Rajeev Raizada: Statement of research interests

Using machine learning to study how the brain represents the world

My research applies machine learning to fMRI data, in order to try to understand how the brain's representations are structured, and how they give rise to behaviour. Recently my work, funded by an NSF CAREER award, has concentrated especially on the question of how the brain represents the meanings of words and sentences. These studies, described in more detail below, have aimed to uncover *how* linguistic information is encoded in the brain, rather than just localising *where* in the brain such processing takes place.

Brain imaging to uncover neural representations, not just "what lit up"

A caricature of fMRI studies, sadly sometimes all too deserved, is that they merely give us pretty pictures of brain activation, with some regions brightly coloured-in to show us "what lit up". Such pictures look nice in newspaper articles, but they do not, in themselves, tell us anything interesting about how the brain is actually processing or representing information.

Fortunately, neuroimaging can go far beyond that. A crucial step has been to apply computational methods from machine learning to brain data, seeking to extract information from activation patterns that are spatially distributed across multiple voxels (a "voxel" is 3D volumetric pixel). This multivariate approach is quite different from how fMRI analysis has traditionally been done: such analyses are univariate, with a statistical model fitting each voxel's response separately.

Those individual voxel activations can only go up or down, so the outputs of standard univariate analysis are best suited for localising the spots where that activation is most intense. In contrast, a multivariate approach examines the relations between distributed multivoxel patterns of activation. These multi-dimensional activations do not merely vary in intensity, but instead have varying degrees of pattern-similarity with each other. This neural similarity-structure provides us with an tool to probe the structure of neural representations.

All this talk of "neural representations", however, requires us to address the following question:

What is a "neural representation", anyway?

This is a deep and difficult question, but in the absence of a complete answer, we can adopt a working operational definition: a neural representation exists whenever information about the world somehow ends up inside someone's head. So, if we can measure patterns of brain activation and extract from those patterns some information about the world, then we have found evidence that this information is neurally represented. As a concrete example, we might be able to measure someone's brain activation while they are looking at either a cat or a chair, and be able to infer from that with above-chance accuracy which of those possible objects they are actually looking at.

Note that we still would not know whether the brain actually *uses* the information that we have been able to extract: that would require finding causal links to behaviour. Nor would we know what "format" the brain uses for representing the information. We would instead only have found evidence that the information is present. Despite these difficult challenges, even just finding that information is extractable from activation patterns is a marked advance beyond just localising activation hot-spots.

A crucial role for machine learning

Human brain data is high-dimensional and noisy. In the case of fMRI, it consists of a time series of brain images, each containing tens of thousands of voxels. Distributed spatial patterns of brain

activation, spread across those thousands of voxels, contain meaningful information about what that person's brain is thinking about, and how that information is being used to guide their behaviour. The challenge is to find those meaningful patterns from all the many thousands of voxels, and amongst all the noise. Fortunately, algorithms from machine learning are able to do precisely that, very imperfectly but nonetheless with above-chance success. My research applies machine learning to fMRI data, in order to try to understand how neural representations are structured, and how they give rise to behaviour.

Linguistic meaning is full of richly structured representations. One illustration is the set of semantic similarity relations between different words. For example, a dog is more similar to a cat than it is to a chair, and correspondingly the meaning of the word 'dog' is more similar to that of 'cat' than it is 'chair'. It turns out that the same similarity relations are found in the brain's representations: when a person reads those three words, the neural patterns produced by 'dog' and 'cat' are more similar to each other than they are to the pattern produced by 'chair'. (The measurement of neural similarity here can be as simple as just the spatial correlation between the multivoxel patterns.) Moreover, these semantic similarity relations can also be captured by and compared against computational semantic models, as is discussed in more detail below.

Similarity-based decoding of meaning across languages

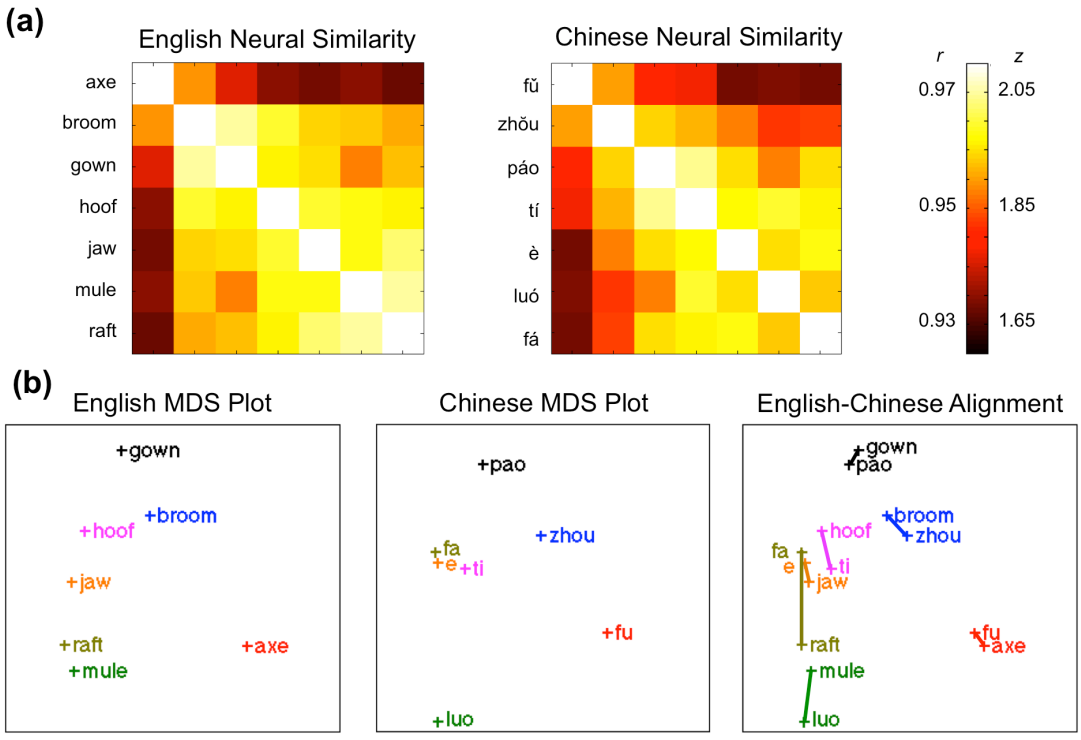

Figure 1: Our demonstration of neurally-mediated cross-language translation between English and Chinese native speakers, achieved by structural alignment of neural similarity spaces. The upper two panels show neural similarity matrices from the English and Chinese groups. The lower panels show multidimensional scaling plots of those neural similarity spaces. The bottom right plot shows the word-pairing that best aligns the neural similarity structures from the two speaker groups. It can be seen that all seven English words are successfully paired with their Chinese translations. From Zinszer et al. (2016).

The fact that neural and semantic similarity correlate with each other is interesting, but it does not in itself allow us to decode neural activity, i.e. to take an unlabeled neural activation pattern and to

infer its representational content. I developed a new similarity-based decoding approach that allows precisely that (Raizada and Connolly, 2012; Anderson et al., 2016b). One illustration of it in action is a study from my lab in which my postdoc and I were able to translate words between two different languages, purely by aligning neural similarity spaces (Zinszer et al., 2016).

In that study, native Chinese-speakers were presented with words in Chinese, and native English speakers read the corresponding English-translation words. The only content in common across the two language groups was purely semantic, as the two languages are as orthographically and phonologically different as can be: Chinese words neither look, nor sound, at all like their English counterparts. In that study, we demonstrated that by matching the neural similarity structure of elicited brain activation across the two groups we were able to deduce the corresponding semantic matches. In other words, we could translate between English and Chinese words using only neural activation as our guide.

From neural decoding of words to decoding of entire sentences

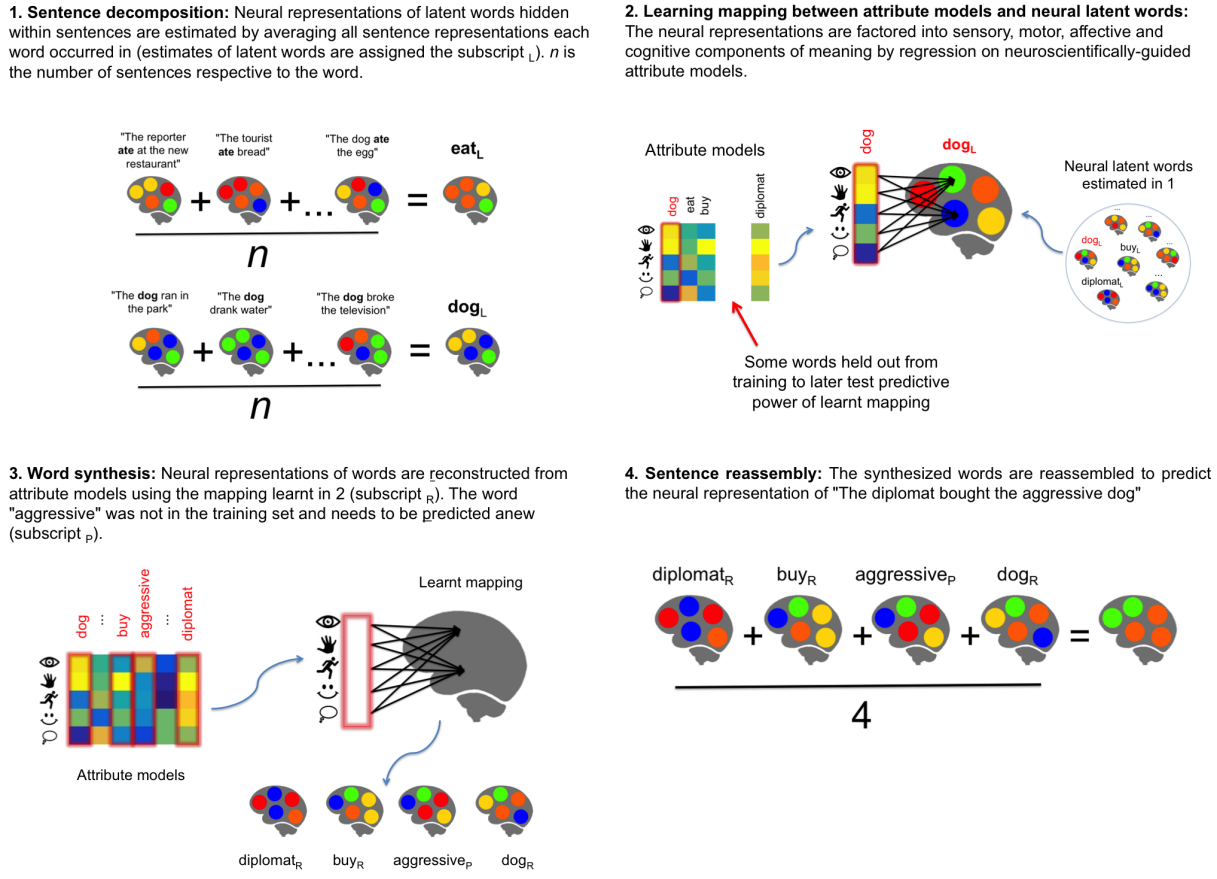

Figure 2: An overview of our approach for neurally decoding entire sentences, from Anderson et al. (2016a).

Building upon our success in neurally decoding individual words, my lab embarked on a project to decode entire sentences. This was obviously a much more challenging problem, as the neural representation of a sentence contains the semantic content of all its constituent words. Although fMRI has the good spatial resolution that is essential for fine-grained decoding of this sort, its temporal resolution is poor, so the measured fMRI signal mixes representations of all the sentence's words together.

It is often worth trying a simple and direct approach first, even if it is too simple to do justice to the full richness of a problem. This is even more true in a field as packed with unknowns as the neural decoding of linguistic meaning. If the simple approach turns out to work to some degree, then refinements can be added in follow-up work. But without an initial partial success, the question of refinements does not even arise.

That is the approach that I took in my sentence decoding work. The overall architecture of the method is illustrated in Figure 2. We found that semantic models of individual words can be arithmetically combined to predict the neural activity associated with sentences (to a first approximation). E.g. the sentence "the cat sat on the mat" can be predicted by combining vector models of 'cat' + 'sit' + 'mat', and the prediction is more accurate when all three words are combined as compared to when made using a subset of the words such as just 'cat' + 'mat'. The resulting paper, Anderson et al. (2016a), was the first published study to decode entire sentences.

Combining different types of semantic models

Quantitative models of word meaning have a long history, both in cognitive psychology and in computer science. There are two main sets of approaches. Both of them represent words as vectors of numbers, with each such number being the strength of a particular semantic feature. However, the two approaches produce those feature vectors in very different ways.

One approach, often called text-based semantic models or distributional semantics, takes a large body, or corpus, of text and calculates how often different words co-occur with each other. For example, the word 'apple' co-occurs more frequently with the word 'eat' than it does with "drive', corresponding to the fact that an important part of the meaning of 'apple' is that it is something to do with eating. With today's powerful computers and huge amounts of available text data, such corpora can be vast, e.g. Google's one trillion word text corpus.

Another approach, more rooted in cognitive psychology, asks people to give their ratings of different aspects of a word's meaning. For example, the participants might be asked to rate the degree to which "apple" strikes them as having various sensory properties: how much does it have properties of flavour (a lot), shape (quite a lot), sound (not much), motion (also not much), and so on. Approaches of this sort are often referred to as feature norms, or experiential semantic models.

Of these two approaches, the text-based approach is much more widely used, not only in computational linguistics, but also in the vast majority of fMRI language decoding studies. It has many points in its favour: the available text corpora are huge, and it avoids the time-consuming business of asking people to give behavioural ratings.

However, there are also some important potential advantages of using experiential semantic features. Unlike in text-based approaches, people assess the properties of words directly, rather than relying on word co-occurrence frequencies. Although co-occurrences contain useful information, they can faces problems distinguishing between semantic similarity and semantic association.

This is best illustrated by an example: the words 'coffee' and 'cup' often co-occur together. Semantic vectors built up out of their co-occurrence frequencies will therefore tend to be similar. However, any person (or computational system) which thinks that 'coffee', and 'cup' mean more or less the same thing will quickly run into trouble at the local Starbucks. In short, those words are semantically associated, but they have different meanings. Many other examples of related but distinct meanings can easily be found: one would not want to confuse 'car' with 'driver', 'needle' with 'thread' or 'doctor' with 'medicine'.

In recently published work (Anderson et al., 2019), we carried out the first neural decoding of sentences that combined both types of semantic model. We found that the text-based and experiential models when used together performed better than either model on its own. More specifically, the text-based model was especially helpful for decoding abstract words from the sentences, and the experiential models aided the decoding of concrete words. This work shows that although computational text-based models are dominant in terms of how widely they are used, there is still much to be gained by asking actual people for their behavioural ratings of word meaning.

Open questions

The work sketched out above shows how model-based fMRI decoding can yield rich insights into how the brain represents linguistic meaning. However, they only scratch the surface of the astonishing richness and complexity of language. A great many questions remain open. Here are a few:

How does the brain represent syntactic structure and word order? Current methods for fMRI decoding are unable to know the difference between "Mary loves John" and "John loves Mary". The brain, however, obviously knows this difference. But how does it represent that?

How might recently developed machine learning approaches be used for neural decoding? Major advances have recently been made in machine translation, sentiment analysis, and other longstanding problems in natural language processing, most notably using recurrent neural network models. Powerful as these approaches might be, they still capture only a tiny fraction of linguistic meaning, and their inner workings are often tantalisingly opaque. With data as noisy as fMRI, simpler models often end up being the best-performing, in large part due to their resistance to overfitting. Can these more complex deep learning language models be used to improve fMRI decoding? More importantly, even if they improve the percentage-correct performance, will they help us better to understand what the brain is actually doing?

The short answer to the above questions is: at present, nobody knows. But it would be very interesting indeed to try to find out.

The rise of open data: important questions attackable with a computer and an internet connection

The studies described above involved designing and performing new fMRI scans, using an in-house fMRI scanner and paying hefty scanning fees. Until recently, this was the only way to be able to carry out fMRI research. However, in recent years, two developing trends have changed this situation dramatically.

First, scientific advances have often come from devising new and ingenious analysis techniques, typically drawing from progress in machine learning. So, it has become possible to ask new and interesting questions, using already existing data.

Second, very rich open databases of fMRI data are now available, freely downloadable for any researcher to explore and analyse. Examples include OpenNeuro.org, the Healthy Brain Initiative, the Adolescent Brain Cognitive Development (ABCD) study, the Alzheimer's Disease Neuroimaging Initiative (ADNI) and more. This last database, ADNI, was used in studies from my own lab (Wang et al., 2019a,b) which applied machine learning to fMRI data to predict Alzheimer's progression.

The existence of these open databases means that new fMRI studies can, for the first time, be carried out without actually needing access to an fMRI scanner. In one sense, of course, this is limiting, as the questions that can be asked are constrained by a data collection protocol that one did not design. However, there is another sense in which such datasets open up questions that simply cannot be asked by individual labs. Many of these datasets, such as ABCD and ADNI, contain data from thousands of

different participants, collected across multiple sites. When effects are subtle and the data is noisy, as is almost always the case in studies of the brain, very large datasets are often the only way to achieve true predictive power. A large class of important questions can be asked *only* from shared open datasets.

This raises exciting possibilities for being able to do innovative research, even when no on-site scanner is available. All that one requires are some computers, an internet connection, and informed and inquiring minds. The computational analysis techniques that are required can be, and indeed have been, learned by undergraduate students taking my courses. I am very excited by the future possibilities of such work.

References

Anderson, A. J., Binder, J. R., Fernandino, L., Humphries, C. J., Conant, L. L., Aguilar, M., Wang, X., Doko, D., and Raizada, R. D. S. (2016a). Predicting neural activity patterns associated with sentences using a neurobiologically motivated model of semantic representation. *Cereb Cortex*, 27(9):4379–4395.

Anderson, A. J., Binder, J. R., Fernandino, L., Humphries, C. J., Conant, L. L., Raizada, R. D. S., Lin, F., and Lalor, E. C. (2019). An integrated neural decoder of linguistic and experiential meaning. *J Neurosci*. Advance Online Publication. https://doi.org/10.1523/JNEUROSCI.2575-18.2019.

Anderson, A. J., Zinszer, B. D., and Raizada, R. D. S. (2016b). Representational similarity encoding for fMRI: Pattern-based synthesis to predict brain activity using stimulus-model-similarities. *NeuroImage*, 128:44–53.

Raizada, R. D. S. and Connolly, A. C. (2012). What makes different people's representations alike: neural similarity-space solves the problem of across-subject fMRI decoding. *Journal of Cognitive Neuroscience*, 24(4):868–877.

Wang, X., Ren, P., Baran, T. M., Raizada, R. D. S., Mapstone, M., Lin, F., and the Alzheimer's Disease Neuroimaging Initiative (2019a). Longitudinal functional brain mapping in supernormals. *Cereb Cortex*, 29(1):242–252.

Wang, X., Ren, P., Mapstone, M., Conwell, Y., Porsteinsson, A. P., Foxe, J. J., Raizada, R. D. S., Lin, F., and and the Alzheimer's Disease Neuroimaging Initiative (2019b). Identify a shared neural circuit linking multiple neuropsychiatric symptoms with Alzheimer's pathology. *Brain Imaging Behav*, 13(1):53–64.

Zinszer, B. D., Anderson, A. J., Kang, O., Wheatley, T., and Raizada, R. D. S. (2016). Semantic structural alignment of neural representational spaces enables translation between English and Chinese words. *J Cogn Neurosci*, 28(11):1749–1759.