[dissemination] Integrating Cognitive Neuroscience Insights into NLP: A New Approach to Understanding Narrative Processing

Avital Hahamy[1,*], Haim Dubossarsky[2,3] and Timothy Behrens[1,4]

[1]*Wellcome Trust Centre for Neuroimaging, UCL Queen Square Institute of Neurology, University College London, 12 Queen Square, London WC1N 3AR, UK*

[2]*School of Electronic Engineering and Computer Science, Queen Mary University of London, Mile End Rd, London E1 4NS*

[3]*Language Technology Lab, University of Cambridge, 9 West Road Cambridge CB3 9DA, UK*

[4]*Wellcome Centre for Integrative Neuroimaging, University of Oxford, John Radcliffe Hospital, Oxford OX3 9DU, UK*

Abstract

This paper describes how a biological neural network comprehends narratives, with the goal of applying these insights to artificial neural networks. To this end, we present our findings, recently published in Nature Neuroscience, detailing a mechanism by which the human brain processes narratives. Our study utilized functional Magnetic Resonance Imaging (fMRI) to monitor brain activity in human participants as they were exposed to narratives. The human brain segments continuous narratives into discrete events that are represented by neural activity. Using a novel fMRI method and a Distributional Semantic Model, we revealed that whenever an event ends, the brain binds the representation of that event with the representations of contextually-relevant past event. This suggests that narrative comprehension is based on the continuous embedding of new events into the narrative context: newly-formed event representations are updated based on prior narrative events that are uploaded from memory. This paper not only summarizes our findings, but also advocates for interdisciplinary collaboration: we aim to inspire the incorporating of cognitive principles into NLP models, which has the potential to improve the way NLP models understand and process narratives.

Keywords

brain, movie, story, fMRI, cognitive, reactivation, events

1. Introduction

What is the cognitive function of narratives? Consider what you know of Albert Einstein. Your knowledge likely forms a narrative, linking pieces of information related to his life and work. Now consider what you did yesterday. This knowledge would again translate into a narrative linking the events of the day. This exemplifies that narratives are more than an efficient manner of transmitting information between people - they also organize our knowledge. In fact, narratives may reflect an important design principle of human intelligence.

*Corresponding author.

a.hahamy@ucl.ac.uk (A. Hahamy); h.dubossarsky@qmul.ac.uk (H. Dubossarsky); timothy.behrens@ndcn.ox.ac.uk (T. Behrens)

0000-0001-5862-851X (A. Hahamy); 0000-0002-2818-6113 (H. Dubossarsky); 0000-0003-0048-1177 (T. Behrens)

The study of narrative processing in the human brain has significant implications for artificial intelligence, particularly in the context of recent advancements in NLP. The design of language models has so far focused on the representation of individual words and the prediction of upcoming words. While these design principles have marked astonishing achievements in both simple language tasks and human-AI interactions, it remains questionable whether NLP models understand text in the same way humans do. Such understanding is tightly related to the ability to represent information in a narrative form, defined as events that are linked together across time in an inherent structure.

Here, we hope to inspire the NLP research of narratives through insights gained from studying the human brain. These insights, recently published in Nature Neuroscience[1], offer a computational framework by which a biological (or artificial) neural network could bind relevant information across time to subserve the understanding of narratives.

2. Related Literature

A narrative, such as this paper, comprises a continuous stream of words. Psychological literature suggests that the brain divides this continuous input into discrete units, called ”events”. This parcellation occurs whenever the brain recognizes a contextual change in the inputs, similar to a ”cut” between movie scenes. These transitions, that mark the end of one event and the beginning of another, are termed “event boundaries”[2]. For example, in reading this paper, an event boundary could occur at the end of each paragraph.

Neuroimaging studies show that each event is represented by a unique pattern of neural activity, termed brain representations[3, 4]. These brain representations can be thought of as a mosaic of activity levels across all “neuroimaging pixels” contained in a brain region (Fig. 1a). These representations are also stable throughout the duration of events in a set of brain regions termed the Default Mode Network[5, 6, 7]. These brain regions are suggested to hold an internal representation of the gist of events, that are stored into memory at each event boundary[8].

But to understand narratives, it is not enough to create a separate representation for each event. We also rely on the ability to interpret each event in light of relevant past events. For example, one cannot understand why Snow White wakes up when a piece of apple drops from her mouth, without connecting this part of the narrative to the part where she eats a poisoned apple. These two events need to be linked together, even though they are not adjacent in time (in between, the dwarfs come home and discover Snow White, then build a glass coffin, time passes and then the prince arrives). This suggests that event representations must incorporate the incoming event-related information with relevant, and possibly remote information stored in previous event representations.

What mechanism may underlie this linking between event representations? The results presented below suggest that at event boundaries, the brain reactivates past event representations that are relevant for understanding the current situation. This can be thought of as if at the end of each event, after all inputs have been acquired from the environment, the neural activity representing relevant past events is being uploaded from memory. This prior knowledge is then integrated with the newly formed event representation, thus supporting the understanding of an event within the context of its containing narrative.

3. Method

We tested whether the human brain reactivates past event representations at event boundaries. To this end, we used functional Magnetic Resonance Imaging (fMRI) datasets that measured brain activity in people watching a movie (n=17 participants)[3] or listening to a story (n=25 participants)[9]. This allowed us to test the inherent reproducibility of our results and their generalizability across different narrative experiences.

We defined event boundaries as the moments in time when scenes/paragraphs (hereafter, events) transitioned, and developed a new fMRI analysis method that allowed us to detect reactivation of past events during these event boundaries. In essence, for each region of the human brain, we measured the brain representations of whole events and of event boundaries (Fig. 1a). Reactivation entails the uploading of prior event representations from memory for updating the current event representation. To detect the expression of past event-representations at event boundaries, we cross-correlated the representation of event boundaries with the representations of events. As an estimate for reactivation, we looked for brain areas that show significantly more correlations between representations of event boundaries and representations of past events (lower triangular part of the similarity matrix) compared to future events (upper triangular part of the similarity matrix) (Fig. 1b).

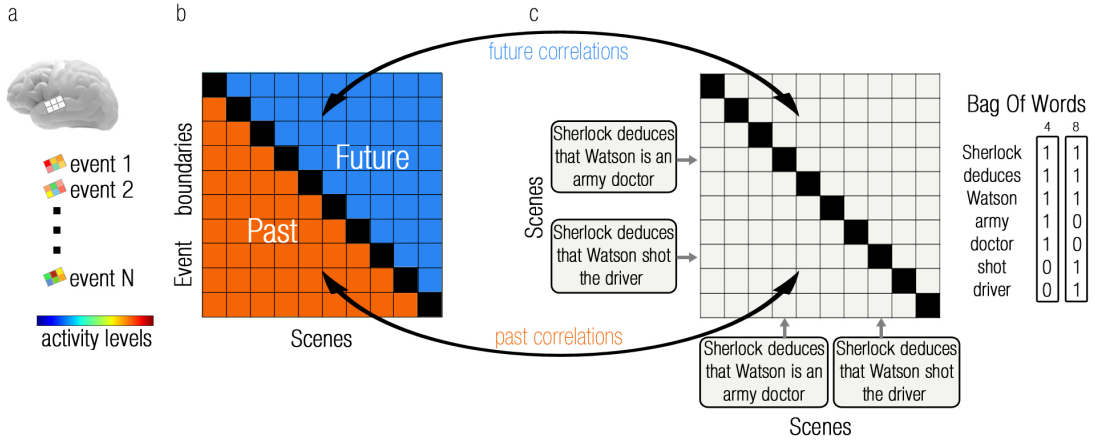

Figure 1: Method schematic. (a) For each brain area (example depicted as a white grid) we extracted representations for each event (depicted as colored grids) and each event boundary. (b) To detect reactivations in each brain area, we correlated the representations at event boundaries with event representations. The lower/upper triangular parts of the resulting similarity matrix held correlations between representations at event boundaries and their preceding/following events (highlighted in orange/blue), respectively. We looked for brain areas that showed more reactivations of past compared to future events by contrasting the two triangular parts of the matrix. (c) To test whether *relevant* past events are preferentially reactivated at event boundaries, we represented events using the Bag Of Words model (illustrated for scenes 4 and 8) and cross-correlated these event representations. We then correlated the past/future part of the neural similarity matrix with the past/future part of the context similarity matrix, respectively. Brain areas than show higher past compared to future correlations between these triangular parts would indicate a selective reactivation of past events.

We next aimed to determine if past events that are relevant for understanding the current

narrative stage are reactivated more than irrelevant ones. We therefore modelled the contextual similarity between all events in each narrative. We used a Bag of Words model (Fig. 1c) as a simple test for this hypothesis, under the assumption that events that share the same context would also have similar word cooccurrences (see Discussion in §5).

The inputs to the Bag of Words model were manual descriptions of scenes (including dialogue transcript) in the movie[10], and the text of the story. To this end, we created count-based vectors for each scene/paragraph in each narrative (excluding stop words), and transformed them into probability vectors. We then used the Jensen-Shannon distance[11, 12] to measure context similarities between each two cooccurrence-vectors.

This context-similarity matrix was correlated with the neural event-boundary X scene similarity matrix, to find brain areas that reactivate contextually-relevant events more that irrelevant events. To test if this context-dependent reactivation is specific for past events, we contrasted the past and future correlation coefficients.

4. Results

Using our novel fMRI method, we found that at event boundaries (whenever an event ends), regions of the Default Mode Network reactivate temporally-remote events (Fig. 2a).

Of these brain regions, the precuneus also reactivated events selectively. Our Bag of Word analysis demonstrated that in this region, the representations of events that were *relevant* to understanding the current event were reactivated more than irrelevant event representations (Fig. 2b).

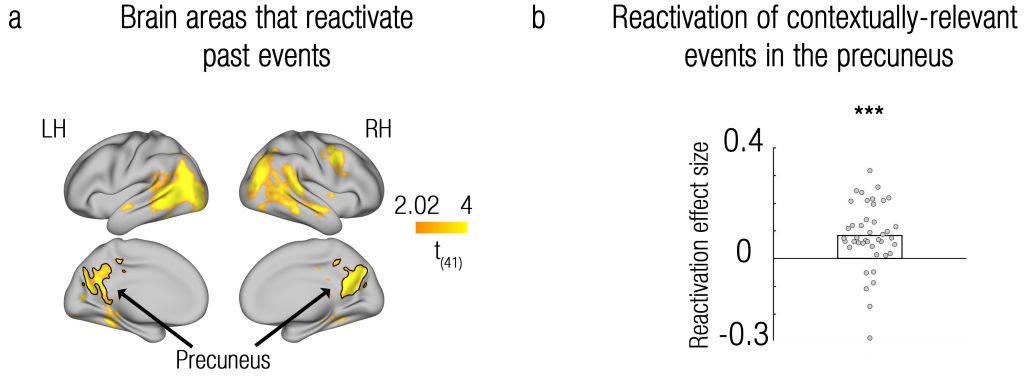

Figure 2: Reactivation of relevant past events at event boundaries. (a) Significant reactivation of past events was found in regions of the Default Mode Network (corrected for multiple comparisons across the entire brain). (b) Using a Bag of Words model, we found that in the precuneus (marked with arrows in (a)), events with semantic context similar to that of the current event were reactivated more than events with different semantic contexts. Single-participant reactivation effect sizes are represented as circles and the group mean is represented as a bar. These results demonstrate that reactivation in the precuneus integrates information that is specifically needed for the understanding of each current narrative stage. All results were replicated across the two datasets, but for simplicity, we present here results aggregated across the datasets. LH, left hemisphere; RH, right hemisphere; *** $p<0.001$.

5. Discussion

Our findings suggest a mechanism by which the human brain understands a narrative. By reactivating representations of past events, new information is integrated with prior knowledge, helping us interpret current events in light of our past experiences. These findings are based on a novel fMRI analysis method supplemented by the simplest language model available. The simplicity of the Bag of Words model also reflects its strength: it is assumption-free, parameter-free and requires no training. Furthermore, our choice of this model was also driven by the assumption that even the most advanced language model currently available cannot mimic the human understanding of narratives. As such ability is still under development, descriptive models, such as the Bag of Words, at least provide output that is easily interpretable at face value (i.e. similar word cooccurrences reflect similar contexts).

While this simple language model was ideally suited for our brain analyses, it is undisputed that this model is immensely inferior in its potential to emulate human cognition compared to state-of-the-art NLP models. In fact, we propose that our new mechanistic understanding of the way the human brain understands narratives could inspire the augmentation of advances language model with similar abilities. It is tempting to conjecture that a similar in-silico mechanism could be implemented, which would allow an automated parcellation of text into events, the drawing of relations between contextually-related events (even using the simple heuristic of similarity in word cooccurrences as a proxy for context similarity) and the updating of event representations based on prior knowledge. We envision that combining advanced language models with this narrative-understanding mechanism will lead to a leap in AI performance and enhanced ability to interact with humans.

Acknowledgments

AH was supported by the European Molecular Biology Organization non-stipendiary Long-Term Fellowship (848-2017), Human Frontier Science Program (LT000444/2018), Israeli National Postdoctoral Award Program for Advancing Women in Science, and the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 789040. HD was supported by the Blavatnik Postdoctoral Fellowship and the research program Change is Key! of the Riksbankens Jubileumsfond (M21-0021). TEJB was supported by a Wellcome collaborator award (214314/Z/18/Z), a Wellcome Trust Senior Research Fellowship (104765/Z/14/Z) and a Principal Research Fellowship (219525/Z/19/Z), together with a James S. McDonnell Foundation Award (JSMF220020372).

References

[1] A. Hahamy, H. Dubossarsky, T. E. Behrens, The human brain reactivates context-specific past information at event boundaries of naturalistic experiences, Nature neuroscience 26 (2023) 1080–1089.

[2] J. M. Zacks, N. K. Speer, K. M. Swallow, T. S. Braver, J. R. Reynolds, Event perception: A mind/brain perspective, Psychological bulletin 133 (2007) 273.

[3] J. Chen, Y. C. Leong, C. J. Honey, C. H. Yong, K. A. Norman, U. Hasson, Shared memories reveal shared structure in neural activity across individuals, Nature neuroscience 20 (2017) 115–125.
[4] A. Zadbood, S. Nastase, J. Chen, K. A. Norman, U. Hasson, Neural representations of naturalistic events are updated as our understanding of the past changes, eLife 11 (2022).
[5] D. Stawarczyk, M. A. Bezdek, J. M. Zacks, Event representations and predictive processing: The role of the midline default network core, Topics in Cognitive Science 13 (2021) 164–186.
[6] A. Ben-Yakov, N. Eshel, Y. Dudai, Hippocampal immediate poststimulus activity in the encoding of consecutive naturalistic episodes., Journal of Experimental Psychology: General 142 (2013) 1255.
[7] C. M. Bird, How do we remember events?, Current Opinion in Behavioral Sciences 32 (2020) 120–125.
[8] G. A. Radvansky, J. M. Zacks, Event perception, Wiley Interdisciplinary Reviews: Cognitive Science 2 (2011) 608–620.
[9] C. H. Chang, C. Lazaridi, Y. Yeshurun, K. A. Norman, U. Hasson, Relating the past with the present: Information integration and segregation during ongoing narrative processing, Journal of Cognitive Neuroscience 33 (2021) 1106–1128.
[10] H. Lee, J. Chen, Narratives as networks: Predicting memory from the structure of naturalistic events, BioRxiv (2021) 441287.
[11] D. M. Endres, J. E. Schindelin, A new metric for probability distributions, IEEE Transactions on Information theory 49 (2003) 1858–1860.
[12] F. Osterreicher, I. Vajda, A new class of metric divergences on probability spaces and its applicability in statistics, Annals of the Institute of Statistical Mathematics 55 (2003) 639–653.