Inspiration through Observation: Demonstrating the Influence of Automatically Generated Text on Creative Writing

Melissa Roemmele
Language Weaver (RWS Group)
Los Angeles, CA, USA
mroemmele@sdl.com

Abstract

Getting machines to generate text perceived as creative is a long-pursued goal. A growing body of research directs this goal towards augmenting the creative writing abilities of human authors. In this paper, we pursue this objective by analyzing how observing examples of automatically generated text influences writing. In particular, we examine a task referred to as sentence infilling, which involves transforming a list of words into a complete sentence. We emphasize "storiability" as a desirable feature of the resulting sentences, where "storiable" sentences are those that suggest a story a reader would be curious to hear about. Both humans and an automated system (based on a neural language model) performed this sentence infilling task. In one setting, people wrote sentences on their own; in a different setting, people observed the sentences produced by the model while writing their own sentences. Readers then assigned storiability preferences to the resulting sentences in a subsequent evaluation. We find that human-authored sentences were judged as more storiable when authors observed the generated examples, and that storiability increased as authors derived more semantic content from the examples. This result gives evidence of an "inspiration through observation" paradigm for human-computer collaborative writing, through which human writing can be enhanced by text generation models without directly copying their output.[1]

Introduction

Creative text generation is a significant focal point at the intersection between computational creativity and natural language processing research. The goal behind much of this research is to understand and simulate human creative writing abilities. There is also increasing interest in using this work to augment human creativity. This objective has become especially visible given recent advancements in systems that can directly interface with human-authored text.

Many existing creative text generation systems can be applied to facilitate human authoring, even if they are not explicitly presented in this way. The clarity of this use case can largely depend on how the system is evaluated. There is no standard design for such evaluations of benefits to human authoring. Much work uses the convention of comparing generated output to human reference output for a given task, either by comparing the features of the text itself or comparing relative human judgments of it. Success by this standard is based on how well the system simulates human writing. One could theorize that the more a system writes like a human, the more it will be able to help other humans write, but further empirical exploration of this is needed. Alternatively, systems that explicitly aim to support human authoring are often evaluated in the context of interactive applications where authors can elicit generated text. Here, the quality of the model can be evaluated according to authors' interaction with the generated output.

In this paper, we focus on an "inspiration through observation" paradigm for human interaction with generated text. In many application settings for text generation, this human interaction is dynamic, with system output changing frequently in direct response to user choices. While discovering the best interaction paradigm is a critical objective of research on authoring support, here we minimize the role of user control over the generated text in order to assess the impact of merely observing the text. Authors see examples of generated text that fulfill a particular authoring objective, and they repeat the same task on their own. We compare human authoring outcomes in the absence and presence of these generated examples. This broad methodology could be applied to probe the ability of any system for aiding authoring, even systems that have not previously been assessed for this use case.

Our exploration of this paradigm focuses on a particular authoring task, *sentence infilling*, and a particular authoring objective, which we term *storiability*. Sentence infilling involves expanding a list of words into a full sentence. In our version of this task, the sentences we elicit can be viewed as story excerpts. The construct of storiability is related to previously discussed ideas such 'storiness' by Bailey (1999), which pertains to the success of a story from a reader's perspective. We define storiability as the degree to which an excerpt (here, a single sentence) alludes to an appealing story. Even though this is a broad definition, we operationalize it through specific instructions in our experiments. Through our experiments we find that observing automatically generated examples of our sentence infilling task helps people

[1] All code associated with our model, dataset synthesis, and authoring experiments is available at github.com/roemmele/InSentive. The data resulting from the authoring experiments is also available upon request by contacting the authors.

better fulfill the storiability authoring objective. This provides evidence for a general inspiration-through-observation framework by which generation systems can improve human authoring.

Background

As artificial intelligence has progressed, so has the development of Creativity Support Tools (CSTs). CSTs are digital applications intended to augment human abilities in creative endeavors like visual and performance art, music, and writing (see Frich et al. (2019) for a review of several applications). CSTs for writing in particular have been boosted by recent advances in natural language generation, making it possible for systems to interface with any unconstrained human-authored text. This includes figurative language like poetry (Kantosalo, 2019) and metaphors (Gero and Chilton, 2019). Advances in story generation (e.g. Fan, Lewis, and Dauphin, 2018; Martin, 2021) have been showcased by the increasing development of CSTs that support authoring in the narrative domain. One design pattern for these systems involves authors querying a generation model for a "suggestion" that can be integrated into their text (Clark et al., 2018; Khalifa, Barros, and Togelius, 2017; Manjavacas et al., 2017; Roemmele and Gordon, 2018b). This enables analysis of what users choose to do with the generated text (e.g. retaining or deleting it) and how their choices are affected by the features of the text (Akoury et al., 2020; Roemmele and Gordon, 2018a; Clark and Smith, 2021).

Human-computer interaction studies have compared people's writing with and without the use of AI-based tools, showing that these tools do change how people write. Existing work has examined the effect of word and phrase predictions for content like image captions (Arnold, Chauncey, and Gajos, 2020), emails (Buschek, Zürn, and Eiband, 2021), and movie reviews (Bhat, Agashe, and Joshi, 2021). For more open-ended creative writing tasks, most research has focused on optimizing and assessing how much people favor the generated content. What is needed is more experimental comparison of how the use of CSTs changes the authoring outcome as perceived by readers. Mizrahi, Yardeni Seelig, and Shahaf (2020) recently pursued this for the specific task of creating neologisms (i.e. new words). In their work, people wrote neologisms before and after observing automatically generated examples. The results showed that observing these examples helped people produce better neologisms in terms of their perceived creativity. In this paper, we follow a similar approach to examine the intervening effect of generated examples for the sentence infilling task.

Sentence Infilling

We focus on the specific task of sentence infilling to evaluate our hypotheses about authoring. Given a sequence of input words (e.g. "he town rain"), which we refer to as a "prompt", the infilling task expands the sequence into a complete sentence by inserting additional words (e.g."he rode his bike to town in the pouring rain."). We created a dataset for this task and trained an automated model on it, as detailed below.

Overview

Text infilling, alternatively known as expansion or elaboration, has recently attracted significant attention for multiple types of corpora (Donahue, Lee, and Liang, 2020; Fedus, Goodfellow, and Dai, 2018; Huang et al., 2020; Shen et al., 2020). There are different configurations of this task based on the length of the text to be infilled. For stories, some work has focused on inserting sentence-length sequences that connect passages (Chandu, Dong, and Black, 2020; Ippolito et al., 2019; Mori et al., 2020). A more constrained version of infilling turns it into a cloze (i.e. fill-in-the-blank) task where infilled segments are single words or short phrases. Our infilling model outputs a single sentence given a sequence of words, but no assumptions are made about the number of words to infill. This design is reflected in existing work applied to creative authoring support (Özbal, Pighin, and Strapparava, 2013; Safovich and Azaria, 2020), but it has yet to be examined how automatically infilled sentences affect human performance of this task.

Dataset

We are not aware of any datasets that mirror the design of our particular infilling task, by which sentences can be generated from any arbitrary sequence of words. However, it is easy to simulate an infilling dataset using existing corpora. Given that the task is framed in the context of storytelling, we obtained 10,000 English-language stories from a variety of genres in the BookCorpus (Kobayashi, 2018). We segmented each story into sentences[2], filtering sentences with less than ten words. To derive pairs of prompts and infilled sentences, we randomly dropped between 60-100% of words in each sentence. We required that the resulting ablated sentence consist of at least 50% content words (i.e. nouns, verbs, adjectives), since function words that convey little semantic meaning (i.e. pronouns, prepositions, determiners) are more frequent in text. The ablated sentences became the prompts used as the source inputs to the model, whereas the corresponding original sentences were the target infilled outputs generated by the model. The mean number of words in the prompts and infilled sentences was 4.86 and 19.19, respectively. These pairs were divided into 34,172,128 training instances, 897,473 validation instances, and 894,484 test instances fully held-out during training.

Model Design

Our infilling model[3] is a Transformer language model (LM) (Vaswani et al., 2017), which is currently a popular architecture for many machine learning approaches to language generation. Figure 1 broadly illustrates the model. Our scheme

[2]All linguistic processing steps used to derive this dataset, including sentence segmentation, word tokenization, and part-of-speech tagging, were performed with the spaCy library: spacy.io

[3]We used the Texar-PyTorch library for implementation: texar-pytorch.readthedocs.io. Additional hyperparameter settings included: maximum epochs = 100, batch size = 32, gradient accumulation over 8 steps, validation every 25,000 steps, early stopping after 25 consecutive rounds of no validation improvement, static learning rate = 0.001, maximum gradient norm = 1.0.

for applying this architecture to infilling is closely related to that described in Donahue, Lee, and Liang (2020), with one main distinction. Their approach uses designated tokens (i.e. [BLANK]) in the input sequences to indicate the position where text should be infilled in the output. Alternatively, we only represent prompt words in the input, without any explicit signal for where text should be infilled between prompt words. As in the cited work, we initialized the model with weights from pretrained GPT-2 (Radford et al., 2019) as a means of embedding general knowledge of English text. GPT-2 has been highlighted for its potential to generate creative text (See et al., 2019; Dathathri et al., 2020). We used the "small" version of GPT-2 (117M parameters) and also the corresponding GPT-2 tokenizer to represent all text as subword tokens. We concatenated each prompt and corresponding infilled sentence together as a single token sequence, using designated tokens to signify the start ({{) and end (}}) of the prompt. To avoid memory errors, we set a limit on the size of the sequences by truncating prompts to the first 25 subword tokens and target sentences to the first 75 tokens. We then fine-tuned the pretrained model for the infilling task by training it on the dataset described above, using the maximum likelihood estimation loss function that is standard for training neural LMs. Our only variation from standard LM training was that we optimized using only the loss for the tokens in the target infilled sentences, and did not compute the loss of the source prompt tokens. This simulates an encoder-decoder scheme which decodes target text from the encoded source input; here the LM functions as both an encoder and decoder, which significantly reduces the number of parameters in the model. We monitored perplexity on the validation items in order to end training when perplexity stopped improving. In inference mode, the model observes a prompt and generates an infilled sentence through a standard LM decoding method. In particular, we autoregressively sample from the LM probability distribution and append the resulting token to the sentence, until the end-of-sequence token (i.e.[EOS]) is generated.

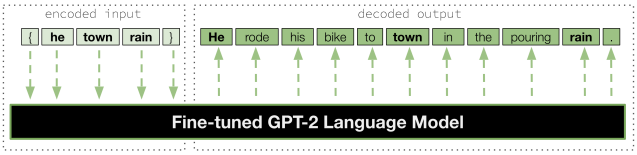

Figure 1: General architecture of sentence infilling model

Authoring Experiment

We next designed a human authoring task that integrates our trained infilling model. To broadly summarize this process detailed in this section: we selected certain prompts from the test partition of our dataset and generated infilled sentences for them. We then elicited human-authored infilled sentences for these same prompts. People produced sentences in two conditions. In the first, they simply wrote sentences for each prompt. In the second, they were shown the sentences generated by our model for the same prompts and wrote new sentences. We explain each of these steps below.

Prompt Selection

We selected prompts from the test set with exactly three words. This particular length value was picked based on intuition. Fewer words approximates unconstrained generation rather than infilling, while more words simulates a constrained fill-in-the-blank task. We excluded prompts derived from dialogue sentences (i.e. those containing quotation marks). Dialogue can pose issues for sentence segmentation (e.g. "he said." may be segmented as a separate sentence from its adjacent quote). We also excluded prompts containing punctuation, numerical digits, named entities[4], or word tokens not recognized in the DistilBERT (described below) tokenizer vocabulary. Finally, we excluded prompts with more than one function word (e.g. pronouns, prepositions, determiners). By applying these constraints, we expected the prompts to give clear semantic cues for the infilled sentences. The resulting selection consisted of 23,005 prompts.

Since the process for deriving prompts involved random ablation of full sentences and the position of the ablated words varied, we theorized that even prompts of the same length require different degrees of infilling to yield grammatical sentences. For example, the prompt "his, body, relax" already resembles English syntax, and thus it would only take a single infilling word to produce a grammatical sentence (e.g. "His body could relax"). In contrast, it is possible but harder for native English speakers to find a single infilled word that could transform the prompt "peculiar, rob, more" into a grammatical sentence. Accordingly, we expected that the difficulty of the task would vary according to the degree of required infilling for a prompt. We designed an approach for automatically scoring this difficulty. For each selected prompt, we scored the probability of each of its word tokens according to the Masked LM configuration of DistilBERT[5] (Sanh et al., 2019). A Masked LM is well-suited for this measure because it is specifically trained on a fill-in-the-blank task to predict the likelihood of words according to their context. We used the average of the prompt token probabilities to represent the inverse difficulty (i.e. easiness) of a prompt. We theorized that high-probability prompts are easier in infill since they are already probable sequences, whereas low-probability prompts require more infilling to become probable. We assigned the difficulty label "easy" to the 10% highest-probability prompts and the label "hard" to 10% lowest-probability prompts, yielding 2,301 prompts for each difficulty level.

Generated Sentences for Prompts

We then applied the trained model to produce infilled sentences for the selected prompts. We generated five infilled sentences per prompt, using the decoding method of nucleus (top-p) sampling with p = 0.7, based on the parameters recommended by DeLucia et al. (2020) for generating narrative text. The generated output followed constraints

[4]As with the data creation, this detection was done with spaCy.

[5]We used the model interface provided by the HuggingFace transformers library: huggingface.co/transformers

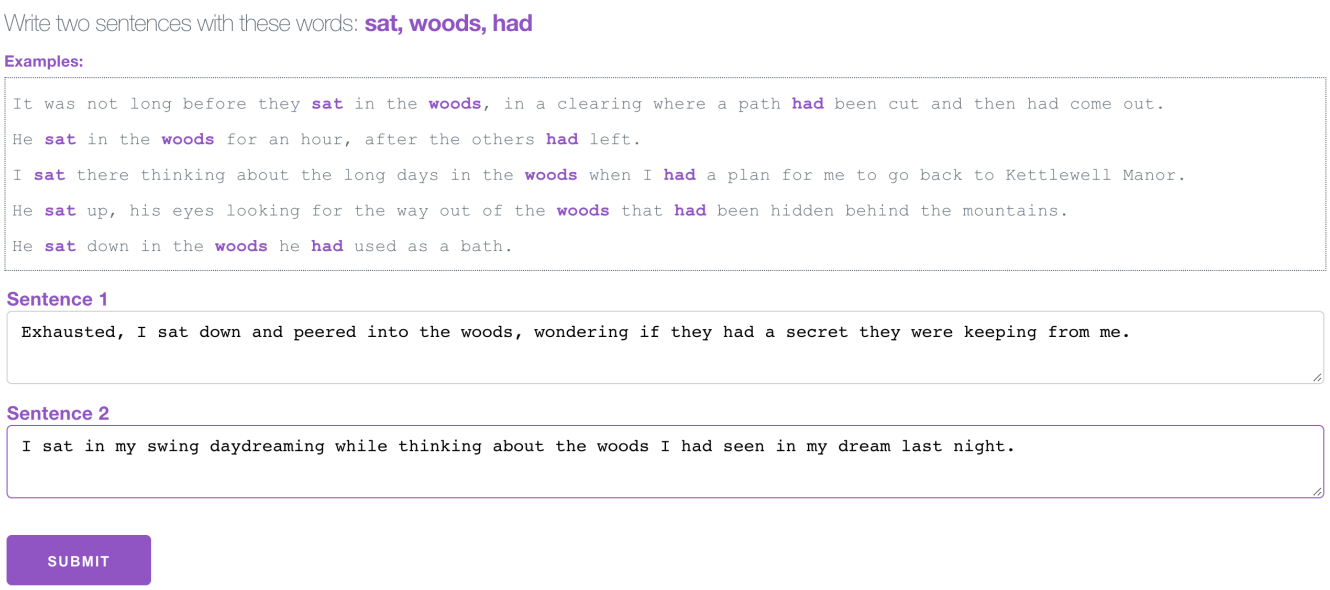

Figure 2: Screenshot of authoring interface for a single prompt in the POST stage. In the PRE stage, the example sentences are not visible.

consistent with the human authoring instructions described below. In particular, the generated sentences had to contain all prompt words in the same order as they appeared in the prompt. Prompt words were allowed to be capitalized in the sentence. Sentences had to consist of at least seven word tokens but no more than fifty. We additionally restricted sentences with quotation marks and missing end-of-sentence punctuation (i.e. by requiring the last character to be non-alphanumeric), since this may signify the sequence is not a complete sentence or combines multiple sentences (e.g. quoted dialogue). We filtered sentences with adjacently repeated words (this is a frequently observed issue with neural LMs). Finally, we promoted the diversity of the five sentence outputs for a given prompt by filtering any sentence with 60% or more of its words already appearing in previously generated sentences for that prompt. All of this criteria was satisfied by continually generating sentences for a prompt until it yielded five acceptable outputs. As a last step that we performed through manual review, we filtered any items where the prompt or generated sentence contained profanity or offensive content. This was done to minimize potential risk of harm to participants in the experiment. The final set consisted of 2,205 easy items and 2,189 hard items.

Human Authoring Task

We then conducted a human authoring task[6] utilizing the selected prompts and generated sentences. Participants were instructed that they would be shown a list of three words (the prompt) and would write two unique sentences containing those words. They were presented with some manually written examples of infilled sentences. The instructions emphasized that they should "try to write sentences that evoke a story someone would be curious to hear", which activates the construct of storiability that we focus on in this work. The authors' sentences were required to obey the same prompt token order, length, and end-of-sentence punctuation constraints as the model output, which we enforced through the user interface. In the first stage of the task (the PRE stage), each author wrote two sentences for five prompts, which were randomly sampled from the "easy" and "hard" categories. In the second stage (the POST stage), authors were again shown the same five prompts and wrote an additional two unique sentences for each. This time, the five generated sentences were shown to them as examples they could reference while writing. Their sentences were required to be different from the examples. Figure 2 shows an example screenshot of the interface for this exercise.

The presence of the generated examples was the only variable that differed between the two stages. In both stages, after submitting the sentences for a single prompt, participants were shown generated text passages (described as "stories") that each began with the sentences they wrote. These passages were generated by the original pretrained GPT-2[7] (not the infilling model). Passages had a maximum length of 75 words, and only the first k complete sentences within this limit were displayed. The instructions informed authors that writing more interesting sentences would yield more interesting stories. However, this component of the task was not an experimental variable, since it was not varied between the two stages. This feedback was simply intended to incentivize authors to write more storiable sentences.

[6]This was implemented as a ReactJS + Flask web application.

[7]Using the model interface provided by HuggingFace transformers; generated using nucleus sampling with p = 0.7

Prompt	Difficulty	PRE Sentences	POST Sentences	GEN Examples
walking and seeing	easy	**1.** The little children enjoyed walking through the zoo and seeing all the different animals. **2.** The boy's favorite activity was walking to the marina and seeing all of the boats in the water.	**1.** After being released from prison for a crime he didn't commit, the old man was thoroughly enjoying walking through the city and seeing how the world had changed. **2.** The woman cried when she saw her little girl walking and seeing for the first time after she got her new glasses.	**1.** She felt the urge to cry, but she kept walking and seeing no sign of it. **2.** He was walking in front of the stove and he looked down on the ground seeing what was going on. **3.** We were walking in and were immediately upon seeing what the neighbors had in store. **4.** She was walking with a friend, and she just happened to be seeing a man, a man, and he was going to kill her. **5.** She could hear men walking up and down the alley, and she didn't know what they were doing, but she couldn't deny seeing the resemblance.
nose pushed see	hard	**1.** The sled dogs nose was in the air as it pushed through the snow to see his owner. **2.** I held my nose and pushed the stinky garbage can to the curb to see if I can catch the garbage man in time.	**1.** The dog, using his big nose, pushed the front door open to see if his owner was home. **2.** The boy held his nose to stifle a sneeze but the involuntary reflex pushed his head forward, watering his eyes and making it hard for him to see.	**1.** The man's nose was being pushed up and down, and as he moved closer to the screen, the image started to dawn on him, and he was shocked to see his father lying on the ground, dying. **2.** He cleared his throat, the same way he had when he had slapped the back of his head and nose, then pushed himself away, but he was careful not to let her see his anger. **3.** When he saw his own nose in the white sordid mess, he pushed off his seat to see it for himself. **4.** He kissed her nose and pushed the sleeve of her shirt back to see what she was thinking. **5.** A stray nose-bleed might be pushed up, but I couldn't see anything out of place.

Table 1: Examples of authoring blocks. Each block consists of sentences written by a single author before (PRE) and after (POST) observing the generated (GEN) example sentences.

We recruited participants for this task through Amazon Mechnical Turk[8] (AMT), a crowdsourcing platform. 23 authors from majority native English-speaking countries were each paid $10 based on an estimated completion time of 45 minutes to 1 hour. The result was a dataset of *authoring blocks*, with each block consisting of a prompt shown to an author, their two sentences written before observing the generated examples (PRE), their two sentences written after the observing the generated examples (POST), and the five generated examples they saw (GEN). Examples of authoring blocks are shown in Table 1. With each author responding to five unique prompts, this yielded 115 blocks. We filtered six blocks where at least one sentence response (PRE or POST) was revealed to actually consist of multiple sentences (since this wasn't straightforward to check through the interface during the task). This ultimately resulted in a set of 109 blocks to be used for evaluation, 53 for easy prompts and 56 for hard prompts.

Evaluation of Authoring Experiment

In line with the objective of the authoring task, we conducted a judgment task to evaluate readers' perceived storiability of the sentences in the authoring blocks. This resembles story generation evaluations where people are asked which one of a set of stories they most prefer reading (e.g. Fan, Lewis, and Dauphin, 2018). For each of the 109 blocks, we gathered all unique combinations of the two PRE sentences, two POST sentences, and the first two of the observed GEN examples in that block, yielding 872 *judgment groups* ($109 * 2 * 2 * 2 = 872$). Thus, each judgment group consisted of a PRE, POST, and GEN sentence aligned to the same prompt and author. We designed a questionnaire targeting the relative storiability of the sentences in each group. Raters were instructed to "imagine that each sentence [in the judgment group] is an excerpt from a story and pick the one that makes you most want to read that story". Only the sentence text itself was shown, and the sentences in each group were randomly ordered. We recruited 16 participants from majority native English-speaking countries through AMT to

[8] mturk.com

Prompt	Difficulty	PRE Sentence	POST Sentence	GEN Sentence
felt meet again	easy	**Jenna felt a spooky sense of deja vu and felt that she was about to meet a familiar stranger yet again.**	Bonnie felt a syrupy sentimentality and nostalgia and wanted to meet Raphael again.	I felt so relieved to meet you again.
regard sorts prevent	hard	**He had no regard for his own safety, a maverick of sorts, which did nothing to help prevent him from oft getting injured.**	In regard to the message, there were all sorts of interpretations that could be made, so she asked for clarification to prevent misunderstandings.	A lower regard may come to any type of treatment that may result in a delay of sorts in order to prevent future evidence of therapy.
servants early life	easy	It's sad when people have servants that have to wake up early and do everything for someone else without having a life of their own.	**They became servants at a very early age after having a difficult life and losing their parents.**	But, yes, there were two excellent servants from a very early age in the village, who could carry the life of an even younger man.
hoping questions few	hard	I was hoping I could find the answer to my homework questions, and after a few minutes I found them by doing a simple Google search.	**She pored her thoughts, fears, and dreams into her diary, hoping that by writing them down, she could answer the vexing questions of life that few people ever really understood.**	They were hoping to avoid answering any questions for a few days.
quickly and joined	easy	There was a bird that quickly fell from the sky and joined with the ground.	The car quickly entered the lane and joined with the traffic.	**The nurse quickly packed up the case and joined him.**
arms awkwardly car	hard	The arms hung awkwardly out the window of the car.	His arms flung awkwardly as the police slammed him up against the car to cuff him.	**Sue wrapped her arms around his neck, pulled him awkwardly out of the car, and then pushed him down the long, steep driveway.**

Table 2: Examples of judgment groups. The bolded sentence in each group was selected by both raters as the most storiable.

rate subsets of 55-56 judgment groups, with each paid $5 for an estimated completion time of 25-30 minutes. There were two raters for each subset, yielding a total of 1,744 responses (848 for authoring blocks with easy prompts and 896 for hard). Examples of judgment groups are shown in Table 2. In these examples the bolded sentence was picked by both of its raters as the most storiable among the group.

For the results described below, we discuss judgments in terms of storiability preferences. In particular, each response is a single data point where the most storiable sentence selected by the rater was labeled as "Preferred" and the other sentences in the judgment group were labeled as "Not Preferred". All data points have equal weight in the analyses.

Results

Human versus Generated Storiability Table 3 shows the normalized distribution of storiability preferences across the PRE, POST, and GEN sentences, along with their raw number of "Preferred" votes. Note that if preferences were randomly distributed across these three sets, each would approximate 0.33 (one-third) of the distribution. The numbers show that people notably preferred human-authored sentences (both PRE and POST) to GEN sentences (statistically significant at $p < 0.05$)[9].

In contrast with human authoring, the infilling model did not receive any explicit instructions about the storiability authoring objective. The model was simply trained to generate sentences that appeared in stories. We can guess that the training sentences observed by the model are not all equally likely to be perceived as storiable. It is possible that this is why raters favored human-authored sentences over the generated ones. However, even generated text designed to mimic human writing objectives often does not meet this standard (e.g Lin et al., 2020), so the difference in preferences is not simple to interpret. The focus of this particular paper is not on comparing the relative quality of human and generated text, but on whether generated text can alter the quality of human writing. Thus, the rest of our analyses concentrate on this question.

Prompt Difficulty Table 4 shows the effect of difficulty on the number of infilling words people used to connect the prompt words. The human-authored sentences for the hard prompts had significantly more infilled words between

[9]Statistical significance for all analyses was determined by two-sample Monte Carlo permutation tests.

Preferred PRE	Preferred POST	Preferred GEN
0.356 (621)	0.365 (636)	0.279 (487)

Table 3: Distribution of storiability preferences

prompt words compared with easy prompts ($p < 0.05$). This validates the expected difference between these conditions, suggesting that hard prompts required more authoring effort.

Difficulty	Infilled Words
easy	3.035
hard	4.317

Table 4: Mean number of words between prompt words in human-authored sentences according to difficulty

Prompt Difficulty and Storiability Table 5 shows the distribution of preferences for PRE and POST sentences grouped by difficulty level. We found that POST sentences had higher storiability than the PRE sentences, but only for hard prompts ($p < 0.05$). Thus, people were more likely to write storiable sentences for these prompts after observing the GEN examples. The result for easy prompts showed a tendency towards the reverse pattern, but the difference in this case was not statistically significant. Based on this result, we focus our subsequent analyses on the items associated with hard prompts. We return to some discussion of this interaction effect regarding difficulty in the next section.

Difficulty	Preferred PRE	Preferred POST
easy	0.384	0.354
hard	0.329	0.375

Table 5: Distribution of storiability preferences for human-authored sentences by difficulty

Influence of Generated Examples The higher preference for the POST sentences suggests that observing the GEN examples had some impact on authors. One could consider other interpretations: for example, maybe authors were simply better at the task in the POST stage after a round of practice in the PRE stage. To investigate this, we first determined whether any influence of the GEN examples could be quantitatively detected in the POST sentences. There are many different features that could be used to quantify this influence. Here we focused on whether authors incorporated semantic content from the examples they observed. We assessed this using a quantitative measure of semantic similarity between sentences based on vector representations given by a pretrained language model. Intuitively, pretrained LMs are expected to produce similar vector representations for sentences with a similar meaning. This representation should transcend the lexical level, so that even sentences with few words in common can have a high similarity score if their respective words in context are synonymous. We computed semantic similarity between the PRE and GEN sentences, and then separately between the POST and GEN sentences. Since the GEN examples were not shown in the PRE condition and thus could have no influence on the PRE sentences, any significant difference in this measure between the PRE and POST sentences can be attributed to authors observing the GEN examples.

We computed the cosine vector similarity between sentences encoded with the DistilBERT[10] LM. For a given prompt, the similarity score for a human-authored sentence h is its maximum similarity over all GEN examples gs for that prompt, i.e. $score(h, gs) = \max_{g \in gs} sim(h, g)$. We select the maximum because there may be one GEN example in particular that most influences a given sentence.

Table 6 shows the mean of this similarity measure for the PRE and POST sentences. POST sentences had higher similarity to GEN sentences ($p < 0.05$), confirming that the GEN examples had semantic influence on the authors' writing.

Condition	Similarity
PRE	0.921
POST	0.923

Table 6: Similarity between human and generated sentences before (PRE) and after (POST) observation of GEN examples

Influence and Storiability After verifying that the difference between the PRE and POST conditions can be attributed to semantic influence from the GEN examples, we examined whether this influence was related to the higher storiability of the POST sentences. Table 7 demonstrates that sentences preferred as more storiable were also more semantically influenced by the GEN examples, as indicated by the higher similarity scores for the Preferred sentences ($p < 0.05$). Thus, by incorporating some degree of content from the GEN examples, people tended to better fulfill the authoring objective. Table 8 gives some examples of judgment groups where semantic influence can be qualitatively observed in the POST sentence. The GEN example with the most influence is shown (i.e. the one most similar to the POST sentence), and we comment on the subjective evidence of their similarity. These results encourage future opportunities for explaining the exact mechanism underlying semantic influence. We discuss this further in the next section.

Judgment	Similarity
Not Preferred	0.922
Preferred	0.925

Table 7: Similarity between POST and GEN sentences (i.e. degree of semantic influence) according to storiability preferences

[10]The same core model used for computing probability scores to determine prompt difficulty, as described earlier. Here, we use the raw hidden states of the model for feature representation instead of the Masked LM probability outputs.

Prompt	PRE Sentence	POST Sentence	Influential GEN Example	Description
shoulders waves color	My shoulders were aching but I was set on diving through the waves, the color of the water getting deeper the further out I went.	Her new hair cut had the length to the shoulders, with waves of a bright pink color all the way down.	His hair was cropped short, flowing down his shoulders, but there were waves of the same color.	Connected prompt words via semantic category of hair
there die capacity	The bouncer thought there was a chance people might die if there was a fire because the club was way over its capacity.	There is no chance you're not going to die, so you have to come to terms with that in some capacity.	It seems, that there is a good chance that I will die in my capacity to forgive and to get on with my life.	Used less literal sense of word "capacity"
meant said store	The child yelled at her sister not understanding what she meant when she said to her that she wanted some comics from the store.	It meant a lot to me when she said she was going to the toy store to get me a game.	It meant a lot to me, because I'd said I'd drop by the store.	Used phrase "it meant a lot to me"
spent wind him	After his run he stood by the beach, spent, as the wind whipped by him.	She spent the day by the water, the wind whipping her hair, aching for him.	She spent the rest of the day in the saddle, keeping the wind from blowing through her hair and reminding her of her promise to get him a hot bath.	Used expanded form of phrase "spent the day" ("spent the rest of the day")
peculiar rob more	I have a peculiar friend named Rob who always wants more excitement.	I felt it was very peculiar that after talking to Rob for only about an hour, I wanted to know more about him.	They felt a peculiar attraction to Rob, but couldn't afford to spend much more time together.	Referred to curiosity about Rob

Table 8: Examples of POST sentences demonstrating semantic influence, with subjective descriptions of how influence is seen. For reference, the PRE sentence without semantic influence is also shown.

Discussion

Observing automatically generated examples of sentence infilling influenced authors to better perform this infilling task on their own. Even though this is a contrived exercise different from conventional forms of creative writing, it still calls upon the same linguistic creativity. A related task is reflected in the real world through popular word games where people produce sentences given word constraints and players rate the interpretability and creativity of the resulting sentences (e.g. Cooper and McNeill, 2005). The task is also applicable to CSTs for writing: for example, a writer might want to brainstorm about potential connections between words they already have in mind, which could be facilitated by a model related to infilling. In contrast to other research on CSTs, this paper focuses less on the interactive capabilities of such systems, like enabling author control over generated output, but our findings are still relevant to interactive applications.

We chose to emphasize the authoring objective of storiability because of our focus on AI-augmented story writing. Storiability is not a one-size-fits-all metric for this research. The quality of a story can be judged on multiple dimensions that are often not consistently defined across different studies, as discussed in Celikyilmaz, Clark, and Gao (2020). Evaluations tend to target both the sensibility of stories (e.g. grammaticality, coherence, plausibility) and their more "creative" aspects (e.g. interestingness, suspensefulness, humorousness). The notion of storiability is more related to the latter group, but does not preclude other dimensions. For example, if a sentence contains grammatical errors, a person may not prefer to read the story associated with that sentence. By operationalizing storiability according to a specific question ("which sentence makes you want to read more?"), we tried to elicit judgments that encompass many ways this objective can be achieved. Future research can examine more specific formulations of this question.

An intriguing finding was the difference in outcomes according to prompt difficulty, such that only sentences for hard prompts displayed more storiability as an effect of observing generated text, with no such pattern for easy items. This points to a broad direction for future work: to examine how the demands of the writing task itself affect authors' interaction with an automated model. For instance, authors' engagement with writing assistance tools varies at different times during a single writing session, as discussed in Huang, Huang, and Huang (2020). This may be due to some parts of the text being harder to write than others, as hinted by the mediating effect of difficulty in our results. Interestingly, a follow-up analysis showed there were no significant differences in POST similarity to GEN examples based on difficulty, meaning that the GEN sentences for easy prompts had just as much semantic influence as those for hard prompts. Thus, this influence was somehow not as helpful in promoting storiability in the easy case. One possibility is that authors were already good at producing storiable sentences for

easy prompts in the PRE stage, so even when they were influenced by the GEN examples, this influence did not additionally benefit the POST sentences. The hard prompts may have been more challenging, giving the GEN examples a larger opportunity to enhance the POST sentences in this case. Because our evaluation did not include pairwise comparisons between sentences for easy and hard prompts, it will require further research to better understand this finding.

Our analysis of semantic influence confirms authors derived certain content from the observed examples. More investigation is needed to understand what type of content was most influential. Authors may have extracted specific words and phrases, as indicated by some of the examples in Table 8, but they did not simply copy or mimic the examples at large; if they had, there would not be a significant difference in storiability between the POST and GEN sentences as reported in Table 3. One thought is that authors utilized an idea conveyed by a GEN sentence, but reformulated the sentence to repair inadequacies such as ill-formed, awkward, or vague wording. It is also possible that the GEN examples revealed a semantic dimension by which the prompt words were related, one that authors did not initially consider in the PRE condition. The first example in Table 8 might convey this: the GEN example connects the prompt words "shoulders", "waves", and "color" through the conceptual dimension of "hair". Perhaps the example triggered the author to recall this particular concept unifying the prompt words, and they emulated it in the POST sentence. One targeted metric for examining influence could focus specifically on modeling this activation of "latent" concepts. Our work quantified influence according to a single measure, but future work could attempt to narrow down the influence of specific linguistic features such as syntactic style (e.g. relative proportion of nouns, verbs, prepositions, etc.), emotional tone (e.g. joyful, sorrowful, fearful), and narrative perspective (e.g. references to pronouns and proper nouns). Existing work has addressed this by examining the strategies authors develop for eliciting precise types of content from generation models; for example, by triggering the model at certain syntactic positions in a sentence (Calderwood et al., 2020). We can use these analyses to guide future systems towards producing content authors find most helpful.

Conclusion

In this paper, we explore the question of how automatically generated text can influence human creative writing. We specifically assessed this question through the authoring task and objective of sentence infilling and storiability, respectively. In accordance with a proposed inspiration-through-observation paradigm by which automated models provide helpful examples of how to fulfill the task, we found that observing generated sentences enhanced reader-judged appeal of human-authored sentences. Our results provide empirical evidence that automated models can intervene in the writing process without necessarily replacing human effort. This invites further exploration of this paradigm for other authoring tasks and objectives. The outcome has the potential to transcend the standard of both human and computer authoring when each function independently.

References

Akoury, N.; Wang, S.; Whiting, J.; Hood, S.; Peng, N.; and Iyyer, M. 2020. STORIUM: A Dataset and Evaluation Platform for Machine-in-the-Loop Story Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6470–6484. Association for Computational Linguistics.

Arnold, K. C.; Chauncey, K.; and Gajos, K. Z. 2020. Predictive text encourages predictable writing. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, 128–138. Association for Computing Machinery.

Bailey, P. 1999. Searching for storiness: Story-generation from a reader's perspective. In *Working Notes of the Narrative Intelligence Symposium*.

Bhat, A.; Agashe, S.; and Joshi, A. 2021. How do people interact with biased text prediction models while writing? In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, 116–121. Association for Computational Linguistics.

Buschek, D.; Zürn, M.; and Eiband, M. 2021. The impact of multiple parallel phrase suggestions on email input and composition behaviour of native and non-native english writers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery.

Calderwood, A.; Qiu, V.; Gero, K. I.; and Chilton, L. B. 2020. How novelists use generative language models: An exploratory user study. In *HAI-GEN+ user2agent@ IUI*.

Celikyilmaz, A.; Clark, E.; and Gao, J. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Chandu, K. R.; Dong, R.-P.; and Black, A. W. 2020. Reading between the lines: Exploring infilling in visual narratives. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 1220–1229. Association for Computational Linguistics.

Clark, E., and Smith, N. A. 2021. Choose your own adventure: Paired suggestions in collaborative writing for evaluating story generation models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, 3566–3575. Association for Computational Linguistics.

Clark, E.; Ross, A. S.; Tan, C.; Ji, Y.; and Smith, N. A. 2018. Creative writing with a machine in the loop: Case studies on slogans and stories. In *23rd International Conference on Intelligent User Interfaces*, 329–340. Association for Computing Machinery.

Cooper, P. E., and McNeill, D. 2005. You've been sentenced! `https://boardgamegeek.com/boardgame/20790/youve-been-sentenced`.

Dathathri, S.; Madotto, A.; Lan, J.; Hung, J.; Frank, E.; Molino, P.; Yosinski, J.; and Liu, R. 2020. Plug and play language models: A simple approach to controlled

text generation. In *International Conference on Learning Representations*.

DeLucia, A.; Mueller, A.; Li, X. L.; and Sedoc, J. 2020. Decoding methods for neural narrative generation. *arXiv preprint arXiv:2010.07375*.

Donahue, C.; Lee, M.; and Liang, P. 2020. Enabling language models to fill in the blanks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2492–2501. Association for Computational Linguistics.

Fan, A.; Lewis, M.; and Dauphin, Y. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 889–898. Association for Computational Linguistics.

Fedus, W.; Goodfellow, I.; and Dai, A. 2018. MaskGAN: Better Text Generation via Filling in the_. In *International Conference on Learning Representations (ICLR)*.

Frich, J.; MacDonald Vermeulen, L.; Remy, C.; Biskjaer, M. M.; and Dalsgaard, P. 2019. Mapping the landscape of creativity support tools in HCI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–18.

Gero, K. I., and Chilton, L. B. 2019. Metaphoria: An algorithmic companion for metaphor creation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12. Association for Computing Machinery.

Huang, Y.; Zhang, Y.; Elachqar, O.; and Cheng, Y. 2020. INSET: Sentence infilling with INter-SEntential transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2502–2515. Association for Computational Linguistics.

Huang, C.-Y.; Huang, S.-H.; and Huang, T.-H. K. 2020. Heteroglossia: In-situ story ideation with the crowd. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*.

Ippolito, D.; Grangier, D.; Callison-Burch, C.; and Eck, D. 2019. Unsupervised hierarchical story infilling. In *Proceedings of the First Workshop on Narrative Understanding*.

Kantosalo, A. 2019. *Human-Computer Co-Creativity: Designing, Evaluating and Modelling Computational Collaborators for Poetry Writing*. Ph.D. Dissertation, University of Helsinki.

Khalifa, A.; Barros, G. A.; and Togelius, J. 2017. Deeptingle. In *8th International Conference on Computational Creativity*.

Kobayashi, S. 2018. Homemade bookcorpus. `https://github.com/BIGBALLON/cifar-10-cnn`.

Lin, B. Y.; Zhou, W.; Shen, M.; Zhou, P.; Bhagavatula, C.; Choi, Y.; and Ren, X. 2020. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Manjavacas, E.; Karsdorp, F.; Burtenshaw, B.; and Kestemont, M. 2017. Synthetic literature: Writing science fiction in a co-creative process. In *Proceedings of the Workshop on Computational Creativity in Natural Language Generation*.

Martin, L. 2021. *Neurosymbolic Automated Story Generation*. Ph.D. Dissertation, Georgia Institute of Technology.

Mizrahi, M.; Yardeni Seelig, S.; and Shahaf, D. 2020. Coming to Terms: Automatic Formation of Neologisms in Hebrew. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 4918–4929. Association for Computational Linguistics.

Mori, Y.; Yamane, H.; Mukuta, Y.; and Harada, T. 2020. Finding and generating a missing part for story completion. In *Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*.

Özbal, G.; Pighin, D.; and Strapparava, C. 2013. BRAINSUP: Brainstorming support for creative sentence generation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1446–1455. Association for Computational Linguistics.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language models are unsupervised multitask learners.

Roemmele, M., and Gordon, A. 2018a. Linguistic features of helpfulness in automated support for creative writing. In *Proceedings of the First Workshop on Storytelling*.

Roemmele, M., and Gordon, A. S. 2018b. Automated Assistance for Creative Writing with an RNN Language Model. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion*. Association for Computing Machinery.

Safovich, Y., and Azaria, A. 2020. Fiction sentence expansion and enhancement via focused objective and novelty curve sampling. In *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence*, 835–843. IEEE.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

See, A.; Pappu, A.; Saxena, R.; Yerukola, A.; and Manning, C. D. 2019. Do massively pretrained language models make better storytellers? In *Proceedings of the 23rd Conference on Computational Natural Language Learning*.

Shen, T.; Quach, V.; Barzilay, R.; and Jaakkola, T. 2020. Blank language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 5186–5198. Association for Computational Linguistics.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*.