Computational Models of Neural Representations in the Human Brain

Tom M. Mitchell

Carnegie Mellon University, Pittsburgh, PA 15213, USA
tom.mitchell@cmu.edu
http://www.cs.cmu.edu/ tom

Extended Abstract

For many centuries scientists have wondered how the human brain represents thoughts in terms of the underlying biology of neural activity. Philosophers, linguists, cognitive scientists and others have proposed theories, for example suggesting that the brain organizes conceptual information in hierarchies of concepts, or that it instead represents different concepts in different local regions of the cortex.

Over the past decade rapid progress has been made on the study of human brain function, driven by the advent of modern brain imaging methods such as functional Magnetic Resonance Imaging (fMRI), which is able to produce three dimensional images of brain activity at a spatial resolution of approximately one millimeter. Using fMRI we have spent several years exploring the question of how the brain represents the meanings of invididual words in terms of patterns of neural activity observed with fMRI. The talk accompanying this abstract will present our results, and the use of machine learning methods to analyze this data and to develop predictive computational models. In particular, we ([1],[2]) have explored the following questions:

- *Can one observe differences in neural activity using fMRI, as people think about different items such as "hammer" versus "house"?* Many researchers have now demonstrated that fMRI does indeed reveal differences in neural activity due to considering different items. We present results [1] showing that it is possible to train a machine learning classifier to discover the different patterns of activity associated with different items, and to use this to classify which of several items a person is considering, based on their neural activity.
- *Are neural representations of concepts similar if the stimulus is a word, versus a line drawing of the object?* We tested this question by asking whether a machine learning classifier trained on fMRI data collected when a person reads words, could successfully distinguish which item they were thinking about when the stimuli were line drawings. The classifier performed nearly as accurately classifying fMRI activity generated by line drawing stimuli as by word stimuli, despite being trained on word stimuli. This result suggests that the neural activity captured by the classifier reflects the semantics of

J.-F. Boulicaut, M.R. Berthold, and T. Horváth (Eds.): DS 2008, LNAI 5255, pp. 26–27, 2008.

the item, and not simply some surface perceptual features associated with the particular form of stimulus.

- *Are neural representations similar across different people?* We tested this question by asking whether a machine learning classifier trained on fMRI data collected from a group of people, could successfully distinguish which item a new person was thinking about, despite the fact that the classifier had never seen data from this new person. These experiments were performed for stimuli corresponding to concrete nouns (i.e., nouns such as "bicycle" and "tomato" which describe physical objects). We found the answer is yes, although accuracies vary by person. This result suggests that despite the fact that invididual people are clearly different, our brains use similar neural encodings of semantics of concrete nouns.
- *Can we discover underlying principles of neural representations sufficient to develop a computational model that predicts neural representations for arbitrary words?* We recently developed a computational model that predicts the neural representation for any concrete noun. While imperfect, this model performs well on the 100 words for which we have data to test it. The model is trained using a combination of fMRI data for dozens of words, plus data from a trillion word text corpus that reflects the way in which people typically use words in natural language. This model represents a new approach to computational studies of neural representations in the human brain.

Acknowledgments. The research summarized here is the product of a large team of collaborators, including Marcel Just, Andy Carlson, Kai-Min Chang, Rebecca Hutchinson, Vicente Maleve, Rob Mason, Mark Palatucci, Francisco Pereira, Indra Rustandi, Svetlana Shinkareva, Wei Wang and others. We are grateful for support from the W.M. Keck foundation and the National Science Foundation.

References

[1] Shinkareva, S., Mason, R., Malave, V., Wang, W., Mitchell, T., Just, M.: Using fMRI Brain Activation to Identify Cognitive States Associated with Perception of Tools and Dwellings. PLoS ONE 3(1) (2008); e1394. doi:10.1371/journal.pone.0001394

[2] Mitchell, T., Shinkareva, S., Carlson, A., Chang, K., Malave, V., Mason, R., Just, M.: Predicting Human Brain Activity Associated with the Meanings of Nouns. Science 320, 1191 (2008)