Punctuation's Semantic Role between Brain and Transformers Models

Zenon Lamprou[1][0000−0003−0042−5036], Frank Pollick[2][0000−0002−7212−4622], and Yashar Moshfeghi[1][0000−0003−4186−1088]

[1] NeuraSearch Laboratory, University of Strathclyde, Glasgow, UK
zenon.lamprou@strath.ac.uk, yashar.moshfeghi@strath.ac.uk
[2] University of Glasgow, Glasgow, UK
frank.pollick@gla.ac.uk

Abstract. Contemporary neural networks intended for natural language processing (NLP) are not designed with specific linguistic rules. It suggests that they may acquire a general understanding of language. This attribute has led to extensive research in deciphering their internal representations. A pioneering method involves an experimental setup using human brain data to explore if a translation between brain and neural network representations can be established. Since this technique emerged, more sophisticated NLP models have been developed. In our study, we apply this method to evaluate four new NLP models aiming to identify the one most compatible with brain activity. Additionally, to explore how the brain comprehends text semantically, we alter the text by removing punctuation in four different ways to understand its impact on semantic processing by the human brain. Our findings indicate that the RoBERTa model aligns best with brain activity, outperforming BERT in accuracy according to our metrics. Furthermore, for BERT, higher accuracy was noted when punctuation was excluded, and increased context length did not significantly diminish accuracy compared to the original results with punctuation.

Keywords: fMRI · Transformers · Explainable AI · Punctuation Symbols

1 Introduction

Continuous advancements in the area of natural language processing (NLP) have produced advanced deep learning models. How these models represent language internally is currently not very well understood because these models often tend to be large and complex. However, what is clear is that these deep learning models don't learn from any explicit rules regarding language, rather they learn generic information about language itself. Some of these models are trained on a large generic corpus and then fine-tuned in a downstream task. These models tend to perform better than models that are trained on just one specific NLP task [12,30,20].

There has been much effort to explain the inner representations of such models. Some of these efforts have focused on designing specific NLP tasks that target specific linguistic information [41,10,26].

Other research has tried to offer a more theoretical explanation regarding what word embeddings can represent [29,7,40]. Toneva and Whebe [37] in their effort to explain the inner representation of four NLP models proposed a novel approach that used brain recordings obtained from functional magnetic resonance (fMRI) and magnetoencephalography (MEG). They hypothesised that by aligning brain data with the extracted features of a model, the mapping from brain to model representation can be learned. This mapping primarily can then offer insights on how these models represent language. They refer to this approach as aligning neural network representation with brain activity.

One of the models tested by [37] was BERT. BERT, since its introduction [12] has generated a great deal of related research. Much of this work has been done to either improve its performance or to fix particular aspects of BERT or to fine-tune BERT to produce a version of BERT expert in a particular subject domain [2,23,19,33,42,24,9]. Toneva and Whebe [37] were able to learn a mapping from the representations of BERT, ELMo [30], USE [6] and T-XL [11] to brain data. However, this alignment was not tested against more sophisticated versions of BERT. We use the same novel approach as [37] to examine four of these more sophisticated versions of the BERT model: RoBERTa [42], DistiliBERT [33], ELECTRA [9] and ALBERT [24].

Moreover, because this novel approach is a proof of concept to show alignment of brain and neural networks we wanted to explore whether removing different punctuation symbols would influence results, potentially to produce a better alignment between brain and neural networks. Research investigating punctuation [27] has shown effects of punctuation on reading behaviour. Punctuation symbols can help the reader skim through text faster, but less is known about the semantic role of punctuation symbols and how they are processed in the brain. There is a substantial amount of research that examines how the brain processes text semantically and syntactically [31,5,34]. In addition there is research [1] that examined which brain areas are responsible for different operations, e.g. a brain area that is responsible for understanding the meaning of complex words. However, even if the goals of these research are different, their approaches share a common fMRI experimental approach, which involves participants reading a set of sentences. When sentences appear to the participants in these experiments, a choice is made whether or not to include punctuation in the pre-processed text. Though there appears to be no general agreement if the punctuation should be included or not.

Contributions: The contributions of this research are divided into two parts that we present in the subsequent sections. Firstly, we examined the degree to which the four different, BERT-related transformer models aligned to brain activity. Secondly, by using the concept of learning a mapping from brain to neural network introduced by [37] we try to use the neural network to obtain an hypothesis about the brain rather than using the brain to make observations

about the network. In particular by using altered sequences, in terms of punctuation, we hypothesise that if we can get better alignment then punctuation might be processed differently for the human brain to comprehend the semantics of language.

2   Related work

A frequent design used in fMRI experiments is to contrast brain activity in two conditions [16]. Often one condition is a specifically controlled baseline and the other condition reflects a process of interest. For example, to capture how the brain semantically processes information, the conditions are often a sentence vs a list of words [16]. However, some research [4,25,38,21,3] has used a narrative to capture how the brain semantically processes information. The work done by Toneva and Wehbe [37] was done in a non-controlled environment and the condition they were trying to capture was how semantically the brain processes text sequences. They used these text sequences to learn a mapping from neural model features to brain features. However this approach didn't take into account how punctuation influences semantic processing but rather how the brain processes the text sequence as a whole.

Previous work [38] managed to align MEG brain activity with a Recurrent Neural Network (RNN) and its word embedding, and that informed sentence comprehension by looking at each word. Other research demonstrated that one could align a Long Short-Term Memory (LSTM) [22] network with fMRI recordings to measure the amount of context in different brain regions. These results provided a proof of concept that one could meaningfully relate brain activity to machine learning models. Other approaches combining brain data and neural networks have tried to identify brain processing effort by using neural networks to learn a mapping from brain activity to neural network representations [15,18].

Previous research has explored the role of punctuation in natural language processing. From a natural language processing view, one of the standard text pre-processing steps is to remove punctuation symbols from the text to reduce the noise in the data. Some researchers have removed punctuation to achieve better results in spam detection [14], or plagiarism detection [8]. Such results have provided mixed results in the sense that in some cases removing punctuation can lead to better results, but not every time. Recent research [13] explored punctuation and differences in performance of MultiGenre Natural Language Inference (MNLI) using BERT and other RNN models. Their results showed that no model is capable of taking into account cases where punctuation is meaningful, and that punctuation is generally insignificant semantically.

From a neuroscience point of view there is limited research that assesses the role of punctuation in semantics and its purpose in brain processing of text. Related to the role of punctuation in the syntax of language, some research has tried to focus on creating multidimensional-features that are specific to syntactical parts of language [31]. Then using these features they tried to model the syntactical representation of punctuation symbols in the text. There is variabil-

ity in the literature in how punctuation is treated. For example, using the same fMRI data but with different research goals, one study preprocessed the text seen by the participants with punctuation [5] and the other without [34]. Furthermore, other research [1] tried to show that the combination of words can create more complex meaning and they tried to find which brain regions are responsible for representing the meaning of these words. In their pre-processing step they decided to remove the punctuation from text before presenting to their participants. This research further illustrates how there is no consensus on how punctuation is processed from the brain semantically and syntactically, and if it is needed to be included in a corpus to achieve better results.

Overall, it can be seen that there is not enough research that evaluates or improves NLP models through brain recordings as proposed by Toneva and Wehbe [37]. Though there is research examining cognition that evaluates whether word embeddings contain relevant semantics [35]. Furthermore, other research [17] tried to create new embeddings that align with brain recordings to asses if these new embeddings aligned better with behavioural measures of semantics. We believe that by using new state-of-the-art models to assess how their representations align with brain activity we can help expand this important and growing research area. We can unravel what training choices might lead to better alignment with brain data by assessing how the training choices of particular models improve or worsen this alignment.

3   Methodology

3.1   Data

The fMRI data were obtained originally as part of previously published research [38] and were made publicly available to use.[3] Their experimental setup involved 8 subjects reading chapter 9 from Harry Potter and the Sorcerer's stone [32] in its entirety. The data acquisition was done in 4 runs. Every word in the corpus appeared for 0.5s on the screen and a volume of brain activity represented by the BOLD signal would be captured every 2s (1 TR). Because of this timing arrangement, in each TR 4 words would be contained in each brain volume. These 4 words are not in random order, they are presented in the order they appear in the Harry Potter book. The data for every subject were preprocessed and smoothed by the original authors and the preprocessed version is the one available online. In their paper they note the use of MEG data as well but in this research we only focus using the fMRI since only these data have been made publicly available.

3.2   Transformer models

In our effort to find a more advanced transformer model that might produce better brain alignment we tested four different models against BERT, which we

---

[3] The data and the original code can be found at this link

used as our baseline. Each model was selected based on its new characteristics compared to BERT.

– **RoBERTa** RoBERTa was one of the first variations of BERT that came after its initial introduction. [42] stated that training neural networks is often expensive and its performance really relies on hyper-parameter choices. By comparing different techniques on the two main tasks that BERT was pre-trained on, Masked Language Modelling (MLM) and Next Sentence Prediction (NSP) they were able to produce a better performing version of BERT. The interest in using RoBERTa to try and understand its inner representation is to understand if choosing different hyper-parameters in a natural language experiment can lead to better representations that are closer to how the brain represents contextual information.

– **DistiliBERT** Transformer model architectures using NLP tend to be large and slow to fine-tune and pre-train. [33] in their pursuit to combat those problems proposed a BERT architecture that is reduced in size by 40% and is 60% percent faster to train compared to the original BERT model. Moreover during their experiments they managed to retain 97% of its language understanding capabilities. We thought that this design choice makes DistiliBERT interesting to test and show that indeed its internal representations are not affected by the reduction in size.

– **ALBERT** ALBERT was firstly introduced by [24] it tried to address the factor that modern day natural language processes are too big and we can easily reach memory limits when using a GPU or TPU. To combat that issue they proposed a two parameter reduction technique. The first technique they used is called factorised embedding parameterization. They hypothesised that by exposing the embedding space to a lower embedding space before exposing it to the hidden space of the model they can reduce the embedding space dimension without compromising the performance. The second technique that ALBERT implements is cross-layer parameter sharing. All the parameters in ALBERT are shared across all its layers. The features of the model in the original research [37] are extracted for each layer independently. The prediction and evaluation on the brain data is done layer by layer using the extracted features of each layer. We hypothesise that if ALBERT's features yield a better accuracy when used on predicting brain data, then this would suggest that the brain has a uniform weighting mechanism across its own layers.

– **ELECTRA** ELECTRA was developed by [9]. Their purpose for this model was to redefine the Masked Language Modelling (MLM) task, which was one of the two tasks that the original BERT model was pre-trained on. Because modern natural language models generally require a large amount of computational power to train, they proposed an alternative to the MLM task where instead of using the [MASK] token to mask a token in the sentence and have the model predict a possible value for that token they replace it with a proposed solution generated from a small generation network. Then the main model is pre-trained by trying to predict if the generated token is

indeed a correct one or if is not correct, similarly to a binary classification task. They hypothesised that this process results in a model with better contextual representations than BERT. This hypothesis makes this model a good candidate to assess its performance when aligned with brain data and determine if its inner representations are closer to how the brain represents contextual information.

3.3   Experimental Procedure

Our experimental procedure can be divided into three main operations. First we extracted the features from the desired models using different sequence lengths. Then we trained a ridge regression model using the extracted features and used the ridge regression to predict the brain data. Finally, we evaluated the predictions using searchlight classification. In the rest of this section we present each of these operations in detail.The code used to run the procedure that follows was made publicly available [4].

Extracting features from models The first step of the experimental procedure was to extract features from different sequences of length S. In the original research [37] the analysis was done using sequence lengths of 4, 5, 10, 15, 20, 25, 30, 35, 40 so we followed the same pattern. The sequences are formed from chapter 9 from Harry Potter and the Sorcerer's Stone book. A dictionary was created that contains an entry for every layer in the model, plus the embedding layer if it exists. The representations of the embeddings layer were extracted first for every word in the chapter. The next step was to get the first sequence of S length and extract the representation of every layer for that sequence S times. Then the representation of every sequence of S length was extracted using a sliding window the same way as the first sequence. The first sequence was extracted S times to ensure that the number of sequences was the same as the number of words in the dataset. A separate script was used for every model we chose. In order to use the pre-trained checkpoints of the models the hugging face library was used so we could download this checkpoints.

Making predictions using the fMRI recordings and the extracted features The second step was to train a ridge regression model to learn a mapping from brain representation to neural network representation. This approach followed some previous work [38,36,39,28,21] that learned a linear function with a ridge penalty to learn the mapping from brain to neural network representation. Although a ridge regression is a relatively simple model we chose to use it since it previously has been demonstrated to be useful [37]. However as a future work, new more complex models can be used with the hope of producing a better mapping. We used this ridge regression model to make predictions using the brain data and the features extracted from the previous step. The predictions were

---

[4]

done for every layer, for every subject and for every model in a cross validation setting. We used a 4-fold cross validation setting for the predictions which followed the same procedure followed by Toneva and Wehbe [37]. Because the fMRI data were gathered in four runs for every fold the fMRI data gathered in the corresponding run were treated as test data for that fold. First the extracted features from the models were loaded and using PCA were reduced to ten dimensions. The next step was to time align the corresponding fMRI data with the corresponding features extracted from the model so we knew the mapping between the features and brain data for the ridge regression to be trained on. A total of 1351 images were recorded for every subject. By removing the edges of every run, to combat the edge effect, the 1351 were reduced to 1211. In order to align the features with the fMRI data we find out in which TR each word was presented to the user. Using that mapping because the sequences are the same number as the words then the sequence in the same index as a word corresponds to the same TR. After the mapping from TR to sequence is complete the t-4,t-3,t-2,t-1 at any given time t were concatenated to t, because by concatenating representation in the previous four time points can give better representation to the model and yield better accuracy at the end [37]. After the previous steps were done we ended up with an array of 1351 values which are the sequences time-stamped. The last part was to remove the edges in every run, and that resulted in 1211 time-stamped sequences. Moving on, time stamping was done and the ridge regression model was trained. Different weights were obtained for every voxel using a set of different lambda values where

$$\lambda = 10^x where -9 \geq x \leq 9 \tag{1}$$

for the ridge regression model. Then for every voxel the lambda value that produced the lowest error score was used to construct the weights of the final model. Then the model generated from this procedure was used to make predictions on the test data.

Evaluating the predictions The final step was to evaluate predictions from the ridge regression model. To do that [37] used a searchlight classification algorithm to binary classify neighbours for every voxel, for every subject. The neighbourhoods have been pre-computed and were provided from the original research. We used the same approach and the same pre-computed neighbourhood for our evaluation method. We used the recorded brain data and a random chunk of 20 TRs was chosen and this corresponds to the correct predictions. From the predicted data the same chunk was chosen and then another random chunk of 20 TRs was chosen from the predicted data to be the wrong chunk. Then using the neighbourhoods the euclidean distance between the 2 chunks from the predicted chunk was calculated. If the distance from the correct chunk was less than the wrong chunk the prediction was marked as correct. This process was done 1000 times for every voxel and at the end the average accuracy for every voxel for every fold was obtained for every subject.

Removing punctuation To be able to assess how the human brain processes punctuation symbols semantically we used the same experimental procedure described above but altering slightly the first part. We used four different scenarios to remove the punctuation characters before inputting the sequences to the models. Then we extracted the features in the same manner as described above but as an input we used the altered text sequences. After the sequences were extracted then the same approach was followed. The four scenarios we tested were:

1. Replace fixation (”+”) symbol with the [”UNK”] token (Removing fixation). This scenario was used to assess if using the fixation symbol in the sequences as text can influence the quality of the extracted features and how brain aligned they are. The [”UNK”] token indicated in this scenario that this is an unknown token for the model.
2. Replace fixation symbol (”+”) with the [”PAD”] token (Padding fixation). This scenario was used to assess if using the fixation symbol in the sequences as text can influence the quality of the extracted features and how brain aligned they are. The [”PAD”] token indicated to the model that this symbol should be ignored when extracting the sequences.
3. Replace [”–”,”...”,”—”] with the [”PAD”] token (Padding all). When looking at the text from the Harry Potter chapter we identified in the text some special characters. We wanted to assess if these special characters can influence the performance when learning the brain-network mapping.
4. Replace [”–”,”...”,”—”,”.”,”?”] with the [”PAD”]token (Padding everything). Motivated from the previous scenario we extended the list of the special characters by adding ”.” and ”?” to the list of tokens to be ignored since they are one of the most commonly used punctuation symbols. Again we wanted to test how this alteration in sequences can affect the mapping.

4   Results and discussion

Comparing the models: We compared the performance of BERT against the models presented in Section 3.2. In this section we discuss the differences of these networks and what the results might show in terms of how the human brain functions. A common observation that can be seen in all the neural networks including BERT is that as the sequence length increases the accuracy of the ridge regression model decreases. All the results discussed below can be seen in Figure 1.

BERT Following the experimental procedure described in Section **??**, and using the same data, we were not able to exactly reproduce the same original results as [37]. Thus, our reproduced baseline, used to compare with the other models was obtained by our own code and is presented in Figure 2 along with the original results. We believe we did not have the same results for three reasons. Firstly, when averaging across subjects and across folds for every layer the exact procedure of the averaging did not appear to be completely described in the

(a) ALBERT

(b) RoBERTa

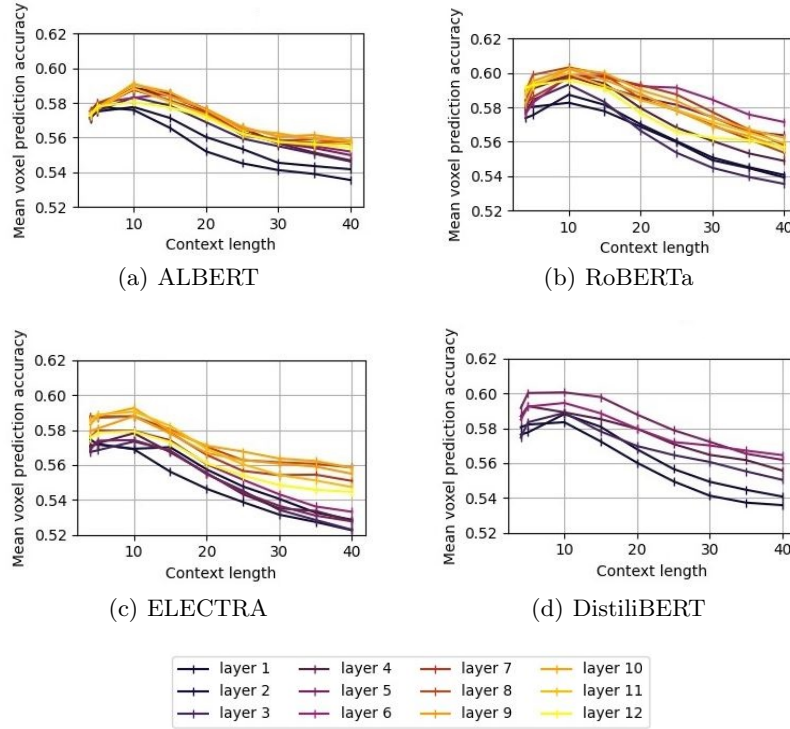(c) ELECTRA

(d) DistiliBERT

Fig. 1: This figure shows the overall accuracy across all subjects for the four different models we investigated.

original paper. Moreover when PCA is used the random state is not provided so we believe that this might be another reason for not replicating the exact same results. Finally our hardware is not the same as used by the original authors so this might also contribute to getting different results. However, we believe the results are accurate. By examining the results in Figure 2 which show the original results of [37] and our reproduced results we can see that qualitatively the overall shape of the graphs are the same but quantitatively there are some differences in the exact quantities. Similarities include that our reproduced results reaches its peak at sequence length 10, the same as the original results. Furthermore, as in the original results the accuracy decreases as the sequence length increases.

ALBERT The first model tested was ALBERT. The interesting characteristic to assess on ALBERT was that all the layers shared the same representations. By looking closer it can be seen that there are not any major differences in the accuracy. This would suggest that the changes made on ALBERT to distinguish it from BERT have not made its representations more brain-aligned than obtained with BERT. The fact that shared weights across the layers does not

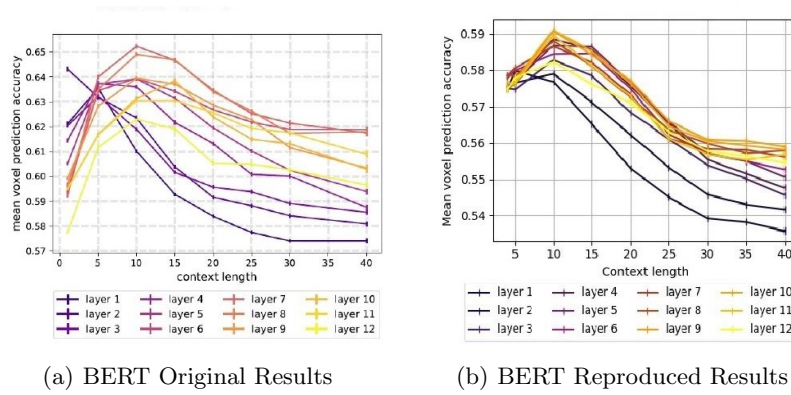(a) BERT Original Results   (b) BERT Reproduced Results

Fig. 2: On the left hand side are the original results reported in [37]. On the right hand side are the reproduced results we obtained when running our code for BERT. Note that the y-axis has a different minimum and maximum value for the two panels, original and reproduced

improve the alignment between brain and neural networks might suggest that the brain has a shared "weights" mechanism but this is something that needs to be investigated further.

RoBERTa The next model tested was RoBERTa. The same trend in the graphs that is present in BERT seems to be present when using RoBERTa. An immediate observation is that RoBERTa, in almost every layer, at its peak achieves a slightly better accuracy than BERT. Taking that into consideration indicates that the decision to choose different hyper parameters when training and choosing the best performing ones leads to inner representations that are slightly better aligned with those of the human brain.

DistiliBERT Moving on to examine the results with DistiliBERT shows that even though DistiliBERT has fewer layer, and is smaller in size than BERT the performance of its layers is better for all the layers compared to the layers of BERT. The increase is not so substantial but it indicates that the inner representation of the DistiliBERT can capture semantic information as good as BERT even though it is smaller in size. It also shows that layers of DistiliBERT are as brain aligned as the ones in BERT.

ELECTRA Looking at the results for ELECTRA we can see that ELECTRA doesn't surpass the baseline results obtained with BERT, even though the same trend in both graphs can be seen. From this result we can argue that the training choices when training ELECTRA have not made the model as brain aligned as BERT.

Results with Removing Punctuation: After using the corpus without any modifications, we wanted to modify the corpus by removing punctuation symbols. In doing so, we wanted to see what this might suggest of how the brain semantically processes punctuation symbols.
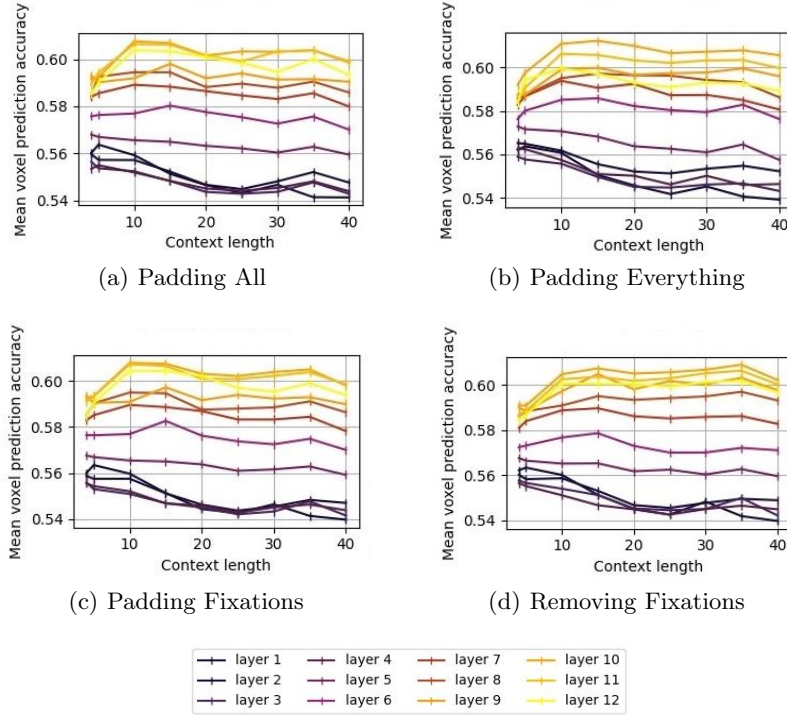

(a) Padding All   (b) Padding Everything

(c) Padding Fixations   (d) Removing Fixations

Fig. 3: The figures presents the results of the 4 different punctuation scenarios.

Our results showed that in all four punctuation-modification scenarios the accuracy increases only on layers 7-12. The increase is not always substantial, and the biggest boost in performance is almost 1.5%. The most interesting thing that can be seen is that removing the punctuation as padding tokens makes the model not lose as much accuracy as the length of the sequence increases. Moreover, layer 6 seems to be acting as a divisor between the starting and ending layers. This resonates with the observation of [37] that the first 6 layers are not as brain aligned as the last 6 layers and that removing the attention mechanism from the 6 first layers might result in better representations.

Taking into account the observations mentioned above and the observation of [37] that some layers of BERT are brain aligned, one can hypothesise that the brain might have limited use of punctuation to understand a sentence semantically. It can also be seen that as the sequence length becomes longer, which

means more information is presented to the model and brain, the performance does not drop as much, revealing the limited contribution of punctuation.

5   Conclusion

We have investigated four different models : ELECTRA, RoBERTa, ALBERT and DistiliBERT to determine which of the models, compared to a baseline obtained with BERT, can output language representations that are more closely related to brain representations of language as indicated by the alignment between model and fMRI brain data. The most aligned models appear to be RoBERTa and DistiliBERT as they outperform our baseline. In addition, using the four different scenarios we wanted to test how the semantic processing of punctuation symbols might vary. By looking at the work done by Toneva and Wehbe [37] and using their proof of concept that a mapping from the neural network features to the brain features can be learned, we tested if we can get a better mapping by removing these punctuation symbols. Indeed in all four cases the performance of the model increased. Also, the drop in accuracy when predicting brain data was not as big as originally obtained when the sequence length increased. These results support a view that the brain might have limited use for punctuation symbols to understand semantically a sentence and that as the context text length gets longer then there is less need for punctuation symbols. We believe that our research can be extended in the future for more transformer models to be analysed with the goal of finding a more brain-aligned model. We also believe that the experimental procedure created by Toneva and Wehbe [37] can be used to evaluate new models and produce models that are more brain aligned. Last, but not least we believe that the matter of how the brain processes punctuation symbols semantically should be investigated more in depth, using different models, in the quest of unravelling the capabilities of the human brain.

References

1. Acunzo, D.J., Low, D.M., Fairhall, S.L.: Deep neural networks reveal topic-level representations of sentences in medial prefrontal cortex, lateral anterior temporal lobe, precuneus, and angular gyrus. NeuroImage **251**, 119005 (May 2022). https://doi.org/10.1016/j.neuroimage.2022.119005
2. Beltagy, I., Lo, K., Cohan, A.: SciBERT: A Pretrained Language Model for Scientific Text. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 3615–3620. Association for Computational Linguistics, Hong Kong, China (Nov 2019). https://doi.org/10.18653/v1/D19-1371
3. Blank, I.A., Fedorenko, E.: Domain-General Brain Regions Do Not Track Linguistic Input as Closely as Language-Selective Regions. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience **37**(41), 9999–10011 (Oct 2017). https://doi.org/10.1523/JNEUROSCI.3642-16.2017

4. Brennan, J., Nir, Y., Hasson, U., Malach, R., Heeger, D.J., Pylkkänen, L.: Syntactic structure building in the anterior temporal lobe during natural story listening. Brain and Language **120**(2), 163–173 (Feb 2012). https://doi.org/10.1016/j.bandl.2010.04.002

5. Caucheteux, C., King, J.R.: Language processing in brains and deep neural networks: computational convergence and its limits. Tech. rep., bioRxiv (Jan 2021). https://doi.org/10.1101/2020.07.03.186288, section: New Results Type: article

6. Cer, D., Yang, Y., Kong, S.y., Hua, N., Limtiaco, N., St. John, R., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Strope, B., Kurzweil, R.: Universal sentence encoder for English. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 169–174. Association for Computational Linguistics, Brussels, Belgium (Nov 2018). https://doi.org/10.18653/v1/D18-2029

7. Chen, Y., Gilroy, S., Maletti, A., May, J., Knight, K.: Recurrent neural networks as weighted language recognizers. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 2261–2271. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018). https://doi.org/10.18653/v1/N18-1205

8. Chong, M., Specia, L., Mitkov, R.: Using natural language processing for automatic detection of plagiarism. In: Proceedings of the 4th International Plagiarism Conference (IPC-2010) (2010)

9. Clark, K., Luong, M.T., Le, Q.V., Manning, C.D.: ELECTRA: Pre-training text encoders as discriminators rather than generators

10. Conneau, A., Kruszewski, G., Lample, G., Barrault, L., Baroni, M.: What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2126–2136. Association for Computational Linguistics, Melbourne, Australia (Jul 2018). https://doi.org/10.18653/v1/P18-1198

11. Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q.V., Salakhutdinov, R.: Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context. arXiv:1901.02860 [cs, stat] (Jun 2019), arXiv: 1901.02860

12. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423

13. Ek, A., Bernardy, J.P., Chatzikyriakidis, S.: How does Punctuation Affect Neural Models in Natural Language Inference. In: Proceedings of the Probability and Meaning Conference (PaM 2020). pp. 109–116. Association for Computational Linguistics, Gothenburg (Jun 2020)

14. Etaiwi, W., Naymat, G.: The Impact of applying Different Preprocessing Steps on Review Spam Detection. Procedia Computer Science **113**, 273–279 (Jan 2017). https://doi.org/10.1016/j.procs.2017.08.368

15. Frank, S.L., Otten, L.J., Galli, G., Vigliocco, G.: The ERP response to the amount of information conveyed by words in sentences. Brain and Language **140**, 1–11 (Jan 2015). https://doi.org/10.1016/j.bandl.2014.10.006

16. Friederici, A.D.: The brain basis of language processing: from structure to function. Physiological Reviews **91**(4), 1357–1392 (Oct 2011). https://doi.org/10.1152/physrev.00006.2011

17. Fyshe, A., Talukdar, P.P., Murphy, B., Mitchell, T.M.: Interpretable semantic vectors from a joint model of brain- and text- based meaning. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 489–499. Association for Computational Linguistics, Baltimore, Maryland (Jun 2014). https://doi.org/10.3115/v1/P14-1046

18. Hale, J., Dyer, C., Kuncoro, A., Brennan, J.R.: Finding Syntax in Human Encephalography with Beam Search. arXiv:1806.04127 [cs] (Jun 2018), arXiv: 1806.04127

19. Hong, W., Ji, K., Liu, J., Wang, J., Chen, J., Chu, W.: GilBERT: Generative Vision-Language Pre-Training for Image-Text Retrieval. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1379–1388. SIGIR '21, Association for Computing Machinery, New York, NY, USA (Jul 2021). https://doi.org/10.1145/3404835.3462838

20. Howard, J., Ruder, S.: Universal language model fine-tuning for text classification. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 328–339. Association for Computational Linguistics, Melbourne, Australia (Jul 2018). https://doi.org/10.18653/v1/P18-1031

21. Huth, A.G., de Heer, W.A., Griffiths, T.L., Theunissen, F.E., Gallant, J.L.: Natural speech reveals the semantic maps that tile human cerebral cortex. Nature **532**(7600), 453–458 (Apr 2016). https://doi.org/10.1038/nature17637

22. Jain, S., Huth, A.G.: Incorporating Context into Language Encoding Models for fMRI. Tech. rep., bioRxiv (Nov 2018). https://doi.org/10.1101/327601, section: New Results Type: article

23. Jia, Q., Li, J., Zhang, Q., He, X., Zhu, J.: RMBERT: News Recommendation via Recurrent Reasoning Memory Network over BERT. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1773–1777. SIGIR '21, Association for Computing Machinery, New York, NY, USA (Jul 2021). https://doi.org/10.1145/3404835.3463234

24. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: ALBERT: A Lite BERT for Self-supervised Learning of Language Representations (Apr 2020)

25. Lerner, Y., Honey, C.J., Silbert, L.J., Hasson, U.: Topographic Mapping of a Hierarchy of Temporal Receptive Windows Using a Narrated Story. Journal of Neuroscience **31**(8), 2906–2915 (Feb 2011). https://doi.org/10.1523/JNEUROSCI.3684-10.2011, publisher: Society for Neuroscience Section: Articles

26. Linzen, T., Dupoux, E., Goldberg, Y.: Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. arXiv:1611.01368 [cs] (Nov 2016), arXiv: 1611.01368

27. Moore, N.: What's the point? The role of punctuation in realising information structure in written English. Functional Linguistics **3**(1), 6 (May 2016). https://doi.org/10.1186/s40554-016-0029-x

28. Nishimoto, S., Vu, A.T., Naselaris, T., Benjamini, Y., Yu, B., Gallant, J.L.: Reconstructing visual experiences from brain activity evoked by natural movies. Current Biology **21**(19), 1641–1646 (Oct 2011). https://doi.org/10.1016/j.cub.2011.08.031

29. Peng, H., Schwartz, R., Thomson, S., Smith, N.A.: Rational Recurrences. arXiv:1808.09357 [cs] (Aug 2018), arXiv: 1808.09357

30. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettle-moyer, L.: Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 2227–2237. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018). https://doi.org/10.18653/v1/N18-1202
31. Reddy, A.J., Wehbe, L.: Can fMRI reveal the representation of syntactic structure in the brain? In: Advances in Neural Information Processing Systems. vol. 34, pp. 9843–9856. Curran Associates, Inc. (2021)
32. Rowling, J.K.: Harry Potter and the Philosopher's Stone, vol. 1. Bloomsbury Publishing, London, 1 edn. (June 1997)
33. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter
34. Shain, C., Kean, H., Lipkin, B., Affourtit, J., Siegelman, M., Mollica, F., Fedorenko, E.: 'Constituent length' effects in fMRI do not provide evidence for abstract syntactic processing. preprint, Neuroscience (Nov 2021). https://doi.org/10.1101/2021.11.12.467812
35. Søgaard, A.: Evaluating word embeddings with fMRI and eye-tracking. In: Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP. pp. 116–121. Association for Computational Linguistics, Berlin, Germany (Aug 2016). https://doi.org/10.18653/v1/W16-2521
36. Sudre, G., Pomerleau, D., Palatucci, M., Wehbe, L., Fyshe, A., Salmelin, R., Mitchell, T.: Tracking neural coding of perceptual and semantic features of concrete nouns. NeuroImage **62**(1), 451–463 (Aug 2012). https://doi.org/10.1016/j.neuroimage.2012.04.048
37. Toneva, M., Wehbe, L.: Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In: Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019)
38. Wehbe, L., Murphy, B., Talukdar, P., Fyshe, A., Ramdas, A., Mitchell, T.: Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. PloS One **9**(11), e112575 (2014). https://doi.org/10.1371/journal.pone.0112575
39. Wehbe, L., Vaswani, A., Knight, K., Mitchell, T.: Aligning context-based statistical models of language with brain activity during reading. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 233–243. Association for Computational Linguistics, Doha, Qatar (Oct 2014). https://doi.org/10.3115/v1/D14-1030
40. Weiss, G., Goldberg, Y., Yahav, E.: On the Practical Computational Power of Finite Precision RNNs for Language Recognition. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 740–745. Association for Computational Linguistics, Melbourne, Australia (Jul 2018). https://doi.org/10.18653/v1/P18-2117
41. Zhu, X., Li, T., de Melo, G.: Exploring semantic properties of sentence embeddings. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 632–637. Association for Computational Linguistics, Melbourne, Australia (Jul 2018). https://doi.org/10.18653/v1/P18-2100
42. Zhuang, L., Wayne, L., Ya, S., Jun, Z.: A robustly optimized BERT pre-training approach with post-training. In: Proceedings of the 20th Chinese National Conference on Computational Linguistics. pp. 1218–1227. Chinese Information Processing Society of China, Huhhot, China (Aug 2021)