Towards Sentence-Level Brain Decoding with Distributed Representations

Jingyuan Sun,[1,2] **Shaonan Wang,**[1,2] **Jiajun Zhang,**[1,2] **Chengqing Zong**[1,2,3]
[1]National Laboratory of Pattern Recognition, CASIA, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
[3]CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing, China
{jingyuan.sun, shaonan.wang, jjzhang, cqzong}@nlpr.ia.ac.cn

Abstract

Decoding human brain activities based on linguistic representations has been actively studied in recent years. However, most previous studies exclusively focus on *word*-level representations, and little is learned about decoding whole *sentences* from brain activation patterns. This work is our effort to mend the gap. In this paper, we build decoders to associate brain activities with sentence stimulus via distributed representations, the currently dominant sentence representation approach in natural language processing (NLP). We carry out a systematic evaluation, covering both widely-used baselines and state-of-the-art sentence representation models. We demonstrate how well different types of sentence representations decode the brain activation patterns and give empirical explanations of the performance difference. Moreover, to explore how sentences are neurally represented in the brain, we further compare the sentence representation's correspondence to different brain areas associated with high-level cognitive functions. We find the supervised structured representation models most accurately probe the language atlas of human brain. To the best of our knowledge, this work is the first comprehensive evaluation of distributed sentence representations for brain decoding. We hope this work can contribute to decoding brain activities with NLP representation models, and understanding how linguistic items are neurally represented.

Introduction

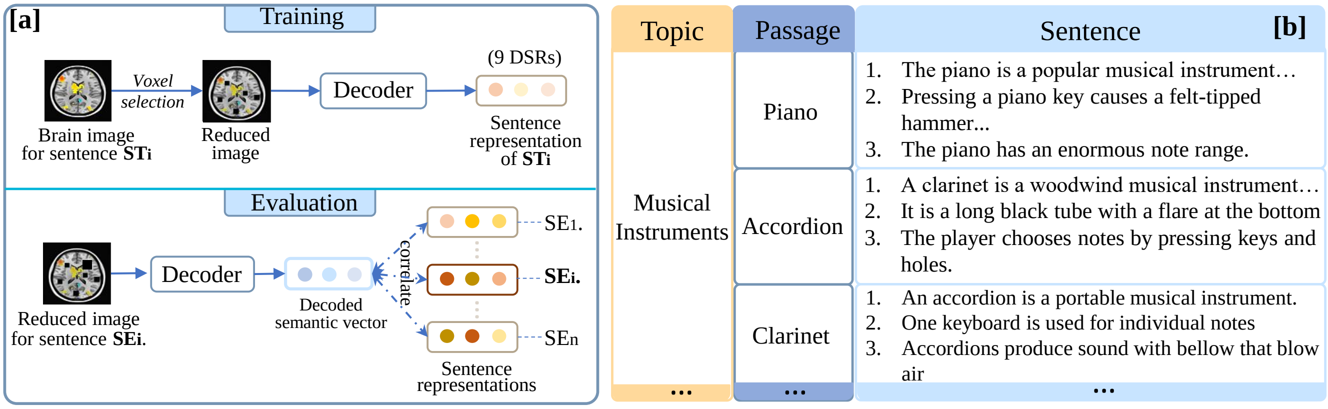
In the past decade, brain imaging technology has been developed to reveal what a person is seeing, perceiving or attending to through analyzing his brain activity patterns (Tong and Pratte 2012). This approach, namely brain decoding, might one day make it possible to read a person's thoughts from noninvasive measurements of brain activations. One important and promising aspect of brain decoding is in the form of language. Much has recently been learned about reconstructing simple linguistic items, such as words and phrases from brain activities measured by functional magnetic resonance imaging (fMRI) (Thirion et al. 2006; Wehbe et al. 2014). But only a few developments have been made in decoding whole sentences (Matsuo et al. 2016).

The key aspect of a sentence decoder is a computational mapping between the brain activities and the sentence stimulus. To build the decoder, it's thus important to find numerical representations of the sentence. Pioneering work (Wehbe et al. 2014) in brain imaging generally uses human-elicited features, which can't fully express fine-gradient differences of sentence meanings. Human-elicited features are also limited in scope to cover the large compositional space of natural language (Nishida and Nishimoto 2017). We instead adopt distributed sentence representation (DSR) models, the currently dominant approach in NLP community. DSR models roughly fall into two categories: unstructured and structured models (Wang, Zhang, and Zong 2017). Unstructured models treat a sentence as a bag of words (Iyyer et al. 2015; Shen et al. 2018), while structured models explicitly catch the sentence structure (Kiros et al. 2015; Logeswaran and Lee 2018), such as word order. We don't yet know which of the two is the better option for brain decoding. More specifically, will structured models more accurately decode the brain activities than unstructured models? If so, under what condition? Is the conclusion consistent with different decoding methods? We will answer these questions to benefit brain decoding from DSR models.

Except for sentence representations, low-dimensional representations of the brain activities are also necessary for brain decoding. The brain images measured by fMRI usually contain hundreds of thousands of voxels, which must be reduced in case of overfitting. Inspired by Pereira et al. (2018), we train regression models to predict sentence representations from the brain images, and keep voxels most informative in the prediction. Other than direct dimensionality reduction, such selection method provides an extra bonus. Through the spatial distribution of the informative voxels on the brain, we can study the relations between the sentence representations and brain areas associated with high level cognitive functions. This may offer some insights for how sentences are represented in the human brain.

In this paper, we explore 9 DSRs, covering both unstructured and structured models, classical baselines and state-of-the-art methods, to represent the sentence stimulus [1]. We then select informative voxels from fMRI images to represent the brain activities. With these representations at hand,

[1] The sentence stimulus are organized in the hierarchy of topic—passage—sentence, allowing for decoding task in different granularities. Fig. 1(b) gives an example of the sentence stimulus



Topic	Passage	Sentence
Musical Instruments	Piano	1. The piano is a popular musical instrument… 2. Pressing a piano key causes a felt-tipped hammer... 3. The piano has an enormous note range.
	Accordion	1. A clarinet is a woodwind musical instrument… 2. It is a long black tube with a flare at the bottom 3. The player chooses notes by pressing keys and holes.
	Clarinet	1. An accordion is a portable musical instrument. 2. One keyboard is used for individual notes 3. Accordions produce sound with bellow that blow air
…	…	…

Figure 1: [a] Training and evaluation of sentence-level brain decoders. The brain images are reduced by voxel selection. The decoders are trained to map the brain imaging data to match the distributed sentence representations. During evaluation, the decoders produce semantic vectors from unseen images to refer the corresponding sentence stimulus. [b] Example of sentence stimulus organized in topic–passage–sentence, taking the musical instrument topic as an instance

we can start the sentence-level brain decoding as shown in Fig.1(a). We use two decoding methods: similarity-based decoding and regression-based decoding. The decoder respectively learns to map the brain images to different sentence representations. The mapping is then evaluated on unseen brain images to reveal corresponding sentence stimulus.

In the evaluation, we find:

(i) Simple unstructured representation models are capable of decoding coarser-grained difference of brain activations. But they lose to the structured models in tasks of finer granularity.

(ii) Performance of representation models may fluctuate in a narrow range across different decoding methods. But one of the supervised structured models, InferSent, consistently outperforms other baselines in nearly all experiments.

(iii) The distributed sentence representations actively probe some of the functional networks of human brain. Sentence representations which better decode the brain activities are shown to have a higher correspondence with the voxels of the language cortex.

Findings of this paper not only demonstrate the weakness and advantages of different sentence representation models in the brain decoding task. They offer a deeper insight of the connection between the two manifestations of mental meanings: the neural activation patterns and the extrinsic linguistic representations. We hope this could boost further research on using NLP representation models in analyzing brain activities.

Related Work

Brain Decoding

The past decade has witnessed considerable progress in the field of brain decoding. Early studies managed to recover simple verbal stimulus, including words (Mitchell et al. 2008; Palatucci et al. 2009; Just et al. 2010; Pereira, Detre, and Botvinick 2011; Handjaras et al. 2016) and phrases (Wehbe et al. 2014; Fyshe 2015; Huth et al. 2016), from brain activation patterns measured by functional magnetic resonance imaging (fMRI). Recent work has shown that sentences may also be decoded (Matsuo et al. 2016; Pereira et al. 2018). These results potentially support a brain-computer interface that could perform “brain reading” (Nishimoto et al. 2011).

In brain decoding, we need to build mapping between stimulus and corresponding brain activities. We call such mapping function a *decoder*. Two types of decoders are most widely used in brain imaging research: the similarity-based decoder (Anderson, Zinszer, and Raizada 2016) and regression-based decoder (Nishida and Nishimoto 2017; Bulat, Clark, and Shutova 2017). Similarity-based decoders re-represent a stimuli by its similarity with other stimulus. The brain activities are re-represented in the same way. The decoding is based on matching the similarity space. Regression-based decoder learns a parametric mapping from the brain activation patterns to the stimulus. The learned mapping then produces quantitative representations of unseen brain activities for further decoding. In this paper, we adopt both the two types of decoders for a comprehensive evaluation. We give a detailed introduction of them in the *Brain Decoding Methods* section.

Distributed Sentence Representations

Sentence representation is an extensively studied field in the NLP community, currently dominated by distributed sentence representation (DSR) models (Wang, Zhang, and Zong 2018). DSR models roughly fall into two classes: unstructured and structured models. Unstructured models generally enjoy minimal parameters and fast training, but may not explicitly account for structural information of a sentence (Iyyer et al. 2015; Shen et al. 2018). In contrast, structured models can catch the sentence structure at the cost of higher computation expense.

Structured models can be further classified into unsupervised and supervised methods. Unsupervised methods (Kiros et al. 2015; Logeswaran and Lee 2018) generally encode a sentence to predict its contexts. So the produced representations catch the lexical co-occurrence patterns of the training corpora. While for supervised models, sentence representations are trained on various supervised tasks, such as natural language inference (NLI) and machine translation (MT). The trained encoders integrate additional task-related semantic information. InferSent (Conneau et al. 2017) and GenSen (Subramanian et al. 2018), state-of-the-art distributed sentence representations, are both supervised structured methods.

Brain Imaging Data

We experiment with the fMRI neural activation data published by Pereira et al. (2018), acquired on a whole-body 3-Tesla Siemens Trio scanner with a 32-channel head coil. We use the brain images corresponding to sentence stimulus. These images are scanned from 5 participants (mean age 27.7, range 21-50), all of them are native English speakers. The details of experimental setup, materials and presentation scripts are available online[2].

Sentence Stimulus

The sentences stimulus are organized in the hierarchy of topic–passage–sentence, as shown in Fig 1(b). In experiments, participants are presented a set of 168 passages, each consisting of 3-4 sentences about a particular concept. The passages cover 48 broad topics (e.g., professions, opera, natural disasters, bone fractures, dreams, etc.) and provide basic information of the corresponding concept in a Wikipedia-style. All passages are presented sentence by sentence. Each sentence is presented for 4s followed by a 4s fixation gap. The entire set of 637 sentences in 168 passages is seen 3 times. The participants are asked to attentively read the sentences they are presented for scanning.

All subjects are scanned three times for every sentence stimulus. The scan is running consistently during the presence of sentence stimulus. Then the acquired data series is corrected by slicing time and motion and concatenated to align the sentence. The fixation gap is used to separate the sentences and distinguish brain activities of language processing with other (noisy) brain activities.

Voxel Selection

The voxel selection method is inspired by Pereira et al. (2018). Formally, regression models are trained on each voxel and its 26 neighbors in 3D to predict each dimension of the sentence representations. The correlation between predicted values and the ground truth sentence representations is then calculated. We take the mean correlation across all dimensions as a voxel's informativeness score. The 5,000 voxels with highest scores are selected. With such method, we select voxels according to how they correspond with the sentence representation. Therefore, through the spatial distribution of the selected voxels on the brain, we can gain some deeper insights of the relation between sentence representations and brain areas associated with different cognitive functions.

Sentence Representation Models

In this paper we evaluate 9 sentence representation models. Both simple and advanced unstructured models, supervised and unsupervised structured models are all included.

Unstructured Models

These models ignore the structure of a sentence, thus are generally easier to train than structured models.

Simple Polling Methods Averaging is almost the simplest way to generate a sentence representation. It returns the element-wise average over word embeddings in a sentence, which can be seen as an average pooling over the sequence. Averaging takes semantic information from every single word, but it also dilutes the most salient features of the sentence. Considering that only a small number of words in a sentence contribute most to its meaning, we also adopt max-pooling as a representation model. It extracts the maximum value along each dimension of the word embeddings, aggregating the most salient features of all dimensions in the sentence representation. Intuitively, features extracted by averaging and max-pooling catch complementary semantic information of a sentence. So we concatenate the two extracted features as the third representation method, motivated by Shen et al (2018).

Advanced Pooling Methods Simple pooling methods enjoy minimal parameters but may only catch limited features of a sentence. Advanced pooling methods are still unstructured but they integrate additional information. FastSent (Hill, Cho, and Korhonen 2016) sums word embeddings in a sentence as its representation to predict the surrounding sentences. SIF[3] (Arora, Liang, and Ma 2016) adapts the naive averaging of word embeddings to weighted averaging. Both methods show improvements over simple pooling methods in downstream tasks. We also test FastSent and SIF in this paper.

Structured Models

Structured representation models are aware of the order and structure of a sentence. How words or phrases affect each other is modeled in the training process but somewhat sacrificing computational efficiency.

Unsupervised Methods Unsupervised structured models generally encode a target sentence to predict its contexts. One typical method is Skip-thought[4] (Kiros et al. 2015). It trains an RNN based encoder-decoder model that reconstructs the contexts of an encoded sentence. Sentences with similar semantic and structure properties can thus be mapped to closer vectors in the representation space. Quick-Thought[5] (Logeswaran and Lee 2018) is an advanced version of skip-thought. It formulates the sentence predicting

[2] https://osf.io/crwz7/wiki/home/

[3] https://github.com/PrincetonML/SIF

[4] https://github.com/ryankiros/skip-thoughts

[5] https://github.com/lajanugen/S2V

as a classification task and achieves impressive acceleration of training speed over skip-thought. Quick-thought is also the state-of-art unsupervised structured model, delivering impressive performance on downstream tasks.

Supervised Methods These methods learn sentence encoders with supervised data of certain NLP tasks. InferSent[6] (Conneau et al. 2017) is a state-of-the-art supervised model trained on the Stanford Natural Language Inference datasets. It is shown to consistently surpass unsupervised models, such as Skip-thought, in a series of downstream tasks. InferSent is trained on one type of supervised task, while another well-performing supervised model, GenSen[7] (Subramanian et al. 2018), is learned in a multi-task manner. The variant of GenSen we test is trained on machine translation and semantic parsing.

Brain Decoding Methods

After the sentence representations are built, the brain decoder establishes an associative mapping between the representation and the imaging data. In this paper we use two brain decoding methods: similarity-based decoding and regression-based decoding.

Similarity-based Decoding

Similarity-based decoding is proposed by Anderson et al. (2016). The first step is to build the similarity-based representation of each sentence embedding and each brain image respectively. Formally, given a set of n sentence representations $\{S_0, ..., S_i, ..., S_n\}$, we calculate similarity-based representation of S_i as

$$R_{S_i} = [corr(S_i, S_0), corr(S_i, S_1), ..., corr(S_i, S_n)], \quad (1)$$

where $corr$ denotes Pearson's correlation. The similarity-based representation of brain images $\{B_0, ..., B_i, ...B_n\}$ can be acquired by analogy. At test time, two of the n similarity-based representations of sentences and their corresponding brain images are chosen for decoding at one time. They are represented by the similarity vectors, R_{S_i}, R_{S_j} for the two sentences and R_{B_i}, R_{B_j} for the corresponding brain images. The decoding is scored a success if $corr(R_{S_i}, R_{B_i}) + corr(R_{S_j}, R_{B_j}) > corr(R_{S_i}, R_{B_j}) + corr(R_{S_j}, R_{B_i})$. This is actually a pairwise matching task, reflecting if the sentence representations match the brain activities in similarity relation.

Regression-based Decoding

Regression-based decoding operates by estimating a semantic vector directly from the voxels, with each dimension predicted by a separate regression model. Let's take Ridge Regression as an example to give a detailed explanation. Formally, we are given the voxel matrix $X \in \mathbb{R}^{N_E \times N_V}$ and sentence representation matrix $Z \in \mathbb{R}^{N_E \times N_D}$ in the training set, where N_E denotes the number of examples, N_V denotes the number of voxels, and N_D denotes the number of dimensions for sentence representation. The regression coefficients W and b are estimated to minimize

$$||WX + b - z_i||_2^2 + \lambda||b||_2^2 \quad (2)$$

for each column z_i in Z, i.e., each dimension of the sentence vectors. λ is the regularization parameter separately set for each dimension with cross-validation. Except for ridge regression, we also test with Lasso-regression and Multilayer Perceptron (MLP) to cover the most widely used regression models in brain decoding. How to evaluate the decoded semantic vectors will be detailed in the following section.

Results

Similarity Based Decoding

As depicted in the *Brain Decoding Methods* section, similarity based decoding is tested on a pairwise matching task. The matching task includes three subtasks in progressively finer granularity, with training and testing sentences coming from:

(i) different topics (e.g., a sentence about a piano vs. a butterfly),

(ii) different passages from the same topic (e.g., a sentence about a dragonfly vs. a butterfly),

(iii) different sentences within the same passage (e.g., two sentences about a piano), for all possible pairs in every subtask.

Fig.2 shows the matching accuracy of different sentence representations under the similarity based decoding. All the tested representations perform significantly above chance level in decoding the sentence stimulus from different topics. The performance is consistent across subjects. Averaging achieves satisfactory performance but doesn't rank the top, even in the pooling based methods. Concatenating maximum and average embeddings achieves minor but consistent improvements over simply averaging and max-pooling. SIF, though taking the form of weighted averaging, performs even worse than naive averaging in some cases.

The pooling-based methods perform almost on a par with the structured model in the first two sub-tasks. From Fig.2[b] we can see that averaging, concatenation and FastSent, the three unstructured models, perform even better than the two unsupervised structured models. However, the advantages of structured models become clear as the tasks getting tough. Especially in the third task of decoding sentences from the same passage, InferSent and GenSen exceed all other baselines by an impressively large margin.

Regression Based Decoding

The regression model is trained and tested on different subsets of the 176 passages (627 sentences) in a 5-fold cross-validation for each participant. We use 141 passages for training and 35 passages for testing in each fold. During each testing round, the mapping function learned in the training round is applied to decode semantic vectors from corresponding brain activations. This then yields decoded vector for every sentence after all folds are done. We evaluate the decoded vectors with the pairwise matching task and a ranking task.

[6] https://github.com/facebookresearch/InferSent

[7] https://github.com/Maluuba/gensen

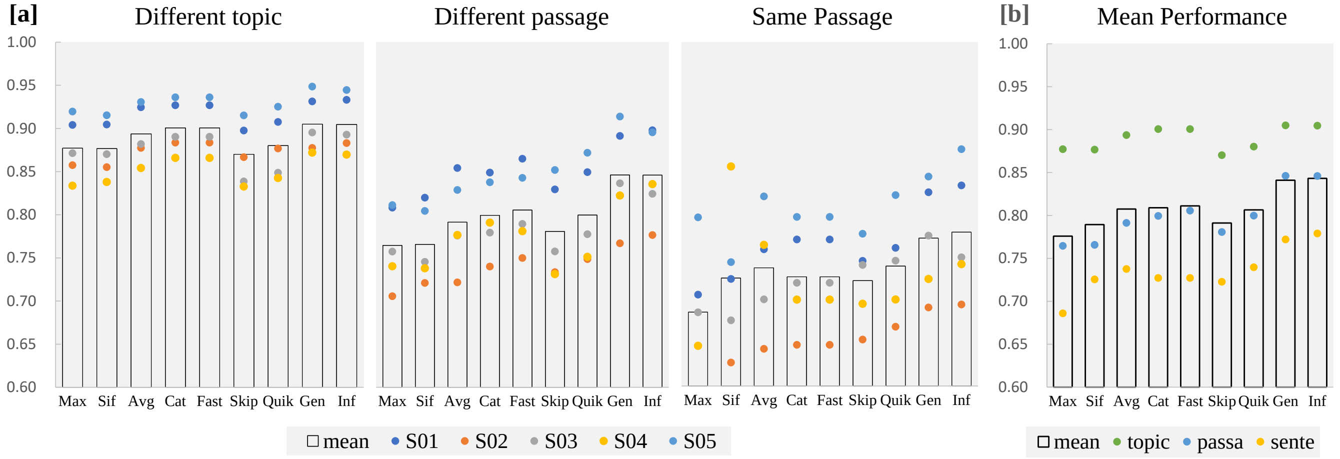


Figure 2: Pairwise matching results for similarity based decoding. We report the results of three subtasks: matching sentences from (i) different topics, (ii) same topic but different passages and (iii) the same passage. We also illustrate the average performance across the three subtasks in the last figure. Each colored dot in the figures denotes the matching accuracy for an individual subject. And dots with the same color refer to a same subject across the subtasks, as depicted by the legend

[a]	Ridge			Lasso			MLP		
	topic	passa.	sente.	topic	passa.	sente.	topic	passa.	sente.
Max	0.88	0.76	0.65	0.88	0.75	0.70	0.83	0.70	0.63
Avg	0.90	0.83	0.73	0.92	0.81	0.78	0.89	0.78	0.67
Cat	0.92	0.83	0.74	0.90	0.81	0.80	0.86	0.74	0.66
Sif	0.89	0.84	0.69	0.91	0.77	0.72	0.84	0.73	0.65
Fast	0.92	0.81	0.74	0.90	0.79	0.77	0.88	0.76	0.67
Skip	0.90	0.82	0.75	0.91	0.80	0.79	0.86	0.81	0.73
Quik	0.91	0.84	0.75	0.91	0.81	0.79	0.90	0.82	0.77
Gen	0.91	0.84	0.78	0.92	0.84	**0.84**	0.91	**0.84**	**0.80**
Inf	**0.94**	**0.90**	**0.83**	**0.93**	**0.86**	**0.84**	**0.92**	**0.84**	0.79

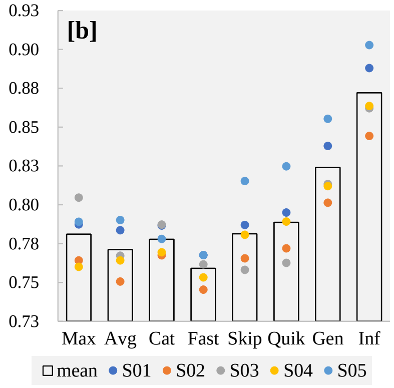


Figure 3: Decoding performance under three regression models: Ridge regression, Lasso regression, and MLP. Table[a]: Average matching accuracy of each tested sentence representation among all subjects. Figure[b]: Ranking accuracy of different sentence representations. Each colored dot in the figure denotes the performance on a certain subject

Pairwise Matching Task In Fig.3(a), we show the matching accuracy of different sentence representations with three regression models. We test with imaging data from every single subject and report the mean accuracy. Generally, representation models deliver consistent performance from the similarity based decoding to the current regression based decoding.

Still, most unstructured models perform nearly as well as structured models in the first two subtasks. Avg and Cat prove strong baselines, fully comparable to the two advanced pooling models. However, the structured models work better in the third subtask, i.e. discriminating sentences from a same passage. Even skip-thought which doesn't perform very well in the first subtask overtakes Avg and Cat here. InferSent surpasses other baselines in nearly all subtasks with the three regression models.

To further understand what kind of of sentence performs better or worse on different methods, we do a case study. Sentence length and word order are of interest. We first analyze the influence of sentence length on decoding performance and show it in Fig.4 , taking several typical representation models. We record the decoding performance (average decoding accuracy among all the subjects) of every single sentence stimuli on different methods. We find the effect of sentence length change on the decoding performance is pretty consistent among the representation models. Generally the accuracy improves when the sentences become longer, especially for the unstructured models. This is within expectation. As sentence become longer, it tends to express more rich and specific meanings. Both qualities of brain images and sentence representations could benefit from that. We then evaluate how much the sentence order matters in

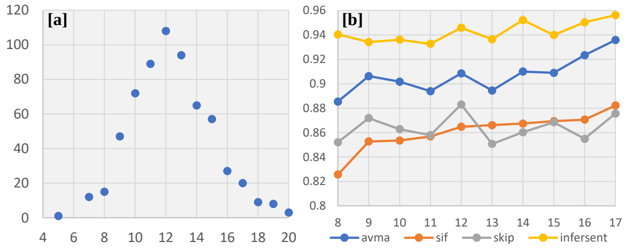


Figure 4: [a] The length distribution of sentences [b] The influence of sentence length to decoding performance. We choose one representative model for each class of methods to display.

the decoding performance for structured models. We don't expect much of that since unsupervised structured models which encode word orders didn't outperform unstructured ones significantly in the previous experiments. We do a semantically identical while syntactic plausible rearrangement of the word order and test the structured models. No significant difference is observed so we won't display it.

Rank Evaluation In the rank evaluation, every decoded vector is correlated with all 637 sentence embeddings and then ranked by the correlation. The score is 1 if the correct sentence embedding ranks top, 0 if it ranks the bottom and in-between otherwise (Pereira et al. 2018). We use the mean ranking score across all decoded vectors as the final score.

Fig.3(b) shows the ranking accuracy. All the tested representation models consistently score above the chance level across the subjects. The three simple pooling embeddings give pretty similar results. Concatenated embedding ranks top in the three and max embedding outperforms the average, but just with small margins. GenSen and InferSent surpass all the simple pooling methods with InferSent ranking the top. Skip-thought's performance ranks the lowest among the structured models, virtually the same as the simple pooling models.

Analysis and Summary

We carry out extensive experiments. All representation models are compared under two decoding frameworks. In regression-based decoding particularly, we test with three different regression models. Throughout all the decoding tasks, we have some common findings to summarize.

In simple pooling based methods, the concatenation of averaging and max-pooling achieves improvements over the single methods. We owe the improvements to the additional information caught by maximum pooling, since averaging dilutes the most salient features of each word in a sentence. Concatenation fares well in both the two brain decoding frameworks, and shows comparable performance to unsupervised structured models in coarse-grained decoding tasks. Given its low computational complexity, it might be considered as a competent alternative in brain semantic decoding. As for the advanced pooling methods, FastSent performs similar with concatenation and slightly better than the naive averaging. This may indicate the benefit of catching co-occurrence patterns of sentences in the representations.

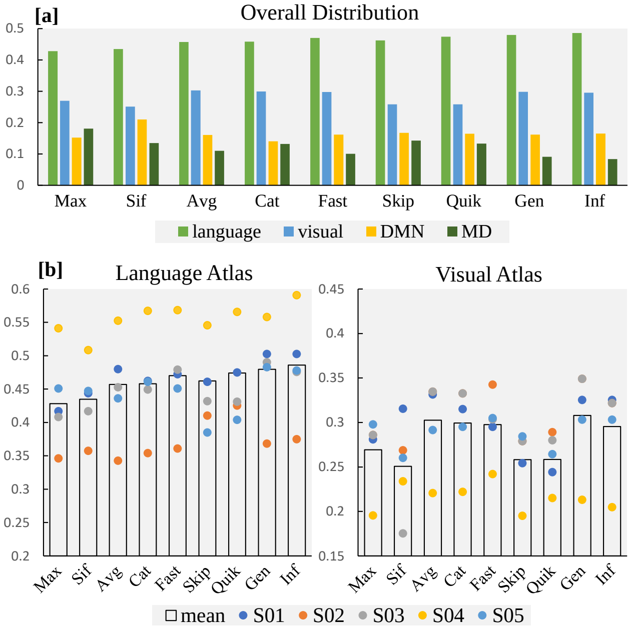


Figure 5: Distribution of informative voxels among brain atlases.[a] Overall distribution [b] Distribution on language and visual atlas of informative voxels on language atlas selected by different sentence representations for every subject. [b] Distribution of informative voxels on language abd visual atlas selected by different sentence representations for every subject.)

The structured models tend to perform better than unstructured models in fine-grained decoding tasks. InferSent, especially, delivers significant improvements over all other baselines in nearly every single experiment. Why? In our experiments, sentences from a same passage tend to use semantically related words to describe one single concept, as shown in Fig. 1(b). This means merely pooling on the the word embeddings is largely possible to produce similar sentence representations. That's where the structured and other auxiliary semantic information come to rescue. But for sentences from different topics or different passages, words are less overlapping. Simply the word embeddings may provide enough semantic features for a distinguishable sentence representation.

Cognitive Insights

How sentences are neurally represented in human brain remains a unsolved problem. We gain some insights through studying the correspondence between sentence representations and functional brain areas. Following Pereira et al. (2018), we pick four brain areas: language atlas (Power et al. 2011), visual atlas (Fedorenko, Behr, and Kanwisher 2011), default mode network (DMN)(Buckner, Andrews-Hanna, and Schacter 2008) and and multi-demand (MD) at-

	Language Atlas			Visual Atlas			DMN			MD		
	topic.	passa.	sente.	topic.	passa.	sente.	topic.	passa.	sente.	topic	passa.	sente.
Max	0.849	0.678	0.554	0.799	0.636	0.497	0.784	0.610	0.568	0.818	0.664	0.516
Avg	0.877	0.712	0.564	0.816	0.656	0.485	0.799	0.616	0.517	0.834	0.676	0.545
Cat	0.882	0.724	0.549	0.825	0.673	0.445	0.809	0.639	0.537	0.842	**0.689**	0.545
Sif	0.850	0.668	0.534	0.778	0.613	0.465	0.750	0.560	0.466	0.795	0.608	0.540
Fast	0.880	0.718	0.557	0.821	0.664	0.465	0.804	0.627	0.527	0.838	0.682	0.545
Skip	0.847	0.673	0.572	0.790	0.632	0.522	0.760	0.620	0.540	0.790	0.631	0.534
Quik	0.847	0.673	0.534	0.810	0.625	0.555	0.780	**0.649**	0.519	0.810	0.614	0.537
Gen	0.860	0.696	0.580	0.819	0.653	0.561	0.793	0.611	0.537	0.813	0.643	0.546
Inf	**0.906**	**0.744**	**0.583**	**0.846**	**0.676**	**0.580**	**0.822**	0.638	**0.575**	**0.856**	0.674	**0.572**

Figure 6: (Table 1) Matching accuracy of sentence representations with voxels constrained on specific brain areas

las (Duncan 2010)[8] . We are particularly interested in the language atlas, since it stores the mappings between linguistic forms and meanings (Power et al. 2011).

In previous experiments, we select informative voxels based on how well they predict the sentence representations, without any spatial constraints over the brain. Nevertheless, voxels themselves belong to different brain areas with high-level cognitive functions. We show how the informative voxels are distributed among these areas, which is actually probing the functional brain areas with sentence representations. As depicted in Fig.5, the informative voxels are not evenly distributed among the atlases. Of all the voxels selected by different sentence representations, about 40% (in average) fall into the language atlas. InferSent, the supervised structured model which consistently leads in previous decoding tasks, selects significantly more language atlas voxels than other methods. Further, most of informative voxels fall outside of the visual atlas, indicating that the information decoded is not primarily visual in nature.

To demonstrate how well different sentence representations decode specific brain area, we further constrain the voxels to that area and re-train the decoders. We show the results in Table 1. The performances are generally consistent with previous experiments. InferSent and GenSen still rank the top. Using voxels in the language atlas leads to better decoding results than other areas. The results may provide some insights of the deeper relationship between the distributed lexical representation and the mental representation of a sentence. InferSent and GenSen capture not only the sentence structure in the representations. The supervised settings allow them to further integrate task-oriented semantic information. For example, sentence representations trained in machine translation may catch the cross-lingual correspondence, trained in NLI better reflect the logic relations among words and phrases. That means these representations are not just statistical mappings of linguistic pattens, but an aggregation of the sentence properties from different points of view. Such supervised representations accurately decode the brain activities and actively probe the language cortices. We thus guess that brain representation of sentence may also be a bindings of multi-source semantic information, but not just a simple reflection of linguistic features.

Conclusion

In this paper, we fully explore different types of distributed representations for sentence-level brain decoding. We conduct the evaluation with the two most widely used decoding methods and multiple tasks to qualify the findings. Empirically, we show the cases where unstructured models can handle and where they fail to structured models, which provides lessons for applying sentence representation models in future decoding work. We also find that the supervised structured models, largely overlooked in previous work, are surprisingly effective in decoding. InferSent, specifically, consistently outperforms the other models in nearly all the tasks. This leads us to recommend supervised structured models to be considered in sentence level brain decoding.

To gain deeper insights of how sentences are neurally represented, we further study the correlation between sentence representations and different brain functional areas. We first show that the informative voxels selected by different sentence representations have roughly consistent distribution patterns, even though the representations themselves might be acquired in very different ways. We are also surprised to find that, without any apriori location constraints, nearly half of the voxels selected by the supervised structured models fall into the language atlas. This means that brain regions active in language processing also highly correspond to these representations. We thus suggest that the way supervised structured models encode a sentence may provide some insights on how human brains represent a sentence.

Acknowledgments

The research work descried in this paper has been supported by the Natural Science Foundation of China under Grant No. 61333018.

References

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.

[8] On average, there are respectively 16193, 12335, 16657 and 32351 voxels in the language atlas, visual atlas, DMN and MD.

Arora, S.; Liang, Y.; and Ma, T. 2016. A simple but tough-to-beat baseline for sentence embeddings. In *Proceedings of the 2016 International Conference on Learning Representations (ICLR)*.

Buckner, R. L.; Andrews-Hanna, J. R.; and Schacter, D. L. 2008. The brain's default network. *Annals of the New York Academy of Sciences* 1124(1):1–38.

Bulat, L.; Clark, S.; and Shutova, E. 2017. Speaking, seeing, understanding: Correlating semantic models with conceptual representation in the brain. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1081–1091.

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 the 2017 conference on empirical methods in natural language processing (EMNLP)*.

Duncan, J. 2010. The multiple-demand (md) system of the primate brain: mental programs for intelligent behaviour. *Trends in cognitive sciences* 14(4):172–179.

Fedorenko, E.; Behr, M. K.; and Kanwisher, N. 2011. Functional specificity for high-level linguistic processing in the human brain. *Proceedings of the National Academy of Sciences* 201112937.

Fyshe, A. 2015. *Corpora and Cognition: The Semantic Composition of Adjectives and Nouns in the Human Brain*. Ph.D. Dissertation, Doctoral dissertation, Air Force Research Laboratory.

Handjaras, G.; Ricciardi, E.; Leo, A.; Lenci, A.; Cecchetti, L.; Cosottini, M.; Marotta, G.; and Pietrini, P. 2016. How concepts are encoded in the human brain: a modality independent, category-based cortical organization of semantic knowledge. *Neuroimage* 135:232–242.

Hill, F.; Cho, K.; and Korhonen, A. 2016. Learning distributed representations of sentences from unlabelled data. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Huth, A. G.; de Heer, W. A.; Griffiths, T. L.; Theunissen, F. E.; and Gallant, J. L. 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* 532(7600):453.

Iyyer, M.; Manjunatha, V.; Boyd-Graber, J.; and Daumé III, H. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, 1681–1691.

Just, M. A.; Cherkassky, V. L.; Aryal, S.; and Mitchell, T. M. 2010. A neurosemantic theory of concrete noun representation based on the underlying brain codes. *PloS one* 5(1):e8622.

Kiros, R.; Zhu, Y.; Salakhutdinov, R. R.; Zemel, R.; Urtasun, R.; Torralba, A.; and Fidler, S. 2015. Skip-thought vectors. In *Advances in neural information processing systems*, 3294–3302.

Logeswaran, L., and Lee, H. 2018. An efficient framework for learning sentence representations. *arXiv preprint arXiv:1803.02893*.

Matsuo, E.; Kobayashi, I.; Nishimoto, S.; Nishida, S.; and Asoh, H. 2016. Generating natural language descriptions for semantic representations of human brain activity. In *Proceedings of the ACL 2016 Student Research Workshop*, 22–29.

Mitchell, T. M.; Shinkareva, S. V.; Carlson, A.; Chang, K.-M.; Malave, V. L.; Mason, R. A.; and Just, M. A. 2008. Predicting human brain activity associated with the meanings of nouns. *science* 320(5880):1191–1195.

Nishida, S., and Nishimoto, S. 2017. Decoding naturalistic experiences from human brain activity via distributed representations of words. *NeuroImage*.

Nishimoto, S.; Vu, A. T.; Naselaris, T.; Benjamini, Y.; Yu, B.; and Gallant, J. L. 2011. Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology* 21(19):1641–1646.

Palatucci, M.; Pomerleau, D.; Hinton, G. E.; and Mitchell, T. M. 2009. Zero-shot learning with semantic output codes. In *Advances in neural information processing systems*, 1410–1418.

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.

Pereira, F.; Detre, G.; and Botvinick, M. 2011. Generating text from functional brain images. *Frontiers in human neuroscience* 5:72.

Power, J. D.; Cohen, A. L.; Nelson, S. M.; Wig, G. S.; Barnes, K. A.; Church, J. A.; Vogel, A. C.; Laumann, T. O.; Miezin, F. M.; Schlaggar, B. L.; et al. 2011. Functional network organization of the human brain. *Neuron* 72(4):665–678.

Shen, D.; Wang, G.; Wang, W.; Min, M. R.; Su, Q.; Zhang, Y.; Li, C.; Henao, R.; and Carin, L. 2018. Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms. In 32^{nd} *AAAI Conference on Artificial Intelligence (AAAI-18)*, 5964–5972.

Subramanian, S.; Trischler, A.; Bengio, Y.; and Pal, C. J. 2018. Learning general purpose distributed sentence representations via large scale multi-task learning. In *Proceedings of the 2018 International Conference on Learning Representations (ICLR)*.

Thirion, B.; Duchesnay, E.; Hubbard, E.; Dubois, J.; Poline, J.-B.; Lebihan, D.; and Dehaene, S. 2006. Inverse retinotopy: inferring the visual content of images from brain activation patterns. *Neuroimage* 33(4):1104–1116.

Tong, F., and Pratte, M. S. 2012. Decoding patterns of human brain activity. *Annual review of psychology* 63:483–509.

Wang, S.; Zhang, J.; and Zong, C. 2017. Learning sentence representation with guidance of human attention. In *International Joint Conference on Artificial Intelligence (IJCAI-17)*, 4137–4143.

Wang, S.; Zhang, J.; and Zong, C. 2018. Empirical exploring word-character relationship for chinese sentence representation. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* 17(3):14.

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.