Comparing Natural Language and Adaptive Querying Approaches For Estimating Semantic Similarity Structure

April Murphy, Chris Cox, Kevin Jamieson, Rob Nowak, Tim Rogers
University of Wisconsin-Madison

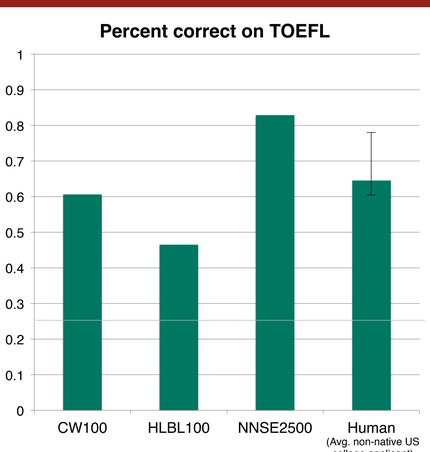
Introduction

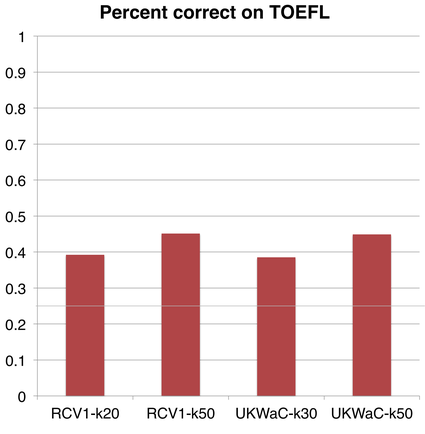
Neuroimaging studies which use multi-voxel pattern analysis to predict neural representations associated with word meanings seek to identify patterns of activation containing categorical information representing equivalence classes such as tools and animals. To accomplish this, a reliable estimate of semantic similarity is required as input to the MVPA model. Because obtaining thousands of individual high-quality human judgments of word similarity is extremely labor-intensive, this problem has been addressed most recently by using natural language processing (NLP) methods which measure word co-occurrence statistics in large text corpora. A challenge with these models, however, is that performance can be difficult to evaluate and may vary widely. Moreover, similarity estimations for abstract words are much worse than for concrete words.

Our study had the following goals:

- Identify the best NLP model based on traditional performance metrics
- Evaluate this model against human judgments on a set of abstract words which lack discrete equivalence relations
- Compare NLP models against a new, efficient method of estimating similarity using human judgments

NLP Methods





We first looked at three publicly-available NLP methods which have claimed strong performance on semantic tasks. Performance was measured using the TOEFL synonym judgment test, a 4-alternative multiple choice test which includes many abstract words and is a common performance metric for semantic models[3].

Example of a TOEFL question:

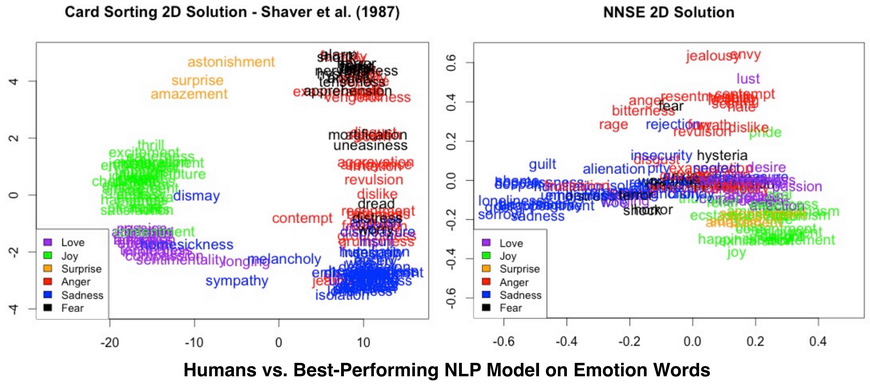
ZENITH

a) Completion	c) Outset
b) Pinnacle	d) Decline

Given that research in lexical semantics[1] has shown that word co-occurrence models are capable of achieving 100% accuracy on the TOEFL, we also ran several models in-house using multiple corpora to determine if we could uncover key optimization parameters to approach the TOEFL benchmark. While all models scored better than chance, none attained optimal performance.

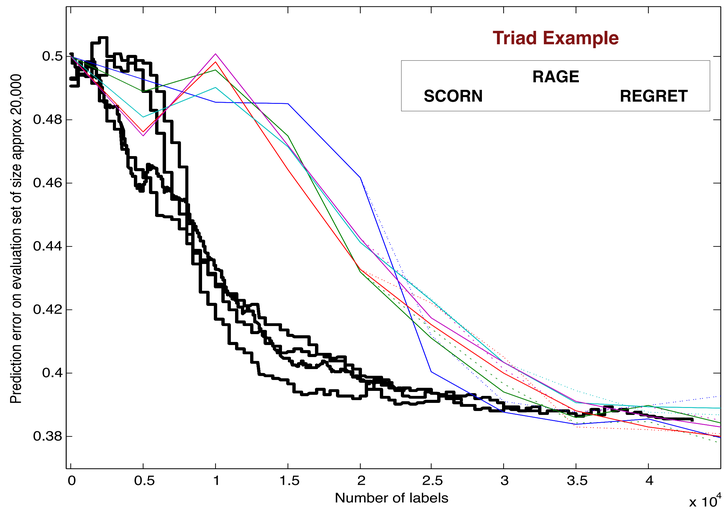
NLP vs. Humans

The best-performing NLP model was NNSE[4], scoring 83% on the TOEFL. To evaluate NNSE against human judgments, we used data from a card-sorting task of emotion words conducted by Shaver et al.[5] Subjects sorted similar words into piles and pairwise comparisons were applied to identify fixed categories using multidimensional scaling. Compared with human subjects in the Shaver et al. task, and despite very high TOEFL performance, NNSE had difficulty estimating emotion word similarity. This suggests that similarity structure associated with emotion knowledge may be particularly difficult to model using NLP, and raises questions over the use of such models for predicting patterns of neural activity for other abstract concepts.

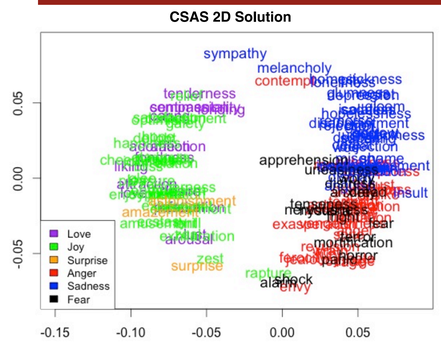


CSAS Method

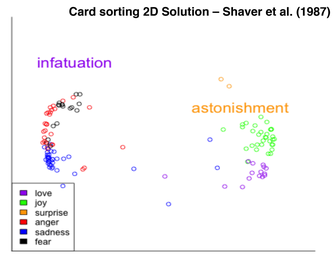
Crowd-Sourced Adaptive Sampling (CSAS) uses a new algorithm to optimize embeddings for non-metric data. Assuming that similarity structure resides in a low-dimensional space, the algorithm learns *d*-dimensional embeddings from human subjects by adaptively selecting queries in the form of simple triads. Triad optimization is based on all prior observed responses, and adjusted on-line to minimize the overall number of queries needed to define the embedding. Prediction error using CSAS triads (black lines) drops rapidly compared with random triads (colored lines).

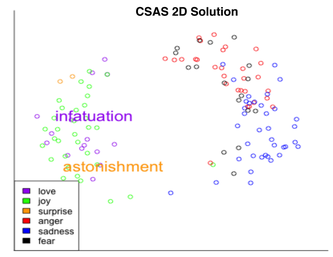


CSAS Validation

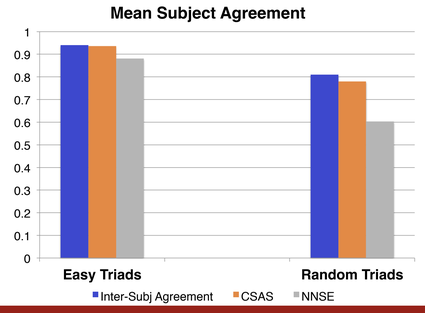


We collected data from 229 subjects using CSAS and evaluated all three methods according to mean subject agreement with the prediction. 33 additional subjects judged a set of random triads, triads where model predictions agreed (“easy”), and where the human data (Shaver et al.) and CSAS models conflicted (“discriminating”).





Average subject agreement with CSAS on conflicting triads was 88%. Subject agreement with CSAS was 94% on easy triads; with NNSE, 88%. Agreement with CSAS on random triads was 78%, compared with 60% with NNSE. Inter-subject agreement was 94% on easy and 81% on random triads.



Summary

Crowd-Sourced Adaptive Sampling (CSAS) provides high-quality estimates of similarity relationships among emotion words, using a relatively low number of observations to reach a solution. Compared with similarity measures from a classic card-sorting task and a state-of-the-art NLP model, humans agreed more with CSAS predictions than either the card-sort or NLP estimates. Given the present results, we suggest that these measures can be appropriated to better guide multi-voxel pattern analyses of neural semantics in future applications.

References

1. Bullinaria, J. A., & Levy, J. P. (2013). Limiting Factors for Mapping Corpus-Based Semantic Representations to Brain Activity. *PloS one*, 8(3), e57191.

2. Jamieson, K. G., & Nowak, R. D. (2011, September). Low-dimensional embedding using adaptively selected ordinal data. In Communication, Control, and Computing (Allerton), 2011 49th Annual Allerton Conference on (pp. 1077-1084). IEEE.

3. Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2), 211.

4. Murphy, B., Talukdar, P. P., & Mitchell, T. (2012). Learning Effective and Interpretable Semantic Models using Non-Negative Sparse Embedding. In *COLING* (pp. 1933-1950).

5. Shaver, P., Schwartz, J., Kirson, D., & O'connor, C. (1987). Emotion knowledge: further exploration of a prototype approach. *Journal of Personality and Social Psychology*, 52(6), 1061.

6. Turian, J., Ratinov, L., & Bengio, Y. (2010, July). Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 384-394). Association for Computational Linguistics.

Contact aprildmurphy@gmail.com for questions and comments