# Contents



# 1 Abstract

I will be speaking how to model the relationship between brain activity and language. I will first give some background on brain data and fMRI in particular, quickly covering Mitchell's 2008 paper on predicting brain activity associated with the meaning of nouns. I will then discuss Pereira et. al's work on generating text from fMRI, and follow up with a paper from 2014 by Mitchell's group at CMU identifying story features predictive of fMRI signal as subjects read Harry Potter. At the end if there is still time, I will speak about some related work on methods of obtaining better semantic features (word vectors) derived from joint embeddings of text and fMRI data, a paper by Fyshe (who is a student of Tom Mitchell). Fyshe introduces the JNNSE algorithm which improves upon the NNSE algorithm features used in the 2014 story-reading paper.

Note: Presentation given on October 22, 2015 to the Alg-ML reading group at Princeton (led by Profs. Sanjeev Arora and Elad Hazan).

# 2 Connecting Brain and Language Representations

## 2.1 Mitchell's 2008 paper [3]

### 2.1.1 Summary

The main contribution of this paper is to present a computational model which predicts fMRI response associated with words for which there is no fMRI data yet available. The theory underlying the model is that the neural semantic representation of concrete nouns is related to the distributional hypothesis of meaning: Basically, brain vectors for concrete nouns should behave similarly to word vectors for those same concrete nouns in a huge corpus. This assumption is basically positing that we learn word meaning based on reading.

The model is trained on a trillion-word text corpus (the Google 5-gram corpus from English web pages) and fMRI data observed after viewing a 58 concrete nouns from 12 semantic categories. For testing, the model predicts fMRI activation for words on a held-out set of size 2 and achieves highly significant accuracies.

They also train competing computational models with different features for encoding meaning of concrete objects in the brain. The best model predicts fMRI activity to the degree that it can match words to their previously unseen fMRI images with accuracy far above chance. Thus there exists a direct predictive relationship between word co-occurence statistics and fMRI patterns associated with thinking about the word.

### 2.1.2 Theoretical Assumptions

There are three key theoretical assumptions made by this paper:

1. The semantic features that distinguish meanings of concrete nouns are reflected by their statistics of their use in a very large text corpus (specifically, for the $n = 25$ co-occurences the authors chose to record).

2. Different spatial patterns of neural activity are associated with different semantic categories of pictures and words.

3. The brain activity observed when thinking about a concrete noun is a linear combination of semantic feature values.

The first assumption is generally known as the distributional hypothesis of meaning, though its use here is more restricted since the authors only use the co-occurences of each concrete noun $w$ with 25 verbs. The authors justify the second assumption by arguing that many linear models are used in the fMRI literature with the assumption that fMRI activation reflects a linear superposition of many sources.

Furthermore this model allows the training data to determine the locations in the brain whose activity is affected by word meaning aspects, rather than making assumptions from neuroscience about which regions of the brain encode which aspects of meaning.

### 2.1.3 Training

Some notation first: Let $n$ be the number of semantic features used to represent a word. Let $m$ be the number of voxels in the brain. There are two steps to training. First, semantic features based on co-occurrence properties are computed from the very large text corpus. The second step learns weights for a linear combination of the semantic features to predict the activation at each voxel. Let $y(w)$ be the $m \times 1$ matrix of voxel activations for a given word $w$, $C$ be an $m \times n$ matrix of coefficients to be learned, and $f(w)$ be the $n \times 1$ semantic feature encoding of word $w$. Then

$$y(w) = Cf(w) = \sum_{i=1}^{n} C_{*,i} f_i(w) \tag{1}$$

Here $C$ is not dependent on a word $w$. We can interpret this equation in terms of the columns $C_{*,i}$ of $C$. By re-writing, we get that $\{C_{*,i}\}_{i=1}^{n}$ is a semantic image feature basis, with each image associated to a different semantic feature.

In this paper, the semantic features are the co-occurence statistics of the input word $w$ with 25 different verbs (accounting for different forms of the verb). The verbs correspond to basic sensory and motor activities, actions performed on objects, and actions involving changes to spatial relationships.

For each voxel $v$, we learn the $1 \times n$ row vector $C_{v,*}$ of $C$ with linear regression. Let the number of different words be $T$. Let $X$ be a $T \times n$ matrix where each row is $f(w_t)$ for $t \in [T]$. Let $y_v$ be a $T \times 1$ vector where each entry $y_v(t)$ is the response for voxel $v$ for word $w_t$. Then, for each $v \in [m]$ we find

$$\text{argmin}_{C_{v,*}} \|y_v - XC_{v,*}^T\|_2^2 + \lambda \|C_{v,*}^T\|_2^2 \tag{2}$$

which is solved by the (Ordinary Least Squares) OLS estimator. If the number of training examples is $< n = 25$, then there is no unique solution. In this case, adding $l_2$ regularization (i.e. ridge regression) gives a unique solution of least norm where $\lambda = 1$. After each $C_{v,*}$ is trained, we have the full predictor matrix $C$ which given a word $w$ and its featurization $f(w)$, we can use to predict the full fMRI response $y(w) = Cf(w)$.

### 2.1.4 Results

There were 60 randomly ordered stimuli (a picture of the object in white over black background) which came from 12 semantic categories (animals, body parts, buildings, etc.). There were only 9 human subjects, of college age. Each word-picture pair was presented 6 times. The representative fMRI response for each word was computed by averaging over the 6 presentations of word-picture pairs. The mean over all 60 presentations (one for each word-picture pair) was then subtracted from each presentation. A separate model was learned for each of the 9 participants.

Evaluation was performed with leave-two-out cross validation. That is, the model was repeatedly trained with 58 out of 60 word-fMRI image pairs, and tested on the remaining two. For testing, first a prediction of the fMRI image was generated for each of the two words, then these predicted fMRI images had to be matched to the correct fMRI image.

This task was executed by comparing cosine similarity of the fMRI image vectors (where only a subset of the voxels were used). The subset of voxels was decided by calculating stability scores for each voxel: For each of the $6 \times 58$ presentations shown, there is a given fMRI voxel matrix. Then they calculated pairwise correlation across the 6 rows in the $6 \times 58$ matrix for each voxel, which assigns higher scores to voxels which exhibit consistent variation across the 58 images presented. Cross validated accuracies for each of the 9 models had a mean of 77% accuracy, which is above chance (they claim an accuracy of 62% is statistically significant based on empirical accuracy distributions for null models).

Another evaluation was performed to test whether the model could distinguish among a more diverse range of words. Here, the model was tested using a leave-one-out test where the model for each individual was trained on 59 words. Then, for 1000 additional words and the held-out word, an fMRI image was predicted. The 1001 words were then ranked by cosine similarity of their predicted fMRI to the true fMRI data for the held-out word. The average percentile rank was 0.72 across participants.

They also manually examined the semantic feature signatures (think of $C_{*,i}$ for semantic feature $f_i(w)$): i.e., whether the predicted activations for various verbs matches the associations. They saw that activity in the gustatory cortex co-occurs with the verb 'eat', activity in motor areas co-occurs with 'push', strong activation in somatosensory cortex co-occurs with 'touch', and 'listen' co-occurs with activation in the language processing regions of the brain.

The authors also checked how accuracy varied over different feature sets. They tested 115 feature sets of 25 randomly drawn words from the 5000 most frequent words in the text corpus excluding the 60 stimulus words and the 500 most frequent words (i.e. containing 'the' and 'have'). The minimum and maximum accuracies of these random feature sets was 0.46 and 0.68, with the average of 0.60 and a standard deviation of 0.04. These results suggest that the hand-picked features do rather better than random.

The success of the 25 sensory-motor specific verbs as a feature set suggests that neural representations of concrete nouns are in part related to sensory-motor features.

I believe Yingyu has used the semantic vectors derived from Prof. Arora's paper earlier this year and used them as regression inputs to learn the fMRI voxel values and achieved similar accuracies on the classification task.

## 2.2 Pereira's 2011 paper [4]

### 2.2.1 Summary

Pereira et. al. propose the inverse problem of the one solved by Mitchell et. al. (2008). Instead of predicting fMRI given a word, given an fMRI response, they generate text using a generative model (LDA). They can generate a probability distribution over words pertaining to left-out novel brain images and that the quality of this distribution is measured quantitatively via a classification task that matches brain images to Wikipedia articles.

The authors use the dataset from Mitchell et. al. (2008) [3], which is fMRI data while subjects looked at both a picture and a word representing a concrete noun (e.g. *house*). From these fMRI images, the authors generated words pertaining to the relevant concept (e.g., *door, window, home*). Then the generated words are matched to corresponding articles

from Wikipedia, providing a way of quantitatively analyzing the results.

### 2.2.2 Theoretical Assumptions

1. Brain images can be directly associated with text.

2. The authors only focus on representations of concrete objects: They do not assess relationships between concepts or representations of abstract concepts.

3. They ignore word order and grammatical structure.

### 2.2.3 A Brief Description of LDA

LDA (Latent Dirichlet Allocation) operates on a word-document co-occurrence matrix, placing documents in low dimensional space by taking advantage of sets of words which appear in multiple documents. Each dimension corresponds to a co-occurrence pattern (a topic word probability distribution). LDA is a generative model and allows you to interpret topic probabilities as the probability that a word came from the distribution of a particular topic. LDA models each Wikipedia article representing concept $w$ as coming from a process where the number of words $N$ and the probabilities of each topic being present $\theta_w$ are drawn. Each word $u$ is drawn by selecting topic $z$ according to probabilities $\theta_w$, and then drawing from $\mathbf{P}\{u|z\}$, the distribution over words given topic $z$. $\theta_w$ is the featurization of each concept; i.e. $f(w) = \theta_w$. Since LDA places the topics in a simplex, the presence of some topics and detract from the presence of others.

Given a concept $w$, we can also induce a probability distribution for words $u$ in $w$:

$$\mathbf{P}\{u|\theta\} = \sum_{i=1}^{|\text{topics in } w|} \mathbf{P}\{u|z_i\}\theta_w^{(i)} \tag{3}$$

### 2.2.4 Training

They follow a series of steps to arrive at their generative text model. Here is the overview:

1. First, from a corpus of 3500 Wikipedia articles about concepts deemed concrete or imageable (including 60 concepts from [3]), the authors created a topic model (latent factor representation using LDA) of each article, which represents the concept the article is about. This topic model effectively takes the place of the semantic features from [3] as an approximation of the mental representation of the concept. The authors ran LDA with the number of topics allowed ranging from 10 to 100 in increments of 10. The result is a representation of each of the 3500 Wikipedia articles in terms of the probabilities of each topic being present: We call these latent factor loadings. Each topic is a probability distribution over words.

2. They use ridge regression to learn a mapping from each topic/concept to a corresponding pattern of brain activation: This is equivalent to learning $C$ before. The only difference from the formulation established from Mitchell et. al (2008) [3] is that

**Table 1 | The top 10 most probable words according to each topic in the 40 topic model used in Figure 2A (topic ordering is slightly different).**

| Topic | Top 10 words |
|---|---|
| 1 | Plant fruit seed grow leaf flower tree sugar produce species |
| 2 | Color green light red white blue skin pigment black eye |
| 3 | Light drink lamp wine beer bottle water produce valve pipe |
| 4 | Drug chemical acid opium cocaine alcohol substance produce form reaction |
| 5 | School university student child education college degree state train Unite |
| 6 | Animal species cat wolf breed hunt dog male wild human |
| 7 | Water metal form temperature carbon process air element iron salt |
| 8 | Vehicle wheel gear car aircraft passenger speed drive truck design |
| 9 | Market party state country price government political trade people economic |
| 10 | Water ice rock river surface form sea ocean wind soil |
| 11 | Species bird egg fish insect female ant live feed bee |
| 12 | Language book write art century form story character word publish |
| 13 | War military force army weapon service submarine soviet world train |
| 14 | Blood cause disease patient treatment infection health risk increase pain |
| 15 | Church bishop pope catholic priest roman soap cardinal religious time |
| 16 | Cell muscle body brain form tissue human organism bone animal |
| 17 | Ship fish boat water vessel sail design build ski bridge |
| 18 | Iron blade steel handle head cut hair metal tool nail |
| 19 | Film image camera digital shotgun movie lens magazine rifle gun |
| 20 | Wear horse woman clothe saddle century dress fashion ride trail |
| 21 | Law state court legal police crime person act Unite criminal |
| 22 | Smoke chocolate light tobacco sign speed cigaret cigar state traffic |
| 23 | key lock switch machine needle tube bicycle type knit design |
| 24 | Card record information service company product datum process program credit |
| 25 | State cross head salute plate model symbol portrait scale circus |
| 26 | Love sexual god woman people pyramid death sex religion evil |
| 27 | Coin gold silver issue currency stamp state dollar value bank |
| 28 | Game play player ball team sport rule football hit league |
| 29 | Fuel engine gas energy power oil hydrogen heat rocket produce |
| 30 | Woman marriage god word christian child term jesus family gender |
| 31 | Fiber sheep wool cotton fabric weave hamlet pig produce silk |
| 32 | City build house store street town state home road bus |
| 33 | Tea tooth pearl kite shoe culture wear tattoo jewelry form |
| 34 | Earth sun star planet moon solar time orbit day comet |
| 35 | Material wood paint build wall structure construction design size window |
| 36 | Human social study people culture theory individual nature behavior term |
| 37 | Power station train signal line locomotive radio steam electric frequency |
| 38 | Food diamond cook meat bread coffee sauce chicken kitchen eat |
| 39 | Measure scale angle formula theory object unit energy line property |
| 40 | Music instrument play string band bass sound note player guitar |

**Figure 1:** A list of topics and their 10 most likely words [4]

the definition of $f(w)$ changes. Instead of using the hand-chosen verb co-occurences, Pereira et. al. use the topic probabilities describing the Wikipedia articles corresponding to the concept $w$, which are the concrete nouns from Mitchell et. al (2008) [3]. The regression inputs are $f(w)$ and output is voxel activation $y_v$. The number of subsampled voxels used here is 1000 as opposed to 500 in [3]. As before, the fMRI images can be decomposed into a set of topic-specific basis images ($\{C_{*,i}\}_{i=1}^{n}$, corresponding to the semantic feature signatures in [3]). At this point we are still predicting fMRI image from featurized words. Now, from before, a probability distribution over a set of topics representing a concept induces a probability distribution on words for that concept, $\mathbf{P}\{u|\theta_w\} = \mathbf{P}\{u|f(w)\}$.

3. For brain images in a test set, the mapping can be used to infer a weighting over latent factors. The generative model from the first step can then be inverted to map from latent factors to text.

   In Mitchell et. al. (2008), learning $f(w)$ given $C$ ($m \times n$ matrix where $m$ is number of voxels and $n$ features) and $y(w)$ would not have allowed us to generate text, since $f(w)$ was just co-occurrences with certain verbs. In principle it could be possible to derive a probability distribution from these co-occurences: The authors would need to tabulate for each of the 25 verbs which words occur within a 5-word window of the verbs; this vector could then be normalized into a probability distribution. In this paper, $f(w)$ is a probability distribution over topics $\theta_w$ which **induces** a probability distribution over words $u$! Thus we can use fMRI images of concepts to produce words from that concept. We simply need to solve the convex optimization problem

$$\begin{aligned} &\text{argmin}_\theta \|y - C\theta\|_2^2 \\ &\text{s.t.} \\ &\theta_i \geq 0 \text{ for all } i \in [n] \\ &\sum_i \theta_i = 1 \end{aligned} \tag{4}$$

$C$ is fixed from the ridge regression and its columns are the basis fMRI images for each concept, $y$ is the new image we want to infer the topic distribution for, and $\theta$ is the distribution to infer. Recall that the number of features $n$ is the number of topics. Let $\theta_y$ be the optimal topic probability distribution for a given novel fMRI image $y$. Now recall that for each topic $z_i$ for $i \in [n]$, we have $\mathbf{P}\{u|z_i\}$ for word $u$ learned from LDA.

$$\mathbf{P}\{u|y\} = \sum_{i=1}^{n} \mathbf{P}\{u|z_i\}\theta_y^{(i)} \tag{5}$$

Note the similarity to Equation (3): The difference here is that $\theta_y$ is inferred. Thus, by solving this convex program, we have inverted our map to estimate topic probability distribution $\theta$ from new unseen concept fMRI images $y$.

### 2.2.5 Results

First we note that the topic representations of the 3500 articles are sparse with respect to topics (though there are multiple topics in the representation of most articles). There is a link online to browse the 3500 concepts and topic distributions in detail.

To objectively evaluate the quality of the generated text, 58 of the concepts are used for training and they test on the held-out 2 concepts. For the 2 concepts, we get the fMRI images and infer the topic probability distributions, $\theta_{y_1}, \theta_{y_2}$. The topic distributions are then matched with corresponding Wikipedia pages by using $\mathbf{P}_{y_1}\{u|\theta_{y_i}\}$ to determine which Wikipedia article is most probable. Note that random chance as accuracy 50% since we are pairing two fMRI images with the two held-out Wikipedia articles. In the majority of cases, classification was accurate.

When the two held-out concepts were in different semantic categories (i.e. *vegetable* and *car*), accuracy was on average over the nine participants was around 0.8, with a max of around 0.9 and a min of around 0.65. When the two held-out concepts were in the same semantic category, average accuracy over the nine participants was around 0.55, with a max of around 0.6 and a min of around 0.48. Note that these values are averaged over using different numbers of topics from 10 to 100, for every possible pair of two held-out concepts. The reason for the inaccuracy suggests text outputs for semantically related concepts are very similar, which is both good and bad: It suggests the model is not fine-grained enough (too bag-of-words-ish), or that the concept-representations (Wikipedia articles) are too similar intrinsically. They also saw that the voxels from the temporal and occiptal cortex voxels were the most stable across the 6 presentations of a concept to the fMRI subjects, suggesting

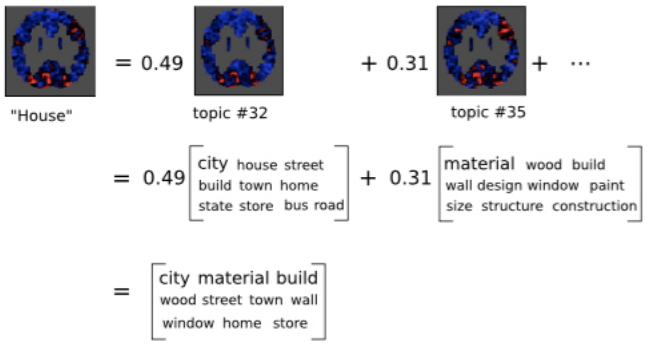


**Figure 2:** Visualization of weighted sums of latent factors [4]

that the learned fMRI basis associated with a topic is related to both semantic (word) and visual (picture) aspects of the topic.

Some takeaways: simple linear models here for both papers, LDA is a relatively simple generative model.

## 2.3 Mitchell's 2014 paper [5]

### 2.3.1 Summary

Previously we only looked at papers which inferred brain state from static words, or bag-of-words document classification from brain state. Now we will look at a paper which develops a model with a more sophisticated understanding of concepts in language.

Mitchell et. al. (2014) predict the fMRI signatures associated with reading arbitrary text passages in a story (namely, Chapter 9 of Harry Potter and the Philosopher's Stone), using understanding of sentences and character relationship as features to predict fMRI signatures. In many ways this paper is an extension of the original 2008 paper [3], removing some of the early assumptions and generalizing from concrete nouns to sentences and story structure. Their model is able to distinguish between which of two story segments (as opposed to which of two concrete nouns) is being read with 74% accuracy.

### 2.3.2 The Model and its Features

This model follows a similar story to [3]. The setup is as follows: For nine individuals, fMRI activity was collected while each individual read the $9^{th}$ chapter of the first Harry Potter book. Reading was performed by having a single word appear at the center of the screen every 0.5 seconds (this format is known as rapid serial visual format (RSVF)). Note that each of the subjects was familiar with the Harry Potter story, had been recently updated on the contents of chapter nine, and had practiced RSVF on an unrelated story to the point where reading in this fashion was considered 'comfortable.' fMRI activity was collected every two seconds (these are called TRs). Thus we have two time series, one of words and one of fMRI activity for every individual. To match the time series up, the word time series was chunked into groups of four words per TR for a time resolution of two seconds.

This model follows a similar pattern as the models from the previous two papers, and we will use the same notation to demonstrate the similarity.

In the original paper [3], we essentially thought of the words $w_t$ as a list of concrete noun examples presented in some order which did not matter. In this paper, the order in which the words presented does matter, and the authors take this into account in their model. Furthermore, each time step now consists of four words rather than one (so that text and fMRI time series are aligned). Thus, features $f_i(w_t)$ are now transformed into features $f_i(\{w_1, w_2, w_3, w_4\}_t)$ since features can be a property of each four-word chunk. For convenience we will refer to this as $f_i(t)$, the feature $i \in [n]$ at time $t \in [T]$.

In the original paper [3], the features were word co-occurrences with 25 different verbs relating to sensory-motor activities. In this paper, the story features attempt to address multiple levels of representation. The types of features can be divided into four categories: visual features, semantic features, syntactic features, and discourse features.

1. Visual features are just the average word length in each TR and the word length variance in each TR.

2. Syntactic features are derived using an automated parser to get parts of speech for each word as well as dependency roles for each word inferred from a parse tree. There are 28 part-of-speech relationships and 17 dependency relationships for a total of 45 binary features indicating if a given part-of-speech or a dependency occurred in a 4-word TR. An additional $46^{th}$ feature is the average position of the words in the TR in the sentence they belong in, numbered starting at 1.

3. Discourse features are derived from manual annotations going through story text. Pronouns are annotated with the character they refer to, and binary features are created for whether or not a certain character (one of 10) shows up in a TR. Frequent physical motions were chosen as well: These come with two values, a binary feature representing the start of the motion and a binary 'sticky' feature representing whether the motion is currently ongoing. Similarly, speech between characters is represented by a feature representing which character is speaking and a sticky binary feature indicating speech by a specific character is ongoing. There are also features for when emotion is mentioned and a corresponding sticky feature indicating an emotion is ongoing. For non-motion verbs (*hearing, knowing, seeing*), only a binary feature indicating that the verb took place is used, since these verbs typically do not last long enough to necessitate a sticky version of the feature.

4. Semantic features are most closely related to the features from [3]. They use non-negative sparse embedding (NNSE) to learn semantic vectors from a massive web corpora on which various dependency and document co-occurrence counts are computed. There are two co-occurrence matrices with different definitions of 'context': document counts are the number of mentions a word has in a particular document, and dependency counts are the number of times a word is in a given dependency parse link (e.g. word $u$ is the subject of the verb "eat"). These dependencies are primarily verb- and adjective-related [1]. The co-occurrence matrices are factored using NNSE. to produce 1000 features of which this paper uses the top 100. Think of these as word vectors.

Since each TR has four words, they need a way to compose these word vectors: Their approach is to simply sum the features within each TR. On pg. 6 is a list of all 195 features.

### 2.3.3 Theoretical Assumptions

1. They assume that each feature has a signature activity in each voxel which is consistently repeated every time the brain encounters this feature (and if a voxel does not encode this feature, the weight is 0).

2. The signature activity is scaled by the value of the feature at the time the feature is presented.

3. Total activation of a voxel is a linear combination of the feature values.

4. There is spherical Gaussian error in voxels with a different variance for each voxel. However, the variance for each voxel remains fixed over time.

5. The activity created by the feature is the convolution of the response signature with the time course of the feature. This convolution makes sense in context of the hemodynamic response function (HDF) of the BOLD (blood-oxygen level dependent) signal, which fMRI measures. The HDR gives the activation curve for each voxel. While canonical HDFs do exist, literature has shown that the HDF is not necessarily uniform across the brain nor is it uniform across people. Thus learning weights for different time dependencies makes sense. Inspecting the weights over $k = 1, 2, 3, 4$ after training for each feature reveals these resemble the characteristic shape of the HDF at different points of the HDF.

6. Putting the last two assumptions together, the activity of voxel $v$ at time $t$ is given by

$$y_v(t) = \sum_{j=1}^{F} \sum_{k=1}^{4} f_j(t-k) c_k^{vj} \tag{6}$$

   Here we adopt the convenient notation that $F = n/4$, and $c_k^{vj}$ is a special indexing of feature coefficients where we let $k$ range from 1 to 4. In practice we stack on 4 additional feature columns per feature to our coefficient matrix $C$ to represent the weights $c_1^{vj}, c_2^{vj}, c_3^{vj}, c_4^{vj}$.

### 2.3.4 Training

First we write down the training objective as for the previous two papers: Note that the definitions are almost identical. Let $X$ be the $T \times n$ matrix such that each row is the featurization of a different time step. Let $C$ be an $m \times n$ matrix where $m$ is the number of voxels in an fMRI scan and $n = 4 \times 195$ is the number of features, where the factor of 4 comes from time shifts. Then each row $C_{v,*}$ is the $n \times 1$ weight vector we learn for a given TR. Let $y_v$ be the $T \times 1$ vector such that each entry is the response of voxel $v$ at each TR

$t$. Let $y(t)$ be the $m \times 1$ vector denoting activation at each voxel at TR $t$. Let $f_i(t)$ be the $i^{th}$ feature at TR $t$ and let $f(t)$ denote the feature vector of the four words at TR $t$.

This paper's model adds noise to the fMRI voxels. Let $\epsilon_v \sim \mathcal{N}(0, \sigma_v^2 I_T)$ be spherical Gaussian noise with zero mean. Let $\epsilon$ be the $m \times 1$ random variable of Gaussian variables $[\epsilon_1, \cdots, \epsilon_m]$. As before, we learn a different $C$ for each subject, and thus every parameter is learned differently for each subject. The only constant parameter is $X$, which represents the featurization of the Harry Potter text over time.

Thus, the equation to predict fMRI from the four words in a given TR $t$ is

$$y(t) = Cf(t) + \epsilon \tag{7}$$

and the least squares regression objective is

$$\text{argmin}_{C_{v,*}} \|y_v - XC_{v,*}^T\|_2^2 + \lambda_v \|C_{v,*}^T\|_2^2 \tag{8}$$

Training separately for each row $C_{v,*}$ gives us the full matrix $C$. Note that this is simply ridge regression again, for which we are guaranteed to get a unique solution (the solution of least norm when $\lambda_v = 1$). $l_2$ regularization gives us the MAP estimator for least squares when we assume Gaussian errors. In practice, cross-validation is used on the training data to find the correct $\lambda_v$s. Ridge regression also results in effective automatic voxel selection since it learns high penalties for noisy voxels and small penalties for good voxels.

#### 2.3.5 Results

They show that the predictions of the trained model are sufficient to distinguish between which of two previously unseen short passages is being read, given only observed fMRI activity. The first test task is analagous to the task from [3]. The trained model predicts the fMRI time series for two held-out story passages. Then it selects the passage such that the predicted fMRI time series is most similar in $l_2$ norm to the held out real fMRI time series. The results are cross-validated across all choices of the two held-out story passages. Random performance on this task is 50%. They attain an accuracy of 74%, which is significant with $p < 10^{-8}$. ($p$-values are determined by assuming the null hypothesis is 50% and then generating sample data and looking at the distribution of predictions for random weights).

The second test the authors run is to identify what type of information is processed by various regions of the brain. First, they effectively partition the brain into $15 \times 15 \times 15$ mm cubes corresponding to $5 \times 5 \times 5$ voxels (note there are typically on the order of $10^5$ voxels, so this is about $10^{-3}$ of the full volume). They test every type of feature at every cube location to determine in which brain regions (the cubes) which types of features yield high classification accuracy. They found that the occipital cortex of the brain were strongly associated with the visual features (word length), as expected.As some examples of other results they find, they also see that the right temporo-parietal cortex is related to sentence length and the presence of dialog. Interestingly, the right temporo-parietal cortex has previously been shown to be more activated for better readers and is related to verbal working memory processes. The imagined physical motion of the story characters is found to activate in the posterior temporal cortex and angular gyrus, which agrees with neuroscientific knowledge. The identity of story characters is distinguishable by activity in the right posterior superior/middle temporal

region, a region that has been found to encode facial identity. They also suggest they have found a partial answer to the question of whether semantic and syntactic properties of language are represented in different locations in the brain: For the semantic and syntactic features they use, there is a large overlap in some areas. They also find regions selectively processing syntax and semantics and that syntactic information is more widely and strongly represented (though this just may be due to the quality of the semantic features versus the synatactic features).

### 2.3.6 Important Takeaways

One key idea of this experiment is the differentiation between fMRI data paired with a specific stimulus (what I will called **supervised** fMRI) and resting-state fMRI data (**unsupervised**). Many studies make use of the supervised setting in order to find the regions of the brain which encode the properties of a given stimulus. In this paper, there is only one stimulus: The text of a chapter in a Harry Potter novel. However, with their diverse featurization (the rich properties of text), it becomes possible to analyze the effect of a wide range of stimuli including characters, motions, and the visual properties of reading on the brain.

### 2.3.7 Aside on NNSE and JNNSE [2]

Let us take another look at the semantic features in the model. They are derived from the Non-Negative Sparse Embedding (NNSE) algorithm, which we will give a brief description of here. Then we will discuss the Joint-NNSE algorithm, which intends to provide a better semantic embedding that uses information about the brain as input.

Let $X \in \mathbb{R}^{w \times c}$ be made from $c$ corpus statistics for $w$ words (i.e. $X$ is a word-context matrix). Then, NNSE produces a low-dimensional, sparse, non-negative latent representation using matrix factorization. The idea behind non-negativity is that you typically describe an object or concept by its positive facets; i.e. you say "an apple is a fruit" and not "an apple is not a tool". Sparsity is common to encourage only the most important features to have high weights. The NNSE objective is given by

$$\begin{aligned} &\text{argmin}_{A,D} \sum_{i=1}^{w} \|X_{i,*} - A_{i,*}D\|_2^2 + \lambda\|A\|_1 \\ &\text{s.t} \\ &D_{i,*}D_{i,*}^T \leq 1 \text{ for all } 1 \leq i \leq r \\ &A_{i,j} \geq 0 \text{ for all } 1 \leq i \leq w, 1 \leq j \leq r \end{aligned} \tag{9}$$

where the algorithm outputs the solution $A \in \mathbb{R}^{w \times r}$ that represents word semantics in $r$-dimensional space while being sparse and non-negative. $D \in \mathbb{R}^{r \times c}$, and note that $D_{i,*}$ and $A_{i,*}$ are row vectors of dimension $1 \times r$. Thus this program factors $X$ to minimize reconstruction error using $l_1$ regularization for sparsity.

The purpose of the Joint-NNSE objective is simply to add an additional data source for a subset of the words in $X$. In the context of [2], the additional data is to be either fMRI or MEG data to encourage $A$ to behave similarly in both the brain and word settings. Here,

first re-order the rows of the corpus data $X$ so that the first $1, \cdots, w'$ rows have associated brain recordings. Then let $Y \in \mathbb{R}^{w' \times v}$ be the data matrix of brain recordings, where $v$ is the number of features associated with the brain data. Then let $D^c \in \mathbb{R}^{r \times c}, D^b \in \mathbb{R}^{r \times v}$. The JNNSE objective is given by

$$\begin{aligned} &\text{argmin}_{A,D^c,D^b} \sum_{i=1}^{w} \|X_{i,*} - A_{i,*}D^c\|_2^2 + \|Y_{i,*} - A_{i,*}D^b\|_2^2 + \lambda\|A\|_1 \\ &\text{s.t.} \\ &D^c_{i,*}\left(D^c_{i,*}\right)^T \leq 1 \text{ for all } 1 \leq i \leq r \\ &D^b_{i,*}\left(D^b_{i,*}\right)^T \leq 1 \text{ for all } 1 \leq i \leq r \\ &A_{i,j} \geq 0 \text{ for all } 1 \leq i \leq w, 1 \leq j \leq r \end{aligned} \tag{10}$$

where again we receive the output $A \in \mathbb{R}^{w \times r}$ is in a low-dimensional space while being sparse, non-negative, and representing word and brain semantics, since we have ensured that words represented in brain space must behave similarly by keeping $A$ fixed across optimizations. Note that JNNSE can handle partially paired data, in comparison to Canonical Correlations Analysis (CCA) which requires fully paired data. In JNNSE, we only seek a solution keeping the transformed form fixed and maximally correlating the data reconstruction instead. In contrast, CCA maximally correlates the transformed form while keeping the input data fixed. This change allows the data to only be partially paired.

Note that neither the NNSE or JNNSE objectives is convex due to the required alternating optimization.

The paper goes on to perform experiments using the data from [3] to demonstrate that JNNSE vectors are more consistent with independent samples of brain activity collected from different subjects for use as semantic features. Here, the authors train a linear predictor of semantic vectors given brain state vector, and use the predicted semantic vectors to see if the model can differentiate between two unseen words. This task inverts [3] and predicts a word from a brain state. Since semantic vectors have a sensible definition of similarity (unlike the verb co-occurrence semantic features from [3]), it makes sense to predict the word associated with the true semantic vector closest to the predicted one. On this prediction task with 50% accuracy, JNNSE achieves around 74% accuracy after cross-validating (this time only using 150 random pairs of hold-out words), which is on average 6% better than when using NNSE. The modality of brain data does not affect the results very much (both fMRI and MEG are good).

# 3 What's Next?

How about the analogue to the previous paper: Predict sentences being read based on fMRI brain state! Interestingly enough, I believe Dr. Pereira is working on this very problem now (as an analogue to his 2011 paper [4]).

# References


[1] Fyshe, A., Talukdar, P., Murphy, B., Mitchell, T. M. Documents and Dependencies: an Exploration of Vector Space Models for Semantic Composition. At cs.cmu.edu/afyshe/papers/conll2013/deps_and_docs.pdf. Accessed October 21, 2015.

[2] Fyshe, A., Talukdar, P., Murphy, B., Mitchell, T. M. Interpretable Semantic Vectors from a Joint Model of Brain- and Text-Based Meaning. At aclweb.org/anthology/P14-1046. Accessed October 21, 2015.

[3] Mitchell, T. M. et al. Predicting Human Brain Activity Associated with the Meanings of Nouns. *Science* **320**, 1191 (2008).

[4] Pereira, F., Detre, G., Botvinick, M. Generating text from functional brain images. *Frontiers in Human Neuroscience* **5**, 72 (2011).

[5] Wehbe, L., Murphy, B., Talukdar, P., Fyshe, A., Ramdas, A., Mitchell, T. M. Simultaneously Uncovering the Patterns of Brain Regions Involved in Different Story Reading Subprocesses. *PLOS One* **9**, 11 (2014).