Differential contributions of episodic and semantic memory to story-telling

Saumya Didwania
Center for Data Science
New York University
sd4469@nyu.edu

Dhruv Saxena
Center for Data Science
New York University
ds6802@nyu.edu

Bogeng Song
Psychology department
New York University
bs4283@nyu.edu

Abstract

Our memory is an integral centrepiece in the process of storytelling however, our intuition leads us to believe that we use different types of memory and memory retrieval for different types of stories. To further examine the relationships, we built a framework of different computational models to better understand the cognitive processes that people use while constructing a type of story.

Our primary dataset, hippoCorpusV2, contains 6,854 diary-like short stories individually labeled into three categories: recall, imagine and retold. We used relevant Machine Learning models and techniques to classify these three types of stories, and extract the corresponding features to compare with our conclusions from human behavior experiments to better understand people's cognitive processes.

We found that our model results and behavioral results are similar, and there are three main characteristics that help us distinguish the three types of stories: the time it takes to build the story, the amount of concrete, specific events mentioned in a story, and detailed, sensory information providing background color to the story. From these results, we can infer that recall stories are based on the direct retrieval of episodic memory, while imagined stories are mainly generated based on general knowledge of semantic memory. While retold stories do retain some details in episodic memory, they also require general knowledge due to the inherent human tendency to forget trivial details.

1. Introduction

Nearly 50 years ago, Tulving [13] proposed that there were two types of memory: semantic and episodic. Semantic memory is a kind of organized knowledge that a person possesses about words or other verbal symbols and their meanings, referents, and relations. Episodic memory, on the other hand, consists of events and episodes that have already occurred. For a long time, researchers segmented these two types of memory as completely disparate systems in humans, but more recently, researchers found that these two types of memory may have more overlap. Researchers discovered that the hippocampus and surrounding medial temporal lobe (MTL) structures, play an important role in both semantic and episodic memory [9] [10] [6], and these results suggest that when humans encode and retrieve specific events, the two types of memory depend on each other and together influence human behavior and performance.

When humans need to recall an event and describe it, although episodic memory is mainly responsible for retrieving relevant event information, some studies have shown that semantic memory is also involved in constructing relevant events. [7]. On the other hand, when we need to imagine a story from a brief description, we need not only general knowledge of semantic memory to construct stories but also episodic memory to extract similar experiences to better describe the story [5]. In addition, when we recall the relevant event again after a lapsed period, our exact memory will be forgotten and changed, and the event we re-describe will be different from the previous description. Recently, some studies have shown that forgetting is a kind of hierarchical representation, and while our low-level details will be preserved, the high-level narrative flow will be forgotten first [3].

Although both kinds of memories play roles in recalling, imagining, and retelling stories, there are still differences between the three types of stories. We can better understand the types of memory and the corresponding cognitive processes that humans use when constructing stories by comparing the similarities and differences between the three kinds of stories and their language content.Additionally, to better understand the effects of forgetting, we can compare three different story types at the same time to try to understand where the differences between our cognitive processes are reflected when people describe different types of stories.

From our review of psychological papers, there were three features that seemed most relevant to understanding the differences between recalling, imagining, and retelling an event: time, concrete events, and sensory details.

1. **Time**: While time since the event being described is important, the time to build the story is also significant [1]. Studies have shown that the time required for direct retrieval of episodic memory is significantly lower than the time required to generate retrieval with general knowledge [1] [12]. For recalled stories, humans will mainly rely on episodic memory as it's easily accessed. This easily accessible memory also allows for humans to recall related events quickly, and supplement their story with necessary details. For retold stories, where related events are recalled again in the future, there will be an additional parameter of "forgetfulness." This effect causes a subject to use semantic memory to supplement the fragmented details of a story, and causes construction time of the story to slightly increase. For the imagined story, especially since a subject does not have direct experience with an event, the subject will rely much more on semantic memory and use their innate knowledge to build a narrative. The time required to retrieve semantic information is the highest of the three groups of storytelling.

2. **Concrete events**: For concrete details such as setting and names, studies have shown that when recalling a past event, we already have an idea of the background and can more easily extract related information [8]. Similarly when retelling an event later on, the latest research [3] has indicated that we forget high-level details first but relative details are retained in memory longer, causing relatively specific details to stay top of mind. Since individuals imaging a story may not understand the background behind an event, they may have more trouble including specific details compared to someone either recalling or retelling a story.

3. **Sensory details**: The final main feature deals with sensory and emotional details but this segment is still undergoing research and conclusions are mixed [8] [2]. When we try to recall an event, the subjects can effectively recall the subjective feelings and emotions, whether postive or negative, at that time. As the time interval from the event increases, the recall of such feelings and emotions will gradually decrease. In the imaginary story, fewer of these subjective feelings and emotions are included since the subjects did not experience those events and related sensory details themselves.

We assume that when we use the model for multi-class classification, the important features of the classification will be consistent with the features we found in the literature review, and we can further examine which additional features are important in distinguishing differences between the three types stories through the results of the model.

2. Method

Our approach for this project was to supplement the initial features of our dataset to incorporate our research on papers about similar topics. After creating proxy variables for the text based details, we proceeded to run multiple multi-class classification algorithms to help segment the three groups and predict the types of stories from both numerical and textual features.

2.1. Dataset

The dataset we used is called hippoCorpusV2 and was created by Sap et al in 2020 [11]. The dataset consists of 6,854 diary-like short stories about salient life events that are categorized into three types: recalled, imagined, and retold. For the recalled stories, participants are asked to recall and write an experience that happened to them in the past six months and then also write a short summary of their event. For the imagined stories, subjects were randomly given a short summary written by a 'recalled story' participant and then told to write a story describing the event as if it had happened to them. The subjects were told to not describe an experience directly related to themselves but rather imagine as if it were to happen. The third group, retold, consisted of participants from the first group that were asked to return for a follow up study. This group was given their original summary and were told to rewrite a story based on that prompt.

Regardless of the participant type, after writing the story, each participant filled out a questionnaire that would include questions to understand their current mental state. They would have to rate themselves on a 5-point Likert scale around topics such as how distracted they were, how draining the task was, etc. The dataset also consists of information around demographics such as age, race, and gender, as well as time information about how long the entire writing process took and the number of days since the event occurred.

The following are snippets from each type of story. As we can see, the Recall and Retold stories seem to have many more specific details and nouns compared to the Imagine story.

1. **Recall story**: *"Then we went to the insect section of the zoo. There were so many of them like spiders, scorpions, and bigger animals like cobras, rattlesnakes' stuff like that. Then we saw the elephants."*

2. **Imagine story**: *"We got to the zoo early before many of the large groups came and grabbed a snack from the snack bar so we could get a head start on seeing our favorite animals."*

3. **Retold story**: *"When we walked in there was the insect/reptile section of the zoo, so we saw many exotic looking spiders, snakes, scorpions etc. I liked it but my girlfriend didn't like insects."*

2.2. Data Preparation

In the initial experiment, less than half of the participants came back to retell their story so in the dataset, we had an uneven number between the groups. To remedy this, we explored both bootstrapping the data to increase the overall dataset and randomly upsampling the 'retold' participants to even out the sample size. Due to the disproportionate classes, we decided on upsampling the 'retold' participants to create balanced classes. We increased the 'retold' participants from 1,319 to 2,779, matching the 'recall' class as shown in the figure below.

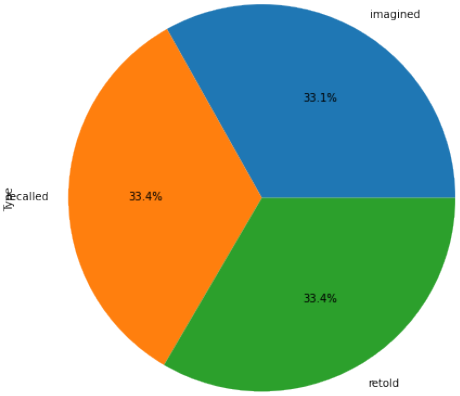

Figure 1. Pie chart showing percentage of memory types after upsampling recall numbers.

Many of the variables collected from the dataset were categorical such as gender and ethnicity in which we grouped together the low frequency items and one-hot encoded the grouped items. To normalize the scale of certain variables such as 'WorkTimeInSeconds', we log transformed the data to make sure that one variable wouldn't greatly affect our results. The TimeSinceEvent variable had already been log transformed for us so we were able to leave that as included.

As the original paper and many psychological papers talked about the importance of language in storytelling, we created a few variables to incorporate text into the models. As mentioned in the introduction, the more detailed and specific the story, the higher likelihood it was to be a recalled story than the other two categories. We created variables to count the number of nouns, adjectives, and verbs, and also created a variable for counting specific named entities. While we weren't able to incorporate any advanced Natural Language Processing to help process the text and understand the specific details, the basic count variables were good proxies to get a base understanding of how textual details may influence the type of story. This procedure was inspired from a narrative ability assessment task, called the Narrative Assessment Protocol. [4]

We hypothesized that these text features and the duration of the subject's recall are the features that contribute the most to the model's classification, and hoped to see how significant those features would be in our classification models to better understand the overall cognitive process.

3. Results

Our first step was to perform descriptive statistics on each variable. For the textual features we were most interested in, we noticed that in the recall story, the stories contained the most nouns, verbs and adjectives. The retold stories contained the second highest concentration of each of those parts of speech and the imagined story contained the least number of nouns, verbs and adjectives.

Memory Type	Nouns	Verbs	Adjectives
Recalled	53.68	26.36	17.45
Retold	51.09	25.63	16.70
Imagined	45.87	24.04	15.05

Table 1. Parts of Speech by Story.

After creating all our variables, we split our data into a training set and testing set. While we would have liked to also create a validation set, we felt that with a limited amount of data, just the two sets would suffice. Since this was a multi-class classification problem, we attempted to supplement the initial models with the OnevsRest Classifier. The OnevsRest classifier allows us to compare each category against the other two categories. Since this problem only had 3 classes, OvR didn't add much computation time to any of our initial models. After testing, the best model was XG Boost which is common with most machine learning competitions. While the XG Boost model does sacrifice some interpretability, it was much more accurate than all the other simpler classification models without much additional training time due to the smaller dataset. Our results are shown in Table 2

Model	Test Accuracy
Logistic Regression	61.2%
Logistic Regression (One vs Rest)	62.0%
SVM	60.9%
SVM (One vs Rest)	60.4%
Naive Bayes	51.6%
Naive Bayes (One vs Rest)	50.4%
KNN	49.9%
KNN (One vs Rest)	48.6%
Random Forest	62.7%
Random Forest (One vs Rest)	67.9%
XG Boost	76.4%

Table 2. Summary of results.

As predicted from the research papers, we found that the time in seconds to complete the task and the number of nouns in the story were the most important features in predicting the type of story as shown in Figure 2.

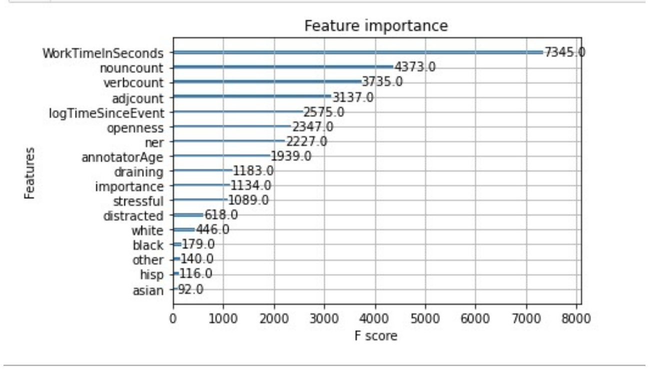

Figure 2. F-scores of the most important features from the best performing model : XGBoost

4. XGBoost

4.1. Hyperparameter Tuning

Once we found XGBoost had a better performance than all the other models/algorithms, we spent most of our efforts on finding the best hyperparameters for XGBoost. Rather than picking random ranges for each parameter, we decided to incorporate sophisticated techniques with the help of HYPEROPT library in python. We shortlisted certain parameters which we believed would influence this model the most, as shown in Table 3. We combined all of these hyperparameters in a space and generated an objective function for evaluating XGBoost's predictions.For each hyperparameter in the space, a uniform random generator was used to generate a value within the range provided. Both the space and objective were given to the fmin() function to run for 100 iterations of different hyperparameter combinations and return the best possible accuracy.

Hyper-parameter	Range	Optimal
max_depth	(3, 18)	8
gamma	(1,9)	3.845
reg_alpha	(40,180)	75
reg_lambda	(0,1)	0.538
colsample_bytree	(0.5,1)	0.759
min_child_weight	(0,10)	3
n_estimators	(300,500)	437

Table 3. Hyperparameter ranges and optimal value.

4.2. XGBoost and human cognition

XGBoost is a highly effective, scalable gradient boosting tree algorithm, and is widely used in many Machine Learning competitions worldwide for predicting and forecasting based on tabular data. However, looking at our results, we can make very interesting correlations between the underlying principle behind (gradient) boosting and human thinking process.

Boosting consists of a sequential process where the outputs of a decision tree are evaluated by simpler trees/functions. An ensemble is added together to make the final prediction. One of the more surprising aspects of the boosting process is the use of weak or base learners, which classify at an accuracy only slightly better than chance. For example, a weak learner used in boosting is called a decision stump, which is effectively a tree with the capability of only one split. Despite the simplicity of its underlying processes, we can see that it produced results which were significantly higher than the ones from more established methods.

A parallel can be drawn between boosting and human decision making. We often make attempts to improve our own memory processes, by which we attempt to improve some aspect of our lives governed by our own cognitions. But instead of making grand, radical changes in our habits and approaches, sometimes its the simple, consistent efforts that we put in, consciously or sub-consciously that prove to be the most effective.

5. Discussion

The experimental results show that most of the off-the-shelf models did well and some, such as Random Forest did extremely well. With random guessing, since our dataset was split evening, we'd expect a 33% accuracy just by guessing one class every time. With a more complicated model such as the XG Boost one, we hit over 75 % accuracy, indicating that there are indeed some significant variables to differentiate between the three-story types. Looking at our feature importance chart, we found that the time the subjects spent describing stories was the most important feature in distinguishing the three story types, consistent with our literature review. Additionally, we can see that the number of nouns, verbs, and adjectives all have relatively high feature importance compared to the initial features included in our dataset. Interestingly, more specific nouns that would be counted by named entity recognition are not as important as just the general number of nouns. The use of nouns indicate a story with more detailed descriptions which is what current research assumes about recalled stories. While we found that the usage of verbs and adjectives are important in the model, there was very little difference between them according to table 1

The results of our model suggest that when we need to recall an event, we directly extract episodic memory to describe this event, and enrich our story with more details. As the time since the event increases, our episodic memory is gradually forgotten, and this forgetting leads to spending more time on writing the stories and including specific details around the event. For imagined stories that we have never experienced, we rely more on semantic memory and use more general verbiage to mask the need for specific details.

5.1. Future Work

Although we achieved a fairly high accuracy rate of 76.4%, we outlined a few things we would want to include in order to potentially increase the accuracy even further. We had to use proxy variables for our Natural Language Processing tasks and could benefit from some advanced algorithms. Instead of counting the number of nouns, verbs, and adjectives, we could explore more specific text features such as the fluency of sentences, the use of metaphors, and specific event recognition. We could use implement deeper text language processing models such as GPT or Bert to extract additional features and improve accuracy. While we stuck to more off-the-shelf machine learning algorithms, we could also try and implement neural networks to find some underlying connections in the data that we can't see on the surface of our features. While we had almost 7,000 rows of data, this project could have also been extended by using supplemental, similar datasets. We could try and test our model around news articles to see if we can outline differences between fact-based articles and those that consist more of an author's opinion.

The subject of understanding human memory has been an open field of exploration in the world of cognitive science and while our results can provide some insights into this topic, we hope to further explore this issue from the perspectives of both cognitive experiments and model development.

References

[1] Donna Rose Addis, Katie Knapp, Reece P Roberts, and Daniel L Schacter. Routes to the past: neural substrates of direct and generative autobiographical memory retrieval. *Neuroimage*, 59(3):2908–2922, 2012. 2

[2] Donna Rose Addis, Ling Pan, Mai-Anh Vu, Noa Laiser, and Daniel L Schacter. Constructive episodic simulation of the future and the past: Distinct subsystems of a core brain network mediate imagining and remembering. *Neuropsychologia*, 47(11):2222–2238, 2009. 2

[3] Nora Andermane, Bárur H Joensen, and Aidan J Horner. Forgetting across a hierarchy of episodic representations. *Current opinion in neurobiology*, 67:50–57, 2021. 1, 2

[4] Maryam Arabpour, Mahbubeh Nakhshab, Stephen Humphry, and Yalda Kazemi. Systematic analysis of language transcripts'(salt) transcribing method and narrative assessment protocol (nap) online coding method: are they interchangeable? *Logopedics Phoniatrics Vocology*, pages 1–9, 2021. 4

[5] Mathieu B Brodeur, Mary O'Sullivan, and Lauren Crone. The impact of image format and normative variables on episodic memory. *Cogent Psychology*, 4(1):1328869, 2017. 1

[6] Lila Davachi. Item, context and relational episodic encoding in humans. *Current opinion in neurobiology*, 16(6):693–700, 2006. 1

[7] Melissa C Duff, Natalie V Covington, Caitlin Hilverman, and Neal J Cohen. Semantic memory and the hippocampus: Revisiting, reaffirming, and extending the reach of their critical relationship. *Frontiers in Human Neuroscience*, 13:471, 2020. 1

[8] Arnaud D'Argembeau and Martial Van der Linden. Phenomenal characteristics associated with projecting oneself back into the past and forward into the future: Influence of valence and temporal distance. *Consciousness and cognition*, 13(4):844–858, 2004. 2

[9] Howard Eichenbaum, Andrew P Yonelinas, and Charan Ranganath. The medial temporal lobe and recognition memory. *Annu. Rev. Neurosci.*, 30:123–152, 2007. 1

[10] Michael D Rugg, Jeffrey D Johnson, and Melina R Uncapher. Encoding and retrieval in episodic memory. *The Wiley handbook on the cognitive neuroscience of memory*, pages 84–107, 2015. 1

[11] Maarten Sap, Eric Horvitz, Yejin Choi, Noah A Smith, and James W Pennebaker. Recollection versus imagination: Exploring human memory and cognition via neural language models. In *Association for Computational Linguistics*, 2020. 2

[12] Daniel L Schacter, Donna Rose Addis, Demis Hassabis, Victoria C Martin, R Nathan Spreng, and Karl K Szpunar. The future of memory: remembering, imagining, and the brain. *Neuron*, 76(4):677–694, 2012. 2

[13] Endel Tulving. 12. episodic and semantic memory. *Organization of memory/Eds E. Tulving, W. Donaldson, NY: Academic Press*, pages 381–403, 1972. 1