Event similarity and word-level salience predict how humans summarize information from complex naturalistic narratives

Claire Sun (csun28@ur.rochester.edu)
Department of Brain and Cognitive Sciences
University of Rochester
500 Wilson Blvd, Rochester, NY 14611

Coraline Rinn Iordan (mci@rochester.edu)
Department of Brain and Cognitive Sciences and Department of Neuroscience
University of Rochester
500 Wilson Blvd, Rochester, NY 14611

Abstract

We continuously encounter multitude sources of complex, multisensory information. To navigate this deluge, we summarize the contents of our experiences into manageable chunks that help us encode them into memory. Here, we seek to identify the features of complex narratives that predict how information will be summarized, and thereby provide insights into the cognitive mechanisms of summarization. We show that (1) transformer models specifically trained for document summarization generate internal features that are relevant for how humans summarize information from similar content; and (2) that the process of summarizing complex narratives can be described partially through its interaction with how narratives are segmented into individual events.

Keywords: summarization; transformer models, LLMs, memory, naturalistic stimuli, event perception

Introduction

Summarization is necessary for effective communication of complex ideas. As such, this process is ubiquitous: every scientific paper has an abstract, every book has a CliffsNotes, and every movie has a synopsis. We constantly summarize daily experiences for ourselves (Jeunehomme & D'Argembeau, 2020) and for sharing with others (Michelmann, Staresina, Bowman, et al., 2019). Recent advances in transformer neural networks (Raffel et al., 2020) approach human-level performance in summarizing longform content (Pilault, Li, Subramanian, & Pal, 2020). Yet, it remains unclear how this task is achieved in human cognition (Musz & Chen, 2022). Here, we sought to identify the features of complex naturalistic narratives that explain how humans summarize them. We hypothesized that, to successfully summarize stories, transformer models must learn a measure of salience for which concepts in the (long form) narratives should be included in a (short form) summary and that this salience measure may also capture human summarization behavior. To test this, we identified the cross-attention (encoder-decoder) layer of a state-of-the-art summarization transformer (T5) as a potential salience measure for individual concepts (tokens) and used it to predict the contents of human-generated summaries. We also asked whether this salience measure can predict which events from a narrative participants decide to include in a summary (Zacks, 2020). Finally, to test how these measures generalize to very brief (constrained) summaries, we replicated our analyses on an additional pilot dataset of succinct summaries.

Materials and Methods

Participants and Procedure

We used the open-source recall/summarization dataset (Musz & Chen, 2022) which comprises 17 participants who viewed an episode of the TV show *Sherlock* and recalled its plot. This dataset was one of the few to provide annotations for which subset of the participants' recall utterances are a summary of the corresponding plot vs. a confabulation, etc. Our pilot summarization dataset comprised 4 adults (all female) from the Rochester community who were asked to view the episode and provide a succinct (300w) summary.

Transformer Architecture and Model Features

We used an open-source implementation of the Long T5 transformer model trained to summarize long sequence data (Guo et al., 2022) and fine-tuned for written narratives using the BookSum dataset (Kryściński, Rajani, Agarwal, Xiong, & Radev, 2021). To investigate whether the model could predict which concepts and actions participants may include in their summaries, we investigated the model cross-attention (encoder-decoder attention) layer, which represents the contextual information shared between individual tokens captured by multiple (12) separate attention heads. Activation profiles in each head are taken as a measure of token saliency.

Event Boundaries and Segmentation

We independently segmented the *Sherlock* episode script using the 48 event boundaries provided in the open-source dataset. Event boundaries were defined as scene boundaries that followed major narrative shifts. Participant recalls/summaries, both from the open-source dataset and our in-lab pilot experiment, were similarly segmented into events.

Results

Narrative token salience predicts summary tokens

To investigate which tokens from the episode script participants include in their summaries, the (event-separated) movie script was preprocessed to remove capitalization, text formatting, punctuation, and to only contain nouns and verbs (i.e., concepts and actions). The resulting text was then processed by the transformer to obtain a matrix of attention/salience values (one value per token across the 12 attention heads of the cross-attention layer). For each head and event, we computed the Pearson correlation between the token salience values and a binary map of the tokens' presence in each participant's summary of that event (Fig. 1).

We found that salience from multiple attention heads (1,2,6,10,11) predicted participants' retelling of information from the episode, regardless of summarization constraints (detailed recall in (Musz & Chen, 2022), "Recall" $p<0.001$; summary subset of recall in the latter, "Summary" $p<0.001$; and our in-lab constrained summaries, "Constrained Summary" $p<0.003$). Additionally, salience from multiple heads (2,7,9) better predicted the more stringent summarization conditions (Summary vs. Recall: $p<0.012$; Constrained Summary vs. Recall, $p<0.001$). This suggests that transformer salience is a useful and predictive measure for characterizing which concepts people recall and which concepts they choose to include in a summary of an event.

Narrative event salience predicts summary events

To further investigate the relationship between the salience measure and event-related summarization of naturalistic nar-

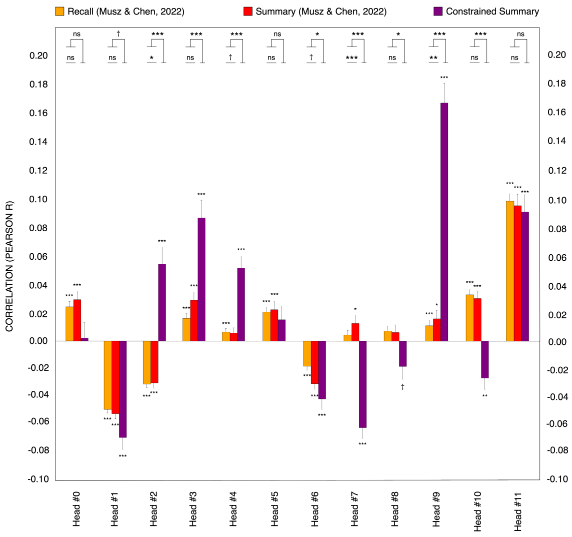

Figure 1: Correlation between salience and summary tokens

ratives, we asked whether stronger attention to particular events can predict which events participants include in their summaries. We computed the Pearson correlation between the maximum token attention weight across each event and a binary map of the event's presence in the summary (Fig. 2).

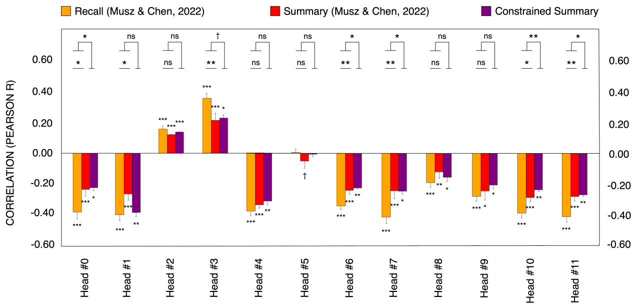

Figure 2: Correlation between salience and summary events

We found that salience from almost all heads (1—4, 6—11) predicted the events included in participants' retelling of the original narrative, regardless of summarization constraints (Recall p<0.001; Summary p<0.003; Constrained Summary p<0.048). Additionally, salience from multiple heads (0,6,7,10) predicted events included in the more stringent summarization conditions (Summary vs. Recall: p=0.022; Constrained Summary vs. Recall, p<0.049). This suggests that transformer salience is a useful measure for characterizing which events people recall and include in their summaries.

Narrative event structure predicts summary events

To further investigate the link between event perception and summarization, we used the Universal Sentence Embedding (USE) package to obtain a semantic vector embedding for each event in the episode script. We then used k-means clustering to identify groups of similar events in the narrative (Fig. 3A shows event groups plotted along the top two PCA dimensions of the embedding space; a hierarchical clustering analysis identified n=5 as the optimal number of event groups).

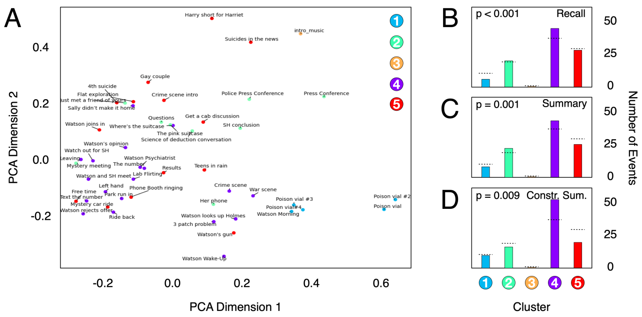

Figure 3: Event structure predicts summary events

Finally, we measured, for each participant's Recall, Summary, or Constrained Summary, whether they had a systematic bias toward a particular event group. Across all conditions, participants showed a consistent bias towards disproportionately including events from the 4th cluster (Fig. 3B,C,D; Chi-square tests: Recall p=0.001; Summary p=0.001; Constrained Summary p=0.009), a cluster that we assessed to comprise the essential (core) events of the narrative.

Discussion

In this study, we showed that transformer-based models are able to learn internal representations that predict how humans summarize complex narratives, both at the concept level and at the event level. Specifically, we identified a salience measure for concepts, actions, and events in a narrative that is directly related to what type of information is recalled later on and/or incorporated in a summary of that narrative. Our work provides evidence that the individual events of a narrative may constitute a key structural element that interact with the process of information summarization. This is further evidenced by the fact that transformer models that are not given any prior knowledge of narrative event structure during training, manage to learn feature representations (e.g., attention/salience) that generate significant predictions that interact with the event-level of a narrative.

In summary, our study constitutes a preliminary step towards identifying the behavioral mechanisms underlying summarization and how they interact other high-level cognitive processes (e.g., event perception). In future work, we aim to investigate the neural mechanisms of this process, i.e., how constructing a summary of a narrative under various constraints (e.g., time, length) unfolds and interacts with event perception across multiple human brain regions and networks (Baldassano, Hasson, & Norman, 2018) during naturalistic perception.

References

Baldassano, C., Hasson, U., & Norman, K. A. (2018). Representation of real-world event schemas during narrative perception. *The Journal of Neuroscience*, *38*(45), 9689–9699.

Guo, M., Ainslie, J., Uthus, D., Ontanon, S., Ni, J., Sung, Y.-H., & Yang, Y. (2022, July). LongT5: Efficient text-to-text transformer for long sequences. In M. Carpuat, M.-C. de Marneffe, & I. V. Meza Ruiz (Eds.), *Findings of the association for computational linguistics: Naacl 2022* (pp. 724–736). Seattle, United States: Association for Computational Linguistics.

Jeunehomme, O., & D'Argembeau, A. (2020). Event segmentation and the temporal compression of experience in episodic memory. *Psychological Research*, *84*(2), 481–490.

Kryściński, W., Rajani, N., Agarwal, D., Xiong, C., & Radev, D. (2021). Booksum: A collection of datasets for long-form narrative summarization. *arXiv preprint arXiv:2105.08209*.

Michelmann, S., Staresina, B. P., Bowman, H., et al. (2019). Speed of time-compressed forward replay flexibly changes in human episodic memory. *Nature Human Behaviour*, *3*, 143–154.

Musz, E., & Chen, J. (2022). Neural signatures associated with temporal compression in the verbal retelling of past events. *Communications Biology*, *5*(1), 489.

Pilault, J., Li, R., Subramanian, S., & Pal, C. (2020, November). On extractive and abstractive neural document summarization with transformer language models. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 9308–9319). Online: Association for Computational Linguistics.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*(140), 1–67.

Zacks, J. M. (2020). Event perception and memory [Journal Article]. *Annual Review of Psychology*, *71*(Volume 71, 2020), 165-191.