Interpretable Semantic Vectors from a Joint Model of Brain- and Text-Based Meaning

Alona Fyshe[1], **Partha P Talukdar**[1], **Brian Murphy**[2], **Tom M Mitchell**[1]
[1]Machine Learning Department, Carnegie Mellon University
[2]School of Electronics, Electrical Engineering and Computer Science
Queen's University Belfast
[afyshe,partha.talukdar,tom.mitchell]@cs.cmu.edu
brian.murphy@qub.ac.uk

Abstract

Vector space models (VSMs) represent word meanings as points in a high dimensional space. VSMs are typically created using a large text corpora, and so represent word semantics as observed in text. We present a new algorithm (JNNSE) that can incorporate a measure of semantics not previously used to create VSMs: brain activation data recorded while people read words. The resulting model takes advantage of the complementary strengths and weaknesses of corpus and brain activation data to give a more complete representation of semantics. Evaluations show that the model 1) matches a behavioral measure of semantics more closely, 2) can be used to predict corpus data for unseen words and 3) has predictive power that generalizes across brain imaging technologies and across subjects. We believe that the model is thus a more faithful representation of mental vocabularies.

1 Introduction

Vector Space Models (VSMs) represent lexical meaning by assigning each word a point in high dimensional space. Beyond their use in NLP applications, they are of interest to cognitive scientists as an objective and data-driven method to discover word meanings (Landauer and Dumais, 1997).

Typically, VSMs are created by collecting word usage statistics from large amounts of text data and applying some dimensionality reduction technique like Singular Value Decomposition (SVD). The basic assumption is that semantics drives a person's language production behavior, and as a result co-occurrence patterns in written text indirectly encode word meaning. The raw co-occurrence statistics are unwieldy, but in the compressed VSM the distance between any two words is conceived to represent their mutual semantic similarity (Sahlgren, 2006; Turney and Pantel, 2010), as perceived and judged by speakers. This space then reflects the "semantic ground truth" of shared lexical meanings in a language community's vocabulary. However corpus-based VSMs have been criticized as being noisy or incomplete representations of meaning (Glenberg and Robertson, 2000). For example, multiple word senses collide in the same vector, and noise from mis-parsed sentences or spam documents can interfere with the final semantic representation.

When a person is reading or writing, the semantic content of each word will be necessarily activated in the mind, and so in patterns of activity over individual neurons. In principle then, brain activity could replace corpus data as input to a VSM, and contemporary imaging techniques allow us to attempt this. Functional Magnetic Resonance Imaging (fMRI) and Magnetoencephalography (MEG) are two brain activation recording technologies that measure neuronal activation in aggregate, and have been shown to have a predictive relationship with models of word meaning (Mitchell et al., 2008; Palatucci et al., 2009; Sudre et al., 2012; Murphy et al., 2012b).[1]

If brain activation data encodes semantics, we theorized that including brain data in a model of semantics could result in a model more consistent with semantic ground truth. However, the inclusion of brain data will only improve a text-based model if brain data contains semantic information not readily available in the corpus. In addition, if a semantic test involves another subject's brain activation data, performance can improve only if the additional semantic information is consistent across brains. Of course, brains differ in shape, size and in connectivity, so additional information encoded in one brain might not translate to an-

[1] For more details on fMRI and MEG, see Section 4.2

other. Furthermore, different brain imaging technologies measure very different correlates of neuronal activity. Due to these differences, it is possible that one subject's brain activation data cannot improve a model's performance on another subject's brain data, or for brain data collected using a different recording technology. Indeed, intersubject models of brain activation is an open research area (Conroy et al., 2013), as is learning the relationship between recording technologies (Engell et al., 2012; Hall et al., 2013). Brain data can also be corrupted by many types of noise (e.g. recording room interference, movement artifacts), another possible hindrance to the use of brain data in VSMs.

VSMs are interesting from both engineering and scientific standpoints. In this work we focus on the scientific question: Can the inclusion of brain data improve semantic representations learned from corpus data? What can we learn from such a model? From an engineering perspective, brain activation data will likely never replace text data. Brain activation recordings are both expensive and time consuming to collect, whereas textual data is vast and much of it is free to download. However, from a scientific perspective, combining text and brain data could lead to more consistent semantic models, in turn leading to a better understanding of semantics and semantic modeling generally.

In this paper, we leverage both kinds of data to build a hybrid VSM using a new matrix factorization method (JNNSE). Our hypothesis is that the noise of brain and corpus derived statistics will be largely orthogonal, and so the two data sources will have complementary strengths as input to VSMs. If this hypothesis is correct, we should find that the resulting VSM is more successful in modeling word semantics as encoded in human judgements, as well as separate corpus and brain data that was not used in the derivation of the model. We will show that our method:

1. creates a VSM that is more correlated to an independent measure of word semantics.
2. produces word vectors that are more predictable from the brain activity of different people, even when brain data is collected with a different recording technology.
3. predicts corpus representations of withheld words more accurately than a model that does not combine data sources.
4. directly maps semantic concepts onto the brain by jointly learning neural representations.

Together, these results suggest that corpus and brain activation data measure semantics in compatible and complimentary ways. Our results are evidence that a joint model of brain- and text-based semantics may be closer to semantic ground truth than text-only models. Our findings also indicate that there is additional semantic information available in brain activation data that is not present in corpus data, and that there are elements of semantics currently lacking in text-based VSMs. We have made available the top performing VSMs created with brain and text data (`http://www.cs.cmu.edu/~afyshe/papers/acl2014/`).

In the following sections we will review NNSE, and our extension, JNNSE. We will describe the data used and the experiments to support our position that brain data is a valuable source of semantic information that compliments text data.

2 Non-Negative Sparse Embedding

Non-Negative Sparse Embedding (NNSE) (Murphy et al., 2012a) is an algorithm that produces a latent representation using matrix factorization. Standard NNSE begins with a matrix $X \in \mathbb{R}^{w \times c}$ made of c corpus statistics for w words. NNSE solves the following objective function:

$$\underset{A,D}{\operatorname{argmin}} \sum_{i=1}^{w} \left\| X_{i,:} - A_{i,:} \times D \right\|^2 + \lambda \left\| A \right\|_1 \tag{1}$$

$$\text{subject to: } D_{i,:} D_{i,:}^T \leq 1, \forall\, 1 \leq i \leq \ell \tag{2}$$

$$A_{i,j} \geq 0,\ 1 \leq i \leq w,\ 1 \leq j \leq \ell \tag{3}$$

The solution will find a matrix $A \in \mathbb{R}^{w \times \ell}$ that is sparse, non-negative, and represents word semantics in an ℓ-dimensional latent space. $D \in \mathbb{R}^{\ell \times c}$ gives the encoding of corpus statistics in the latent space. Together, they factor the original corpus statistics matrix X in a way that minimizes the reconstruction error. The L_1 constraint encourages sparsity in A; λ is a hyperparameter. Equation 2 constrains D to eliminate solutions where A is made arbitrarily small by making D arbitrarily large. Equation 3 ensures that A is non-negative. We may increase ℓ to give more dimensional space to represent word semantics, or decrease ℓ for more compact representations.

The sparse and non-negative representation in A produces a more interpretable semantic space, where interpretability is quantified with a behavioral task (Chang et al., 2009; Murphy et al., 2012a). To illustrate the interpretability of NNSE, we describe a word by selecting the word's top scoring dimensions, and selecting the top scoring words in those dimensions. For example, the word chair has the following top scoring dimensions:

1. chairs, seating, couches;
2. mattress, futon, mattresses;
3. supervisor, coordinator, advisor.

These dimensions cover two of the distinct meanings of the word chair (furniture and person of power).

NNSE's sparsity constraint dictates that each word can have a non-zero score in only a few dimensions, which aligns well to previous feature elicitation experiments in psychology. In feature elicitation, participants are asked to name the characteristics (features) of an object. The number of characteristics named is usually small (McRae et al., 2005), which supports the requirement of sparsity in the learned latent space.

3 Joint Non-Negative Sparse Embedding

We extend NNSEs to incorporate an additional source of data for a subset of the words in X, and call the approach Joint Non-Negative Sparse Embeddings (JNNSEs). The JNNSE algorithm is general enough to incorporate any new information about the a word w, but for this study we will focus on brain activation recordings of a human subject reading single words. We will incorporate either fMRI or MEG data, and call the resulting models JNNSE(fMRI+Text) and JNNSE(MEG+Text) and refer to them generally as JNNSE(Brain+Text). For clarity, from here on, we will refer to NNSE as NNSE(Text), or NNSE(Brain) depending on the single source of input data used.

Let us order the rows of the corpus data X so that the first $1 \ldots w'$ rows have both corpus statistics and brain activation recordings. Each brain activation recording is a row in the brain data matrix $Y \in \mathbb{R}^{w' \times v}$ where v is the number of features derived from the recording. For MEG recordings, $v =$sensors $\times$ time points$= 306 \times 150$. For fMRI $v =$ grey-matter voxels $=\simeq 20,000$ depending on the brain anatomy of each individual subject. The new objective function is:

$$\underset{A,D^{(c)},D^{(b)}}{\operatorname{argmin}} \sum_{i=1}^{w} \left\| X_{i,:} - A_{i,:} \times D^{(c)} \right\|^2 + \sum_{i=1}^{w'} \left\| Y_{i,:} - A_{i,:} \times D^{(b)} \right\|^2 + \lambda \left\| A \right\|_1 \tag{4}$$

$$\text{subject to:} \quad D^{(c)}_{i,:} D^{(c)^T}_{i,:} \leq 1, \forall\, 1 \leq i \leq \ell \tag{5}$$

$$D^{(b)}_{i,:} D^{(b)^T}_{i,:} \leq 1, \forall\, 1 \leq i \leq \ell \tag{6}$$

$$A_{i,j} \geq 0,\ 1 \leq i \leq w,\ 1 \leq j \leq \ell \tag{7}$$

We have introduced an additional constraint on the rows $1 \ldots w'$, requiring that some of the learned representations in A also reconstruct the brain activation recordings (Y) through representations in $D^{(b)} \in \mathbb{R}^{\ell \times v}$. Let us use A' to refer to the brain-constrained rows of A. Words that are close in "brain space" must have similar representations in A', which can further percolate to affect the representations of other words in A via closeness in "corpus space".

With A or D fixed, the objective function for NNSE(Text) and JNNSE(Brain+Text) is convex. However, we are solving for A and D, so the problem is non-convex. To solve for this objective, we use the online algorithm of Section 3 from Mairal et al. (Mairal et al., 2010). This algorithm is guaranteed to converge, and in practice we found that JNNSE(Brain+Text) converged as quickly as NNSE(Text) for the same ℓ. We used the SPAMS package[2] to solve, and set $\lambda = 0.025$. This algorithm was a very easy extension to NNSE(Text) and required very little additional tuning.

We also consider learning shared representations in the case where data X and Y contain the effects of known *disjoint* features. For example, when a person reads a word, the recorded brain activation data Y will contain the physiological response to viewing the stimulus, which is unrelated to the semantics of the word. These signals can be attributed to, for example, the number of letters in the word and the number of white pixels on the screen (Sudre et al., 2012). To account for such effects in the data, we augment A' with a set of n fixed, manually defined features (e.g. word length) to create $A'_{percept} \in \mathbb{R}^{w \times (\ell + n)}$. $D^{(b)} \in \mathbb{R}^{(\ell+n) \times v}$ is used with $A'_{percept}$, to reconstruct the brain data

[2] SPAMS Package: http://spams-devel.gforge.inria.fr/

Y. More generally, one could instead allocate a certain number of latent features specific to X *or* Y, both of which could be learned, as explored in some related work (Gupta et al., 2013). We use 11 *perceptual* features that characterize the non-semantic features of the word stimulus (for a list, see supplementary material).

The JNNSE algorithm is advantageous in that it can handle partially paired data. That is, the algorithm does not require that every row in X also have a row in Y. Fully paired data is a requirement of many other approaches (White et al., 2012; Jia and Darrell, 2010). Our approach allows us to leverage the semantic information in corpus data even for words without brain activation recordings.

JNNSE(Brain+Text) does not require brain data to be mapped to a common average brain, which is often the case when one wants to generalize between human subjects. Such mappings can blur and distort data, making it less useful for subsequent prediction steps. We avoid these mappings, and instead use the fact that similar words elicit similar brain activation *within* a subject. In the JNNSE algorithm, it is this closeness in "brain space" that guides the creation of the latent space A. Leveraging intra-subject distance measures to study inter-subject encodings has been studied previously (Kriegeskorte et al., 2008a; Raizada and Connolly, 2012), and has even been used across species (humans and primates) (Kriegeskorte et al., 2008b).

Though we restrict ourselves to using one subject per JNNSE(Brain+Text) model, the JNNSE algorithm could easily be extended to include data from multiple brain imaging experiments by adding a new squared loss term for additional brain data.

3.1 Related Work

Perhaps the most well known related approach to joining data sources is Canonical Correlation Analysis (CCA) (Hotelling, 1936), which has been applied to brain activation data in the past (Rustandi et al., 2009). CCA seeks two linear transformations that maximally correlate two data sets in the transformed form. CCA requires that the data sources be paired (all rows in the corpus data must have a corresponding brain data), as correlation between points is integral to the objective. To apply CCA to our data we would need to discard the vast majority of our corpus data, and use only the 60 rows of X with corresponding rows in Y. While CCA holds the input data fixed and maximally correlates the transformed form, we hold the transformed form fixed and seek a solution that maximally correlates the reconstruction ($AD^{(c)}$ or $A'D^{(b)}$) with the data (X and Y respectively). This shift in error compensation is what allows our data to be only partially paired. While a Bayesian formulation of CCA can handle missing data, our model has missing data for $> 97\%$ of the full $w \times (v + c)$ brain and corpus data matrix. To our knowledge, this extreme amount of missing data has not been explored with Bayesian CCA.

One could also use a topic model style formulation to represent this semantic representation task. Supervised topic models (Blei and McAuliffe, 2007) use a latent topic to generate two observed outputs: words in a document and a categorical label for the document. The same idea could be applied here: the latent semantic representation generates the observed brain activity and corpus statistics. Generative and discriminative models both have their own strengths and weaknesses, generative models being particularly strong when data sources are limited (Ng and Jordan, 2002). Our task is an interesting blend of data-limited and data-rich problem scenarios.

In the past, various pieces of additional information have been incorporated into semantic models. For example, models with behavioral data (Silberer and Lapata, 2012) and models with visual information (Bruni et al., 2011; Silberer et al., 2013) have both shown to improve semantic representations. Other works have correlated VSMs built with text or images with brain activation data (Murphy et al., 2012b; Anderson et al., 2013). To our knowledge, this work is the first to integrate brain activation data into the construction of the VSM.

4 Data

4.1 Corpus Data

The corpus statistics used here are the downloadable vectors from Fyshe et al. (2013)[3]. They are compiled from a 16 billion word subset of ClueWeb09 (Callan and Hoy, 2009) and contain two types of corpus features: dependency and document features, found to be complimentary for most tasks. Dependency statistics were derived by dependency parsing the corpus and compil-

[3] `http://www.cs.cmu.edu/~afyshe/papers/conll2013/`

ing counts for all dependencies incident on the word. Document statistics are word-document co-occurrence counts. Count thresholding was applied to reduce noise, and positive pointwise-mutual-information (PPMI) (Church and Hanks, 1990) was applied to the counts. SVD was applied to the document and dependency statistics and the top 1000 dimensions of each type were retained. We selected the rows corresponding to noun-tagged words (approx. 17000 words).

4.2 Brain Activation Data

We have MEG and fMRI data at our disposal. MEG measures the magnetic field caused by many thousands of neurons firing together, and has good time resolution (1000 Hz) but poor spatial resolution. fMRI measures the change in blood oxygenation that results from differential neural activity, and has good spatial resolution but poor time resolution (0.5-1 Hz). We have fMRI data and MEG data for 18 subjects (9 in each imaging modality) viewing 60 concrete nouns (Mitchell et al., 2008; Sudre et al., 2012). The 60 words span 12 word categories (animals, buildings, tools, insects, body parts, furniture, building parts, utensils, vehicles, objects, clothing, food). Each of the 60 words was presented with a line drawing, so word ambiguity is not an issue. For both recording modalities, all trials for a particular word were averaged together to create one training instance per word, with 60 training instances in all for each subject and imaging modality. More preprocessing details appear in the supplementary material.

5 Experimental Results

Here we explore several variations of JNNSE and NNSE formulations. For a comparison of the models used, see Table 1.

5.1 Correlation to Behavioral Data

To test if our joint model of Brain+Text is closer to semantic ground truth we compared the latent representation A learned via JNNSE(Brain+Text) or NNSE(Text) to an independent behavioral measure of semantics. We collected behavioral data for the 60 nouns in the form of answers to 218 semantic questions. Answers were gathered with Mechanical Turk. The full list of questions appear in the supplementary material. Some example questions are:“Is it alive?”, and “Can it bend?”. Mechanical Turk users were asked to respond to each question for each word on a scale of 1-5. At least 3 respondents answered each question and the median score was used. This gives us a semantic representation of each of the 60 words in a 218-dimensional behavioral space. Because we required answers to each of the questions for all words, we do not have the problems of sparsity that exist for feature production norms from other studies (McRae et al., 2005). In addition, our answers are ratings, rather than binary yes/no answers.

For a given value of ℓ we solve the NNSE(Text) and JNNSE(Brain+Text) objective function as detailed in Equation 1 and 4 respectively. We compared JNNSE(Brain+Text) and NNSE(Text) models by measuring the correlation of all pairwise distances in JNNSE(Brain+Text) and NNSE(Text) space to the pairwise distances in the 218-dimensional semantic space. Distances were calculated using normalized Euclidean distance (equivalent in rank-ordering to cosine distance, but more suitable for sparse vectors). Figure 1 shows the results of this correlation test. The error bars for the JNNSE(Brain+Text) models represent a 95% confidence interval calculated using the standard error of the mean (SEM) over the 9 person-specific JNNSE(Brain+Text) models. Because there is only one NNSE(Text) model for each dimension setting, no SEM can be calculated, but it suffices to show that the NNSE(Text) correlation does not fall into the 95% confidence interval of the JNNSE(Brain+Text) models. The SVD matrix for the original corpus data has correlation 0.4279 to the behavioral data, also below the 95% confidence interval for all JNNSE models. The results show that a model that incorporates brain activation data is more faithful to a behavioral measure of semantics.

5.2 Word Prediction from Brain Activation

We now show that the JNNSE(Brain+Text) vectors are more consistent with independent samples of brain activity collected from different subjects, even when recorded using different recording technologies. As previously mentioned, because there is a large degree of variation between brains and because MEG and fMRI measure very different correlates of neuronal activity, this type of generalization has proven to be very challenging and is an open research question in the neuroscience community.

The output A of the JNNSE(Brain+Text) or NNSE(Text) algorithm can be used as a VSM, which we use for the task of word prediction from

Table 1: A Comparison of the models explored in this paper, and the data they operate upon.

Model Name	Section(s)	Text Data	Brain Data	Withheld Data
NNSE(Text)	2, 5	✓	x	-
NNSE(Brain)	2, 5.2.1, 5.3	x	✓	-
JNNSE(Brain+Text)	3, 5	✓	✓	-
JNNSE(Brain+Text): Dropout task	5.2.2	✓	✓	subset of brain data
JNNSE(Brain+Text): Predict corpus	5.3	✓	✓	subset of text data

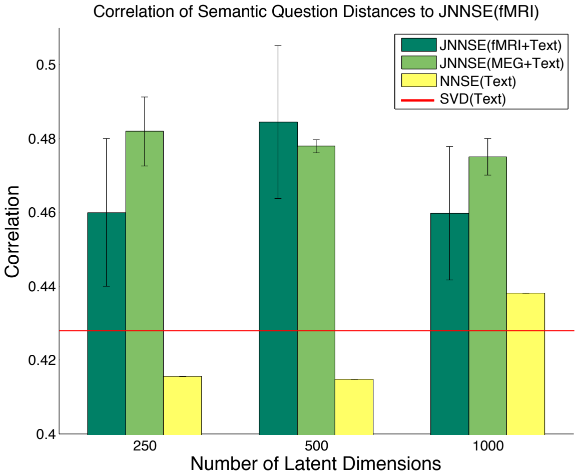


Figure 1: **Correlation of JNNSE(Brain+Text) and NNSE(Text) models with the distances in a semantic space constructed from behavioral data.** Error bars indicate SEM.

fMRI or MEG recordings. A JNNSE(Brain+Text) created with a particular human subject's data is never used in the prediction framework with that same subject. For example, if we use fMRI data from subject 1 to create a JNNSE(fMRI+Text), we will test it with the remaining 8 fMRI subjects, but all 9 MEG subjects (fMRI and MEG subjects are disjoint).

Let us call the VSM learned with JNNSE(Brain+Text) or NNSE(Text) the *semantic vectors*. We can train a weight matrix W that predicts the semantic vector $\mathbf{a}$ of a word from that word's brain activation vector $\mathbf{x}$: $\mathbf{a} = W\mathbf{x}$. W can be learned with a variety of methods, we will use L_2 regularized regression. One can also train regressors that predict the brain activation data from the semantic vector: $\mathbf{x} = W\mathbf{a}$, but we have found this to give lower predictive accuracy. Note that we must *re-train* our weight matrix W for each subject (instead of re-using $D^{(b)}$ from Equation 4) because testing always occurs on a different subject, and the brain activation data is not inter-subject aligned.

We train ℓ independent L_2 regularized regressors to predict the ℓ-dimensional vectors $\mathbf{a} = \{a_1 \dots a_\ell\}$. The predictions are concatenated to produce a *predicted* semantic vector: $\hat{\mathbf{a}} = \{\hat{a}_1, \dots, \hat{a}_\ell\}$. We assess word prediction performance by testing if the model can differentiate between two unseen words, a task named *2 vs. 2 prediction* (Mitchell et al., 2008; Sudre et al., 2012). We choose the assignment of the two held out semantic vectors $(\mathbf{a}^{(1)}, \mathbf{a}^{(2)})$ to predicted semantic vectors $(\hat{\mathbf{a}}^{(1)}, \hat{\mathbf{a}}^{(2)})$ that minimizes the sum of the two normalized Euclidean distances. 2 vs. 2 accuracy is the percentage of tests where the correct assignment is chosen.

The 60 nouns fall into 12 word categories. Words in the same word category (e.g. screwdriver and hammer) are closer in semantic space than words in different word categories, which makes some 2 vs. 2 tests more difficult than others. We choose 150 random pairs of words (with each word represented equally) to estimate the difficulty of a typical word pair, without having to test all $\binom{60}{2}$ word pairs. The same 150 random pairs are used for all subjects and all VSMs. Expected chance performance on the 2 vs. 2 test is 50%.

Results for testing on fMRI data in the 2 vs. 2 framework appear in Figure 2. JNNSE(fMRI+Text) data performed on average 6% better than the best NNSE(Text), and exceeding even the original SVD corpus representations while maintaining interpretability. These results generalize across brain activity recording types; JNNSE(MEG+Text) performs as well as JNNSE(fMRI+Text) when tested on fMRI data. The results are consistent when testing on MEG data: JNNSE(MEG+Text) or JNNSE(fMRI+Text) outperforms NNSE(Text) (see Figure 3).

NNSE(Text) performance decreases as the number of latent dimension increases. This im-

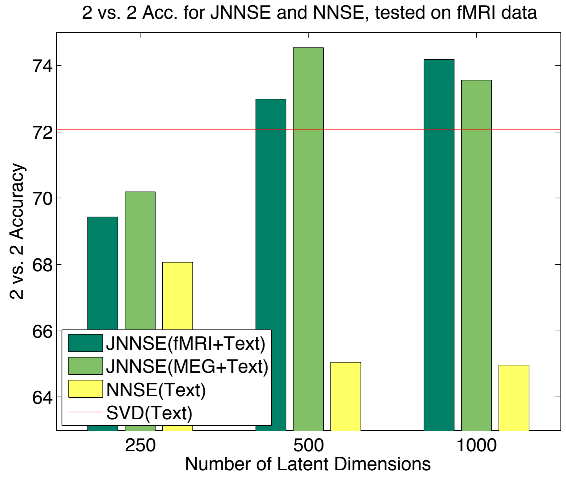


Figure 2: **Average 2 vs. 2 accuracy for NNSE(Text) and JNNSE(Brain+Text), tested on fMRI data.** Models created with one subject's fMRI data were not used to compute 2 vs. 2 accuracy for that same subject.

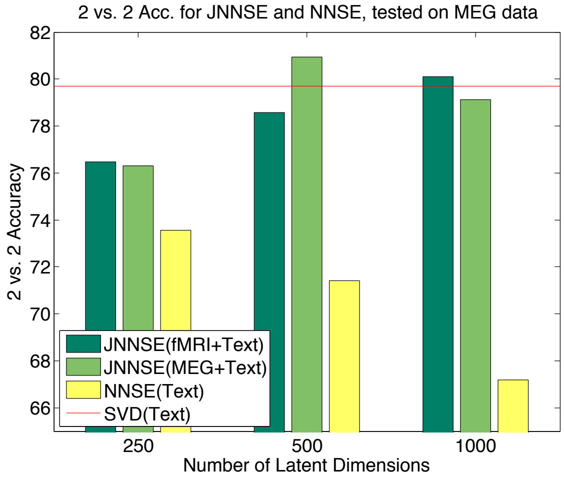


Figure 3: **Average 2 vs. 2 accuracy for NNSE(Text) and JNNSE(Brain+Text), tested on MEG data**. Models created with one subject's MEG data were not used to compute 2 vs. 2 accuracy for that same subject.

plies that without the regularizing effect of brain activation data, the extra NNSE(Text) dimensions are being used to overfit to the corpus data, or possibly to fit semantic properties not detectable with current brain imaging technologies. However, when brain activation data is included, increasing the number of latent dimensions strictly increases performance for JNNSE(fMRI+Text). JNNSE(MEG+Text) has peak performance with 500 latent dimensions, with $\sim 1\%$ decrease in performance at 1000 latent dimensions. In previous work, the ability to decode words from brain activation data was found to improve with added latent dimensions (Murphy et al., 2012a). Our results may differ because our words are POS tagged, and we included only nouns for the final NNSE(Text) model. We found that with the original $\lambda = 0.05$ setting from Murphy et al. (Murphy et al., 2012a) produced vectors that were too sparse; four of the 60 test words had all-zero vectors (JNNSE(Brain+Text) models did have any all-zero vectors). To improve the NNSE(Text) vectors for a fair comparison, we reduced $\lambda = 0.025$, under which NNSE(Text) did not produce any all-zero vectors for the 60 words.

Our results show that brain activation data contributes additional information, which leads to an increase in performance for the task of word prediction from brain activation data. This suggests that corpus-only models may not capture all relevant semantic information. This conflicts with previous studies which found that semantic vectors culled from corpus statistics contain all of the semantic information required to predict brain activation (Bullinaria and Levy, 2013).

5.2.1 Prediction from a Brain-only Model

How much predictive power does the corpus data provide to this word prediction task? To test this, we calculated the 2 vs. 2 accuracy for a NNSE(Brain) model trained on brain activation data only. We train NNSE(Brain) with one subject's data and use the resulting vectors to calculate 2 vs. 2 accuracy for the remaining subjects. We have brain data for only 60 words, so using $\ell \geq 60$ latent dimensions leads to an under-constrained system and a degenerate solution wherein only one latent dimension is active for any word (and where the brain data can be perfectly reconstructed). The degenerate solution makes it impossible to generalize across words and leads to performance at chance levels. An NNSE(MEG) trained on MEG data gave maximum 2 vs. 2 accuracy of 67% when $\ell = 20$. The reduced performance may be due to the limited training data and the low SNR of the data, but could also be attributed to the lack of corpus information, which provides another piece of semantic information.

5.2.2 Effect on Rows Without Brain Data

It is possible that some JNNSE(Brain+Text) dimensions are being used exclusively to fit brain activation data, and not the semantics represented in both brain and corpus data. If a particular dimension j is solely used for brain data, the sparsity constraint will favor solutions that sets $A_{(i,j)} = 0$ for $i > w'$ (no brain data constraint), and $A_{(i,j)} > 0$ for some $0 \leq i \leq w'$ (brain data constrained). We found that there were no such dimensions in the JNNSE(Brain+Text). In fact for the $\ell = 1000$ JNNSE(Brain+Text), all latent dimensions had greater than $\sim 25\%$ non-zero entries, which implies that all dimensions are being shared between the two data inputs (corpus and brain activation), and are used to reconstruct both.

To test that the brain activation data is truly influencing rows of A not constrained by brain activation data, we performed a *dropout* test. We split the original 60 words into two 30 word groups (as evenly as possible across word categories). We trained JNNSE(fMRI+Text) with 30 words, and tested word prediction with the remaining 8 subjects and the other 30 words. Thus, the training and testing word sets are disjoint. Because of the reduced size of the training data, we did see a drop in performance, but JNNSE(fMRI+Text) vectors still gave word prediction performance 7% higher than NNSE(Text) vectors. Full results appear in the supplementary material.

5.3 Predicting Corpus Data

Here we ask: can an accurate latent representation of a word be constructed using only brain activation data? This task simulates the scenario where there is no reliable corpus representation of a word, but brain data is available. This scenario may occur for seldom-used words that fall below the thresholds used for the compilation of corpus statistics. It could also be useful for acronym tokens (lol, omg) found in social media contexts where the meaning of the token is actually a full sentence.

We trained a JNNSE(fMRI+Text) with brain data for all 60 words, but withhold the corpus data for 30 of the 60 words (as evenly distributed as possible amongst the 12 word categories). The brain activation data for the 30 withheld words will allow us to create latent representations in A for withheld words. Simultaneously, we will learn a mapping from the latent representation to the corpus data ($D^{(c)}$). This task cannot be performed with a NNSE(Text) model because one cannot learn a latent representation of a word without data of some kind. This further emphasizes the impact of brain imaging data, which will allow us to generalize to previously unseen words in corpus space.

Table 2: Mean rank accuracy over 30 words using corpus representations predicted by a JNNSE(MEG+Text) model trained with some rows of the corpus data withheld. Significance is calculated using Fisher's method to combine p-values for each of the subject-dependent models.

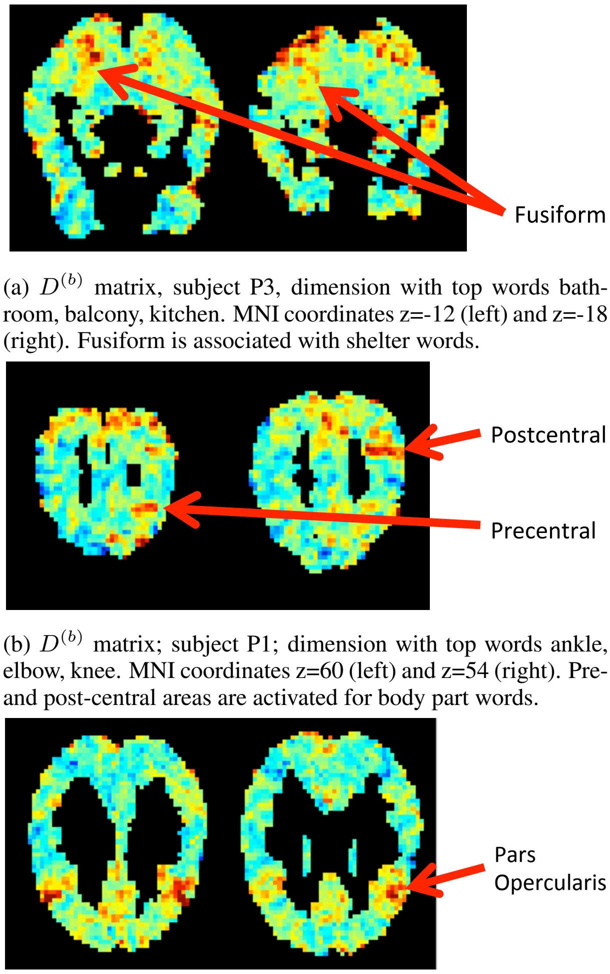
Latent Dim size	Rank Accuracy	p-value
250	65.30	$< 10^{-19}$
500	67.37	$< 10^{-24}$
1000	63.47	$< 10^{-15}$

We use the latent representations in A for each of the words without corpus data and the mapping to corpus space $D^{(c)}$ to predict the withheld corpus data in X. We then rank the withheld rows of X by their distance to the predicted row of X and calculate the mean rank accuracy of the held out words. Results in Table 2 show that we can recreate the withheld corpus data using brain activation data. Peak mean rank accuracy (67.37) is attained at $\ell = 500$ latent dimensions. This result shows that neural semantic representations can create a latent representation that is faithful to unseen corpus statistics, providing further evidence that the two data sources share a strong common element.

How much power is the remaining corpus data supplying in scenarios where we withhold corpus data? To answer this question, we trained an NNSE(Brain) model on 30 words of brain activation, and then trained a regressor to predict corpus data from those latent brain-only representations. We use the trained regressor to predict the corpus data for the remaining 30 words. Peak performance is attained at $\ell = 10$ latent dimensions, giving mean rank accuracy of 62.37, significantly worse than the model that includes both corpus and brain activation data (67.37).

5.4 Mapping Semantics onto the Brain

Because our method incorporates brain data into an interpretable semantic model, we can directly map semantic concepts onto the brain. To do this, we examined the mappings from the latent space to the brain space via $D^{(b)}$. We found that the most interpretable mappings come from models where the perceptual features had been scaled down (divided by a constant factor), which encourages more of the data to be explained by the semantic features in A. Figure 4 shows the mappings ($D^{(b)}$) for dimensions related to shelter, food and body parts. The red areas align with areas of the brain previously known to be activated by the corresponding concepts (Mitchell et al., 2008; Just et al., 2010). Our model has learned these mappings in an unsupervised setting by relating semantic knowledge gleaned from word usage to patterns of activation in the brain. This illustrates how the interpretability of JNNSE can allow one to explore semantics in the human brain. The mappings for one subject are available for download (`http://www.cs.cmu.edu/~afyshe/papers/acl2014/`).



(a) $D^{(b)}$ matrix, subject P3, dimension with top words bathroom, balcony, kitchen. MNI coordinates z=-12 (left) and z=-18 (right). Fusiform is associated with shelter words.

(b) $D^{(b)}$ matrix; subject P1; dimension with top words ankle, elbow, knee. MNI coordinates z=60 (left) and z=54 (right). Pre- and post-central areas are activated for body part words.

(c) $D^{(b)}$ matrix; subject P1; dimension with top scoring words buffet, brunch, lunch. MNI coordinates z=30 (left) and z=24 (right). Pars opercularis is believed to be part of the gustatory cortex, which responds to food related words.

Figure 4: **The mappings ($D^{(b)}$) from latent semantic space (A) to brain space (Y) for fMRI and words from three semantic categories.** Shown are representations of the fMRI slices such that the back of the head is at the top of the image, the front of the head is at the bottom.

6 Future Work and Conclusion

We are interested in pursuing many future projects inspired by the success of this model. We would like to extend the JNNSE algorithm to incorporate data from multiple subjects, multiple modalities and multiple experiments with non-overlapping words. Including behavioral data and image data is another possibility.

We have explored a model of semantics that incorporates text and brain activation data. Though the number of words for which we have brain activation data is comparatively small, we have shown that including even this small amount of data has a positive impact on the learned latent representations, including for words without brain data. We have provided evidence that the latent representations are closer to the neural representation of semantics, and possibly, closer to semantic ground truth. Our results reveal that there are aspects of semantics not currently represented in text-based VSMs, indicating that there may be room for improvement in either the data or algorithms used to create VSMs. Our findings also indicate that using the brain as a semantic test can separate models that capture this additional semantic information from those that do not. Thus, the brain is an important source of both training and testing data.

Acknowledgments

This work was supported in part by NIH under award 5R01HD075328-02, by DARPA under award FA8750-13-2-0005, and by a fellowship to Alona Fyshe from the Multimodal Neuroimaging Training Program funded by NIH awards T90DA022761 and R90DA023420.

References

Andrew J Anderson, Elia Bruni, Ulisse Bordignon, Massimo Poesio, and Marco Baroni. 2013. Of words , eyes and brains : Correlating image-based distributional semantic models with neural representations of concepts. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing*.

David M Blei and Jon D. McAuliffe. 2007. Supervised topic models. In *Advances in Neural Information Processing Systems*, pages 1–22.

Elia Bruni, Giang Binh Tran, and Marco Baroni. 2011. Distributional semantics from text and images. In *Proceedings of the EMNLP 2011 Geometrical Models for Natural Language Semantics (GEMS)*.

John A Bullinaria and Joseph P Levy. 2013. Limiting factors for mapping corpus-based semantic representations to brain activity. *PloS one*, 8(3):e57191, January.

Jamie Callan and Mark Hoy. 2009. The ClueWeb09 Dataset.

Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M Blei. 2009. Reading Tea Leaves : How Humans Interpret Topic Models. In *Advances in Neural Information Processing Systems*, pages 1–9.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

Bryan R Conroy, Benjamin D Singer, J Swaroop Guntupalli, Peter J Ramadge, and James V Haxby. 2013. Inter-subject alignment of human cortical anatomy using functional connectivity. *NeuroImage*, 81:400–11, November.

Andrew D Engell, Scott Huettel, and Gregory McCarthy. 2012. The fMRI BOLD signal tracks electrophysiological spectral perturbations, not event-related potentials. *NeuroImage*, 59(3):2600–6, February.

Alona Fyshe, Partha Talukdar, Brian Murphy, and Tom Mitchell. 2013. Documents and Dependencies : an Exploration of Vector Space Models for Semantic Composition. In *Computational Natural Language Learning*, Sofia, Bulgaria.

Arthur M Glenberg and David a Robertson. 2000. Symbol Grounding and Meaning: A Comparison of High-Dimensional and Embodied Theories of Meaning. *Journal of Memory and Language*, 43(3):379–401, October.

Sunil Kumar Gupta, Dinh Phung, Brett Adams, and Svetha Venkatesh. 2013. Regularized nonnegative shared subspace learning. *Data Mining and Knowledge Discovery*, 26(1):57–97.

Emma L Hall, Siân E Robson, Peter G Morris, and Matthew J Brookes. 2013. The relationship between MEG and fMRI. *NeuroImage*, November.

Harold Hotelling. 1936. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377.

Yangqing Jia and Trevor Darrell. 2010. Factorized Latent Spaces with Structured Sparsity. In *Advances in Neural Information Processing Systems*, volume 23.

Marcel Adam Just, Vladimir L Cherkassky, Sandesh Aryal, and Tom M Mitchell. 2010. A neurosemantic theory of concrete noun representation based on the underlying brain codes. *PloS one*, 5(1):e8622, January.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. 2008a. Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2(November):4, January.

Nikolaus Kriegeskorte, Marieke Mur, Douglas A Ruff, Roozbeh Kiani, Jerzy Bodurka, Hossein Esteky, Keiji Tanaka, and Peter A Bandettin. 2008b. Matching Categorical Object Representations in Inferior Temporal Cortex of Man and Monkey. *Neuron*, 60(6):1126–1141.

TK Landauer and ST Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 1(2):211–240.

Julien Mairal, Francis Bach, J Ponce, and Guillermo Sapiro. 2010. Online learning for matrix factorization and sparse coding. *The Journal of Machine Learning Research*, 11:19–60.

Ken McRae, George S Cree, Mark S Seidenberg, and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, 37(4):547–59, November.

Tom M Mitchell, Svetlana V Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L Malave, Robert A Mason, and Marcel Adam Just. 2008. Predicting human brain activity associated with the meanings of nouns. *Science (New York, N.Y.)*, 320(5880):1191–5, May.

Brian Murphy, Partha Talukdar, and Tom Mitchell. 2012a. Learning Effective and Interpretable Semantic Models using Non-Negative Sparse Embedding. In *Proceedings of Conference on Computational Linguistics (COLING)*.

Brian Murphy, Partha Talukdar, and Tom Mitchell. 2012b. Selecting Corpus-Semantic Models for Neurolinguistic Decoding. In *First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 114–123, Montreal, Quebec, Canada.

Andrew Y. Ng and Michael I. Jordan. 2002. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in neural information processing systems*, volume 14.

Mark Palatucci, Geoffrey Hinton, Dean Pomerleau, and Tom M Mitchell. 2009. Zero-Shot Learning with Semantic Output Codes. *Advances in Neural Information Processing Systems*, 22:1410–1418.

Rajeev D S Raizada and Andrew C Connolly. 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.

Indrayana Rustandi, Marcel Adam Just, and Tom M Mitchell. 2009. Integrating Multiple-Study Multiple-Subject fMRI Datasets Using Canonical Correlation Analysis. In *MICCAI 2009 Workshop: Statistical modeling and detection issues in intra- and inter-subject functional MRI data analysis*.

Magnus Sahlgren. 2006. *The Word-Space Model Using distributional analysis to represent syntagmatic and paradigmatic relations between words*. Doctor of philosophy, Stockholm University.

Carina Silberer and Mirella Lapata. 2012. Grounded models of semantic representation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1423–1433.

Carina Silberer, Vittorio Ferrari, and Mirella Lapata. 2013. Models of Semantic Representation with Visual Attributes. In *Association for Computational Linguistics 2013*, Sofia, Bulgaria.

Gustavo Sudre, Dean Pomerleau, Mark Palatucci, Leila Wehbe, Alona Fyshe, Riitta Salmelin, and Tom Mitchell. 2012. Tracking Neural Coding of Perceptual and Semantic Features of Concrete Nouns. *NeuroImage*, 62(1):463–451, May.

Peter D Turney and Patrick Pantel. 2010. From Frequency to Meaning : Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Martha White, Yaoliang Yu, Xinhua Zhang, and Dale Schuurmans. 2012. Convex multi-view subspace learning. In *Advances in Neural Information Processing Systems*, pages 1–14.