Computational Estimation of Lexical Semantic Norms: A New Framework

Bryor Snefjella & Idan Blank (UCLA)

The meanings stored in our mental lexicon are vast and varied, and include many dimensions (or "semantic features") such as a word's emotional attributes (e.g., valence, arousal), functional properties (e.g., usefulness), sensorimotor attributes (e.g., size, color), etc. (Binder et al., 2016; Lynott et al., 2019; Warriner et al., 2013). Over the past 60 years, the space of semantic features has been steadily increasing, yet the study of meaning has struggled with data sparsity throughout. Whereas English speakers know approximately 40,000 words, most semantic features have available behavioural ratings ("semantic norms") for merely 1,000−10,000 words (Fig. 1), despite massive online crowdsourcing efforts at considerable cost. The default methodological solution is to limit statistical analyses only to the subset of words for which all semantic features of interest are available; any words with partial data are simply excluded. This precarious practice, known as listwise deletion or complete case analysis, is known to damage statistical power and can bias data analysis (Rubin, 2004).

A recent alternative to complete case analysis in the field of lexical semantics replaces expensive survey methods with "efficient" computational methods which have been shown to predict semantic norms with high accuracy (c.f. Hollis et al., 2017). This task is performed in two steps. First, a representation of words as high-dimensional vectors ("word embeddings") is automatically generated from corpus co-occurrence data; then, the vector features are used as predictors in a machine learning algorithm that is trained on a small set of words for which norms have been empirically collected. This model then predicts the missing semantic norms based on those words' embeddings. Such "extrapolated semantic norms" are now publicly shared and their use in statistical inference, in place of empirical norms, is an emerging practice.

Herein, we argue that both complete case analyses and norm extrapolation are statistically problematic. First, we show that words lacking empirical semantic norms are a non-random selection from the lexicon, making complete case analysis an unwise default practice. This problem has gone unacknowledged when semantic norms are used to predict behavior (e.g., lexical decision times) in megastudies, so the semantic effects discovered therein may have yielded biased results. Second, we claim that while norm extrapolation has been construed as a *prediction* problem, it should be conceived of as a *missing data* problem. To demonstrate the far-reaching statistical implications of this reframing, we draw upon principles of analysis of partially observed data, simulations, and empirical data.

Given the pattern of missing data and the misguided framing of the statistical problem at hand, deficiencies in current semantic norm extrapolation methods include (1) overconfidence, due to "forgetting" of the uncertainty in the imputation model; (2) biased statistical inference, particularly when testing hypotheses involving nonlinearities or interactions; and (3) inefficiency, due to a failure to take into account all relevant sources of information, and not accounting for missing data in variables other than semantic norms (namely, dependent variables in analyses, such as reading times). Practical solutions to these issues are offered by the technique of multiple imputation (Rubin, 2004). Our specific analysis pipeline uses a combination of LASSO variable selection (Tibshirani, 1996) and a model-based multiple imputation method (SMCFCS, Bartlett and Morris, 2015) embedded within multiple imputation by chained equations (MICE, van Buuren and Groothuis-Oudshoorn, 2011). We use simulation evidence to show these methods in concert can accommodate high-dimensional imputation with an analysis model potentially involving nonlinearities and interactions, and restore unbiased estimation with close to nominal confidence interval coverage. We also revisit theorized effects of words' connotations of danger and usefulness (Wurm, 2007) in lexical decision, where our method yields qualitatively different results (Fig. 2) than the existing, naive extrapolation methods. Surprisingly, our results further indicate that given the particular nature of missing data, a proper implementation of semantic norm extrapolation via multiple imputation should in fact be preferred over the de-facto default use of complete case analysis in lexical semantics.

Figure 1

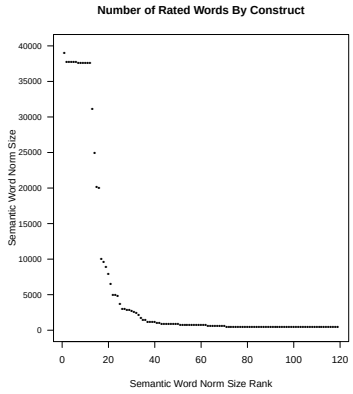
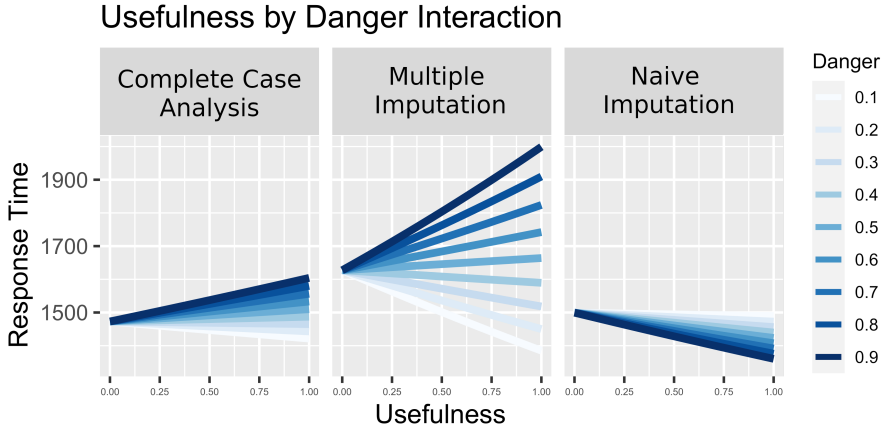


Figure 2



Leftmost Panel (Fig. 1): Number of words normed for 118 semantic features in the lexical semantics literature, ranked by number of words normed. Only a handful of semantic features have measurements matching the size of an average English speaker's lexicon.

Right Panels (Fig. 2): Interaction of word danger and usefulness on lexical decision response times in the English Crowdsourcing Project, as analyzed by a complete case analysis (left panel), after multiple imputation of danger and usefulness norms (middle panel), and after a naive imputation of danger and usefulness norms (right panel). The multiple imputation shows the predicted usefulness by danger interaction with correct functional shape, where high danger, high usefulness words yield slowed responses, but low danger, high usefulness words speed responses. This interaction is flipped and insignificant when danger and usefulness norms are imputed naively. A complete case analysis using empirical danger and usefulness norms shows an insignificant interaction of reduced magnitude.

References

Bartlett, J. W., & Morris, T. P. (2015). Multiple imputation of covariates by substantive-model compatible fully conditional specification. *The Stata Journal*, *15*(2), 437–456.

Binder, J. R., Conant, L. L., Humphries, C. J., Fernandino, L., Simons, S. B., Aguilar, M., & Desai, R. H. (2016). Toward a brain-based componential semantic representation. *Cognitive neuropsychology*, *33*(3-4), 130–174.

Hollis, G., Westbury, C., & Lefsrud, L. (2017). Extrapolating human judgments from skip-gram vector representations of word meaning. *Quarterly Journal of Experimental Psychology*, *70*(8), 1603–1619.

Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2019). The lancaster sensorimotor norms: Multidimensional measures of perceptual and action strength for 40,000 english words. *Behavior Research Methods*, 1–21.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys* (Vol. 81). John Wiley & Sons.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, *45*(3), 1–67. https://www.jstatsoft.org/v45/i03/

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, *45*(4), 1191–1207.

Wurm, L. H. (2007). Danger and usefulness: An alternative framework for understanding rapid evaluation effects in perception? *Psychonomic Bulletin & Review*, *14*(6), 1218–1225.