Word Embeddings and the Brain

Ravid Dimant and Nicolas Zaknun

Faculty of Data and Decision Sciences, Technion, Israel

Git Repository: 096222-Language-Computation-and-Cognition-Project

Abstract

This project aims to expand on the work of Pereira et al. (2018), which focused on decoding words from fMRI data. It consists of three parts: a structured task, semi-structured tasks, and an open-ended task. During the second and third parts, we focused on comparisons between the model used in the article and the GPT-1 and GPT-2 language models. We conducted various analyses, and in this paper, we will present the results obtained.

1 Introduction

Pereira et al. (2018) conducted a groundbreaking study that utilized fMRI data to decode words based on their neural activity patterns. Their analysis, known as Analysis 1, employed GloVe embeddings to bridge the gap between language and neural signals. In this project, we aim to expand on their work by exploring additional facets of word embeddings and their relationship with brain activity.

In the structured task, we seek to replicate Pereira et al.'s Analysis 1 using a different type of static word embeddings, namely Word2Vec. We examined the similarities and differences between analyzes 1, 2, and 3 in that paper and performed various analyses; including testing the GloVe-based decoder model we trained in Homework Assignment 3, question 3, on the datasets from analyses 2 and 3. Additionally, we explored the success of the decoder in predicting sentences across different topics.

In the semi-structured tasks, we conducted a more detailed analysis by training decoder models using the dataset from analysis 2. We employed both the sentence representations used in the paper, obtained through the GloVe model, and the sentence representations extracted from contextualized word embedding models. For this task, we focused on GPT-1 and GPT-2 language models. Continuing with these language models, our objective was to predict human neural signals from the embedding vector representations of the sentences.

Finally, in the open-ended task, out of personal interest, we focus on comparing GPT-1 and GPT-2 with GloVe embeddings to examine their abilities in recognizing topics. By analyzing their performance in decoding sentences related to different broad topics, we can evaluate the effectiveness of these models in capturing semantic information and predicting topic-related sentences.

Through these structured, semi-structured, and open-ended tasks, this project aims to enhance our understanding of the relationship between word embeddings and brain activity.

2 Structured Task - Sentence decoding

Repeating the Analysis of Homework Assignment 3 with Word2Vec

In Homework Assignment 3, question 3, we were tasked with splitting the data from experiment 1 from Pereira et al. (2018), into training and test sets using 18-fold cross-validation. We then trained the decoder and decoded semantic vectors. Then, for each fold we evaluated the accuracy of the decoded vectors using the average accuracy rank. The results from that task are shown in Fig.13 in the appendix. We performed this analysis once again using another type of static word embedding. We chose to do it with Word2Vec Google News 300, and compared the results to those obtained with GloVe, as you can see in Fig.1.

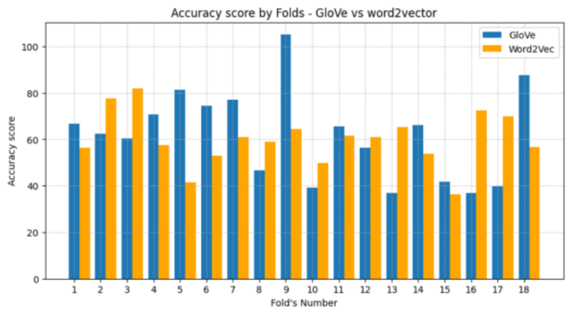

Fig.1: Accuracy scores per fold in EXP1 – Glove vs. Word2Vec

We can see that in some of the folds there are some changes, such as folds 2,3,16 and especially 9, but even though, the average score is nearly maintained: 59.922 in word2vec model comparing to the one in GloVe: 61.911. In addition, the standard deviation in GloVe is even higher: 18.786 comparing to 11.423 in our new model. Therefore, we say that greater variability in GloVe's predictions makes Word2Vec better than him, but to a small extent.

Similarities and Differences Points in Pereira et al. (2018)

Afterwards, we were asked to describe the similarities and differences between analyzes 1, 2, and 3 in Pereira et al. (2018).

Similarities points between analyzes 1, 2, and 3:

- In all three analyses, semantic information is decoded using a decoder that has been trained on brain activation patterns.
- The decoders in all three analyzes use ridge regression for predicting each dimension of the semantic vectors.
- Participants in various experiments provided the brain imaging data used to train the decoders.
- The decoders were evaluated using cross-validation techniques.
- Pairwise classification tasks and rank accuracy classification tasks were used to evaluate the decoders' performance.

Points of difference between analyzes 1, 2, and 3:

- Analysis 1 focuses on single concept decoding using three different paradigms: sentences, pictures, and word clouds. The decoding performance is evaluated separately for each paradigm, as well as for the average of the three paradigms. analyzes 2 and 3, on the other hand, involve sentence decoding using independent sets of stimuli in two separate experiments.
- Analysis 1 evaluates decoding performance by comparing decoded vectors to text-derived semantic vectors, whereas analyzes 2 and 3 evaluate performance by distinguishing sentences at different levels of granularity.
- While analyzes 2 and 3 each have 8 and 6 participants, analysis 1 has more - 16 participants.
- In analysis 1 they trained the decoder and evaluated its ability to generate concepts from provided brain data, whereas in experiments 2 and 3 they used a decoder that had already been trained on specific concepts and tested if it could decode brain-data for sentences.
- The only significant differences between experiments 2 and 3 are the actual stimuli used and the addition of a narrative passage as a stimulus. However, experiment 1 is different from 2 and 3 in that it used single words as the stimulus rather than passages.

Testing the Trained GloVe Decoder on datasets from analyses 2 & 3

In this part, we used the GloVe based decoder model that we trained to test it on the datasets from analyses 2 and 3. Each dataset contains sentence representations and the corresponding neural data from an individual subject. We compared the accuracy scores from all the experiments as shown in Fig.2.

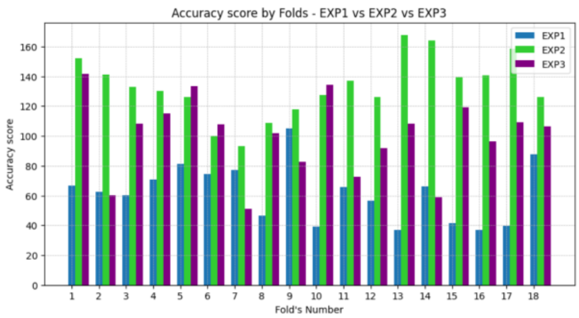

Fig.2: Accuracy scores per fold with GloVe – EXP1 vs. EXP2 vs. EXP.3

These results are not surprising. The results from Experiment 1 indicate that it differs from Experiments 2 and 3 (combined) to a greater extent than the difference between Experiments 2 and 3 (separated). That's because we got lower accuracy scores in each fold in experiment 1 compared to the other. The mean accuracy score in experiment 1 is 61.911 which is lower than the higher mean results from experiments 2 and 3: 132.775 and 100.021 accordingly. This, in full accordance with the differences we saw before between analyzes 1, 2, and 3 in Pereira et al. (2018). According to the article, we saw that experiment 1 differs to a higher degree from experiments 2 and 3 together than the level of difference between experiment 2 and 3.

Analyzing the Performance of the GloVe Decoder from a Topic Perspective

In this part, we analyzed the accuracy scores from the previous section, to identify the topics where the decoder was successful in predicting the sentences. The datasets that we are using contain sentence representations that are related to specific passages, and every passage is related to a specific broad topic. This means that we can look at the accuracy scores for each topic to see how well the decoder performed on different topics. We presented the accuracy rank for each topic in tow tables – one for each experiment.

Topic ID	Topic Name	Accuracy Rank
21	tool	255.5
2	appliance	245.5
3	bird	244.5
5	building_part	244.5
9	drink_non_alcoholic	226.5
11	fish	225.5
6	clothing	221.5
19	place	214.5
14	human	200.5
22	vegetable	197.5
24	weapon	193.5
17	landscape	193.5
1	animal	183.5
16	kitchen_utensil	181.5
8	disaster	177.5
23	vehicles_transport	175.5
20	profession	173.5
7	crime	171.5
18	music	171.5
15	insect	156.5
4	body_part	155.5
10	dwelling	145.5
12	fruit	135.5
13	furniture	129.5

Table 1: GloVe Decoder Performance by Topic in EXP2

Topic ID	Topic Name	Accuracy Rank
21	skiing	238.5
20	rock_climibg	228.0
19	pyramid	217.5
5	castle	207.5
18	polar_bear	197.5
15	owl	187.5
16	painter	177.5
14	opera	167.0
11	infection	156.5
4	bone_fracture	146.0
23	taste	136.0
3	blindness	126.5
24	tuxedo	116.5
22	stress	106.5
17	pharmacist	96.5
12	law_school	86.5
6	computer_graphics	76.5
1	astronaut	66.5
13	lawn_mower	56.5
10	ice_cream	46.5
9	hurricane	26.5
8	gambling	26.5
7	dreams	16.5
2	beekeeping	6.0

Table 2: GloVe Decoder Performance by Topic in EXP3

From the results in Table1 and Table2, we see that in experiment 2 the decoder performs well on the topics: 'tool', 'appliance', 'bird', 'building_part' and 'drink_non_alcoholic', but less succeeds on the topics: 'furniture', 'fruit', 'dwelling', 'body_part' and 'insect'. Regarding experiment 3, the decoder performs well on the topics: 'skiing', 'rock_climibg', 'pyramid', 'castle' and 'polar_bear', but less succeeds on the topics: 'beekeeping', 'dreams', 'gambling', 'hurricane' and 'ice_cream'.

3 Semi-Structured Tasks

Training Decoder Models

In this section, we aim to train a decoder model on the dataset from analysis 2, employing two different approaches for sentence representations. First, we utilize the same sentence representations used in the paper, namely, the GloVe model. Second, we extract sentence representations from a contextualized word embedding model. Especially because these days when "Chat-GPT" by 'OpenAI' has become a 'rising star' all over the world, from personal interest, we focus on GPT-1 and GPT-2 as our chosen contextualized word embedding models. The comparison of these three models is shown in Fig.3.

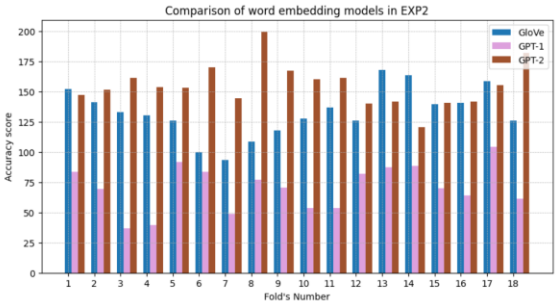

Fig.3: Accuracy scores per fold in EXP2 – GloVe vs. GPT-1 vs. GPT-2

We can observe that the sentence representation model used in the paper, the GloVe mode, outperforms GPT-1 but is surpassed by GPT-2.

Furthermore, it is evident that GPT-2 significantly outperforms GPT-1, as reflected in its higher accuracy scores. The mean accuracy score of GPT-2 surpasses that of GPT-1, with a value of 155.156 compared to 70.394. In addition, GPT-2 exhibits lower standard deviation in its results compared to GPT-1, with values of 17.721 and 18.752, respectively. Kudos to OpenAI for this improvement.

Building a Brain-encoder for Analysis of Neural Encoding

In this task, we built a brain-encoder model that predicts human neural signals based on the embedding vector representations of sentences. We chose to do so with the dataset from analysis 2 of Pereira et al. (2018). For each voxel in the dataset, we trained a distinct linear regression model and computed the R^2 score for each model. This will allow us to examine how many voxels are significantly associated with the information embedded in the word vectors, and how well those voxels are predicted.

We decided that voxel is significant if it's t-test provided P-value < 0.05 and $R^2 > 0.1$. According to Fig.14 in the appendix, we found that there are 0 number of significant voxels for each model. In first sight it was a little bit strange to us, but afterwards, we decided to explore more in order to explain that. Therfore, we looked at the R^2 scores and the P-values obtained from t-tests for each model.

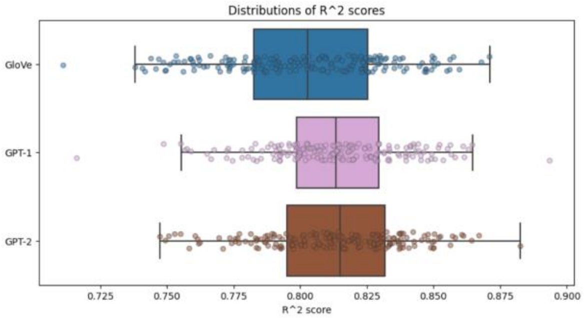

Fig.4: Distributions of R^2 scores per model

From Fig.4 and Fig.5, we can see that there is not much difference between GPT-1 and GPT-2 when measuring how well the regression model fits the observed data. To confirm this, we calculated confidence intervals with 90% confidence level for the mean R^2 score. For GPT-1, the confidence interval is: (0.8101787002379147, 0.8162488888173407), and for GPT-2, it is: (0.8098750320143159, 0.8162793182319829).

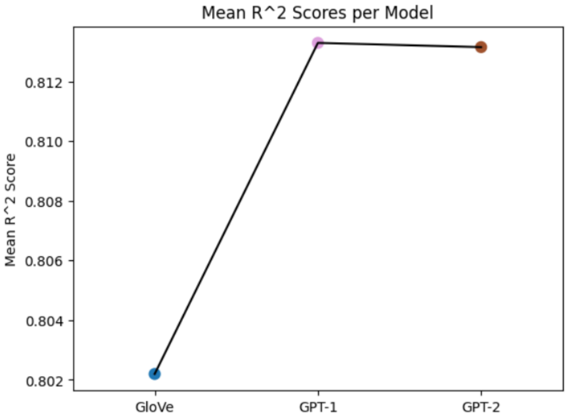

Fig.5: Mean R^2 scores per model

Both models achieve high mean R^2 scores – they are respectively close to 1, indicating that they fit the data in experiment 2 very well. Additionally, the GloVe model performs similarly but to a lesser extent. The confidence interval for the mean R^2 in GloVe is (0.7983735710722256, 0.8058356144969091).

In our t-tests, the null hypothesis claims that there is no significant difference between the mean values of the predicted voxel data and the actual voxel data. From Fif.16 to Fig.17 in the appendix, we see that the P-values for GPT-1 and GPT-2 are very close to 1. We calculated confidence interval for the mean P-values for the three model:
GloVe:
(0.9999999999999394, 0.9999999999999504)
GPT-1:
(0.9999991504442773, 0.999999400564874)
GPT-2:
(0.9999991421180766, 0.9999993792018596)

In this case, the p-values are larger than our significance level, which is 0.05 and close to 1, indicating extremely strong evidence in favor of the null hypothesis. Thus, we do not reject the null hypothesis and state that there isn't a significant difference between the mean values of the predicted voxel data and the actual voxel data. Therfore and according to what we defiened as significant voxel - there are no significantly voxels associated with the information embedded in the word vectors as we saw in Fig.14. Moreover, Fig.15 shows the same results for GloVe and therefore there are 0 significantly voxels for that model as well.

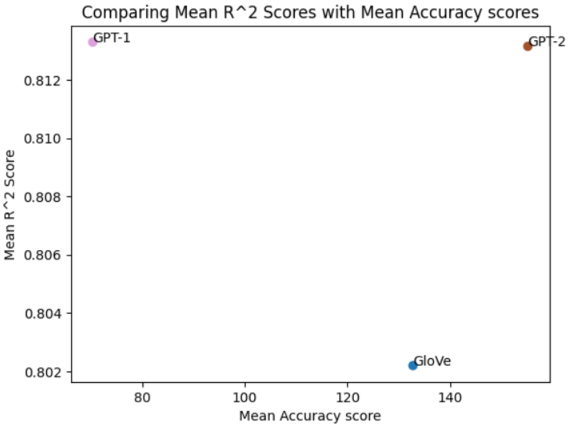

Fig.6: Comparison between mean $R^2$and mean accuracy score

Furthermore, we made a comparison as you can see in Fig.6, between all the 3 models referring to both R^2 scores and accuracy scores. We saw earlier that GPT-2 has better accuracies score, but now we see in that plot that GPT-2 is better in his R^2than Glove but a little bit smaller than GPT-1 but to a very small extent. These results are reflected also in the confidence intervals for the mean R^2scores we saw before. Even though GPT-1 has a little bit larger R^2scores than GPT-2, we believe that GPT-2 is better. The p-values from the t-tests and the high R^2 scores both provide evidence to reject the null hypothesis and accept that the predicted voxel data significantly differs from the actual voxel data.

4 Open-ended Task

Introduction

The development of the GPT family of language models has captured our keen interest. In the last year, OpenAI's introduction of CHAT-GPT sent shockwaves through the world. Today, it is nearly impossible to find any of our associates who have not been acquainted with this groundbreaking revolution. As a result, we decided to delve deeper into examining the differences between GPT-1 and GPT-2.

Our assumption is that GPT-2 is better at identifying topics from sentences than GPT-1. We would like to confirm that assumption and see how much better GPT-2 is than GPT-1. We also wanted to combine the conclusions of the two previous parts of the project and explore the language models in depth.

To maintain continuity in our project analyses, we have decided to proceed with Analysis 2 as outlined in the article. Consequently, we have continued working with the same dataset to ensure consistency in our data exploration.

Analyzing the data

For each model: GloVe, GPT-1, and GPT-2 we created:

- Clustering plots in 2 and 3 dimensions.
- Similarity matrix.
- Word cloud for words within a cluster.

For clustering, we firstly conducted PCA for reducing the dimension for all the vectors from the dataset we used with so far.

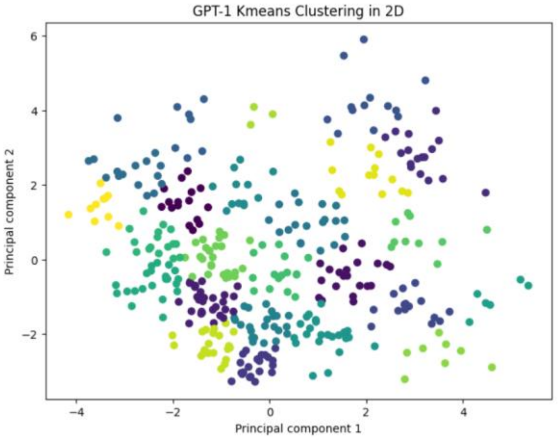

Fig.7: GPT-1 clustering in 2 dimensions.

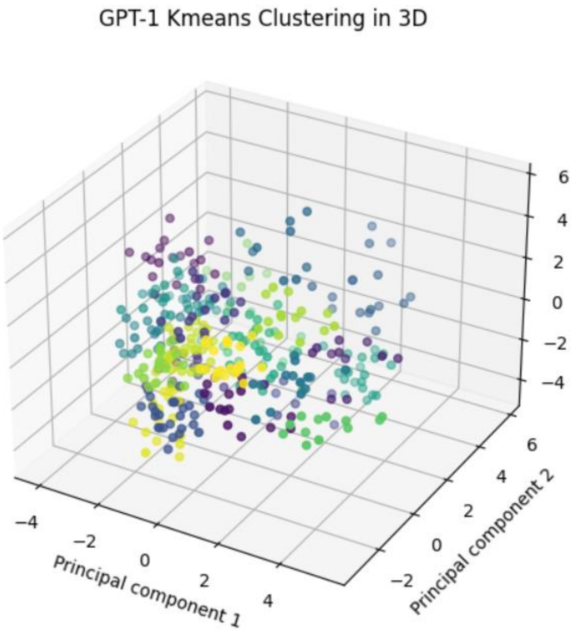

Fig.8: GPT-1 clustering in 3 dimensions.

We reduced all the vectors to 2 and 3 dimensions and then, performed the K-Means algorithm for clustering. We set the number of clusters in the algorithm to 24 out of hope we will see 24 clusters for 24 topics as we saw in the structed task. Afterwards we showed the clustering plots for each dimension as shown in Fig.7 and Fig.8.

Except getting nice looking plots, we did not find any meaning in them. That is, altough we got clusters for close points in the graph we wanted to seek an explenation for the clustering result and moreover, to understand which model condected the best clustering. I.E. – which model knows to identify similiar topics in given sentences from our dataset. Another nice attempt for catching any meaning in this analisis, is the creation of a similarity matrix for each model as shown in Fig.20.

Finally, we decided to focus on the topics within a cluster. For this visuzalization we chose to create 'word clouds' for the topics in every cluster for each one of the three model. We checked which topics each model put in every cluster and were amazed by the results.

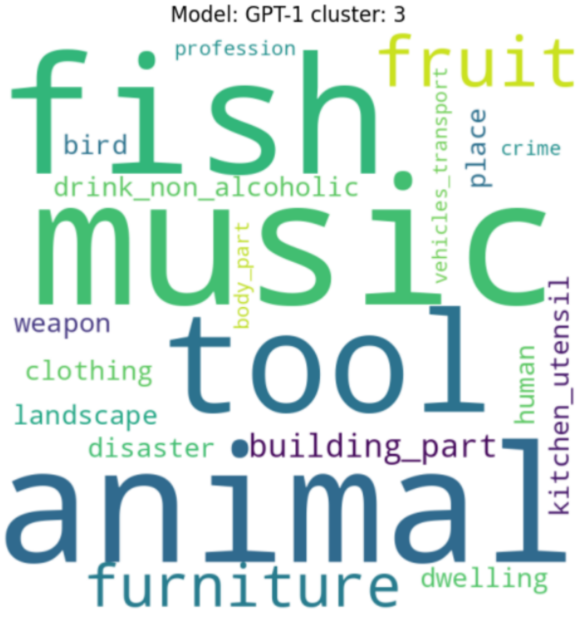

Fig.9: Word cloud for the 3rd cluster obtained by GPT-1

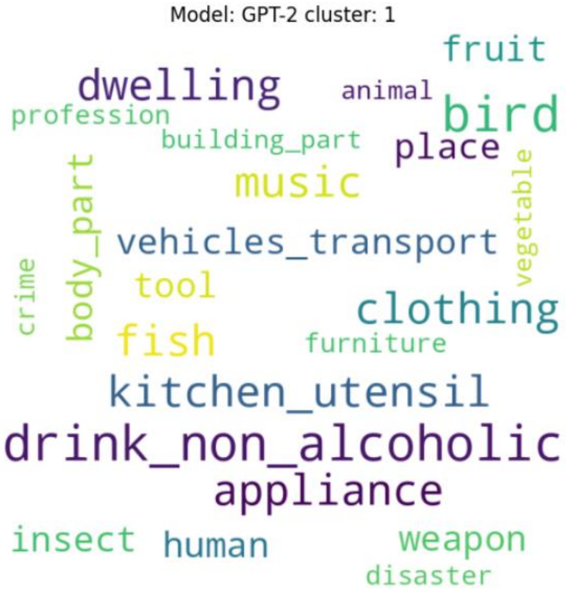

Fig.10: Word cloud for the 1st cluster obtained by GPT-2

Results

In Fig.9 and Fig.10 we can find our answer for the question – which model cluster topics better. Fig.9 describes the topics that GPT-1 captured such as 'fish' and 'fruits'. In Fig.10 we see that GPT-2 captured more topics with the same related field then GPT-1 (in this example- 'Food'). That's because in the last figure we can see that the topic 'vegetables' appear next to 'fruit' and 'fish' topics. We found that the topic 'vegetables' can be found in other cluster generated from the clustering with GPT-1 as shown in Fig.21. GPT-1 and GPT-2 make a good clustering after all – we can find other topics from the same related field in Fig.10 and in Fig.9 such as: 'weapon', 'disaster' and 'crime'.

We decided to compare between the model's clustering according to their Silhouette scores.

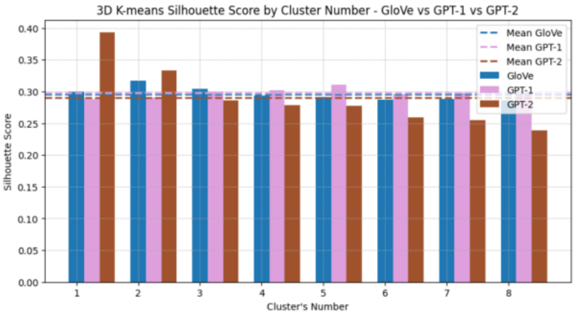

Fig.10: Silhouette score by number of clusters with PCA - Glove vs. GPT-1 vs. GPT-2

Firstly, we compared the means of the Silhouette scores between number of clusters. After performing PCA and reducing the dimension of the data to 3, the means were:

GloVe: 0.2959

GPT-1: 0.2979

GPT-2: 0.2901

These results are very close to each other, and it seems that we lost the option to compare between the ability of identifying topics because we reduce dimensions aggressively. We than thought to ourselves that it effected GPT-2 more than GPT-1 and GloVe and wanted to check it.

As shown in Fig.11, We discovered that GPT-2 has better mean Silhouette score without Principal component analysis Over the other models. In fact, GPT-2 gets better Silhouette score no matter how many clusters we use and that is a major discover to us.

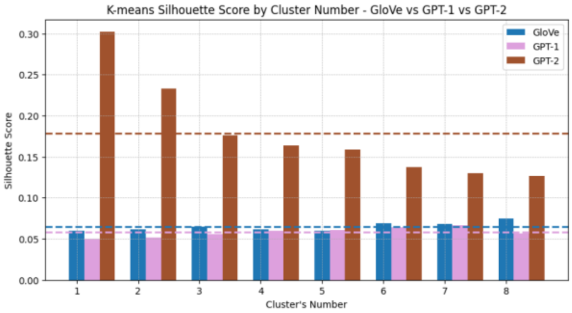

Fig.11: Silhouette score by number of clusters without PCA - Glove vs. GPT-1 vs. GPT-2

The Silhouette score per model over cluster's number is:

GloVe: 0.0649

GPT-1: 0.0582

GPT-2: 0.1784

We additionally calculated confidence intervals with 90% confidence level for the mean silhouette score, once again as before to make it clear:

GloVe:

(0.06242774622456167, 0.068365079159642)

GPT-1:

(0.0550716191202432, 0.0611590780317783)

GPT-2:

(0.152063667069117, 0. 2226721879652755)

We can see than the limits of the confidence intervals for the mean silhouette score obtained by GPT-2 are larger than those of the other models. Thus, we infer that GPT-2 indeed classify topics better than GPT-1 and GloVe.

5 Conclusions

In this report, we conducted an extensive exploration of neural decoding and encoding, specifically examining sentence representations and their connection to the brain. In the structured task, we evaluated the performance of two static word embeddings, namely GloVe and Word2Vec, in sentence decoding. These embeddings yielded nearly identical accuracy results, but we concluded that Word2Vec is slightly better. Moreover, we conducted a thorough comparison between our analysis and Pereira et al.'s (2018) research, which focused on the development of a universal brain decoder. Subsequently, we identified dissimilarities in the accuracy scores when comparing the two analyses.

Moving on to the semi-structured tasks, we trained decoder models using dataset according to Pereira et al.'s analyses 2. The results consistently indicated that GPT-2 outperformed both GloVe and GPT-1 in terms of average rank accuracy. Additionally, we constructed a brain-encoder model to predict neural signals based on sentence embeddings, and all three models demonstrated high $R^2$and P-values scores.

In the open-ended section, we verified that GPT-2 exhibits substantial superiority over both GloVe and GPT-1 in identifying topics in sentences, providing firsthand evidence of OpenAI's progress in developing language models. In summary, this report offers valuable insights into the utilization of word embeddings for sentence decoding, the efficacy of diverse decoder and encoder models, and the investigation of GPT-1, GPT-2, and GloVe in the context of topic identification in sentences. These discoveries significantly enhance our comprehension of language processing within the brain and shed light on the potential applications of word embeddings in tasks related to neural decoding and encoding.

6 Acknowledgments

We would like to thank the 'Language Computation and Cognition' course staff for full responsiveness and support even in stressful moments.

7 References

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S. J., Kanwisher, N., Botvinick, M., & Fedorenko, E. (2018). Toward a universal decoder of linguistic meaning from brain activation. Nature communications, 9 (1), 1–13.

Huth, A. G., De Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. Nature, 532 (7600), 453–458.

8 Appendix

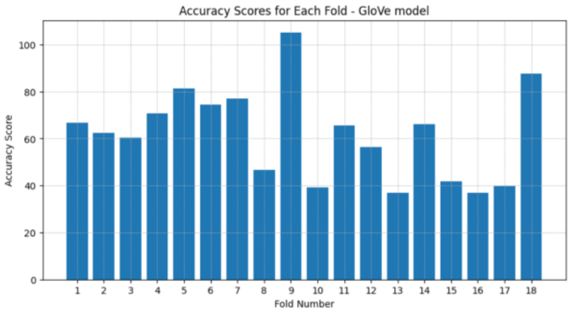

Fig.13: Accuracy scores per fold in EXP1 - Glove

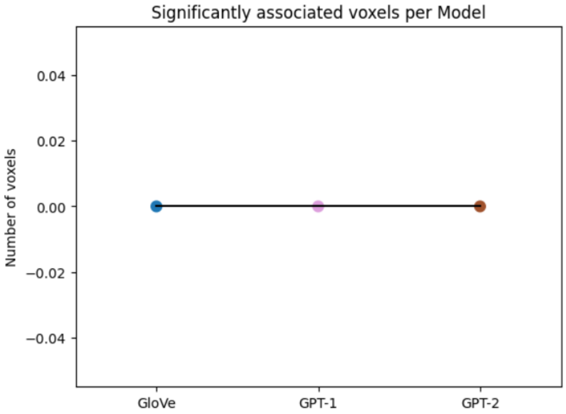

Fig.14: Number of significantly associated voxels per model

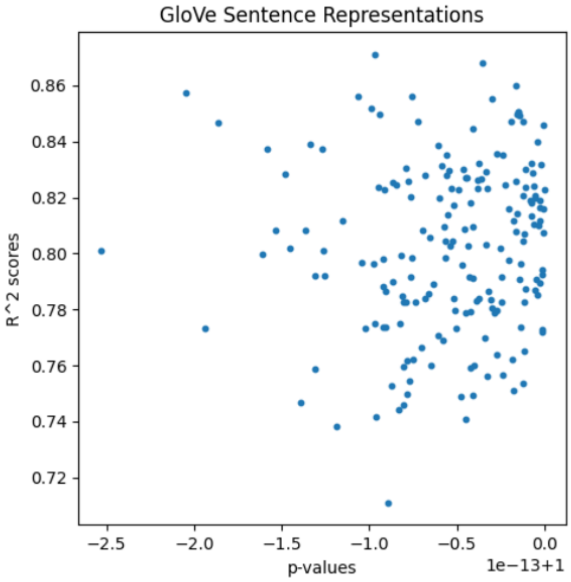

Fig.15: R^2scores as function of P-values in GloVe

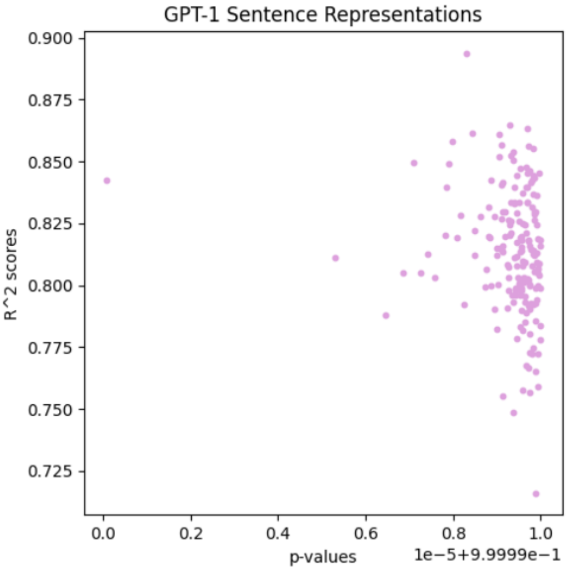

Fig.16: R^2scores as function of P-values in GPT-1

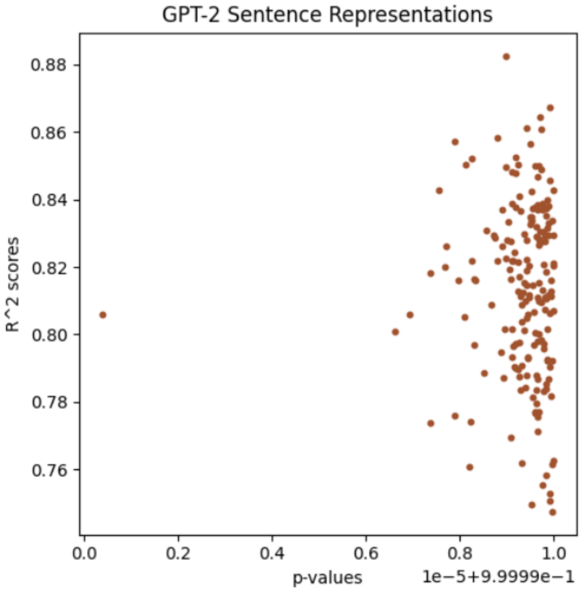

Fig.17: R^2scores as function of P-values in GPT-2

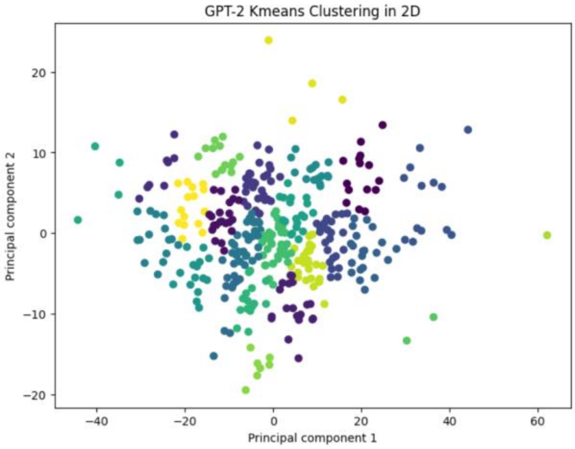

Fig.18: GPT-2 clustering in 2 dimensions.

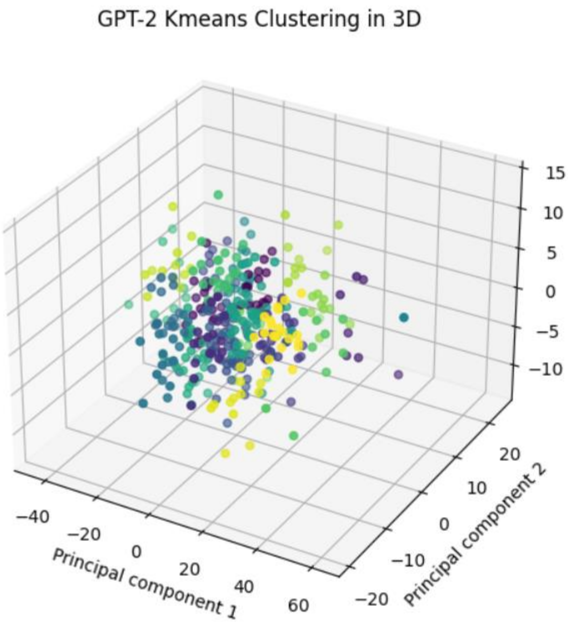

Fig.19: GPT-1 clustering in 2 dimensions.

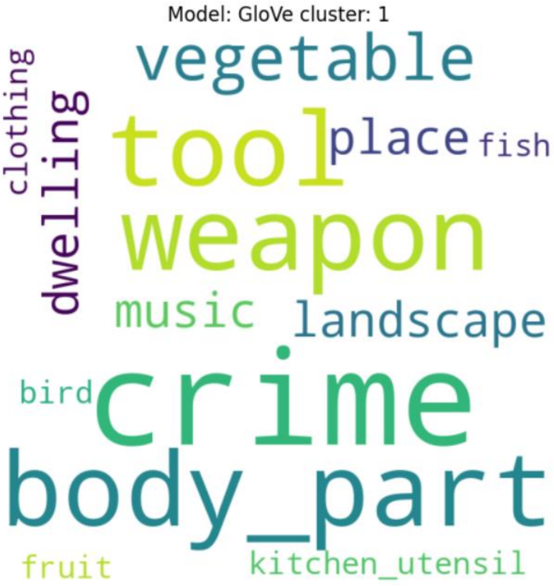

Fig.21: Word cloud for the 1st cluster obtained by GPT-1

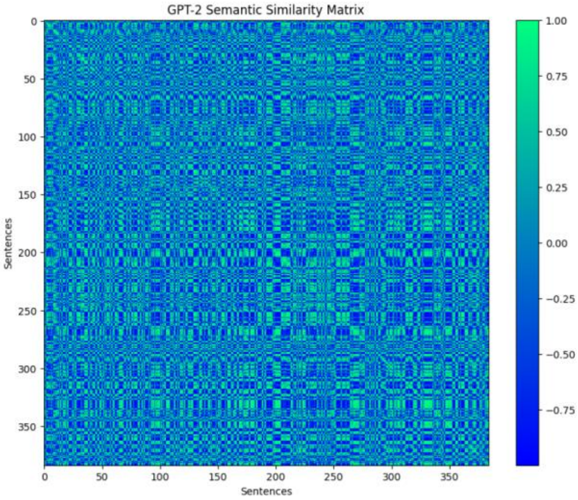

Fig.20: Similarity matrix – GPT-2