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

Melissa Roemmele
Language Weaver (RWS Group)
Los Angeles, CA
melissa@roemmele.io

1 Overview

A growing body of research on creative text generation directs this goal towards augmenting the creative writing abilities of human authors. Much of this work is evaluated in the context of dynamic interactive applications that emphasize users' choices of what to do with the generated text. But does simply observing automatically generated examples of an authoring task affect writers when they perform the same task? To investigate this, we focus on the task of *sentence infilling*, which involves transforming a list of words into a complete sentence. We emphasize the authoring objective of "storiability", 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]

2 Sentence Infilling

Given a sequence of input words (e.g. "he town rain"), which we refer to as a "prompt", the sentence infilling task expands the prompt into a complete sentence by inserting additional words without changing the order of the prompt words (e.g."he rode his bike to town in the pouring rain."). Recent work has explored variations of this infilling task for creative text [e.g. Ippolito et al., 2019, Mori et al., 2020, Donahue et al., 2020, Safovich and Azaria, 2020]. Our infilling model is a Transformer language model (LM) initialized with pretrained GPT-2 weights [Radford et al., 2019]. It is related to the model described in Donahue et al. [2020], with the distinction that instead of the model observing special tokens indicating where text should be infilled, the prompts for our model do not explicitly represent where to make insertions. We trained the model on a dataset of prompt-sentence pairs that we derived from 10K English-language stories in the BookCorpus [Kobayashi, 2018]. For a given sentence in this corpus, we randomly ablated a subset of its tokens. The ablated sentence became a prompt and its original form became the corresponding infilled sentence. Applying this to all sentences in the corpus resulted in $\sim$36M pairs, with $\sim$34M used for training, $\sim$1M used for validation during training, and the final $\sim$1M held out as a test set. We used the standard LM procedure of optimizing MLE loss to train the model, but we simulated an encoder-decoder scheme by only computing the loss for the target infilled sentences while omitting the loss of the source prompt

[1] All code associated with our model, dataset synthesis, and authoring experiments, as well as the data resulting from the experiments, is available at github.com/roemmele/InSentive.

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

tokens. For the experiment described below, we used the convention of autoregressive decoding to generate an infilled sentence for a prompt.

3 Authoring Experiment

We conducted an experiment where people wrote infilled sentences for a selection of three-word prompts from our test set. When selecting prompts, we hypothesized that certain prompts would be harder for people to infill than others, which could influence the role of the generated text in the authoring outcome. To examine this, we computed the average probability of the tokens in a prompt according to the DistilBERT LM [Sanh et al., 2019]. Prompts with high probability scores already resemble complete sentences, so we theorized they would yield fewer infilled tokens and thus require less authoring effort. Accordingly, we assigned the difficulty label "easy" to the 10% highest-probability prompts and the label "hard" to the 10% lowest-probability prompts. We applied our trained model to generate five infilled sentences for each of these prompts, using the decoding method of nucleus (top-p) sampling with p = 0.7. We utilized these prompts and generated sentences in the human authoring task. In this task, participants were instructed that they would be shown a list of words (the prompt) and would write two unique sentences containing those words. The instructions emphasized that they should "try to write sentences that evoke a story someone would be curious to hear", which operationalizes the construct of storiability that we emphasize as the authoring objective. 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. 23 English-speaking authors recruited from Amazon Mechnical Turk (AMT) participated in this task. The result was a dataset of 109 *authoring blocks* balanced between easy and hard prompts. Each block consisted of a prompt shown to the author, the two sentences they wrote before observing the generated examples (PRE), the two sentences they wrote after the observing the generated examples (POST), and the five generated examples they saw (GEN). An example of an authoring block is shown in Table 1 in the appendix.

4 Outcome

We then conducted a judgment task to evaluate readers' perceived storiability of the sentences in the authoring blocks. We gathered *judgment groups* from the blocks, where each judgment group consisted of a randomly ordered PRE, POST, and GEN sentence aligned to the same prompt and author. Raters observed these judgment groups and were told to "imagine that each sentence [in the group] is an excerpt from a story and pick the one that makes you most want to read that story". This instruction is consistent with the objective the authors were originally given. 16 AMT participants rated subsets of judgment groups, yielding a total of 1,744 responses. For our analysis, we labeled the sentence selected by the rater in each group as "Preferred" and the other sentences in the same group as "Not Preferred". The resulting distribution of preferences (Table 2) indicated that while raters dramatically preferred human-authored sentences over the generated ones, they favored the sentences people wrote after observing the generated examples (POST) compared with those written before (PRE). However, this pattern was only significant[2] for prompts with a "hard" difficulty level (Table 3). Hard items contained significantly more infilled tokens compared to easy items, which validates the difference between these conditions (Table 4). Together this suggests that observing the generated text was more impactful when more authoring effort was required, and thus we focused our subsequent analyses on the hard items. We used DistilBERT to measure the vector cosine similarity between the authors' sentences and the corresponding generated examples they saw. We found that the similarity between the POST and GEN sentences was significantly higher than the similarity between the PRE and GEN sentences (Table 5). This confirms that authors were influenced by the content in the GEN examples. Moreover, POST sentences preferred as more storiable were also significantly more likely to be influenced by the GEN examples (Table 6). This ultimately shows that these examples helped authors better fulfill the storiability authoring objective. Future work can explore this "inspiration through observation" paradigm for other authoring tasks and objectives.

[2]For all results, statistical significance was marked by a two-sample Monte Carlo permutation test at $p < 0.05$

5 Ethics

Our generation model is based on GPT-2, which can sometimes produce text deemed offensive for various reasons [e.g. Gehman et al., 2020]. To protect participants from this, we manually filtered items from the authoring task where we considered the prompt or generated sentence to be offensive.

References

Daphne Ippolito, David Grangier, Chris Callison-Burch, and Douglas Eck. Unsupervised hierarchical story infilling. In *Proceedings of the First Workshop on Narrative Understanding*, 2019.

Yusuke Mori, Hiroaki Yamane, Yusuke Mukuta, and Tatsuya Harada. 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*, 2020.

Chris Donahue, Mina Lee, and Percy Liang. Enabling language models to fill in the blanks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2492–2501. Association for Computational Linguistics, July 2020.

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

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Sosuke Kobayashi. Homemade bookcorpus. github.com/BIGBALLON/cifar-10-cnn, 2018.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, 2020.

A Appendix

Prompt: nose, pushed, see (*difficulty*=hard)

PRE Sentences	POST Sentences	GEN Examples
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: Example of an authoring block. A block consists of sentences written by a single author before (PRE) and after (POST) observing the generated (GEN) example sentences.

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

Table 2: Distribution of storiability preferences

Difficulty	Preferred PRE	Preferred POST
easy	0.384	0.354
hard	0.329	0.375

Table 3: Distribution of storiability preferences for human-authored sentences by difficulty

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

Condition	Similarity
PRE	0.921
POST	0.923

Table 5: Similarity between human and generated sentences before (PRE) and after (POST) observation of GEN examples

Judgment	Similarity
Not Preferred	0.922
Preferred	0.925

Table 6: Similarity between POST and GEN sentences (i.e. degree of semantic influence) according to storiability preferences