Instructions for *ACL Proceedings

Anonymous ACL submission

Abstract

This project explores the relationship between word embeddings and brain activity using fMRI data, focusing on decoding neural signals associated with linguistic stimuli. We compared the effectiveness of static embeddings (GloVe) and contextual embeddings (BERT) in predicting neural responses, finding that while BERT captures more contextual nuances, it results in fewer significant voxels compared to GloVe, potentially indicating a higher threshold for significance. Additionally, we conducted an open-ended analysis to examine gender biases in the fMRI data, analyzing both the decoded representations and the ground-truth vectors for masculine and feminine concepts. Our results reveal that the GloVe model strongly reflects expected gender biases, whereas BERT and other models show more varied outcomes. These findings enhance our understanding of how different embedding models can be used to decode linguistic meaning and reflect societal biases in brain signals.

The code for this project can be found on GitHub. [1]

1 Introduction

Decoding linguistic meaning from brain activation patterns marks a significant advancement in our understanding of how language is processed in the human brain. Previous research has often been limited by a focus on concrete nouns and small, constrained sets of stimuli, which restricts the ability to generalize findings across diverse linguistic contexts. To overcome these limitations, Pereira et al. (2018) introduced an innovative approach that decodes semantic vectors from fMRI data, enabling models to generalize across both concrete and abstract concepts.

Building on this work, our project examines the effectiveness of different word embeddings—specifically, static embeddings like GloVe and contextual embeddings like BERT—in decoding neural signals associated with linguistic stimuli. By comparing these embeddings, we aim to assess how well each model captures the complex and nuanced representations of language in the brain.

In addition to the primary analysis, we also explore potential biases present in the neural representations of gendered concepts. This involves analyzing both the decoded representations and the ground-truth vectors for concepts traditionally perceived as masculine or feminine. By investigating how these biases manifest in different embedding models, we aim to shed light on the intersection between language processing in the brain and societal biases reflected in linguistic representations. Our findings contribute to a deeper understanding of the neural encoding of language and highlight the importance of considering biases in cognitive neuroscience research.

2 Data

The data used in this study are derived from three separate fMRI experiments as detailed in Pereira et al. (2018). The first experiment involved individual word stimuli representing 180 distinct concepts. The second and third experiments used sentences related to various topics, with 384 and 243 sentences respectively. The stimuli in Experiments 2 and 3 were developed to cover a broad range of semantic categories, ensuring that the decoding system's generalization capabilities could be rigorously tested.

The fMRI data were collected from multiple participants, with each word or sentence being presented multiple times to average out noise and obtain a stable neural response. The brain imaging data were processed and represented as vectors, which were then used to train and test the decoding models.

[1] https://github.com/TkuiTku100/LCC_Project

3 Experiments and Results

3.1 Structured and Semistructured Tasks

3.1.1 Sentence Decoding

We re-executed the analysis from Homework Assignment 3, Question 3 (see section 2.1 for methodology), using the Word2Vec "word2vec-google-news-300" embedding from the Python library "gensim". The results, shown in Fig. 1 and Fig. 2, revealed no significant difference between the Word2Vec and GloVe embeddings. Both methods yielded similar Rank Accuracy per fold on the FMRI data from Experiment 1, with averages of 60.9 for GloVe and 61.1 for Word2Vec. This suggests that both static embeddings are equally effective for brain decoding tasks within the scope of our analysis.

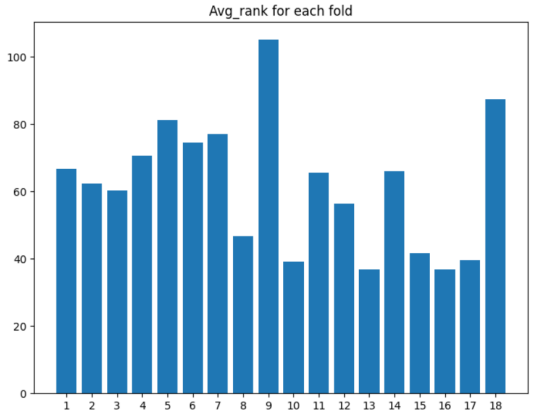

Figure 1: Average rank per fold using glove.

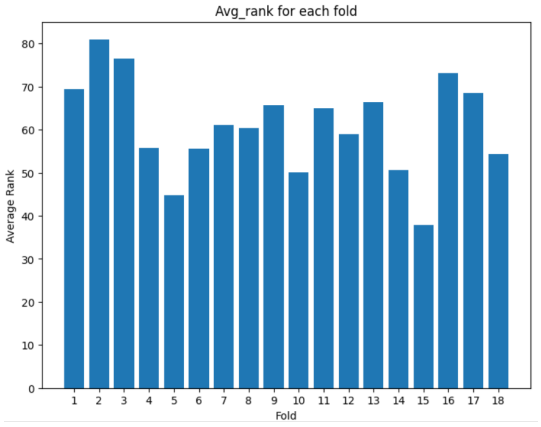

Figure 2: Average rank per fold using word2vec.

3.1.2 Comparison of Analyses

All of the analyses were focused on evaluating the model's ability to create semantic vectors from stimuli representing various concepts. However, there were significant differences between the analyses.

In Analysis 1, the stimuli consisted of individual concepts, specifically single words. This differs from Analyses 2 and 3, where the stimuli were full sentences. The data used in Analysis 1 comprised 180 words, which served as the foundation for training the decoder model across all experiments. In Analysis 1, the model was evaluated using cross-validation to assess its ability to decode words.

In contrast, Analyses 2 and 3 utilized the entire dataset to train the decoder, with the experiments designed to test the model's ability to generalize from single words to sentences. Although Analyses 2 and 3 were more similar to each other than to Analysis 1, there were still some key differences between them.

The dataset for Experiment 2 included 384 sentences, while Experiment 3 contained 243 sentences.Due to the larger number of sentences, Experiment 2 required more time per participant, taking 7 minutes and 34 seconds on average, compared to 4 minutes and 56 seconds for Experiment 3.

Both experiments involved sentences from 24 topics, but the specific topics differed between the two experiments. Despite these differences, both experiments produced similar results in pair-wise comparisons for different topics, the same topic, and same passage comparisons. However, the rank accuracy in Analysis 2 was slightly lower (indicating better performance) than in Analysis 3.

3.1.3 Performance on exp. 2 and exp. 3 data

We used the decoder trained on Word2Vec embeddings from Experiment 1 to decode unseen data from Experiments 2 and 3. Rank accuracy was the performance metric, with results shown in Fig. 3 and Fig. 4. Experiments 2 and 3 contain 384 and 243 sentences, respectively.

The model performed well on both datasets, comparable to its performance on the training data (Experiment 1). However, the decoder's performance was slightly lower, with results closer to the random guess threshold. This increased noise, as indicated by the random threshold being within one standard deviation of the mean rank, is expected due to decoding individual sentences rather than folds.

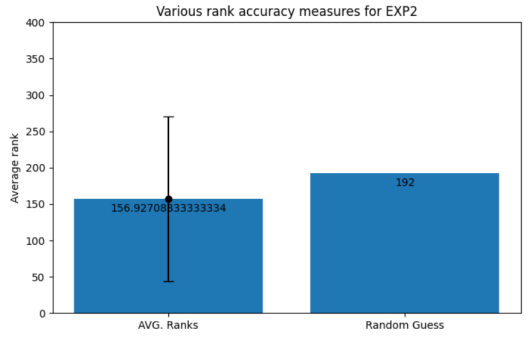

Figure 3: Performance of decoder trained on data from Exp. 1 on datasets from Exp. 2.

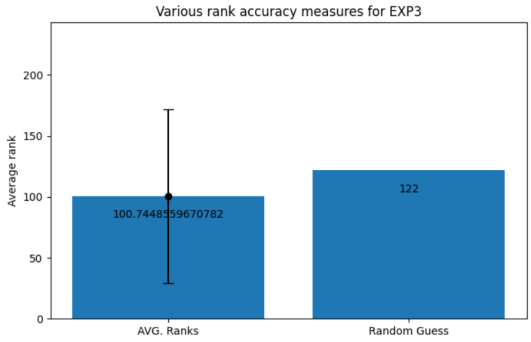

Figure 4: Performance of decoder trained on data from Exp. 1 on datasets from Exp. 3.

3.1.4 Performance on exp. 2 and exp. 3 concepts

we aggregated the results from decoding each sentence in the datasets from Experiments 2 and 3, grouping them by the concepts represented by each sentence. The aggregated results are displayed in Figure 3.

In Figure 5, we can observe that the decoder performed best on the following concepts in Experiment 2: "body part," "human," "drink non alcoholic," "dwelling," and "appliance." Conversely, the concepts where the decoder struggled the most included: "vegetable," "animal," "vehicles transport," "insect," and "profession."

In Figure 6, we can observe that the decoder showed the highest accuracy for the dataset in Experiment 3 for concepts such as "dreams," "stress," "castle," "opera," and "bone fracture." The concepts with the poorest performance included "beekeeping," "owl," "lawn mower," "pharmacist," and "skiing."

Overall, the model demonstrated strong performance across both datasets, achieving better-than-random rank accuracy for the majority of the concepts. As previously noted in Section 1.3, the model generally performed better on the dataset from Experiment 2 compared to Experiment 3, indicating more effective decoding in the former dataset.

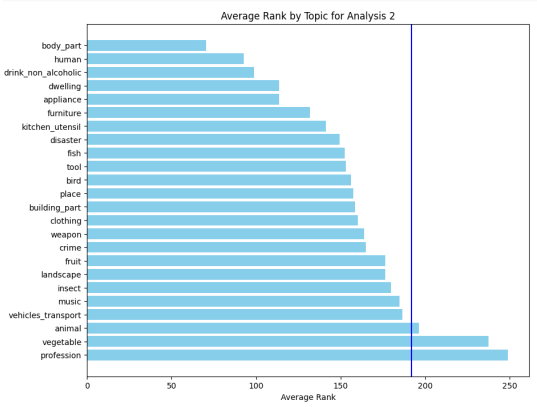

Figure 5: Performance of decoder trained on data from Exp. 1 on datasets from Exp. 2 per concept.

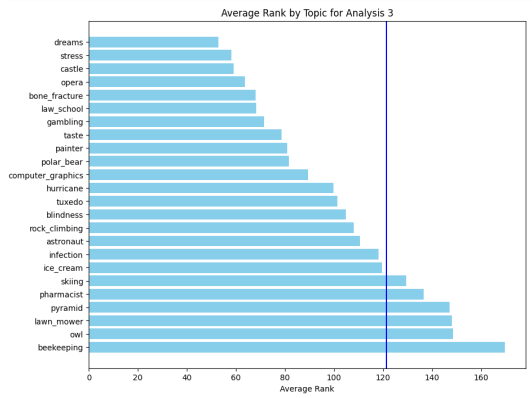

Figure 6: Performance of decoder trained on data from Exp. 1 on datasets from Exp. 3 per concept.

3.1.5 Semi-Structured Part: Neural Encoding

We trained two decoders: one using GloVe embeddings, as in Pereira et al. (2018), and another using BERT embeddings. To generate BERT embeddings for Experiment 2, we encoded the sentences into vectors using a BERT model from the "transformers" library. Each sentence was represented by averaging the embeddings of its words, resulting in 384 vectors.

We divided the Experiment 2 data into 16 folds, training a decoder on 15 folds and testing on the remaining one. The results, summarized in Figure 7, showed that BERT embeddings, which account for sentence context, provided an advantage over static embeddings like GloVe, summarized in Figure 8, with most concepts better than the

random guess, as shown in the blue vertical line.

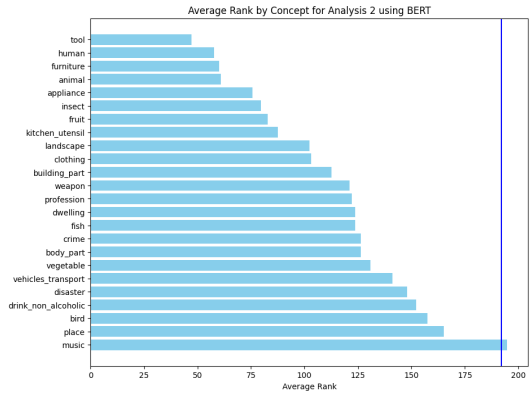

Figure 7: Performance of BERT on Exp. 2 data.

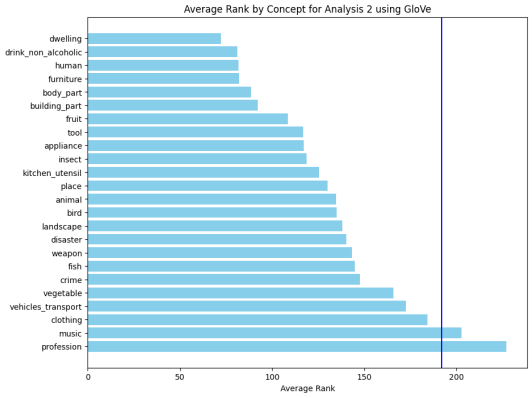

Figure 8: Performance of GloVe on Exp. 2 data.

3.1.6 Voxel Linear Regression

We performed a linear regression on each voxel to assess its relevance in determining the concept a person was thinking about. Voxels with a P-value above 0.05 were classified as noisy and uninformative. This analysis was computationally intensive, taking several hours per run, and required multiple attempts across different machines.

For the BERT model, approximately 20,000 out of 180,000 voxels were found to be significant (P-value < 0.05), with an R^2 of 0.843. In contrast, the GloVe model identified around 29,700 significant voxels, also with an R^2 of 0.843.

The BERT model yielded fewer significant voxels, which we attribute to its ability to capture contextual differences in meaning. For instance, the word "bank" can refer to a financial institution or a riverbank, each invoking different brain regions. BERT captures these distinctions, making it more challenging for voxels to be classified as significant. Conversely, GloVe, lacking contextual understanding, likely includes more noise. We believe BERT sets a higher threshold for significance, making it more precise in identifying relevant voxels.

3.2 Open-Ended Part: Bias Analysis

In this open-ended analysis, we focused on examining gender bias in fMRI data by analyzing both the representations created using the models from the structured and semi-structured tasks, as well as directly on the raw fMRI data. To conduct this analysis, we selected five concepts that are typically regarded as more masculine ("fight", "money", "gun", "beer", "big") and five regarded as more feminine ("dressing", "emotionally", "clothes", "hair", "marriage").

3.2.1 first experiment

In the first experiment, we compared biases in the decoded vectors to biases in the ground-truth vectors. For this purpose, we trained a decoder model similar to the one used in the structured and semi-structured tasks across three representation models: GloVe, Word2Vec, and BERT. Each decoder was trained on the entire dataset of Experiment 1, excluding the specific word being analyzed. For instance, when analyzing the bias for "fight," the decoder model was trained on all concepts except "fight."

We measured bias by comparing the Euclidean distance of the decoded representation to the model's representation of "man" versus "woman" (bias(word) = dis(word, man) - dis(word, woman)). A negative bias indicates that the word is closer to "man," while a positive bias indicates it is closer to "woman." The larger the absolute value, the stronger the bias. We then compared these biases to those in the ground-truth representations.

3.2.2 first experiment analysis

The results of the experiment can be shown in Figures 9, 10 where we present the mean bias across representation models for the masculine and feminine concepts, respectively. These results are interesting and somewhat surprising because, when looking at the actual bias for the model, calculated by the ground truth vector, only the GloVe vectors strongly exhibit the expected behavior in terms of bias. A possible explanation for this is that gender bias is a well-known issue, and representation models may use regularization techniques to minimize or reduce gender bias in vectors.

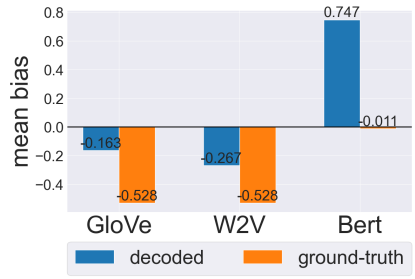

Figure 9: mean bias over representation models for masculine concepts

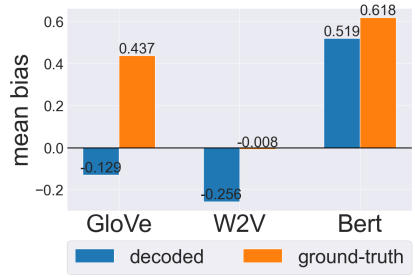

Figure 10: mean bias over representation models for feminine concepts

Still, we observe that Word2Vec finds the masculine concepts closer to "man" than the feminine concepts, while BERT finds the feminine concepts closer to "woman" than the masculine concepts. It is possible that Word2Vec's regularization moves feminine concept vectors closer to "man" and BERT's moves masculine concept vectors closer to "woman," which would explain why Word2Vec's mean bias for both groups of concepts is negative while BERT's is positive.

Considering this, we can see the expected gender bias in the actual bias of all the models. However, when focusing on the decoded vectors, it appears that the decoded representations do not exhibit the bias we expected. For GloVe and Word2Vec, we found that the bias for the feminine concepts is slightly larger than for the masculine concepts, but the difference is not as significant as in the actual biases. As for BERT, we observed a larger bias for the masculine concepts, which is the opposite of what we would expect.

These discrepancies could be attributed to two main factors:

1. The decoder is not complex enough: the training is done only on 180 concepts which them self can be biased (e.g. more masculine than feminine concepts) and we use a pretty simple decoder as we don't train a big NN which can cause the decoded representation be more general and not catch enough complexities.

2. The FMRI data is less biased than the model representation vectors: assuming the decoder creates a semantic vector representation of the FMRI scans, getting less bias for the decoded vectors than the ground-truth might mean the representation in the brain scan themselves are less bias.

To follow up on these reasons we created to additional experiments.

3.2.3 reason (1) follow-up

To test this assumption, we conducted an additional experiment using the GloVe representation model, similar to the first bias experiment we performed. However, this time the training data included data from Experiment 2 and Experiment 3. These experiments expanded our training data to 807 representations across 228 concepts. Adding the data from Experiment 2 and Experiment 3 helps to incorporate further context found in sentences.

concepts	mean bias
masculine	-0.182
feminine	-0.211

Table 1: mean bias for decoded vectors using GloVe model and extended data

In this experiment, we still observed the opposite of the expected bias, with masculine concepts showing a higher mean bias than feminine ones. However, the values are still not very different, which suggests a lower overall bias in the decoded representations.

This experiment alone is not enough to completely refute the first assumption, but it was the best we could do with our computational resources within the scope of this project. Further possible follow-up experiments and extensions will be discussed in Section 4.

3.2.4 reason (2) follow-up

To follow up on this assumption, we focused on the raw fMRI data rather than the decoded representations. A limitation of using the raw data is that there is no clear representation for "woman" or "man" as concepts, so we can't compute the bias as we did in previous experiments. To address this, we decided to use a different method of evaluating bias by comparing the pairwise distance of masculine concepts to themselves and to feminine concepts,

and vice-versa. If there is gender bias in the fMRI data, we would expect each group of concepts to be closer to themselves rather than to words from the other group.

The problem with this experiment is that there are additional confounders between the groups beyond just gender bias. For example, "dressing" and "clothes" are both related to putting on clothes. To address this issue, we looked at each concept separately, as "money," for instance, doesn't necessarily correlate with the other masculine concepts.

concept	mean dis fem concepts	mean dis masc concepts
fight	3345.658	**2466.842**
money	3435.128	**2459.257**
gun	4103.323	**3442.922**
beer	3518.037	**2655.002**
big	3294.925	**2476.460**

Table 2: mean distance for every masculine concept to the other feminine and masculine concept, lower distance in bold

concept	mean dis fem concepts	mean dis masc concepts
dressing	3221.437	**3220.172**
emotionally	**2981.194**	3234.452
clothes	**3190.208**	3882.923
hair	**2964.654**	3294.799
marriage	**3239.829**	4064.725

Table 3: mean distance for every feminine concept to the other feminine and masculine concept, lower distance in bold

The results presented in 2 and 3 align with the expected biases of the groups, as the feminine concepts are generally closer to other feminine concepts, and the masculine concepts are closer to other masculine concepts.

One outlier is "dressing," which may be due to the context given to "dressing" in the original experiment (e.g., ranch dressing compared to dressing up). Not knowing the exact context of the stimuli is a limitation in our analysis and experiments.

This experiment yields interesting results that warrant further research. Possible ways to expand on this study and follow-up experiments are discussed in Section 4.

4 Discussion and Conclusions

In conclusion, both structured and semi-structured analyses reveal the strengths and limitations of static versus contextual embeddings in decoding and predicting brain activity. In the open-ended part we researched and analysed biases in the FMRI data and representation models that can help further understand biases in the human mind and create a base for future work combining these topics and expending our analysis for deeper understanding of the human brain and capabilities of decoding it using simple decoder models Future work could explore more advanced models to further enhance our understanding of how language is represented in the brain.

References

Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. 2018. Toward a universal decoder of linguistic meaning from brain activation. *Nature communications*, 9(1):963.