Zoom meeting ID: 548 112 2840
Passcode: 905374
Or join by URL:
https://upenn.zoom.us/j/5481122840?pwd=YzJyNjdYUkpWY1o2d0lKRFlFOXUvZz09

If neither work, feel free to email Wanling at wanlingz@sas.upenn.edu

Learning Other People's Preference

Wanling Zou, Sudeep Bhatia
Department of Psychology, University of Pennsylvania, Philadelphia, PA, USA

Abstract

Machine learning (ML) algorithms have been successful and sometimes even more accurate in predicting people's preferences than humans. In a preregistered online study, we studied how humans learn preferences of others by showing participants a target's food preference ratings and asking them to predict the target's preference ratings for other foods. We found that Word2Vec word embeddings combined with ML algorithms outperformed human predictions and participants who relied more on their own preferences made less accurate predictions. We also showed that ML algorithms could be handicapped to predict participant predictions (rather than target preferences) by incorporating participant preference.

Background

- Word embeddings
  - They are vectors derived from the distributional structure of words and concepts in large natural language datasets (e.g. a collection of books, Wikipedia, product reviews, and news articles) which represent natural linguistic environment.
  - The distributional patterns of words correspond to their meanings or knowledge representations (Günther et al., 2019; Mandera et al., 2017).
  - Thus, word embeddings provide rich knowledge representations for naturalistic stimuli and have been widely used in modeling high-level judgment (Richie et al., 2019)
  - We used a pretrained Word2Vec model trained on a large corpus of Google news articles (Mikolov et al., 2013) that has 300-dimensional vectors representing 3M words.
- Machine learning algorithm
  - Ridge regression takes each dimension of the Word2Vec vectors as a feature and learns coefficients on each feature by minimizing $\sum_i \left(y_i - \sum_j \boldsymbol{x}_{ij}\boldsymbol{w}_j\right)^2 + \lambda \sum_j \boldsymbol{w}_j{}^2$, where y are target's preference ratings for food items, $\boldsymbol{x}$ are word vectors of the stimuli (i.e. food items), and $\boldsymbol{w}$ are coefficients.
  - Additional features, such as human forecaster's self bias, can be combined with the Word2Vec vectors to jointly predict the target's preference.

Methods

- Study 0 Stimuli collection
  - 5 participants rated their preference for 150 food items from -100 (least prefer) to 100 (most prefer).
  - Theses participants were the targets in Studies 1 and 2.
- Study 1 Food preference (preregistered)
  - 50 participants, randomly assigned with one target

Methods (Cont.)

- Training phase
  - Participants viewed a list of 75 food items with a target's preference ratings collected from Study 0.
  - These 75 food items were selected at random. All participants viewed the same 75 food items but with ratings from different targets.

The table below shows a previous participant's preference ratings of food from -100 (least prefer) to 100 (most prefer). You will have access to this table when you predict this person's preference for other food items in the next session.

Food	Preference
goulash	-20
kidney	17
vegetables	28
⋮	⋮

- Test phase
  - Participants predicted the target's preference ratings for the unseen 75 food items, with access to the training items.

How much would this person prefer the following food?

You may review the previous table here.

Least prefer -100 -80 -60 -40 -20 0 20 40 60 80 100 Most prefer

beans

- Indicating personal preference
  - Participants rated their own preference for all 150 food items, which was considered as self bias and added as an additional feature in ridge regression.

How much do **you** prefer the following food?

Least prefer -100 -80 -60 -40 -20 0 20 40 60 80 100 Most prefer

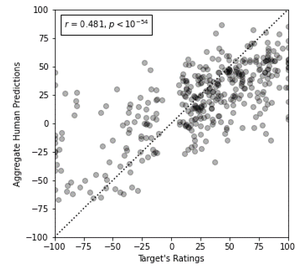
crispbread

Results

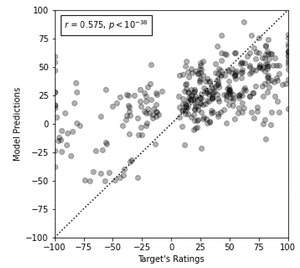
- Human accuracy
  - Each dot represents an average human prediction for a target's preference rating of one food item vs. that target's actual preference rating of the same food item.
- Machine accuracy
  - Each dot represents a model prediction for a target's preference rating of one food item vs. that target's actual preference rating of the same food item.

Results (Cont.)

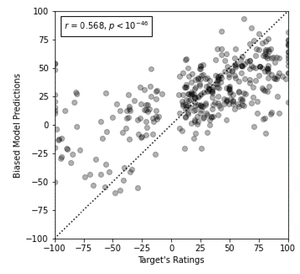
- Human accuracy ($r = 0.481$, $p < 10^{-54}$)



- Machine accuracy without self bias ($r = 0.575$, $p < 10^{-38}$)



- Machine accuracy with self bias ($r = 0.568$, $p < 10^{-46}$)



- Adding participants' own preference ratings as an additional feature makes machine less accurate in predicting the target's preference ratings.
- We replicated this finding in the domain of risk perception, where we asked participants to learn a target person's dread-inducing level of common risk sources.
- We also tested the effect of training size, where we presented participants with 25, 50, or 75 items. The results are available by contacting Wanling Zou at wanlingz@sas.upenn.edu

Günther, F., Rinaldi, L., & Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science, 14*, 1006–1033.
Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language, 92*, 57–78.
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).
Richie, R., Zou, W., & Bhatia, S. (2019). Predicting high-level human judgment across diverse behavioral domains. *Collabra: Psychology, 5*(1): 50. doi: https://doi.org/10.1525/collabra.282.