Probing Brain Activation Patterns by Dissociating Semantics and Syntax in Sentences

Shaonan Wang,[1,2] **Jiajun Zhang,**[1,2] **Nan Lin,**[3,4] **Chengqing Zong**[1,2,5]
[1]National Laboratory of Pattern Recognition, Institute of Automation, CAS
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
[3]CAS Key Laboratory of Behavioural Science, Institute of Psychology
[4]Department of Psychology, University of Chinese Academy of Sciences
[5]CAS Center for Excellence in Brain Science and Intelligence Technology
{shaonan.wang, jjzhang, cqzong}@nlpr.ia.ac.cn; linn@psych.ac.cn

Abstract

The relation between semantics and syntax and where they are represented in the neural level has been extensively debated in neurosciences. Existing methods use manually designed stimuli to distinguish semantic and syntactic information in a sentence that may not generalize beyond the experimental setting. This paper proposes an alternative framework to study the brain representation of semantics and syntax. Specifically, we embed the highly-controlled stimuli as objective functions in learning sentence representations and propose a disentangled feature representation model (DFRM) to extract semantic and syntactic information in sentences. This model can generate one semantic and one syntactic vector for each sentence. Then we associate these disentangled feature vectors with brain imaging data to explore brain representation of semantics and syntax. Results have shown that semantic feature is represented more robustly than syntactic feature across the brain including the default-mode, frontoparietal, visual networks, etc.. The brain representations of semantics and syntax are largely overlapped, but there are brain regions only sensitive to one of them. For instance, several frontal and temporal regions are specific to the semantic feature; parts of the right superior frontal and right inferior parietal gyrus are specific to the syntactic feature.

Knowing the meanings of individual words (semantics) and understanding how these words can combine (syntax) to create new, complex meanings are core components of our language knowledge. Neuroimaging studies have evidenced that our brains process semantic information differently from syntactic information (Dapretto and Bookheimer 1999; Friederici, Opitz, and Von Cramon 2000; Matchin et al. 2019). However, at present, there is still a poor understanding of whether and where this distinction is represented at the neural level, especially outside of controlled experimental settings.

Due to the complexity of language processing in the brain, most existing neuroimaging studies employ the hypothesis-based method and investigate the brain language representations with highly-controlled stimuli. Generally, they design different experimental conditions and employ the subtraction method to explore the brain process for each language feature. For instance, subtracting neuron activations of random word-lists from natural sentences to study syntactic representations in the brain; subtracting neuron activations of jabberwocky sentences (which are created by replacing the content words including nouns, verbs, adjectives, and adverbs in the sentences by pronounceable nonwords) from natural sentences to study semantic representations in the brain (Fedorenko, Nieto-Castanon, and Kanwisher 2012). This paradigm with controlled stimuli is extensively applied in neurosciences, making precise conclusions with carefully designed materials. However, it is not clear whether fake sentences would introduce agnostic factors to brain processing and whether this paradigm can generalize beyond the experimental setting (Hasson and Honey 2012; Wehbe 2015).

This paper proposes an alternative framework to study the brain representation of semantics and syntax. Drawing on the ideas of controlled-stimuli methods, we propose a disentangled feature representation model (DFRM) that can extract the semantic and syntactic information in a sentence, generating one semantic and one syntactic vector for each sentence. To disassociate semantics from syntax in a sentence, we utilize word average encoder and Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber 1997) encoder to represent semantic and syntactic information respectively. Then by introducing objective functions of learning paraphrases and jabberwocky sentences, the semantic and syntactic vectors tend to accumulate semantic and syntactic knowledge respectively. To improve the quality of syntactic vectors, we also utilize the word position loss to capture more syntactic information.

Based on the learned disentangled feature vectors, we explore the brain representation of semantics and syntax by associating these vectors with brain imaging data. We find that both semantic and syntactic features are involved in several frontal and temporal regions, with the semantic feature distributed more robustly and widely across the brain. There are also specific brain regions only sensitive to the semantic or syntactic features. For instance, several frontal and temporal regions are specific to the semantic feature; parts of the right superior frontal and right inferior parietal gyrus are specific to the syntactic feature.

To summarize, our main conclusions include:

- We propose a novel framework to study brain language representation by using computational models to disassociate semantics and syntax in sentences, which can easily be extended to explore cognitive questions of brain representations of visual, auditory, emotional features, etc.
- We propose a disentangled feature representation model that can, to a certain extent, separate semantic and syntactic information in a sentence efficiently and generate a semantic and a syntactic vector for a sentence.
- From the computational perspective, our results provide new evidence for the brain representations of semantics and syntax in a sentence, hopefully helping promote related neuroscience studies.

Related Work

Language representation in the brain

The relation between semantic and syntactic representation in the brain is core to the human language understanding mechanism. Neuroimaging studies of sentence processing reliably activate a set of brain regions in the frontal and temporal lobes (Mazoyer et al. 1993; Pallier, Devauchelle, and Dehaene 2011). However, the specific role each region plays in sentence processing, particularly concerning semantics and syntax, remains unclear (Dapretto and Bookheimer 1999; Friederici 2012; Fedorenko, Nieto-Castanon, and Kanwisher 2012).

The advent of computational approaches has allowed us to supplement the hypothesis-driven science with data-driven science. A pioneering study by Mitchell et al. (2008) uses corpus-derived word representations, in which each dimension corresponds to a specific semantic feature, to predict the neural activations when subjects are exposed to a stimulus word. Through analyzing regression weights in the trained model, they give detailed brain representations of each semantic feature. Recent studies have shown that such a method can be applied to natural sentences and stories (Wehbe et al. 2014; Anderson et al. 2016; Pereira et al. 2018; Sun et al. 2019).

Different from these pioneering works, this paper focus on the question of how the brain represents the semantics and syntax in sentences. The main difficulty of such research lies in the complex relationships between semantics and syntax, thereby making it challenging to disentangle the two features from each other. To solve this problem, we propose a disentangled feature representation model to encode semantic and syntactic information into continuous vectors separately. The resulting feature vectors can capture more abundant information and better portray the relationships between sentences than traditional discrete features (Le and Mikolov 2014; Kiros et al. 2015).

Computational sentence representation

High-quality sentence representation is fundamental for machines to discriminate and understand sentence meanings. Recently, neural-network-based sentence representation models have shown a significant advantage over traditional methods (Le and Mikolov 2014; Kiros et al. 2015; Hill, Cho, and Korhonen 2016; Conneau et al. 2017; Wang, Zhang, and Zong 2017) However, most of these works focus on learning sentence representations that are effective in downstream tasks.

To improve the quality of sentence representations or achieve the goal of generating texts based on the controlled syntactic structure, there has been a surge of recent work on learning the disentangled semantics and syntax representations in various NLP applications, including sentence representation (Chen et al. 2019b), sentiment and style transfer (Zhao et al. 2018), text generation (Iyyer et al. 2018; Chen et al. 2019a), etc.

In contrast to these works, our goal is to disentangle semantic and syntactic information in sentences at the utmost, to explore the brain semantics and syntax processing by the learned vectors. The most similar work to our method is the vMF-Gaussian Variational Autoencoder (VGVAE) model proposed by Chen et al. (2019b). The VGVAE model is a generative model with two latent variables, with one represents the semantics of the sentence and the other to represent its syntax. This model learns the semantic variable by exploiting the loss of learning aligned paraphrastic sentences and learns the syntactic variable using the loss of learning word-order information. Inspired by neurosciences studies, we introduce jabberwocky stimuli, sentences with the same syntactic structure, to learn syntactic information. This constraint is introduced as the jabberwocky loss in our DFRM method, further disentangling the semantic and syntactic information in sentences.

Probing Method

To investigate how semantic and syntactic features associated with each brain region, we propose a new framework as illustrated in Figure 1, which includes two steps: 1) encoding sentences' semantic and syntactic information into continuous vectors respectively, and 2) exploring the relationships between disentangled feature vectors and brain activations.

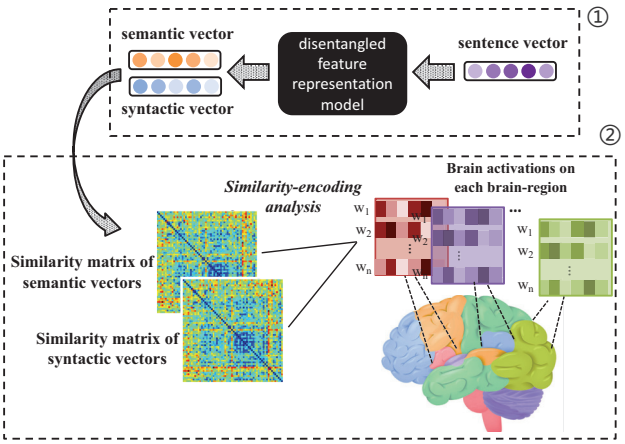

Figure 1: Architecture of the probing method, including (1) building disentangled feature representation model, and (2) conducting similarity-encoding analysis.

Disentangled feature representation model

To extract the semantic and syntactic information in sentences and encode them into vectors, we propose a disentangled feature representation model (DFRM) which is a generative model with neuropsychology-inspired objective functions. Our model is based on the vMF-Gaussian Variational Autoencoder (VGVAE) model (Chen et al. 2019b). The VGVAE model uses two latent variables to extract semantic and syntactic information from sentences respectively. We follow Chen et al. (2019b) and use the von Mises-Fisher (vMF) distribution for the semantic variable and the Gaussian distribution for the syntactic variable.

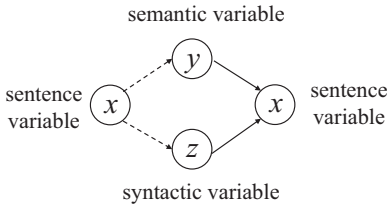

Figure 2: Graphical model of VGVAE. Dashed lines indicate inference model. Solid lines indicate generative model.

As shown in Figure 2, VGVAE assumes that a sentence is generated by independent semantic and syntactic variables, i.e., y and z. Formally, following the conditional independence assumption in the graphical model, the joint probability can be factorized as:

$$\begin{aligned} p_\theta(x,y,z) &= p_\theta(y)p_\theta(z)p_\theta(x|y,z) \\ &= p_\theta(y)p_\theta(z)\prod_{t=1}^{T} p_\theta(x_t|x_{1:t-1},y,z), \end{aligned} \tag{1}$$

where x_t is the tth word of x and T is the sentence length. The probability $p_\theta(x_t|x_{1:t-1},y,z)$ is given by a softmax over a vocabulary of size V.

When applying neural variational inference, VGVAE uses a factorized approximated posterior $q_\phi(y|x)q_\phi(z|x) = q_\phi(y,z|x)$ with objective function of maximizing a lower bound of marginal log-likelihood:

$$\begin{aligned} \log p_\theta(x) \geq\ & \mathbb{E}_{\substack{y\sim q_\phi(y|x)\\ z\sim q_\phi(z|x)}}[\log p_\theta(x|z,y) - \log\frac{q_\phi(z|x)}{p_\theta(z)} \\ & - \log\frac{q_\phi(y|x)}{p_\theta(y)}] \\ =\ & \mathbb{E}_{\substack{y\sim q_\phi(y|x)\\ z\sim q_\phi(z|x)}}[\log p_\theta(x|z,y)] - KL(q_\phi(z|x)||p_\theta(z)) \\ & - KL(q_\phi(y|x)||p_\theta(y)), \end{aligned} \tag{2}$$

where $q_\phi(y|x)$ and $q_\phi(z|x)$ follows a vMF distribution and a Gaussian distribution respectively. The prior $p_\theta(y)$ and $p_\theta(z)$ follows the uniform distribution vMF$(\cdot, 0)$ and a standard Gaussian distribution respectively.

To further guide the two latent variables to encode semantic and syntactic information separately, we employ paraphrase and jabberwocky losses and propose a DFRM method. We also adopt word position loss used in Chen et al. (2019b) to enhance syntactic information accumulation. As shown in Figure 3, given a sentence x_1, we have its paraphrase sentence x_2 and jabberwocky sentence x_{1j}. We also generate the jabberwocky sentence x_{2j} of sentence x_2.

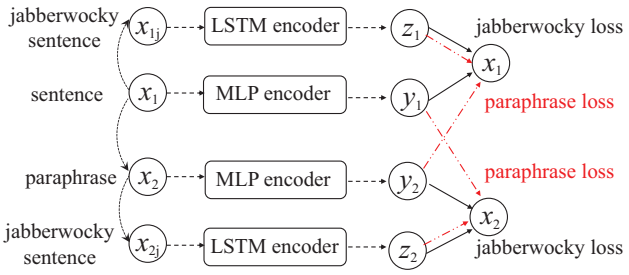

Figure 3: Architecture of our DFRM method. Solid lines on the right denotes jabberwocky loss while red dashed lines denote paraphrase loss.

The DFRM utilizes two different encoders, a word average encoder (i.e., Multilayer Perceptron, MLP) and an LSTM encoder, to represent semantic and syntactic information respectively in a sentence. Specifically, the MLP encoder only utilizes word semantic information, simply averaging all word embeddings in a sentence and then feeding the results into one feed-forward layer. The LSTM encoder utilizes sentence structure information, employing a recurrent neural network to sequentially encode each word in a sentence. The final sentence representation by LSTM is the averaging of all hidden embeddings. In this way, the semantic variable y contains mainly word semantic information, while the syntactic variable z contains both semantic and syntactic information.

To exclude semantic information and accumulate syntactic information for the syntactic variable z, we introduce **Jabberwocky loss (JLoss)**. The key assumption is that for sentence x and its jabberwocky sentence x_j, the syntactic information is the same and only the semantic information varies. For instance, by generating sentence x_1 with semantic variable y_1 and syntactic variable z_1, variable z_1 will learn the syntactic information of x_1 and abandon the semantic information of x_{1j} which are useless to generate x_1. To impose such constraints, JLoss is defined as:

$$\begin{aligned} & \mathbb{E}_{\substack{y_1\sim q_\phi(y|x_1)\\ z_1\sim q_\phi(z|x_1)}}[-\log p_\theta(x_1|y_1,z_1)] + \\ & \mathbb{E}_{\substack{y_2\sim q_\phi(y|x_2)\\ z_2\sim q_\phi(z|x_2)}}[-\log p_\theta(x_2|y_2,z_2)]. \end{aligned} \tag{3}$$

We have also tried a constraint of directly minimizing the distance between two semantic variables of paraphrases x_1 and x_2 and that of two syntactic variables of jabberwocky sentences x and x_j, but gets worse performance.

To enhance the semantic varibale to learn more semantic information, we adopt **paraphrase loss (PLoss)**. Assume that for a paraphrase pair (x_1, x_2), the semantic information

(which is encoded in y_1 and y_2) is equivalent and only the syntactic information (which is encoded in z_1 and z_2) varies. Therefore, a pair of variables y_2 and z_1 can be used to generate sentence x_1, while the other pair of variables y_1 and z_2 cab be used to generate sentence x_2 in the training phase. Similar to JLoss, PLoss is defined as:

$$\begin{aligned}&\mathbb{E}_{\substack{y_2\sim q_\phi(y|x_2)\\ z_1\sim q_\phi(z|x_1)}}[-\log p_\theta(x_1|y_2,z_1)]+\\ &\mathbb{E}_{\substack{y_1\sim q_\phi(y|x_1)\\ z_2\sim q_\phi(z|x_2)}}[-\log p_\theta(x_2|y_1,z_2)],\end{aligned} \tag{4}$$

To guide the syntactic vectors to capture more syntactic information, we also employ the **word position loss (WPLoss)** on both syntactic encoder and decoder as in (Chen et al. 2019b). The WPLoss is parameterized by a three-layer feedforward neural network $f(\cdot)$ with input from the concatenation of all hidden vectors of LSTM encoder or decoder. We then attempt to predict a one-hot-vector representing the position i. Specifically, we define WPLoss as:

$$\mathbb{E}_{z\sim q_\phi(z|x)}[-\sum_i \log \mathrm{softmax}(f([e_i;z]))_i], \tag{5}$$

where the $\log \mathrm{softmax}(\cdot)_i$ indicates the probability at position i and e_i is the embedding vector at input position i.

To train the DFRM, we minimize the summation of the above PLoss, JLoss and WPLoss.

Similarity-encoding analysis

To explore the relationships between computational vectors and brain activations, we adopt the similarity-encoding analysis (Anderson, Zinszer, and Raizada 2016) method, which consists of three steps as follows (as shown in the second box in Figure 1).

(1) For each sentence, we have one semantic and one syntactic vector computed by the DFRM. For each feature vector, we calculate their Pearson correlation coefficient for each sentence pair in a set of n sentences, resulting in a similarity matrix with a size of $n \times n$. Finally, we get two representational similarity matrices in which each row vector represents semantic or syntactic similarity of one sentence with other n-1 sentences.

(2) For each sentence representation in the brain, we can divide them into m brain region vectors. Assume that if a specific brain region encodes the same information as a specific feature vector, then the similarity relation of the feature vector and that of the brain-region vector is the same. Therefore, we can predict each brain-region vectors by multiplying the above semantic or syntactic similarity matrix with corresponding brain region vectors. Consequently, we get m predicted brain-region matrix with a size of $n \times p$ one for each brain-region (p denotes vector dimension).

(3) We use the Pearson correlation coefficient to calculate the similarities between the predicted brain-region vectors and the real brain-region vectors. The higher correlation score means that the semantic or syntactic information is more encoded in the specific brain region.

Experiments

Experimental setup

We randomly sample 500,000 paraphrase pairs from ParaNMT-50M (Wieting and Gimpel 2018) as our training set. For jabberwocky sentences, we adjust the original definition to eliminate out-of-vocabulary words. Specifically, we replace nouns, verbs, adjectives, and adverbs in sentences with randomly selected words with the same POS tag in vocabulary.

To evaluate the quality of the resulting semantic and syntactic vectors, we adopt the semantic textual similarity (STS) task and the syntactic similarity (SS) task. For the STS task, we the STS benchmark dataset and a dataset containing the concatenation of STS tasks from 2012 to 2016 which are from http://ixa2.si.ehu.es/stswiki/index.php/STSbenchmark. This task is to evaluate the degree of semantic similarity between sentence pairs in the test set. For the SS task, we use the part-of-speech (POS) and syntactic parser similarity datasets proposed in Chen et al. (2019b). The SS task is to predict a parse tree for each sentence in the test set by finding its nearest neighbor in the training set based on the cosine similarity of the computational vectors. Results are evaluated with F1 for syntactic parser set and accuracy for POS set.

Models are implemented with Pytorch and parameters are trained for 20 epochs, with each epoch consisting of multiple batches optimized with Adam. Same with the baseline VGVAE model, the dimensions of word embeddings, MLP, and LSTM hidden layers of DFRM are all set to 100.

Baselines

We adopt random word embeddings and several commonly used pre-trained embeddings as baselines, including GloVe (Pennington, Socher, and Manning 2014), InferSent (Conneau et al. 2017), ELMo (Peters et al. 2018) and BERT (Devlin et al. 2019). For all models, we average the word vectors or hidden vectors of each time step to form sentence representations. We also benchmark simple word averaging (WORDAVG) model and bidirectional LSTM averaging (BLSTMAVG) model, both have shown superior performance when trained on the paraphrases datasets (Wieting and Gimpel 2018). For a fair comparison, we use the same training dataset to retrain the VGVAE model in which data and code are from https://github.com/mingdachen/disentangle-semantics-syntax.

Brain activation data

Our experiments are conducted on the dataset from Pereira et al. (2018) which is publicly available at https://osf.io/crwz7/. The dataset includes preprocessed functional activation data that is gathered from 5 participants (P01, M02, M05, M07, M15) while exposed to sentence stimuli in two functional Magnetic Resonance Imaging (fMRI) experiments. In experiment 1, participants are presented with a set of 96 text passages, each consisting of 4 sentences describing basic information of a particular concept, spanning a broad range of content areas from 24 broad topics, with 4 passages per topic. In experiment 2, sentence stimuli is a set

	STS benchmark test (% Pearson correlation ↑)		Averaged STS tests (% Pearson correlation ↑)		Constituent Parsing (F1 ↑)		POS Tagging (%Acc. ↑)	
Random	39.7		42.5		19.2		12.9	
GloVe	41.0		47.9		27.3		23.9	
InferSent	67.8		61.0		28.0		25.1	
ELMo	57.7		60.3		30.4		27.8	
BERT	54.9		59.0		28.6		25.8	
WordAvg	71.9		64.8		25.5		21.4	
LSTMAvg	71.4		64.4		25.7		21.6	
	sem var.	syn var.	sem var.	syn var.	sem var.	syn var.	sem var.	syn var.
VGVAE	65.7	32.2	57.6	28.4	25.1	26.7	20.9	22.6
VGVAE+PLoss	72.5	24.0	**66.3**	28.5	24.2	29.0	19.6	26.3
VGVAE+WPLoss	69.4	8.5	61.1	18.9	24.4	35.7	19.8	33.2
VGVAE+JLoss	55.2	17.0	48.3	24.2	23.6	32.3	18.5	32.3
VGVAE+PLoss+JLoss	72.5	11.4	66.0	22.9	24.2	34.2	19.2	34.4
VGVAE+PLoss+WPLoss	71.2	15.0	65.9	22.6	24.0	34.6	19.3	32.7
DFRM (VGVAE+all)	**73.0**	8.5	65.9	18.4	24.2	**40.0**	19.5	**38.6**

Table 1: Semantic and syntactic evaluation results. Results are bold if they are highest in the STS tasks with the semantic variable (sem var.) or highest in the SS tasks with the syntactic variable (syn var.).

Query sentence	Neighbor sentences by semantic var.	Neighbor sentences by syntactic var.
a cook is making food .	there is a cook preparing food .	a kid is playing keyboard .
the dog is chasing the geese .	one dog is chasing the other .	the cat is licking a bottle .
you can do it , too .	yes , you can do it .	you should prime it first .
it makes absolutely no difference .	i do n't think it makes much difference .	this is a big problem .
but the economy has n't shown signs of sustainable growth .	the economy , nonetheless , has yet to exhibit sustainable growth .	but the north korean nuclear crisis has dominated his time in office .

Table 2: Examples of most similar sentences to particular query sentences calculated by the semantic or syntactic variables.

of 72 passages, each consisting of 3 or 4 sentences about a particular concept. Different from experiment 1, the materials include first-/third-person narratives. The passages span a broad range of content areas from 24 broad topics, unrelated to that in experiment 1, with 3 passages per topic.

All passages are presented sentence by sentence. Each sentence is presented for 4 seconds followed by a fixation gap of 4 seconds. The entire set of 637 sentences is seen 3 times. The participants are asked to attentively read the sentences they are presented for scanning. The details of the experimental setup, materials and presentation scripts are available at https://osf.io/crwz7/wiki/home/.

In the probing experiments, we use region-of-interest-based (ROI-based) analyses in which ROI is defined by the Gordon Parcellation (Gordon et al. 2014). As shown in Figure 4, this parcellation consists of 333 cortical patches (ROI0-ROI332) with different cognitive networks.

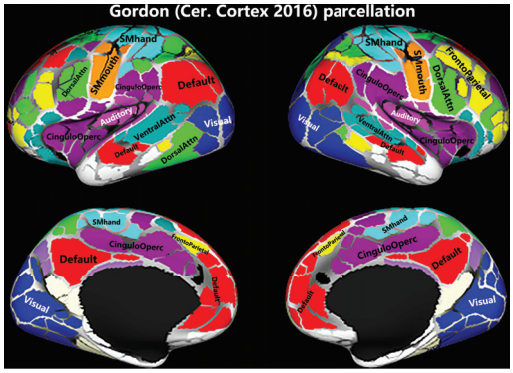

Figure 4: A diagram of the Gordon parcellation.

Results and Analysis

DFRM results

As shown in Table 1, among baseline pre-trained models, InferSent shows the strongest performance overall, followed by ELMo and BERT. The WordAvg and LSTMAvg baselines, which are simply trained on paraphrases, obtain strong performance on the STS tasks. Moreover, all these models have relative worse performance on the SS tasks.

We can also see that the baseline VGVAE model learns different semantic and syntactic variables that are effective on the STS tasks and the SS tasks respectively. Adding PLoss further increases the gap between semantic and syntactic variables. Interestingly PLoss not only strengthens the performance of the semantic variable on the semantic similarity tasks but also improves the performance of the syntactic variable on the syntactic similarity tasks, even though this loss is only imposed on the semantic variable. This finding suggests that by pushing the semantic variable to learn semantic information encoded by paraphrases, the syntactic variables are forced to capture complementary syntactic information. Furthermore, adding WPLoss and JLoss can both strengthen the ability of a syntactic variable to accumu-

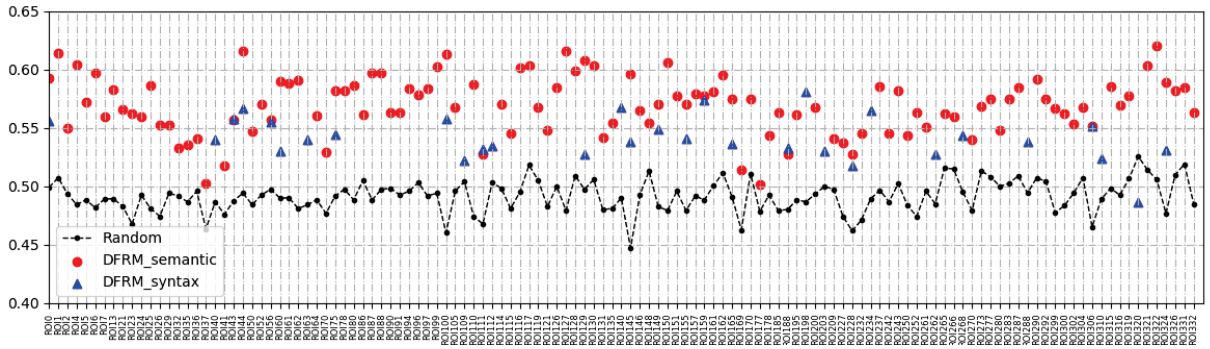

Figure 5: Probing results of DFRM and random baseline on fMRI experiment 1. The x(y)-coordinate denotes the ROIs(similarity encoding score). Only statistically significant DFRM results are shown which correspond to 112 red dots and 31 blue triangles.

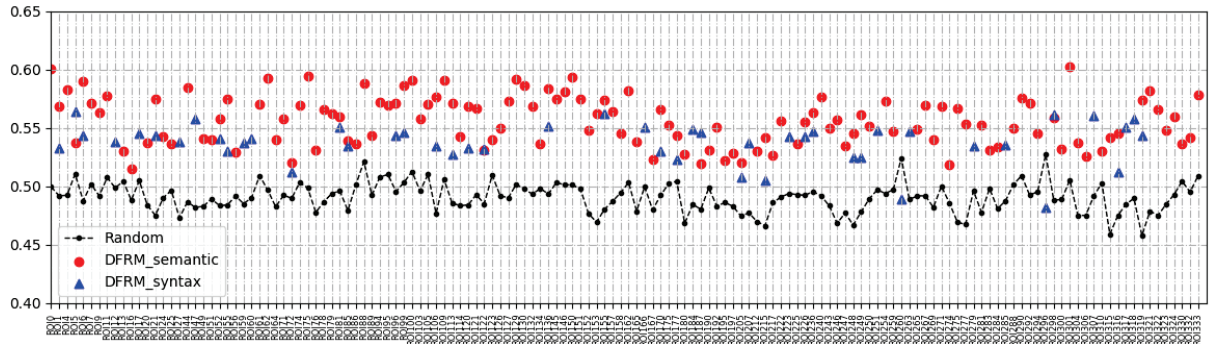

Figure 6: Probing results of DFRM and random baseline on fMRI experiment 2. The x(y)-coordinate denotes the ROIs(similarity encoding score). Only statistically significant DFRM results are shown which correspond to 125 red dots and 48 blue triangles.

late syntactic information of sentences. Our DFRM, which utilizes all three losses, obtains the highest quality and best disentangles semantic and syntactic feature vectors. This observation not only indicates that WPLoss and JLoss encode complementary syntactic information but also illustrates that the DFRM successfully disentangle and encode semantic and syntactic information respectively.

To qualitatively evaluate the learned semantic and syntactic vectors by DFRM, we find the nearest neighbor sentences to test set examples (2,551 sentences in total) by computing cosine similarity in terms of the semantic and syntactic vectors respectively. We show five representative examples in Table 2. It is evident that neighbor sentences calculated by the semantic variable are semantically similar to the query sentence. However, neighbor sentences calculated by the syntactic variable are mostly semantically unrelated but have similar sentence structures. For instance, the query sentence *"a cook is making food"* has the same meaning with *"there is a cook preparing food"* calculated by the semantic variable, and has the same sentence structure with *"a kid is playing keyboard"* calculated by the syntactic variable.

Taken together, we conclude that the proposed DFRM can effectively disentangle the semantic and syntactic information from sentences and encode them into semantic and syntactic vectors respectively. Note that we do not claim to entirely disentangle semantics and syntax in distinct representations. Instead, our goal is to generate vectors that maximally separate the semantic and syntactic information in a sentence.

Probing results

Using the proposed probing method, we return to the central question originally posed. That is, whether and where is the semantic and syntactic information encoded in different brain regions? Based on the semantic and syntactic sentence vectors generated by DFRM, we use the similarity-encoding-analysis method to show the relationships between brain-region and semantic (or syntactic) feature.

Figures 5 and 6 show the averaging results over five subjects on two fMRI experiments respectively. We only show statistically significant results that are higher than random (p-value < 0.05). For fMRI experiment 1, we get 112 and 31 significant ROIs (in which 18 are overlapped) for the semantic and syntactic features respectively. For fMRI experiment 2, we get 125 and 48 significant ROIs (in which 29 are overlapped) for the semantic and syntactic features respectively. These ROIs are distributed across the whole brain, including the default-mode, cingulo-opercular, frontoparietal, smhand networks, etc.. Both semantic and syntactic features are effective in several frontal and temporal regions. Specifically, semantic features are most correlated with angular, cingulum, fusiform, insula and precuneus

gyrus on both left and right head, plus several frontal and temporal regions (i.e., IFGoperc.L, ORBinf.L, ORBinf.R, IFGtriang.L, IFGtriang.R, MFG.L, MFG.R, ORBmid.L, ORBmid.R, SFGdor.L, SFGdor.R, SFGmed.L, ORBmid.L, ORBmid.R, ITG.L, ITG.R, MTG.L, MTG.R, STG.R, TPOmid.R). Syntactic features are most correlated with parts of the superior frontal, superior temporal, middle cingulum, cuneus, inferior parietal, precentral gyrus on right head, plus inferior frontal opercular part, middle frontal, middle temporal gyrus on left head. In addition, the following ROIs are sensitive to the semantic feature only, including right angular, anterior cingulum, right middle cingulum, fusiform, insula, left inferior parietal, right postcentral, precuneus gyrus, plus several frontal and temporal regions (i.e., ORBinf.L, ORBinf.R, IFGtriang.L, IFGtriang.R, MFG.L, ORBmid.L, ORBmid.R, ORBsup.L, ORBsup.R, SFGmed.R, SFGdor.R, ITG.L, ITG.R, MTG.L, MTG.R, STG.R, TPOmid.R) (which corresponds to ROI 4, 7, 13, 24, etc.). The parts of the right superior frontal, right inferior parietal, right cuneus, and left precentral gyrus (which corresponds to ROI 12, 17, 27, 40, etc.) are only sensitive to the syntactic feature.

In general, the above results agree with previous neuroscience findings (Fedorenko, Nieto-Castanon, and Kanwisher 2012; Matchin et al. 2019), further evidencing that linguistic representations are organized in a distributed fashion throughout the language system including most parts of frontal and temporal areas, in which semantic information is represented more robustly than syntactic information across the brain. Furthermore, from the computational perspective, our results provide new evidence for the relationships between brain regions and language features by finding candidate brain regions only sensitive to semantic or syntactic features.

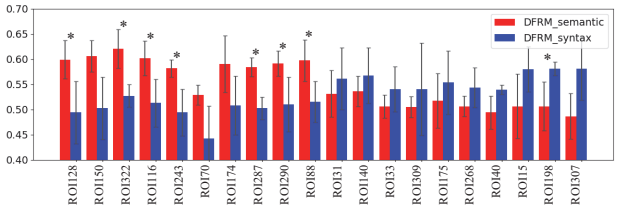

Figure 7: The averaged probing results on ROIs that have the largest gaps between the semantic and syntactic feature on fMRI experiment 1. The bar denotes standard deviation over 5 subjects and the asterisk denotes significantly different results with $p < 0.05$.

To show clearly the differences between semantic and syntactic features, we minus the probing results of the semantic and syntactic features across all ROIs and show the largest 10 ROIs and smallest 10 ROIs in Figures 7 and 8. For fMRI experiment 1, as Figure 7 clearly shows, semantic features are more involved than syntactic features on left inferior temporal, right anterior cingulum, left medial parts of the orbital frontal, right insula, right rectus, right middle temporal, left middle occipital gyrus (i.e., ROI 128, 322, 116, 243, 287, 188). The syntactic features are more involved than semantic features on the right precentral gyrus (i.e., ROI 198). For fMRI experiment 2, as shown in Figure 8, semantic features are more connected to left inferior frontal operculum, right inferior temporal, left middle frontal, right superior temporal, left inferior temporal, right superior temporal, left middle occipital gyrus (i.e., ROI 75, 301, 153, 330, 130, 331, 88, 11) than syntactic features. The syntactic features are more connected to left middle frontal gyrus (i.e., ROI 147) than semantic features.

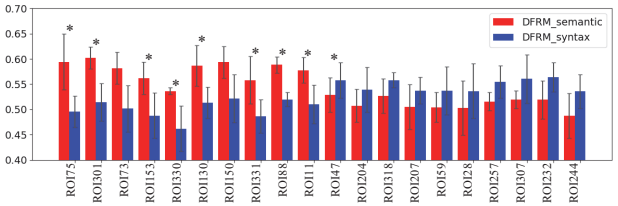

Figure 8: The averaged probing results on ROIs that have the largest gaps between the semantic and syntactic feature on fMRI experiment 2. The bar denotes standard deviation over 5 subjects and the asterisk denotes significantly different results with $p < 0.05$.

Moreover, the probing results in specific brain regions between fMRI experiments 1 and 2 are divergent. The reason is probably because of the significant differences in language understanding among different subjects (refer to the deviation values in Figures 7 and 8). Therefore, it is necessary for future researches on brain language processing to collect much larger datasets which include more participants.

Conclusion and Future Work

Our principal motivation is to understand better whether and where the semantic and syntactic information in sentences are represented in the human brain. The difficulty with such research lies in the intertwined relationships between different kinds of sentence features. To solve this problem, we propose a DFRM method that disentangles semantic and syntactic information in sentences, generating a semantic and a syntactic vector for each sentence. Subsequently, we can explore the brain representation of semantics and syntax by associating disentangled feature vectors with brain activation data. We find that the semantic feature is represented more robustly than the syntactic feature. The brain representations of semantics and syntax are largely overlapped in several frontal and temporal regions, but there are also brain regions only sensitive to one of them. This work corroborates and extends previous findings, highlighting the value of introducing the latest NLP models in studying brain language comprehension.

Future work can move beyond sentence-level analysis, conducting studies of brain language processing from character to discourse level. Moreover, the proposed method can be extended to explore other feature representations in the brain, such as visual, auditory, emotional representations by exploiting the objective function of classifying, or generating corresponding features.

Acknowledgments

The research work has been funded by the Intelligent Science and Technology Project No. 115200S001, the Beijing Municipal Science and Technology Project No. Z181100008918017 and the Natural Science Foundation of China under Grant No. U1836221.

References

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

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

Chen, M.; Tang, Q.; Wiseman, S.; and Gimpel, K. 2019a. Controllable paraphrase generation with a syntactic exemplar. *arXiv preprint arXiv:1906.00565*.

Chen, M.; Tang, Q.; Wiseman, S.; and Gimpel, K. 2019b. A multi-task approach for disentangling syntax and semantics in sentence representations. In *Proceedings of NAACL*, 2453–2464.

Conneau, A.; Kiela, D.; Schwenk, H.; Barrault, L.; and Bordes, A. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of EMNLP*, 670–680.

Dapretto, M., and Bookheimer, S. Y. 1999. Form and content: dissociating syntax and semantics in sentence comprehension. *Neuron* 24(2):427–432.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*.

Fedorenko, E.; Nieto-Castanon, A.; and Kanwisher, N. 2012. Lexical and syntactic representations in the brain: an fmri investigation with multi-voxel pattern analyses. *Neuropsychologia* 50(4):499–513.

Friederici, A. D.; Opitz, B.; and Von Cramon, D. Y. 2000. Segregating semantic and syntactic aspects of processing in the human brain: an fmri investigation of different word types. *Cerebral cortex* 10(7):698–705.

Friederici, A. D. 2012. The cortical language circuit: from auditory perception to sentence comprehension. *Trends in cognitive sciences* 16(5):262–268.

Gordon, E. M.; Laumann, T. O.; Adeyemo, B.; Huckins, J. F.; Kelley, W. M.; and Petersen, S. E. 2014. Generation and evaluation of a cortical area parcellation from resting-state correlations. *Cerebral cortex* 26(1):288–303.

Hasson, U., and Honey, C. J. 2012. Future trends in neuroimaging: Neural processes as expressed within real-life contexts. *NeuroImage* 62(2):1272–1278.

Hill, F.; Cho, K.; and Korhonen, A. 2016. Learning distributed representations of sentences from unlabelled data. In *Proceedings of NAACL*, 1367–1377.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Iyyer, M.; Wieting, J.; Gimpel, K.; and Zettlemoyer, L. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of NAACL*, 1875–1885.

Kiros, R.; Zhu, Y.; Salakhutdinov, R. R.; Zemel, R.; Urtasun, R.; Torralba, A.; and Fidler, S. 2015. Skip-thought vectors. In *Proceedings of NIPS*, 3294–3302.

Le, Q. V., and Mikolov, T. 2014. Distributed representations of sentences and documents. In *Proceedings of ICML*, 1188–1196.

Matchin, W.; Brodbeck, C.; Hammerly, C.; and Lau, E. 2019. The temporal dynamics of structure and content in sentence comprehension: Evidence from fmri-constrained meg. *Human brain mapping* 40(2):663–678.

Mazoyer, B. M.; Tzourio, N.; Frak, V.; Syrota, A.; Murayama, N.; Levrier, O.; Salamon, G.; Dehaene, S.; Cohen, L.; and Mehler, J. 1993. The cortical representation of speech. *Journal of Cognitive Neuroscience* 5(4):467–479.

Pallier, C.; Devauchelle, A.-D.; and Dehaene, S. 2011. Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences* 108(6):2522–2527.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP*, 1532–1543.

Pereira, F.; Lou, B.; Pritchett, B.; Ritter, S.; Gershman, S. J.; Kanwisher, N.; Botvinick, M.; and Fedorenko, E. 2018. Toward a universal decoder of linguistic meaning from brain activation. *Nature communications* 9(1):963.

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. In *Proceedings of NAACL*, 2227–2237.

Sun, J.; Wang, S.; Zhang, J.; and Zong, C. 2019. Towards sentence-level brain decoding with distributed representations. In *Proceedings of AAAI*, volume 33, 7047–7054.

Wang, S.; Zhang, J.; and Zong, C. 2017. Learning sentence representation with guidance of human attention. 4137–4143.

Wehbe, L.; Murphy, B.; Talukdar, P.; Fyshe, A.; Ramdas, A.; and Mitchell, T. 2014. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS one* 9(11):e112575.

Wehbe, L. 2015. *The Time and Location of Natural Reading Processes in the Brain*. Ph.D. Dissertation.

Wieting, J., and Gimpel, K. 2018. Paranmt-50m: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of ACL*, 451–462.

Zhao, J.; Kim, Y.; Zhang, K.; Rush, A.; and LeCun, Y. 2018. Adversarially regularized autoencoders. In *Proceedings of ICML*, 5897–5906.