Neural language representations predict outcomes of scientific research

James P. Bagrow[1,2,*], Daniel Berenberg[3,2], and Joshua Bongard[3,2]

[1]Department of Mathematics & Statistics, University of Vermont, Burlington, VT, United States
[2]Vermont Complex Systems Center, University of Vermont, Burlington, VT, United States
[3]Department of Computer Science, University of Vermont, Burlington, VT, United States
[*]Corresponding author. Email: james.bagrow@uvm.edu, Homepage: bagrow.com

May 17, 2018

Abstract   Many research fields codify their findings in standard formats, often by reporting correlations between quantities of interest. But the space of all testable correlates is far larger than scientific resources can currently address, so the ability to accurately predict correlations would be useful to plan research and allocate resources. Using a dataset of approximately 170,000 correlational findings extracted from leading social science journals, we show that a trained neural network can accurately predict the reported correlations using only the text descriptions of the correlates. Accurate predictive models such as these can guide scientists towards promising untested correlates, better quantify the information gained from new findings, and has implications for moving artificial intelligence systems from predicting structures to predicting relationships in the real world.

1   Introduction

One of the most important applications of machine learning is its ability to replace data that are difficult, expensive or dangerous to collect with predictions generated using more amenable, economical or ethical data. Examples of this include replacing unavailable socioeconomic indicators for regions that are difficult to survey with predictions made from satellite imagery [1], predicting pneumonia mortality and hospital readmission [2, 3], inferring cancer risk from medical imagery or histology results [4, 5, 6]. Predictions learned from data can also guide the scientific discovery process. Some examples include predicting novel chemical reactions from previously collected experimental data [7], classifying quasar candidates and estimating stellar parameters from photometric data [8, 9], or particle discovery from high-energy physics experiments [10].

Correlational findings underpin significant portions of the scientific enterprise, and inform statistical methods, study design and meta-analyses [11]. Reported correlations are often the primary outcome of an experimental or empirical study and capture a field's ability to measure and explain the relationships they study. Findings are often presented
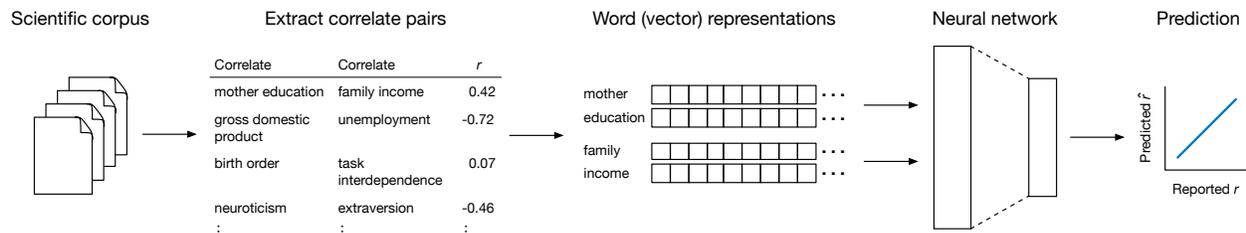
1

Figure 1: Predicting correlational findings from written descriptions only. Using a dataset of correlate pairs extracted from published manuscripts [11, 12] and pretrained vector representations encompassing semantic knowledge of individual words [17], we train a recurrent neural network to predict correlations as a function of correlate text.

as numeric tables along with *metadata*, the written descriptions of the correlates. How much information is present in these written descriptions? Can we quantify the relationships captured by the scientific record by predicting the correlation from the correlates only? The ability to make actionable predictions would be invaluable to deriving new information from metadata contained within the scientific record, identifying correlations unreported in the literature, and thus perhaps providing a guide for future studies.

Our research process aiming to address these questions is illustrated in Fig. 1. We study a longitudinal dataset of approximately 170,000 correlate pairs, extracted from 30 years of research findings published in leading social science journals [11, 12]. These data provide the text descriptions of each correlate and the reported correlation. We then train a recurrent neural network to predict a correlation $\hat{r}$ given only the ordered text sequences describing two correlates. While predictions can be made directly from words (tokens) taken from the correlates, the current practice in natural language processing is to embed tokens into a learned high-dimensional vector space [13, 14, 15]. These vector representations of the tokens allow increased flexibility, for example by handling typos, and capture language syntax and semantics in a computationally useful manner [16]. The recurrent neural network then learns relationships between sequences of these vectors that relate to the reported correlation. Full details on text processing, network architecture, hyperparameters, and training are given in the Supporting Material.

Our neural network uses vector representations of words that are computed in an unsupervised manner from large-scale text corpora [14]. As such, these representations will reflect any biases latent to those text, such as racial or gender bias [18, 19]. This is especially problematic as most training corpora are gathered from uncurated text sources, typically web crawls known to be biased in a number of dimensions [19]. Given these biases, it is not clear how well such representations can support predictions of objective research findings. Therefore, with these issues in mind, we utilize a pretrained vector representation called ConceptNet Numberbatch. Numberbatch is an ensemble of multiple, state-of-the-art representations that has been further enhanced in two important ways: one, the representations are endowed with information from the ConceptNet knowledge graph, a long-running project to codify objective relationships between entities [20]; and two, Numberbatch is explicitly trained to perform well at
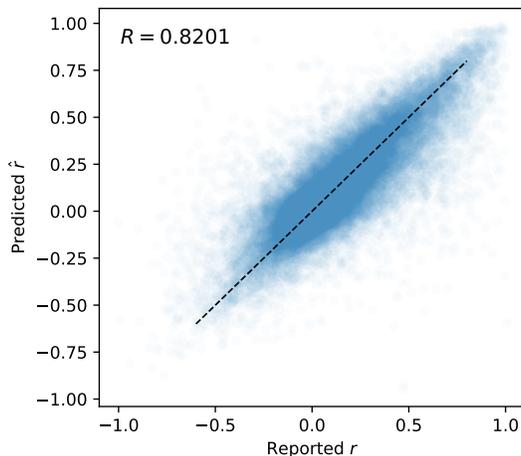
Figure 2: Comparison of reported and predicted correlations for correlate pairs reserved for testing. Predicted values $\hat{r}$ show a strong correlation with reported values r: $R(r, \hat{r}) \approx 0.8201$. The dashed line indicates $r = \hat{r}$ and provides a guide for the eye.

natural language tasks while also minimizing a number of bias indicator scores, such as the implicit association test. Numberbatch is currently the most competitive and least biased vector representation available to researchers, making it the most suitable choice for our task [17].

2   Results

To evaluate our ability to predict correlational findings, we trained the neural network on a random 80% of the reported findings and reserved 20% for testing purposes. After training, the neural network was asked to predict the held-out correlations. As shown in Fig. 2, the model achieves accurate predictions, giving a correlation $R \approx 0.8201$ between reported r and predicted $\hat{r}$.

Given the accuracy demonstrated in Fig. 2, it is important to ask if the predictive model is learning meaningful relationships between correlates or if it is simply memorizing features implicit to the training corpus. Memorization harms the ability of an algorithm to make accurate predictions. While held-out or test data is the gold standard for evaluating predictive performance, and we used test data in Fig. 2, it is worthwhile to examine this issue further. To do so, we introduce a comparative baseline predictive model: for any pair of correlates c_i and c_j we simply predict the mean correlation reported for any correlate pairs in the corpus that contain either c_i or c_j. If the neural network is only memorizing the corpus, we expect this mean-value baseline to achieve comparable predictive performance. Instead, the baseline performs significantly worse, with $R \approx 0.54$ (see SM). This indicates that the neural network is learning meaningful representations and possesses predictive performance beyond this basic mean-value model.

The ability to make accurate predictions of correlational findings has many potential applications. Here we discuss
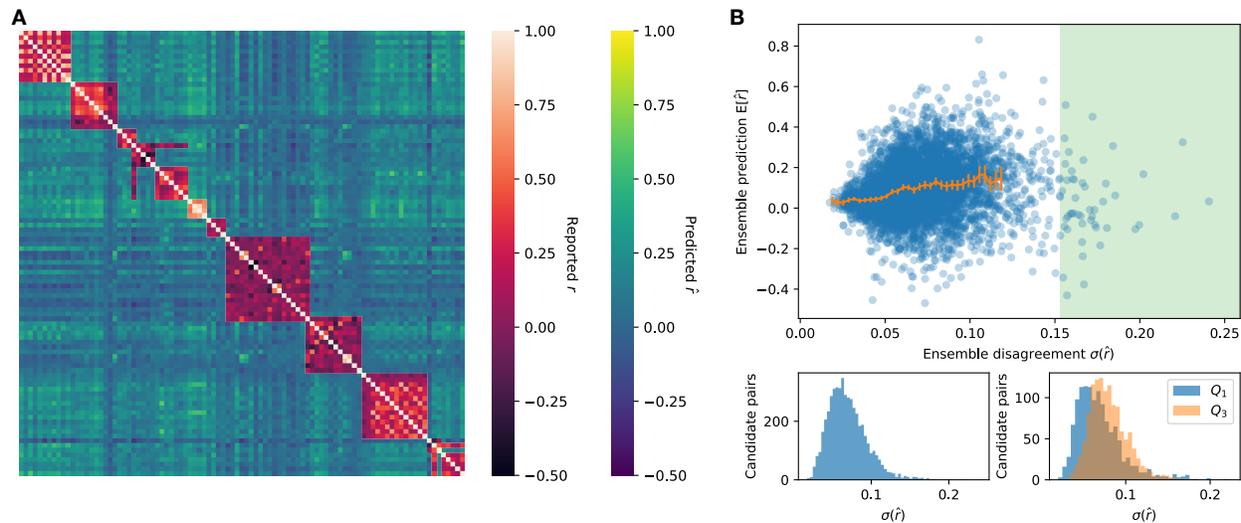
Figure 3: Applications of correlation prediction. (**A**) Infilling a multi-paper correlation table. Reported correlations for 10 papers are shown along the diagonal in one color scale, predictions for unreported correlations in another. Here 88.2% of the table has been infilled by the predictive model. (**B**) Finding candidate correlate pairs worth further investigation. Candidate pairs with disagreement across an ensemble of predictive models are most worth testing as they will lowering that disagreement will yield the most information. The highlighted region denotes the 1% most uncertain candidate pairs. We find higher disagreement for candidate pairs with more positive predicted correlation.

two: (i) infilling a partial correlation table to interrelate a large set of correlates and (ii) discovering untested correlate pairs that are worth further investigation.

Figure 3A shows a correlation table connecting ten papers published in 2010. The blocks along the diagonal of the table denote the reported findings of the individual papers; note that the third and tenth paper have some gaps as they do not report all correlations. The table entries outside the blocks of reported findings and within the gaps of incomplete papers are those infilled by our predictive model. These entries constitute 88.2% of the table. From this infilled table, we see that the predictive model is able to find meaningful connections across papers, in particular the first three papers and the last two papers show interesting correlations. These infilled relationships cannot be explained by the mean-value baseline model only (see SM). Researchers interested in topics bridging existing research can now receive preliminary guidance towards interesting convergence points.

Second, we ask if the model can find untested correlate pairs that are worth devoting resources into testing. The ability to guide research ahead of time is valuable as the space of possible relationships is very large. Indeed, most correlate pairs in our corpus are untested: of the 21,736 unique correlates recorded in the corpus (postprocessing), only 149,374 unique pairs have been tested, leaving 99.94% untested. Of course, many of these untested pairs will not be worthwhile or suitable to investigate, or have been tested in publications outside our corpus, yet there remain many correlate pairs to be studied and any guidance can help maximize limited research time.

Correlate pairs most worth further research are those that will provide the most novel information. A classic approach to measuring information gain is from the prediction disagreement across an ensemble of predictive models: observations where the members of the ensemble give mostly the same prediction provide less novel information than observations where members give diverse predictions. The latter case highlights valuable observations to learn from because the ensemble is more uncertain and gaining access to the correlation the ensemble is most uncertain about will yield the most novel information. Seeking labels for training data via ensemble disagreement is an active learning technique known as *Query by Committee* [21, 22].

We apply ensemble disagreement to highlight untested correlate pairs that may yield the most information when tested. We trained an ensemble of $N = 50$ neural networks (see SM for details) and then performed a simple random search: correlates were paired at random into $n = 5000$ candidate pairs and sent to each member of the network ensemble. For each candidate correlate pair we measured the ensemble prediction $E[\hat{r}]$ and the ensemble disagreement $\sigma(\hat{r})$, where $E[\cdot]$ and $\sigma(\cdot)$ denote the sample mean and sample standard deviation, respectively, taken over the predictions of the N predictive models.

Figure 3B shows the results of this search. Correlate pairs with low ensemble disagreement are well understood by the model ensemble, while those with high disagreement are not. Query-by-committee argues that those correlate pairs that have high disagreement are most worth investigation, even if the predicted correlation is low, as those are most likely to provide novel information gain if tested. We highlight the 1% most uncertain candidate pairs in Fig. 3B. Examining these candidate correlate pairs (see SM), several interesting pairs appear (paraphrased here): 'Gross domestic product' and 'job search behavior (unemployed group)' had a predicted correlation of $\hat{r} = E[\hat{r}] \pm 1.96 \sigma(\hat{r}) / \sqrt{N} = -0.37 \pm 0.046$. Two more pairs with strong predicted correlations were 'Manipulation check' and 'Hardworking' ($\hat{r} = 0.45 \pm 0.048$) and 'Overall handshake' and 'Feels superior' ($\hat{r} = 0.34 \pm 0.043$). These pairs are quite sensible: manipulation checks [23] are a common approach to filter out study participants who are not performing a task well, and handshakes and body language are often associated with feelings of superiority. Another pair, 'Standard Occupational Classification' and 'Perceived interactional justice' ($\hat{r} = -0.43 \pm 0.044$) may highlight some research directions relating how people can interact positively with co-workers. Lastly, the pair 'Organizational embeddedness' and 'Novice completion time' ($\hat{r} = -0.04 \pm 0.061$) shows a marginal correlation effect but the high degree of disagreement may demonstrate the need for more research on how organizational structure influences onboarding of new employees.

Interestingly, there is a modest but statistically significant trend between ensemble disagreement and ensemble prediction. As the distribution of correlations reported in the literature is skewed in favor of positive correlations, we anticipated either no relationship between $E[\hat{r}]$ and $\sigma(\hat{r})$, or more uncertainty among the lower and negative correlations. However, the positive linear trend was significant ($R(E[\hat{r}], \sigma(\hat{r})) = 0.164, p < 10^{-30}$) and examining the candidate

pairs in the lower quartile Q_1 of $E[\hat{r}]$ compared with those in the upper quartile Q_3, we see a significant difference in disagreement (Mann-Whitney U test, $U = 531538, p < 10^{-43}$). We illustrate these distributions in Fig. 3B.

Our Query-by-committee style application is based upon a simple random search of the large space of untested correlate pairs. It is unlikely that many strong correlations will be found from randomly pairing correlates. Indeed, most predicted correlations were modest, with the distribution of $E[\hat{r}]$ peaked at ≈ 0 (Fig. 3B). We anticipate that more effective searches of the space of untested correlations can be utilized. That said, a number of plausible candidate correlate pairs were still found and this naive computational search is inexpensive and scalable relative to the resources required to conduct direct scientific research.

3 Discussion

The ability to predict correlational relationships from written descriptions has implications for the larger research problem of artificial intelligence. Previous research has shown ways for AI systems to synthesize and predict physical relationships [24, 25, 26]. This now extends to social, economic, and other characteristic phenomena, and our work can help to further ground AI systems in the real world.

The volume of published findings continues to grow and automated tools are increasingly necessary to help scientists navigate the scientific record. Our findings underscore the importance of data curation and standardized reporting formats. As more published findings become computationally accessible, due to better data curation and advances in natural language processing, more of the scientific record becomes available for planning and executing research. Research areas that follow reporting standards will benefit the most from computational tools examining their publication record as standardization simplifies the process of extracting large training corpora.

Lastly, we caution that a predictive model such as the one proposed here can fruitfully serve as guidance to researchers conducting scientific investigations, but it is no replacement for those investigations. Proposing and falsifying scientific hypotheses remains the gold standard of science and cannot be replaced by these models. Instead, predictive models complement experiments and empirical findings by codifying the current state of the scientific record and providing helpful tools for researchers to handle the growth of this record.

Acknowledgments

1447634.

References

[1] N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon, "Combining satellite imagery and machine learning to predict poverty," *Science*, vol. 353, no. 6301, pp. 790–794, 2016. 1

[2] G. F. Cooper, C. F. Aliferis, R. Ambrosino, J. Aronis, B. G. Buchanan, R. Caruana, M. J. Fine, C. Glymour, G. Gordon, B. H. Hanusa, *et al.*, "An evaluation of machine-learning methods for predicting pneumonia mortality," *Artificial intelligence in medicine*, vol. 9, no. 2, pp. 107–138, 1997. 1

[3] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1721–1730, ACM, 2015. 1

[4] D. C. Cireşan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Mitosis detection in breast cancer histology images with deep neural networks," in *International Conference on Medical Image Computing and Computer-assisted Intervention*, pp. 411–418, Springer, 2013. 1

[5] A. Cruz-Roa, A. Basavanhally, F. González, H. Gilmore, M. Feldman, S. Ganesan, N. Shih, J. Tomaszewski, and A. Madabhushi, "Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks," in *Medical Imaging 2014: Digital Pathology*, vol. 9041, p. 904103, International Society for Optics and Photonics, 2014. 1

[6] J. Xu, X. Luo, G. Wang, H. Gilmore, and A. Madabhushi, "A deep convolutional neural network for segmenting and classifying epithelial and stromal regions in histopathological images," *Neurocomputing*, vol. 191, pp. 214–223, 2016. 1

[7] P. Raccuglia, K. C. Elbert, P. D. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier, and A. J. Norquist, "Machine-learning-assisted materials discovery using failed experiments," *Nature*, vol. 533, no. 7601, p. 73, 2016. 1

[8] G. T. Richards, R. C. Nichol, A. G. Gray, R. J. Brunner, R. H. Lupton, D. E. V. Berk, S. S. Chong, M. A. Weinstein, D. P. Schneider, S. F. Anderson, *et al.*, "Efficient photometric selection of quasars from the Sloan Digital Sky Survey: 100,000 $z < 3$ quasars from Data Release One," *The Astrophysical Journal Supplement Series*, vol. 155, no. 2, p. 257, 2004. 1

[9] P. R. Fiorentin, C. Bailer-Jones, Y. S. Lee, T. C. Beers, T. Sivarani, R. Wilhelm, C. A. Prieto, and J. Norris, "Estimation of stellar atmospheric parameters from SDSS/SEGUE spectra," *Astronomy & Astrophysics*, vol. 467, no. 3, pp. 1373–1387, 2007. 1

[10] P. Baldi, P. Sadowski, and D. Whiteson, "Searching for exotic particles in high-energy physics with deep learning," *Nature communications*, vol. 5, p. 4308, 2014. 1

[11] F. A. Bosco, H. Aguinis, K. Singh, J. G. Field, and C. A. Pierce, "Correlational effect size benchmarks," *Journal of Applied Psychology*, vol. 100, no. 2, p. 431, 2015. 1, 2

[12] F. A. Bosco, P. Steel, F. L. Oswald, K. Uggerslev, and J. G. Field, "Cloud-based meta-analysis to bridge science and practice: Welcome to metabus," *Personnel Assessment and Decisions*, vol. 1, no. 1, p. 2, 2015. 2

[13] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th international conference on Machine learning*, pp. 160–167, ACM, 2008. 2

[14] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, pp. 3111–3119, 2013. 2

[15] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014. 2

[16] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *Journal of machine learning research*, vol. 3, no. Feb, pp. 1137–1155, 2003. 2

[17] R. Speer, J. Chin, and C. Havasi, "ConceptNet 5.5: An open multilingual graph of general knowledge," in *AAAI Conference on Artificial Intelligence*, pp. 4444–4451, 2017. 2, 3

[18] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings," in *Advances in Neural Information Processing Systems*, pp. 4349–4357, 2016. 2

[19] A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science*, vol. 356, no. 6334, pp. 183–186, 2017. 2

[20] R. Speer and C. Havasi, "Representing general relational knowledge in ConceptNet 5," in *LREC*, pp. 3679–3686, 2012. 2

[21] H. S. Seung, M. Opper, and H. Sompolinsky, "Query by committee," in *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 287–294, ACM, 1992. 5

[22] A. Krogh and J. Vedelsby, "Neural network ensembles, cross validation, and active learning," in *Advances in neural information processing systems*, pp. 231–238, 1995. 5

[23] D. M. Oppenheimer, T. Meyvis, and N. Davidenko, "Instructional manipulation checks: Detecting satisficing to increase statistical power," *Journal of Experimental Social Psychology*, vol. 45, no. 4, pp. 867–872, 2009. 5

[24] J. Bongard, V. Zykov, and H. Lipson, "Resilient machines through continuous self-modeling," *Science*, vol. 314, no. 5802, pp. 1118–1121, 2006. 6

[25] J. Bongard and H. Lipson, "Automated reverse engineering of nonlinear dynamical systems," *Proceedings of the National Academy of Sciences*, vol. 104, no. 24, pp. 9943–9948, 2007. 6

[26] M. Schmidt and H. Lipson, "Distilling free-form natural laws from experimental data," *Science*, vol. 324, no. 5923, pp. 81–85, 2009. 6