Exploring Semantic Representation in Brain Activity Using Word Embeddings

Yu-Ping Ruan[1], **Zhen-Hua Ling**[1] **and Yu Hu**[1,2]
[1]National Engineering Laboratory for Speech and Language Information Processing
University of Science and Technology of China, Hefei, China
[2]iFLYTEK Research, Hefei, China
ypruan@mail.ustc.edu.cn, zhling@ustc.edu.cn, yuhu@iflytek.com

Abstract

In this paper, we utilize distributed word representations (i.e., word embeddings) to analyse the representation of semantics in brain activity. The brain activity data were recorded using functional magnetic resonance imaging (fMRI) when subjects were viewing words. First, we analysed the functional selectivity of different cortex areas by calculating the correlations between neural responses and several types of word representations, including skip-gram word embeddings, visual semantic vectors, and primary visual features. The results demonstrated consistency with existing neuroscientific knowledge. Second, we utilized behavioural data as the semantic ground truth to measure their relevance with brain activity. A method to estimate word embeddings under the constraints of brain activity similarities is further proposed based on the semantic word embedding (SWE) model. The experimental results show that the brain activity data are significantly correlated with the behavioural data of human judgements on semantic similarity. The correlations between the estimated word embeddings and the semantic ground truth can be effectively improved after integrating the brain activity data for learning, which implies that semantic patterns in neural representations may exist that have not been fully captured by state-of-the-art word embeddings derived from text corpora.

1 Introduction

Recently, the topic of exploring semantic representation in human brain has attracted the attention of researchers from both neuroscience and computational linguistics fields. In these studies, concepts are represented in terms of neural activation patterns in the brain that can be recorded by functional magnetic resonance imaging (fMRI) (Haxby et al., 2001). It has been found that the semantic space shared among different individuals is distributed continuously across the cortex (Huth et al., 2012). A recent study proposed an efficient way to measure and visualize the semantic selectivity of different cortex areas (Huth et al., 2016).

Similar to the distributed semantic representation in the brain, describing the meaning of a word using a dense low-dimensional and continuous vector (i.e., word embedding) is currently a popular approach in computational linguistics (Hinton et al., 1986; Turney et al., 2010). Word embeddings are commonly estimated from large text corpora utilizing statistics concerning the co-occurrences of words (Mikolov et al., 2013a; Mikolov et al., 2013b; Pennington et al., 2014). To investigate the correlation between word embeddings and the brain activity involved in viewing words, Mitchell et al. (2008) designed a computational model to predict brain responses using hand-tailored word embeddings as input. Further, Fyshe et al. (2014) proposed a joint non-negative sparse embedding (JNNSE) method to combine fMRI data and textual data to estimate word embeddings. This work improved the correlation between word embeddings and human behavioural data, which lends support to the view that fMRI data can provide additional semantic information that may not exist in textual data.

Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 669–679,
Austin, Texas, November 1-5, 2016.

The factors that can influence the activities of cortex areas are diverse. Recent studies show that visual semantic features such as bag-of-visual-words (BoVW) are significantly correlated with the fMRI data captured when viewing words (Anderson et al., 2013). The primary visual features derived using Gabor wavelets can be used to determine the images presented to the subjects from their recorded brain activity (Kay et al., 2008; Naselaris et al., 2009). Some other research work also indicates that visual experiences (Nishimoto et al., 2011) and speech information (Ryali et al., 2010) can affect neural responses in cortex areas.

In this paper, we first study the semantic representation of words in brain activity by correlation analysis (Anderson et al., 2013; Carlson et al., 2014). Then, we calculate the correlations between subjects' neural responses when viewing words and three types of word representations: skip-gram word embeddings, primary visual features, and visual semantic vectors. The goal of doing this is to investigate whether these representations can account for the brain data and the functional selectivity of different cortex areas. Then, we utilize behavioural data as the semantic ground truth to measure the semantic relevance of brain activity. A method of estimating word embeddings within the constraints of similar brain activities is proposed. This method is based on the semantic word embedding (SWE) model (Liu et al., 2015) which develops from the skip-gram model. It aims at verifying whether textual data and brain activity data can be complementary to derive word embeddings that are more consistent with human judgement.

The contributions of this study are twofold. First, this study involved a comprehensive correlation analysis on brain activity data and state-of-the-art skip-gram word embeddings at both whole-brain and brain lobe levels. Primary visual features and visual semantic vectors are also introduced as auxiliary representations to better understand the functional selectivity across the cortex. Some results of this analysis are interpretable using existing neuroscience knowledge. Second, to our knowledge, this study marks the first attempt to integrate brain activity data into the skip-gram model for estimating word embeddings. The experimental results show that the correlation between the estimated word embeddings and the behavioural measure of semantics can be effectively improved after integrating brain activity data for learning.

2 Related work

The correlation between brain data and word vectors has been studied in previous work. The experiments in Carlson et al. (2014) adopted brain activity data for correlation analysis from only the ventral temporal pathway, not from the whole brain. Anderson et al. (2013) performed correlation analysis using the voxels of the whole brain and compared the HAL-based textual semantic model (Lund and Burgess, 1996) with the BoVW-based visual semantic model (Sivic and Zisserman, 2003; Csurka et al., 2004) in terms of these two model's ability to account for the patterns found in the neural data. However, the experiments in Anderson et al. (2013) failed to detect differential interactions of semantic models with brain areas. In this paper, considering the popularity of word embedding estimation approaches based on neural networks in recent years, we adopt skip-gram word embeddings (Mikolov et al., 2013a) for correlation analysis. To our knowledge, this is the first time that the association between skip-gram word embeddings and brain activity data have been studied. Furthermore, our work improves on the voxel selection strategy used in Anderson et al. (2013), leading to more interpretable results when demonstrating the functional selectivity of brain areas.

To our knowledge, the first and only attempt to integrate brain activity data into the acquisition of textual word embedding is the JNNSE method (Fyshe et al., 2014). In this method, word embeddings were estimated as latent representations using matrix factorization. The objective functions contained additional constraints for reconstructing brain activity data. In this paper, we adopt the SWE model (Liu et al., 2015) to incorporate brain activity knowledge into word embedding estimation. The SWE model was developed from the skip-gram model. In SWE, semantically related knowledge is converted into inequality constraints for learning word embeddings. The experimental results show that our proposed method using SWE can improve

the semantic consistency between word embeddings and human judgements.

3 From Skip-Gram to SWE

3.1 Skip-gram model

The skip-gram model (Mikolov et al., 2013b) adopts a neural network structure to derive the distributed representation of words from textual corpus. The word vectors are learned based on the distributional hypothesis (Harris, 1954; Miller and Charles, 1991), which assumes that words with similar contexts tend to have similar semantic meanings. For a sequence of training data of T words, denoted as $\{w_1, w_2, w_3, \cdots, w_T\}$, the skip-gram model is trained to maximize the following objective function

$$Q = \frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j}|w_t), \quad (1)$$

where w_t and w_{w+j} are the central word and neighbouring words in a context window respectively, and c denotes the size of the context window. The conditional probability $p(w_{t+j}|w_t)$ in Eq.(1) is calculated as

$$p(w_{t+j}|w_t) = \frac{\exp(\mathbf{w}_{t+j}^{(2)} \cdot \mathbf{w}_t^{(1)})}{\sum_{k=1}^{V}\exp(\mathbf{w}_k^{(2)} \cdot \mathbf{w}_t^{(1)})}, \quad (2)$$

where $\mathbf{w}_t^{(1)}$ and $\mathbf{w}_k^{(2)}$ denote row vectors in the matrices $\mathbf{W}^{(1)}$ and $\mathbf{W}^{(2)}$ respectively, and V is the vocabulary size of the corpus. The matrix $\mathbf{W}^{(1)}$ stores the word vectors of input central words, and the matrix $\mathbf{W}^{(2)}$ stores the word vectors of predicted neighbouring words. The optimization of the objective function Q is solved by the stochastic gradient descent (SGD) method (Mikolov et al., 2013b). Finally, the learned matrix $\mathbf{W}^{(1)}$ is used as the estimated word embeddings of all words in the vocabulary.

3.2 Semantic word embedding (SWE)

The skip-gram model learns word embeddings based on the distributional hypothesis; however, this hypothesis still has some limitations. For example, antonyms often appear in similar contexts although they have opposite meanings. The semantic word embedding (SWE) model (Liu et al., 2015) has been proposed to address this issue by incorporating external semantic knowledge into the text-based learning process for word embeddings.

In this method, semantic knowledge is represented as a set of ranking inequalities. Each inequality contains a triplet (i, j, k) of three words $\{w_i, w_j, w_k\}$ with a similarity relation

$$similarity(w_i, w_j) > similarity(w_i, w_k), \quad (3)$$

which can be notated in simplified form as $s_{ij} > s_{ik}$. Then, the learning method of SWE is defined as the following constrained optimization problem

$$\{\mathbf{W}^{(1)}, \mathbf{W}^{(2)}\} = arg\max_{\mathbf{W}^{(1)},\mathbf{W}^{(2)}} Q(\mathbf{W}^{(1)}, \mathbf{W}^{(2)}),$$
$$s.t.\ s_{ij} > s_{ik}, \forall (i, j, k) \in S, \quad (4)$$

where function Q is defined in Eq. (1) and S denotes the inequality set. Then, the constrained optimization problem in Eq. (4) is simplified into an unconstrained problem by introducing a penalty term into the objective function of the skip-gram model. The penalty term is defined as

$$D = \sum_{(i,j,k)\in S} f(i, j, k), \quad (5)$$

where $f(i, j, k) = \max(0, s_{ik} - s_{ij})$ is a Hinge loss function. Finally, the object function to be maximized in SWE can be written as follows:

$$Q' = Q - \beta \cdot D, \quad (6)$$

where β is a parameter to control the contribution of the penalty term. Similar to the skip-gram model, the Q' function in the SWE model is optimized using SGD to estimate word embeddings. The detailed formulae can be found in Liu et al. (2015).

3.3 Integrating brain activity into SWE

In the implementation of the SWE model in Liu et al. (2015), the ranking inequalities were collected using hypernym-hyponym and synonym-antonym relationships extracted from WordNet (Fellbaum and others, 1998). In this paper, the SWE model is utilized as a tool to explore the semantic relevance of brain activity by examining the performance of the estimated word embeddings after integrating brain-activity-related knowledge. Therefore, we construct

the ranking inequalities in Eq. (3) using brain activity data. When a subject is viewing a word, the neural response in the cortex is captured using fMRI and further stored as a vector. After collecting the fMRI data for a set of words, the inequalities in Eq. (3) can be constructed by using a similarity measure on the neural response vectors of word pairs. Here, we adopt Pearson correlation as the similarity measure. The details will be introduced in Section 5.1.

4 Data

4.1 Brain data

The fMRI data used in our experiments was recorded and preprocessed by Mitchell et al. (2008). It includes the recorded data of 9 subjects. To record the data, each of 60 concrete nouns was presented visually to each subject with a textual label and a simple line drawing. The subjects were asked to think about the properties of the objects indicated by the words during fMRI scanning. This procedure repeated 6 times, and the stimuli of the 60 nouns were presented in a random order in each run. More details about the data acquisition and preprocessing procedures can be found in Mitchell et al. (2008) and its supplement materials. Finally, an fMRI vector measuring the neural response at all voxels across the cortex was created for each word and each subject.

4.2 Behavioural data

The behavioural data collects human judgements on the semantic similarity between word pairs. The approach to behavioural data collection in our experiment is similar to the one used in the WordSim-353 dataset (Finkelstein et al., 2001). For the 60 concrete nouns used in Section 4.1, we obtained $C_{60}^2 = 1,770$ word pairs. Then, we asked 15 participants to score the semantic similarity of each word pair on a scale from 0 to 10, in which "0" signified that the two words were totally unrelated and "10" signified that the two words were highly related or had identical meanings. This collection procedure was conducted on the Amazon Mechanical Turk[1] crowdsourcing platform. We tested the average Spearman correlation coefficient among the scores given by different annotators and found that it was approximately 0.4873 with a p-value of 1.1e-02. After gathering the scores for all the word pairs, the highest and lowest scores for each word pair were discarded, and the average of the remaining 13 scores was calculated as the similarity score for each word pair[2].

To verify the reliability of the above data collection process, we also added 15 word pairs from the WordSim-353 dataset into our 1,770 word pairs during score collection. Then, we calculated the similarity scores of these 15 word pairs using the collected scores and compared them with the scores in the WordSim-353 dataset using Spearman correlation analysis. The correlation coefficient was 0.8451 with a p-value of $2.7e-04$. This high correlation verifies the reliability of our behavioural data collection.

5 Experiments

5.1 Correlations between brain activity and word vectors

We calculated the correlations between the fMRI vectors and the different types of word representations to investigate whether these representations can account for the brain activity and the functional selectivity of different cortex areas. We adopted the method of representational similarity analysis (Kriegeskorte et al., 2008) in our experiments. For a specific word representation, we calculated the cosine similarity for each word pair in a set of n words, resulting in a similarity vector with a total length of C_n^2. For the fMRI data[3], we constructed a similarity vector for each subject using the Pearson correlation coefficients between pairs of fMRI vectors (Anderson et al., 2013). Then, the 9 vectors of the 9 subjects were averaged to obtain an overall similarity vector in the fMRI space (Anderson et al., 2013). Finally, the Spearman rank correlation coefficient between the similarity vectors given by the fMRI data and each word representation was calculated together with a p-value for significance

[1] http://www.mturk.com/

[2] The behavioural data are available at http://home.ustc.edu.cn/~ypruan/work/emnlp2016/behaviour_data/

[3] Before using the fMRI data, we first regularized its mean value to 0 and variance to 1.

analysis. The p-value was calculated using a permutation test under a positive hypothesis with the word pair labels randomly shuffled 10,000 times. Empirically, two similarity vectors are considered to be correlated when $p < 0.05$, and they are considered significantly correlated when $p < 0.01$.

5.1.1 Word vectors

Three types of word representations, i.e., skip-gram word embeddings, visual semantic vectors, and primary visual features, were used in the correlation analysis. Some details about the acquisitions of these three word representations will be introduced in the following paragraphs.

Skip-gram word embeddings The Wikipedia text corpus[4], containing 130 million words, was adopted to train our skip-gram word embeddings, and the hierarchical softmax scheme was followed. The dimension of word embedding was 200. The window size, learning rate, and negative sampling number were set to 8, 0.05, and 8, respectively. The model was trained for one iteration using a single execution thread.

Visual semantic vectors On one hand, distributed word representations are usually learnt from text corpora. On the other hand, visual perception also contributes to semantic cognition according to some neuroscience research (Louwerse, 2011), and it has been utilized to complement the semantic representation learned from texts (Bruni et al., 2012). One approach to constructing visual semantic vectors is to first extract the low-level visual features from images and then convert them into higher-level semantic representations using the bag-of-visual-words (BoVW) (Grauman and Leibe, 2011) model. In our experiments, we built the BoVW representations from ImageNet (Deng et al., 2009) using the VSEM[5] toolkit. Due to coverage limitations, only 57 of the 60 concrete nouns in the fMRI data could be found in ImageNet[6] and each noun has approximately 1000 image samples. Similar to Anderson et al. (2013), we adopted the Scale Invariant Feature Transform (SIFT) (Lowe, 2004) to extract lower-level visual features; however, we did not use the "object" box to discriminate "object" and "context" areas during the extraction. Then, we clustered the SIFT features into 1000 classes to construct the visual vocabulary, and each image was divided into 8 regions. Thus, the BoVW representation of an image was a vector of 8000 dimensions. The BoVW vectors of all images in ImageNet corresponding to the same word were averaged to obtain the BoVW representation of that word. Finally, we transformed the BoVW representation matrix of the 57 nouns to nonnegative point-wise mutual information (PMI) association scores (Church and Hanks, 1990) to obtain the final visual semantic vectors.

Primary visual features As introduced in Section 4.1, a line drawing of each word was presented to subjects together with the textual label when collecting the fMRI data (Mitchell et al., 2008). This presentation led to neural responses in visual cortices that may be irrelevant to semantic representation. Because the receptive fields of simple cells in the primary visual cortex of mammalian brains can be modelled by Gabor functions (Marĉelja, 1980; Daugman, 1985), we adopted Gabor wavelets to extract the primary visual features from the line drawings of the 60 nouns and further analysed their correlations with fMRI data. The original resolution of the image stimuli used in Mitchell et al. (2008) was 500 x 500 pixels. These images were converted to 64 x 64 pixels after trimming the black borders and downsampling. The Gabor wavelet filter bank was designed using an open source tool (Haghighat et al., 2015). The number of scales and orientations were set to 5 and 8, respectively. Thus, we represented the primary visual features of each noun as a vector of 163,840 dimensions. The singular value decomposition (SVD) technique was employed to reduce the dimension of each vector to 60.

5.1.2 Correlation analysis at the whole-brain level

The fMRI recording measures the neural responses of more than 20,000 voxels across the cortex. To perform dimensionality reduction, we selected 500 voxels from all voxels for each subject according to

[4] http://mattmahoney.net/dc/enwik9.zip

[5] http://clic.cimec.unitn.it/vsem/

[6] The three missing words are *arm*, *eye* and *saw*.

word representation	rho (p-value)
skip-gram	0.0065 (4.0e-01)
BoVW	**0.3515** (0.0e-00)
Gabor	**0.3924** (0.0e-00)

Table 1: Spearman's rank correlation coefficients (**rho**) between different word representations and whole-brain fMRI data for 57 nouns and their corresponding p-values.

Lobe	Proportion (%)
frontal	5.89
temporal	6.96
parietal	10.13
occipital	58.40
other	18.62

Table 2: The proportions of the regional distributions of the 500 selected voxels.

the stability of the voxel responses across 6 runs of fMRI recordings. This selection strategy was the same as the one used in Mitchell et al. (2008) and Anderson et al. (2013). The correlation analysis followed the method described at the beginning of Section 5.1. Table 1 shows the results, where *skip-gram*, *BoVW*, and *Gabor* denote the skip-gram word embeddings, visual semantic vectors, and primary visual features introduced above, respectively.

As Table 1 shows, the visual semantic vectors and primary visual features are significantly correlated with the fMRI vectors at the whole-brain level; however, the skip-gram word embeddings are not correlated with the fMRI data. To investigate the reason for this lack of correlation, we analysed the distribution of the 500 selected voxels across the four brain lobes (i.e., frontal, temporal, parietal and occipital) using the automated anatomical labeling scheme (Tzourio-Mazoyer et al., 2002). From the results shown in Table 2, we can find that most of the selected voxels are located in the occipital lobe although it is the smallest of the four main lobes in the human brain. The occipital lobe occupies most of the anatomical area of the visual cortex and is considered to be the visual processing centre of the mammalian brain. This unbalanced distribution led to the conclusion that the semantic information related to skip-gram word embeddings is not well represented by the 500 selected voxels. Thus, an alternative strategy to select stable voxels at the brain lobe level for correlation analysis was necessary.

5.1.3 Correlation analysis at the brain lobe level

As an alternative approach, rather than selecting the 500 most stable voxels from the whole-brain data as in (Mitchell et al., 2008; Anderson et al., 2013), we selected the 100 most stable voxels at each of the four main brain lobes independently for this experiment. Then, the correlations between the fMRI vectors measuring different lobes and word representations were calculated and are shown in Table 3.

From this table, we can observe the association differences of different word representations with brain lobe level activities. First, the primary visual features (*Gabor*) are highly correlated with the occipital fMRI data and are uncorrelated with the other three lobes. This is reasonable considering that the primary visual cortex (V1) is located in the occipital lobe. Second, the skip-gram word embeddings are significantly correlated with the fMRI data of all brain lobes except the occipital lobe. Previous neuroscience research has revealed that the frontal, temporal, and parietal lobes all play important roles in semantic cognition, including high-level and abstract knowledge processing (Miller et al., 2002), integration of lexical information (Hagoort, 2005), speech comprehension (Hickok and Poeppel, 2007), and knowledge retrieval (Binder et al., 2009). This indicates that the skip-gram word embeddings can partly account for the semantic processing in the cortex and contain little visual information about words. Third, the visual sematic vectors (*BoVW*) are significantly correlated with all four brain lobes. It has been found that the temporal lobe plays a key role in both the formation of long-term visual memories (Smith and Kosslyn, 2007) and in the recognition of visual stimuli and objects (Chao et al., 1999; Kanwisher and Yovel, 2006). The parietal lobe is relevant to high-level vision and is part of the dorsal visual stream correlated with spatial cognition (Sack, 2009; Vannini et al., 2004). This indicates that the visual sematic vectors used in our experiment may contain not only low-level but also high-level and semantically related visual information.

	Frontal	Temporal	Parietal	Occipital
Skip-gram	**0.1450** (0.0e+00)	**0.1483** (0.0e+00)	**0.2317** (0.0e+00)	-0.0385 (9.4e-01)
BoVW	**0.0601** (8.2e-03)	**0.2053** (0.0e+00)	**0.2750** (0.0e+00)	**0.3120** (0.0e+00)
Gabor	-0.0823 (1.0e+00)	-0.0879 (1.0e+00)	0.0111 (3.4e-01)	**0.5116** (0.0e+00)

Table 3: Spearman's rank correlation coefficients (**rho**) between different word representations and the fMRI data at four main brain lobes and their corresponding p-values.

fMRI data	rho (p-value)
whole brain	**0.1266** (0.0e+00)
frontal lobe	0.0160 (2.5e-01)
temporal lobe	**0.0694** (1.7e-03)
parietal lobe	**0.0698** (1.6e-03)
occipital lobe	**0.0814** (4.0e-04)

Table 4: Spearman's rank correlation coefficients (**rho**) between the behaviour data and the fMRI data of different brain lobes.

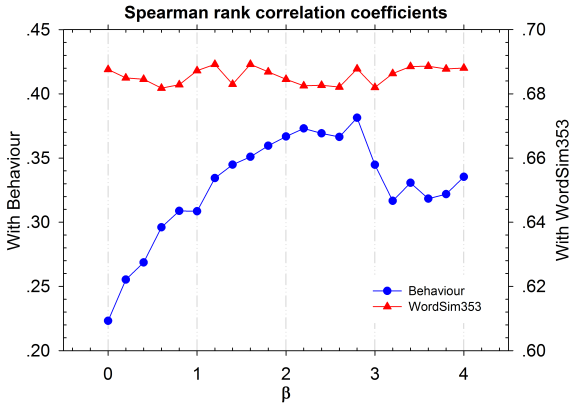

Figure 1: Spearman's rank correlation coefficients between the estimated word embeddings with different β values and the behavioural data of two datasets.

5.2 Correlations between brain activity and behavioural data

After analysing the correlation between brain activity and the three types of word vectors in the previous experiments, we further examined the correlations between brain activity and the behavioural data introduced in Section 4.2. Here, the behavioural data were used as the semantic ground truth to evaluate the semantic relevance of the brain activity and word embeddings. The results are shown in Table 4. In this subsection, the fMRI data at the whole-brain and brain lobe levels adopted the voxel selection strategies introduced in Sections 5.1.2 and 5.1.3, respectively. As Table 4 shows, the behavioural data are significantly correlated with the fMRI data of the whole brain and the occipital lobe, and they are also correlated with the fMRI data of the temporal and parietal lobes.

Furthermore, we utilized the SWE model introduced in Section 3.2 to explore the semantic relevance of brain activity by examining the performance of the estimated word embeddings after integrating brain activity related knowledge. The inequality set used in Eq. (3) was created using the fMRI data, where the similarity score s_{ij} was calculated as the Pearson correlation coefficient between the fMRI vectors of the i-th and the j-th words. For the 60 nouns (a total of 12 categories with 5 words in each category), we produced $12 \times 3 \times C_5^3 = 360$ intra-category inequalities and $3 \times C_{12}^3 = 660$ inter-category inequalities. To collect the inter-category inequalities, we first used the label words of each category and averaged the fMRI vectors of the 5 words belonging to each category to obtain the fMRI data for these label words. Then, the inter-category inequalities were produced from the triplets of these label words. The text corpus and parameter settings we used to train SWE were the same as those used for training the skip-gram word embeddings as described in Section 5.1.1. The penalty term β in Eq. (6) was tuned through experiments.

We evaluated the word embeddings estimated with brain activity constraints using the collected behavioural data for the 60 nouns and the WordSim-353 dataset. WordSim-353 is a behavioural dataset containing semantic similarity scores for 353 word-pairs (Finkelstein et al., 2001). We checked to ensure these word-pairs have no overlap with the 60 nouns used in our experiments. The purpose of using the WordSim-353 dataset is to explore the effects of utilizing the brain data of the 60 nouns on other words for which we had no brain data.

	60 nouns	WordSim353
skip-gram	0.2232	0.6876
SWE (whole brain)	0.3814	0.6878
SWE (frontal)	0.3173	0.6822
SWE (temporal)	0.3613	0.6890
SWE (parietal)	0.3516	0.6706
SWE (occipital)	0.3348	0.6803
JNNSE	0.3006	0.1795

Table 5: Spearman's rank correlation coefficients between different word embeddings and the behavioural data of the two datasets.

The performance of the word embeddings estimated using the SWE model and the whole-brain fMRI data are shown in Figure 1. In this figure, the SWE model becomes a conventional skip-gram model when $\beta = 0$. The correlation coefficient between the skip-gram word embeddings and the behavioural data of the 60 nouns was 0.2232. As β was increased, this correlation coefficient increased significantly. The maximum correlation efficient was 0.3814 when $\beta = 2.8$. This result implies that textual data and brain activity data can be used in a complementary fashion to derive word embeddings that are more consistent with human judgements. On one hand, semantic patterns may exist in neural representations that have not been fully captured by state-of-the-art word embeddings derived from text corpora. On the other hand, we can see that the variation of the correlation coefficients for the WordSim-353 dataset with different β values is small. This indicates that our SWE training didn't negatively affect the word embeddings without fMRI observations.

Furthermore, we produced ranking inequalities using the fMRI data measuring each brain lobe to estimate word embeddings under the SWE framework. The correlations between the learned word embeddings and the behavioural data of the two datasets were calculated and are shown in Table 5. For each SWE model in this table, the value of β was tuned to obtain the highest correlation on the 60 nouns. Comparing the correlation coefficients of the different models on the 60 nouns, we can see that the fMRI data at all brain lobes can contribute to learning more semantically related word embeddings using the SWE model. The improvement from using the fMRI data of the temporal lobe is the most significant among the four lobes, but the highest correlation coefficient is achieved when utilizing the fMRI data of whole brain.

Finally, we compared the performance of our SWE models with the JNNSE model proposed by Fyshe et al. (2014) on the two datasets. The word embeddings estimated by the JNNSE model utilized either fMRI or magnetoencephalography (MEG) measures of the 60 nouns. We adopted the best JNNSE word embeddings reported by the authors[7] for these comparisons, and the results are shown in the last row of Table 5[8]. As Table 5 shows, the performance of the JNNSE word embeddings on the WordSim-353 dataset is not as good as those of the skip-gram and SWE results. Examining the correlation coefficients on the 60 nouns with brain activity data, we can see that the JNNSE model achieves better performance than the skip-gram model, but is still below that of the SWE models. It should be noted that it is unfair to directly compare the SWE models and the JNNSE model because they used different training corpora and word embedding dimensions. Moreover, the β values of the SWE models were tuned to achieve the best performance on these 60 nouns. Here, the motivation behind introducing the JNNSE model as a reference is to help readers better understand the effects of integrating brain data into SWE training. These experimental results demonstrate that the SWE model is an effective model structure for integrating external knowledge into the estimation of word embeddings.

6 Conclusion

This study utilized word embeddings to investigate the semantic representations in brain activity as measured by fMRI. First, the functional selectivity of different cortex areas is explored by calculating the correlations between neural responses and three types of word vectors: skip-gram word embeddings, visual semantic vectors, and primary visual features.

[7] http://www.cs.cmu.edu/~afyshe/papers/acl2014/

[8] Because there were 32 word-pairs in the WordSim-353 dataset that were not covered by the vocabulary of the JNNSE word embeddings, the value 0.1795 in the last row of Table 5 was calculated using only 321 word-pairs.

Experimental results demonstrate the differences between the associations of different word vectors with brain-lobe-level brain activities. The skip-gram word embeddings are significantly correlated with the fMRI data of all brain lobes except the occipital lobe. Furthermore, we utilized behavioural data as the semantic ground truth to measure its relevance to brain activity. The SWE model was employed to explore the semantic relevance of brain activity by examining the performances of word embeddings after integrating brain-activity-related knowledge into their estimations. Experimental results show that whole-brain fMRI data are significantly correlated with human judgement with respect to semantic similarity. The correlations between the estimated word embeddings and the human-assigned similarity scores are effectively improved after integrating brain activity data into SWE training.

The experiments in this paper provide information about how semantic features correlate with brain activities, laying foundations for further investigations of higher-level semantic processing in the human brain. Furthermore, our experiments with SWE modelling show the potential of applying fMRI data to obtain better word embeddings. Although this approach is still far from being a practical engineering application due to issues such as the high costs and low signal-to-noise ratio of fMRI recordings and the diversity among individuals, it provides us with an alternative method for verifying the semantic relevance of brain activities and with evidence for recognizing the limitations of estimating word embeddings using only text corpora.

Acknowledgements

This work was supported in part by the Science and Technology Development of Anhui Province, China (Grant No. 2014z02006), the Fundamental Research Funds for the Central Universities (Grant No. WK2350000001) and the CAS Strategic Priority Research Program (Grant No. XDB02070006). The authors also want to thank Quan Liu for his help and wonderful suggestions during the experiments.

References

[Anderson et al.2013] 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 *EMNLP*, pages 1960–1970.

[Binder et al.2009] Jeffrey R Binder, Rutvik H Desai, William W Graves, and Lisa L Conant. 2009. Where is the semantic system? a critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, 19(12):2767–2796.

[Bruni et al.2012] Elia Bruni, Jasper Uijlings, Marco Baroni, and Nicu Sebe. 2012. Distributional semantics with eyes: Using image analysis to improve computational representations of word meaning. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 1219–1228. ACM.

[Carlson et al.2014] Thomas A Carlson, Ryan A Simmons, Nikolaus Kriegeskorte, and L Robert Slevc. 2014. The emergence of semantic meaning in the ventral temporal pathway. *Journal of cognitive neuroscience*, 26(1):120–131.

[Chao et al.1999] Linda L Chao, James V Haxby, and Alex Martin. 1999. Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. *Nature neuroscience*, 2(10):913–919.

[Church and Hanks1990] Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

[Csurka et al.2004] Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. 2004. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, pages 1–2. Prague.

[Daugman1985] John G Daugman. 1985. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *JOSA A*, 2(7):1160–1169.

[Deng et al.2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE.

[Fellbaum and others1998] Christiane Fellbaum et al. 1998. Wordnet: An electronic lexical database mit press. *Cambridge MA*.

[Finkelstein et al.2001] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In

Proceedings of the 10th international conference on World Wide Web, pages 406–414. ACM.

[Fyshe et al.2014] Alona Fyshe, Partha P Talukdar, Brian Murphy, and Tom M Mitchell. 2014. Interpretable semantic vectors from a joint model of brain-and text-based meaning. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2014, page 489. NIH Public Access.

[Grauman and Leibe2011] Kristen Grauman and Bastian Leibe. 2011. Visual object recognition. *Synthesis lectures on artificial intelligence and machine learning*, 5(2):1–181.

[Haghighat et al.2015] Mohammad Haghighat, Saman Zonouz, and Mohamed Abdel-Mottaleb. 2015. Cloudid: Trustworthy cloud-based and cross-enterprise biometric identification. *Expert Systems with Applications*, 42(21):7905–7916.

[Hagoort2005] Peter Hagoort. 2005. On broca, brain, and binding: a new framework. *Trends in cognitive sciences*, 9(9):416–423.

[Harris1954] Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

[Haxby et al.2001] James V Haxby, M Ida Gobbini, Maura L Furey, Alumit Ishai, Jennifer L Schouten, and Pietro Pietrini. 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430.

[Hickok and Poeppel2007] Gregory Hickok and David Poeppel. 2007. The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8(5):393–402.

[Hinton et al.1986] Geoffrey E Hinton, James L Mcclelland, and David E Rumelhart. 1986. Distributed representations, parallel distributed processing: explorations in the microstructure of cognition, vol. 1: foundations.

[Huth et al.2012] Alexander G Huth, Shinji Nishimoto, An T Vu, and Jack L Gallant. 2012. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6):1210–1224.

[Huth et al.2016] Alexander G Huth, Wendy A de Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458.

[Kanwisher and Yovel2006] Nancy Kanwisher and Galit Yovel. 2006. The fusiform face area: a cortical region specialized for the perception of faces. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 361(1476):2109–2128.

[Kay et al.2008] Kendrick N Kay, Thomas Naselaris, Ryan J Prenger, and Jack L Gallant. 2008. Identifying natural images from human brain activity. *Nature*, 452(7185):352–355.

[Kriegeskorte et al.2008] Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. 2008. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:4.

[Liu et al.2015] Quan Liu, Hui Jiang, Si Wei, Zhen-Hua Ling, and Yu Hu. 2015. Learning semantic word embeddings based on ordinal knowledge constraints. *Proceedings of ACL, Beijing, China*.

[Louwerse2011] Max M Louwerse. 2011. Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, 3(2):273–302.

[Lowe2004] David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110.

[Lund and Burgess1996] Kevin Lund and Curt Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28(2):203–208.

[Marĉelja1980] S Marĉelja. 1980. Mathematical description of the responses of simple cortical cells*. *JOSA*, 70(11):1297–1300.

[Mikolov et al.2013a] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[Mikolov et al.2013b] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

[Miller and Charles1991] George A Miller and Walter G Charles. 1991. Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1):1–28.

[Miller et al.2002] Earl K Miller, David J Freedman, and Jonathan D Wallis. 2002. The prefrontal cortex: categories, concepts and cognition. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 357(1424):1123–1136.

[Mitchell et al.2008] 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*, 320(5880):1191–1195.

[Naselaris et al.2009] Thomas Naselaris, Ryan J Prenger, Kendrick N Kay, Michael Oliver, and Jack L Gallant. 2009. Bayesian reconstruction of natural images from human brain activity. *Neuron*, 63(6):902–915.

[Nishimoto et al.2011] Shinji Nishimoto, An T Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and

Jack L Gallant. 2011. Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, 21(19):1641–1646.

[Pennington et al.2014] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543.

[Ryali et al.2010] Srikanth Ryali, Kaustubh Supekar, Daniel A Abrams, and Vinod Menon. 2010. Sparse logistic regression for whole-brain classification of fMRI data. *NeuroImage*, 51(2):752–764.

[Sack2009] Alexander T Sack. 2009. Parietal cortex and spatial cognition. *Behavioural brain research*, 202(2):153–161.

[Sivic and Zisserman2003] Josef Sivic and Andrew Zisserman. 2003. Video google: A text retrieval approach to object matching in videos. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 1470–1477. IEEE.

[Smith and Kosslyn2007] EE Smith and SM Kosslyn. 2007. *Cognitive Psychology: Mind and Brain*. Pearson Prentice Hall, Upper Saddle River, NJ.

[Turney et al.2010] Peter D Turney, Patrick Pantel, et al. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37(1):141–188.

[Tzourio-Mazoyer et al.2002] Nathalie Tzourio-Mazoyer, Brigitte Landeau, Dimitri Papathanassiou, Fabrice Crivello, Olivier Etard, Nicolas Delcroix, Bernard Mazoyer, and Marc Joliot. 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the mni mri single-subject brain. *Neuroimage*, 15(1):273–289.

[Vannini et al.2004] Patrizia Vannini, Ove Almkvist, Anders Franck, Tomas Jonsson, Umberto Volpe, Maria Kristoffersen Wiberg, Lars-Olof Wahlund, and Thomas Dierks. 2004. Task demand modulations of visuospatial processing measured with functional magnetic resonance imaging. *Neuroimage*, 21(1):58–68.