, London, UK). For each run, we realigned functional images to a reference slice within that run to correct for head motion and applied slice-timing correction. Next, resulting functional images were registered to the structural image collected for each participant and normalized to the standard space using the Montreal Neurological Institute (MNI) template. In the last step, we applied spatial smoothing to the functional images using an 8-mm full-width half-maximum (FWHM) Gaussian kernel.

We used the general linear model (GLM) for statistical analysis. We included regressors for each trial epoch, all of which were convolved with a canonical double-gamma hemodynamic response (HRF) function. To answer our main question regarding the relationship of the BOLD signal to semantic distance during inference, we included a regressor corresponding to the onset of the second (inferred) characteristic for each of the three categories. The regressor for the presentation of the second characteristic was modeled as a boxcar with a duration of 2.5 seconds, spanning the time from the onset of the second characteristic to the onset of the response scale. Cosine similarity values for pairs of characteristics obtained from word embeddings analysis were added as a parametric modulator on the regressor. Response time was entered into each GLM as a parametric modulator. Six nuisance regressors for head motion and AR(1) model of serial autocorrelation were included in the GLM. We used cluster-level FWE correction to correct for multiple comparisons. All results are reported at corrected P<.05.

Results

First, we asked whether cosine similarity between different pairs of characteristics in word embeddings is associated with humans' perceptions of diagnosticity. In order to answer this question, we correlated cosine similarity values with informativeness ratings obtained from independent raters. This analysis revealed a positive association between these two measures, showing that characteristics that are closer together (versus farther apart) in the semantic space are also perceived to be more (versus less) diagnostic of one another (Figure 2).

Next, we asked if and where the semantic distance, captured by cosine (dis)similarity in word embeddings,

relates to brain activation during inference. In order to answer this question, we conducted a whole-brain analysis in which we examined the parametric effect of inverse cosine similarity values (reflecting semantic distance) on brain activation. Consistent with our prediction, this revealed a cluster in the DMPFC (Figure 3A). Additionally, this analysis revealed a cluster in right angular gyrus (AG) as well as left inferior frontal gyrus (IFG) that showed increased activation as a function of semantic distance (Table 1).

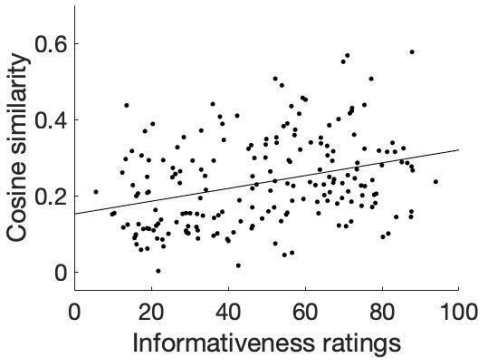

Figure 2: The association between informativeness ratings and cosine similarity. Each dot represents a pair of attribute words (e.g., smart, friendly; wooden, smooth). Informativeness ratings and cosine similarity values showed a significant positive correlation (r = .32, p < .0001), such that greater semantic distance was associated with lower perceived diagnosticity of the given attribute for the inferred attribute.

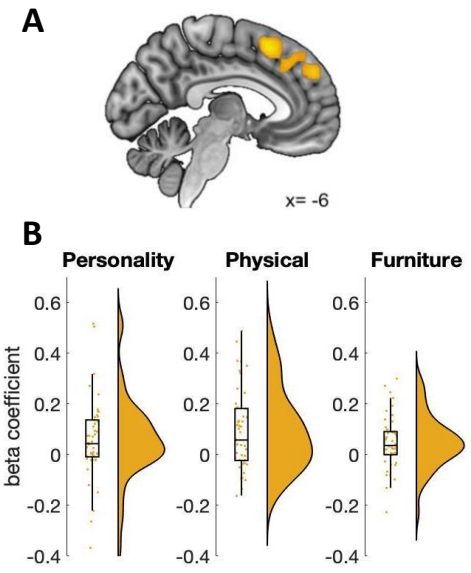

Figure 3: The parametric response to inverse cosine similarity values. (A) Our whole brain analysis revealed a cluster in the DMPFC that correlated with cosine similarity values. (B) The comparison of parameter estimates extracted from the whole-brain cluster in panel A across three categories showed no significant differences (F(2,90) = 1.30, p > .05), suggesting that the parametric effect of cosine similarity did not depend on stimulus category.

Region	*x, y, z*	Voxels	P_{FWE}
Dorsal medial prefrontal cortex	-6, 18, 54	2238	<.001
Right angular gyrus	44, -56, 28	558	.002
Left inferior frontal gyrus	-52, 20, 10	992	<.001

Table 1: Peak MNI coordinates and number of voxels for clusters showing a positive response to semantic distance.

In order to understand whether the parametric effect of cosine similarity on DMPFC activation dependend on the domain in which inference was made, we extracted parameter estimates from the DMPFC cluster that came out of the whole-brain analysis. A comparison across three categories showed no significant differences (F(2,90) = 1.30, p > .05; Figure 3B), suggesting that the effect of semantic distance on DMPFC activation during inference is independent of the domain of inference.

Discussion

In this study, we were interested in understanding how people generalize from known information to unknown information across social and nonsocial contexts. We aimed to examine whether we can use a common objective measure of semantic relatedness obtained using word embeddings to predict people's perceptions of diagnosticity within pairs of attributes describing humans and objects. We next asked whether this measure of semantic relatedness extracted from aggregate human text corpora is represented in the individual brain.

Our results indicate that objective measures of semantic distance between pairs of attributes predict people's ratings of diagnosticity between them, showing how these subjective estimates of diagnosticity are reflected in the distribution of words in the natural language data.

Consistent with our hypothesis, our fMRI results show that DMPFC activation positively tracks measures of semantic distance. In other words, DMPFC activation increased as a function of the distance one needs to traverse in semantic space to infer one attribute based on another. We also observed a positive parametric effect of semantic distance in

the right AG and the left IFG. These two brain regions are important for semantic cognition and semantic retrieval (Badre et al., 2005; Binder & Desai, 2011). The relationship to semantic uncertainty in our study makes contact with the observation that the right AG and the left IFG show increased activation under increased semantic processing demands, notably including selection among competing alternatives in semantic memory (Diveica, Koldewyn, & Binney, 2023; Kuhnke et al., 2023; Thompson-Schill et al., 1997).

The positive parametric effect we observed in the DMPFC is consistent with emerging evidence regarding the contributions of DMPFC to social and nonsocial cognition. Emerging research points to the possibility that DMPFC activation may be especially elevated when reasoning under reducible uncertainty (Berkay & Jenkins, 2023), where people need to make predictions about unknown states based on the information available to them. Additionally, the DMPFC is engaged during tasks thought to evoke imagination and mental simulation, which are candidate processes through which novel inferences may be made based on integrating information available in the environment with information from past experiences (Jenkins & Mitchell, 2010. Together with these findings, our results point to the possibility that the DMPFC may play a role in uncertainty reduction, possibly enabling people to go beyond what is directly observable to make inferences about what is unknown across both social and nonsocial contexts.

References

Addis, D. R., Pan, L., Vu, M. A., Laiser, N. & Schacter, D. L. (2009). Constructive episodic simulation of the future and the past: Distinct subsystems of a core brain network mediate imagining and remembering. *Neuropsychologia 47*, 2222–2238.

Andrews-Hanna J. R. (2012). The brain's default network and its adaptive role in internal mentation. *Neuroscientist*, *18*(3), 251–27

Badre, D., Poldrack, R. A., Paré-Blagoev, E. J., Insler, R. Z., & Wagner, A. D. (2005). Dissociable controlled retrieval and generalized selection mechanisms in ventrolateral prefrontal cortex. Neuron, 47(6), 907-918.

Berkay, D., & Jenkins, A. C. (2023). A role for uncertainty in the neural distinction between social and nonsocial thought. *Perspectives on Psychological Science, 18*(2), 491-502.

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

Binder, J. R., & Desai, R. H. (2011). The neurobiology of semantic memory. *Trends in Cognitive Sciences, 15*(11), 527-536.

Buckner R. L., Carroll D. C. (2007). Self-projection and the brain. *Trends in Cognitive Sciences*, *11*(2), 49–57.

Diveica, V., Koldewyn, K., & Binney, R. J. (2021). Establishing a role of the semantic control network in social cognitive processing: A meta-analysis of functional neuroimaging studies. *NeuroImage, 245*, 118702.

Green, A. E., Kraemer, D. J., Fugelsang, J. A., Gray, J. R., & Dunbar, K. N. (2010). Connecting long distance: semantic distance in analogical reasoning modulates frontopolar cortex activity. *Cerebral Cortex, 20*(1), 70-76.

Hassabis D., Maguire E. A. (2009). The construction system of the brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1263–1271.

Plate, R. C., Ham, H., & Jenkins, A. C. (2022). Exploration is Higher in Social Contexts at the Cost of Rewards. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 44, No. 44).

Jenkins, A. C., & Mitchell, J. P. (2010). Mentalizing under uncertainty: dissociated neural responses to ambiguous and unambiguous mental state inferences. *Cerebral Cortex, 20*(2), 404-410.

Kuhnke, P., Chapman, C. A., Cheung, V. K., Turker, S., Graessner, A., Martin, S., ... & Hartwigsen, G. (2023). The role of the angular gyrus in semantic cognition: a synthesis of five functional neuroimaging studies. *Brain Structure and Function*, *228*(1), 273-291.

Lenci, A. (2018). Distributional models of word meaning. *Annual Review of Linguistics, 4*, 151-171.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems, 26*.

Spreng R. N., Mar R. A., Kim A. S. N. (2009). The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: A quantitative meta-analysis. *Journal of Cognitive Neuroscience*, *21*(3), 489–510.

Thompson-Schill, S. L., D'Esposito, M., Aguirre, G. K., & Farah, M. J. (1997). Role of left inferior prefrontal cortex in retrieval of semantic knowledge: a reevaluation. *Proceedings of the National Academy of Sciences of the United States of America*, *94*(26), 14792–14797.

Xia, A., Solomon, S. H., Thompson-Schill, S. L., & Jenkins, A. C. (2023). Constructing complex social categories under uncertainty. *Cognition, 234*, 105363.