Plan, Write, and Revise: an Interactive System for Open-Domain Story Generation

Seraphina Goldfarb-Tarrant[1,2], **Haining Feng**[1], **Nanyun Peng**[1]
[1] Information Sciences Institute, University of Southern California
[2] University of Washington
serif@uw.edu, haining@usc.edu, npeng@isi.edu

Abstract

Story composition is a challenging problem for machines and even for humans. We present a neural narrative generation system that interacts with humans to generate stories. Our system has different levels of human interaction, which enables us to understand at what stage of story-writing human collaboration is most productive, both to improving story quality and human engagement in the writing process. We compare different varieties of interaction in *story-writing*, *story-planning*, and *diversity controls* under time constraints, and show that increased types of human collaboration at both planning and writing stages results in a 10-50% improvement in story quality as compared to less interactive baselines. We also show an accompanying increase in user engagement and satisfaction with stories as compared to our own less interactive systems and to previous turn-taking approaches to interaction. Finally, we find that humans tasked with collaboratively improving a particular characteristic of a story are in fact able to do so, which has implications for future uses of human-in-the-loop systems.

1 Introduction

Collaborative human-machine story-writing has had a recent resurgence of attention from the research community (Roemmele and Swanson., 2017; Clark and Smith, 2018). It represents a frontier for AI research; as a research community we have developed convincing NLP systems for some generative tasks like machine translation, but lag behind in creative areas like open-domain storytelling. Collaborative open-domain storytelling incorporates human interactivity for one of two aims: to improve human creativity via the aid of a machine, or to improve machine quality via the aid of a human. Previously existing approaches treat the former aim, and have shown that storytelling systems are not yet developed enough to help human writers. We attempt the latter, with the goal of investigating at what stage human collaboration is most helpful.

Swanson and Gordon (2009) use an information retrieval based system to write by alternating turns between a human and their system. Clark and Smith (2018) use a similar turn-taking approach to interactivity, but employ a neural model for generation and allow the user to edit the generated sentence before accepting it. They find that users prefer a full-sentence collaborative setup (vs. shorter fragments) but are mixed with regard to the system-driven approach to interaction. Roemmele and Swanson. (2017) experiment with a user-driven setup, where the machine doesn't generate until the user requests it to, and then the user can edit or delete at will. They leverage user-acceptance or rejection of suggestions as a tool for understanding the characteristics of a helpful generation. All of these systems involve the user in the *story-writing* process, but lack user involvement in the *story-planning* process, and so they lean on the user's ability to knit a coherent overall story together out of locally related sentences. They also do not allow a user to control the novelty or "unexpectedness" of the generations, which Clark and Smith (2018) find to be a weakness. Nor do they enable iteration; a user cannot revise earlier sentences and have the system update later generations. We develop a system[1] that allows a user to interact in all of these ways that were limitations in previous systems; it enables involvement in planning, editing, iterative revising, and control of novelty. We conduct experiments to understand which types of interaction are most effective for improving stories and for making users satisfied and engaged.

We have two main interfaces that enable hu-

[1]The live demo is at http://cwc-story.isi.edu, with a video at https://youtu.be/-hGd2399dnA. Code and models are available at https://github.com/seraphinatarrant/plan-write-revise.

Proceedings of NAACL-HLT 2019: Demonstrations, pages 89–97
Minneapolis, Minnesota, June 2 - June 7, 2019.

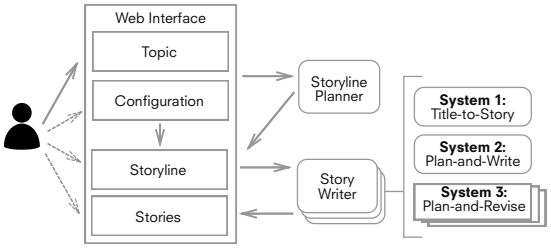


Figure 1: Diagram of human-computer interaction mediated by the the demo system. The dotted arrows represent optional interactions that the user can take. Depending on the set-up, the user may choose to interact with one or all story models.

man interaction with the computer. There is *cross-model* interaction, where the machine does all the composition work, and displays three different versions of a story written by three distinct models for a human to compare. The user guides generation by providing a topic for story-writing and by tweaking decoding parameters to control novelty, or *diversity*. The second interface is *intra-model* interaction, where a human can select the model to interact with (potentially after having chosen it via *cross-model*), and can collaborate at all stages to jointly create better stories. The full range of interactions available to a user is: select a model, provide a topic, change diversity of content, collaborate on the planning for the story, and collaborate on the story sentences. It is entirely user-driven, as the users control how much is their own work and how much is the machine's at every stage. It supports revision; a user can modify an earlier part of a written story or of the story plan at any point, and observe how this affects later generations.

2 System Description

2.1 System Overview

Figure 1 shows a diagram of the interaction system. The dotted arrows represent optional user interactions.

Cross-model mode requires the user to enter a *topic*, such as "the not so haunted house", and can optionally vary the *diversity* used in the STORYLINE PLANNER or the STORY WRITER. *Diversity* numbers correspond directly to softmax temperatures, which we restrict to a reasonable range, determined empirically. The settings are sent to the STORYLINE PLANNER module, which generates a storyline for the story in the form of a sequence of phrases as per the method of Yao et al. (2019). Everything is then sent to the STORY WRITER, which will return three stories.

Intra-model mode enables advanced interactions with one story system of the user's choice. The STORYLINE PLANNER returns either one storyline phrase or many, and composes the final storyline out of the combination of phrases the system generated, the user has written, and edits the user has made. These are sent to the STORY WRITER, which returns either a single sentence or a full story as per user's request. The process is flexible and iterative. The user can choose how much or little content they want to provide, edit, or re-generate, and they can return to any step at any time until they decide they are done.

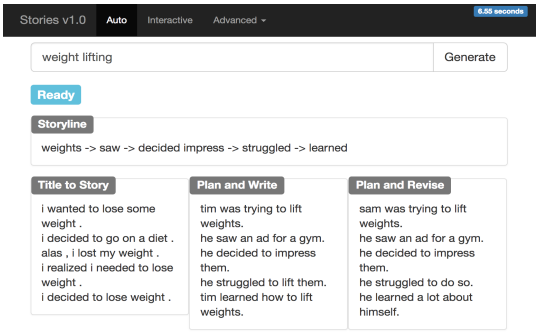
Pre-/Post-processing and OOV handling To enable interactive flexibility, the system must handle open-domain user input. User input is lowercased and tokenized to match the model training data via spaCy[2]. Model output is naively detokenized via Moses (Koehn et al., 2007) based on feedback from users that this was more natural. User input OOV handling is done via WordNet (Miller, 1995) by recursively searching for hypernyms and hyponyms (in that order) until either an in-vocabulary word is found or until a maximum distance from the initial word is reached.[3] We additionally experimented with using cosine similarity to GloVe vectors (Pennington et al., 2014), but found that to be slower and not qualitatively better for this domain.

2.2 Web Interface

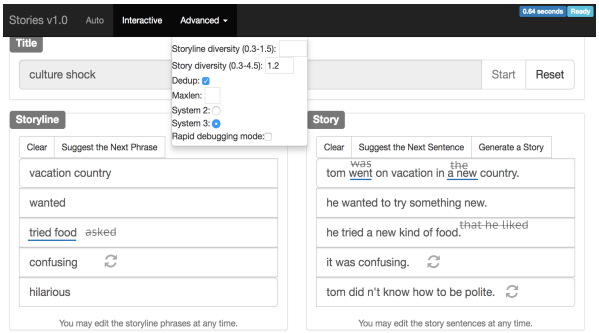
Figure 2 shows screenshots for both the *cross-model* and *intra-model* modes of interaction. Figure 2a shows that the *cross-model* mode makes clear the differences between different model generations for the same topic. Figure 2b shows the variety of interactions a user can take in *intra-model* interaction, and is annotated with an example-in-action. User inserted text is underlined in blue, generated text that has been removed by the user is in grey strike-through. The *refresh* symbol marks areas that the user re-generated to get a different sentence (presumably after being unhappy with the first result). As can be seen in this example, minor user involvement can result in a significantly better story.

[2] spacy.io

[3] *distance* is difference og levels in the WordNet hierarchy, and was set empirically to 10.



(a) cross-model interaction, comparing three models with advanced options to alter the storyline and story diversities.



(b) intra-model interaction, showing advanced options and annotated with user interactions from an example study.

Figure 2: Screenshots of the demo user interface

2.3 Model Design

All models for both the STORYLINE PLANNER and STORY WRITER modules are conditional language models implemented with LSTMs based on Merity et al. (2018). These are 3-stacked LSTMs that include weight-dropping, weight-tying, variable length back propagation with learning rate adjustment, and Averaged Stochastic Gradient Descent (ASGD). They are trained on the ROC dataset (Mostafazadeh et al., 2016), which after lowercasing and tokenization has a vocabulary of 38k. Storyline Phrases are extracted as in Yao et al. (2019) via the RAKE algorithm (Rose et al., 2010) which results in a slightly smaller Storyline vocabulary of 31k. The STORYLINE PLANNER does decoding via sampling to encourage creative exploration. The STORY WRITER has an option to use one or all three systems, all of which decode via beamsearch and are detailed below.

The *Title-to-Story* system is a baseline, which generates directly from topic.

The *Plan-and-Write* system adopts the static model in Yao et al. (2019) to use the storyline to supervise story-writing.

Plan-and-Revise is a new system that combines the strengths of Yao et al. (2019) and Holtzman et al. (2018). It supplements the Plan-and-Write model by training two discriminators on the ROC data and using them to re-rank the LSTM generations to prefer increased *creativity* and *relevance*.[4] Thus the decoding objective of this system becomes $f_\lambda(x, y) = log(P_{lm}(y|x)) + \sum_k \lambda_k s_k(x, y)$ where P_{lm} is the conditional language model probability of the LSTM, s_k is the discriminator scoring function, and λ_k is the learned weight of that discriminator. At each timestep all live beam hypotheses are scored and re-ranked. Discriminator weights are learnt by minimizing Mean Squared Error on the difference between the scores of gold standard and generated story sentences.

3 Experiments

We experiment with six types of interaction: five variations created by restricting different capabilities of our system, and a sixth turn-taking baseline that mimics the interaction of the previous work (Clark and Smith, 2018; Swanson and Gordon, 2009). We choose our experiments to address the research questions: What type of interaction is most engaging? Which type results in the best stories? Can a human tasked with correcting for certain weaknesses of a model successfully do so? The variations on interactions that we tested are:

1. Machine only: no human-in-loop.
2. Diversity only: user can compare and select models but only diversity is modifiable.
3. Storyline only: user collaborates on storyline but not story.
4. Story only: user collaborates on story but not storyline.
5. All: user can modify everything.
6. Turn-taking: user and machine take turns writing a sentence each (user starts). user can edit the machine-generations, but once they move on to later sentences, previous sentences are read-only.[5]

We expand experiment 5 to answer the question of whether a human-in-the-loop interactive sys-

[4] Holtzman et al. (2018) use four discriminators, but based on ablation testing we determined these two to perform best on our dataset and for our task.

[5] This as closely matches the previous work as possible with our user interface. This model does not use a storyline.

tem can address specific shortcomings of generated stories. We identify three types of weaknesses common to generation systems – *Creativity*, *Relevance*, and *Causal & Temporal Coherence*, and conduct experiments where the human is instructed to focus on improving specifically one of them. The targeted human improvement areas intentionally match the *Plan-and-Revise* discriminators, so that, if successful, the "human discriminator" data can assist in training the machine discriminators. All experiments (save experiment 2, which lets the user pick between models) use the *Plan-and-Revise* system.

3.1 Details

We recruit 30 Mechanical Turk workers per experiment (270 unique workers total) to complete story writing tasks with the system.[6] We constrain them to ten minutes of work (five for writing and five for a survey) and provide them with a fixed *topic* to control this factor across experiments. They co-create a story and complete a questionnaire which asks them to self-report on their engagement, satisfaction, and perception of story quality.[7] For the additional focused error-correction experiments, we instruct Turkers to try to improve the machine-generated stories with regard to the given aspect, under the same time constraints. As an incentive, they are given a small bonus if they are later judged to have succeeded.

We then ask a separate set of Turkers to rate the stories for overall quality and the three improvement areas. All ratings are on a five-point scale. We collect two ratings per story, and throw out ratings that disagree by more than 2 points. A total of 11% of ratings were thrown out, leaving four metrics across 241 stories for analysis.

4 Results

User Engagement Self-reported scores are relatively high across the board, as can be seen in Table 1, with the majority of users in all experiments saying they would like to use the system again. The lower scores in the *Diversity only* and *Storyline only* experiments are elucidated by qualitative comments from users of frustration at the inability to sufficiently control the generations with influence over only those tools. *Storyline only* is lowest for *Use Again*, which can be explained by the model behavior when dealing with unlikely storyline phrases. Usually, the most probable generated story will contain all storyline phrases (exact or similar embeddings) in order, but there is no mechanism that strictly enforces this. When a storyline phrase is uncommon, the story model will often ignore it. Many users expressed frustration at the irregularity of their ability to guide the model when collaborating on the storyline, for this reason.

Experiment	E	Q	S	Use Again
Diversity only	3.77	2.90	3.27	1.40
Storyline only	4.04	3.36	3.72	1.27
Story only	**4.50**	3.17	3.60	1.60
All	4.41	3.55	3.76	1.55
All + Creative	4.00	3.27	3.70	**1.70**
All + Relevant	4.20	3.47	3.83	1.57
All + C-T	4.30	**3.77**	**4.30**	1.53
Turn-taking	4.31	3.38	3.66	1.52

Table 1: User self-reported scores, from 1-5. E: Entertainment value, Q: Quality of Story, S: Satisfaction with Story. Note that the final column *Use Again* is based on converting "no" to 0, "conditional" to 1, and "yes" to 2.

Experiment	Overall	Creative	Relevant	C-T
Machine	2.34	2.68	2.46	2.54
Diversity only	2.50	2.96	2.75	2.81
Storyline only	3.21	3.27	3.88	3.65
Story only	3.70*	**4.04***	3.96*	**4.24***
All	3.54	3.62	3.93*	3.83
All + Creative	**3.73***	3.96*	3.98*	3.93*
All + Relevant	3.53*	3.52	4.05	3.91*
All + C-T	3.62*	3.88*	4.00*	3.98*
Turn-taking	3.55*	3.68	**4.27***	3.81

Table 2: Results for all experiments, from 1-5. Best scores per metric are bolded, scores not significantly different ($\alpha = 0.1$, per Wilcoxon Signed-Rank Test) are starred. C-T stands for Causal-Temporal Coherence, the **+** experiments are the extensions where the user focuses on improving a particular quality.

Users were engaged by collaboration; all experiments received high scores on being entertaining, with the collaborative experiments rated more highly than *Diversity only*. The pattern is repeated for the other scores, with users being more satisfied and feeling their stories to be higher quality for all the more interactive experiments. The *Turn-taking* baseline fits into this pattern; users prefer it more than the less interactive *Diversity only* and *Storyline only*, but often (though not always) less than the more interactive *Story only, All, All+* experiments. Interestingly, user perception of the quality of their stories does not align well with independent rankings. Self-reported quality is low

[6] We enforce uniqueness to prevent confounding effects from varying levels of familiarity with the demo UI

[7] Text of questionnaire and other Mechanical Turk materials are included in Appendix C

in the *Story only* experiment, which contrasts with it being highest rated independently (as discussed below). Self-reported scores also suggest that users judge their stories to be much better when they have been focusing on causal-temporal coherence, though this focus carries over to a smaller improvement in independent rankings. While it is clear that additional interactivity is a good idea, the disjunct between user perception of their writing and reader perception under different experiment conditions is worthwhile to consider for future interactive systems.

Story Quality As shown in Table 2, human involvement of *any kind* under tight constraints helps story quality across all metrics, with mostly better results the more collaboration is allowed. The exception to this trend is *Story only* collaboration, which performs best or close to best across the board. This was unexpected; it is possible that these users benefited from having to learn to control only *one* model, instead of both, given the limited time. It is also possible that being forced to be reliant on system storylines made these users more creative.

Turn-taking Baseline The turn-taking baseline performs comparably in overall quality and relevance to other equally interactive experiments (*Story only*, *All, All+*). It achieves highest scores in relevance, though the top five systems for relevance are not statistically significantly different. It is outperformed on creativity and causal-temporal coherence by the strong *Story only* variation, as well as the *All, All+* systems. This suggests that local sentence-level editing is sufficient to keep a story on topic and to write well, but that creativity and causal-temporal coherence require some degree of global cohesion that is assisted by iterative editing. The same observation as to the strength of *Story only* over *All* applies here as well; turn-taking is the least complex of the interactive systems, and may have boosted performance from being simpler since time was constrained and users used the system only once. Thus a turn-based system is a good choice for a scenario where users use a system infrequently or only once, but the comparative performance may decrease in future experiments with more relaxed time constraints or where users use the system repeatedly.

Targeted Improvements The results within the *All* and *All +* setups confirm that stories can be improved with respect to a particular metric. The diagonal of strong scores displays this trend, where the creativity-focused experiment has high creativity, etc. An interesting side effect to note is that focusing on *anything* tends to produce better stories, reflected by higher overall ratings. *All + Relevance* is an exception which does not help creativity or overall (perhaps because relevance instantly becomes very high as soon a human is involved), but apart from that *All +* experiments are better across all metrics than *All*. This could mean a few things: that when a user improves a story in one aspect, they improve it along the other axes, or that users reading stories have trouble rating aspects entirely independently.

5 Conclusions and Future Work

We have shown that all levels of human-computer collaboration improve story quality across all metrics, compared to a baseline computer-only story generation system. We have also shown that flexible interaction, which allows the user to return to edit earlier text, improves the specific metrics of creativity and causal-temporal coherence above previous rigid turn-taking approaches. We find that, as well as improving story quality, more interaction makes users more engaged and likely to use the system again. Users tasked with collaborating to improve a specific story quality were able to do so, as judged by independent readers.

As the demo system has successfully used an ensemble of collaborative discriminators to improve the same qualities that untrained human users were able to improve even further, this suggests promising future research into human-collaborative stories as training data for new discriminators. It could be used both to strengthen existing discriminators and to develop novel ones, since discriminators are extensible to arbitrarily many story aspects.

Acknowledgments

We thank the anonymous reviewers for their feedback, as well as the members of the PLUS lab for their thoughts and iterative testing. This work is supported by Contract W911NF-15- 1-0543 with the US Defense Advanced Research Projects Agency (DARPA).

References

Anne Spencer Ross Chenhao Tan Yangfeng Ji Clark, Elizabeth and Noah A. Smith. 2018. Creative writing with a machine in the loop: Case studies on slogans and stories. In *23rd International Conference on Intelligent User Interfaces (IUI)*.

Ari Holtzman, Jan Buys, Maxwell Forbes, Antoine Bosselut, David Golub, and Yejin Choi. 2018. Learning to write with cooperative discriminators. In *Proceedings of ACL*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2018. Regularizing and optimizing lstm language models. In *Proceedings of the Sixth International Conference on Learning Representations (ICLR)*.

George A. Miller. 1995. Wordnet: A lexical database for english. In *Communications of the ACM Vol. 38*.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of NAACL-HLT*, pages 839–849.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Andrew S. Gordon Roemmele, Melissa and Reid Swanson. 2017. Evaluating story generation systems using automated linguistic analyses. In *SIGKDD 2017 Workshop on Machine Learning for Creativity*.

Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. Automatic keyword extraction from individual documents. In *Text Mining: Applications and Theory*.

Reid Swanson and Andrew S. Gordon. 2009. Say anything: A demonstration of open domain interactive digital storytelling. In *Joint International Conference on Interactive Digital Storytelling*.

Lili Yao, Nanyun Peng, Weischedel Ralph, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In *Proceedings of the thirty-third AAAI Conference on Artificial Intelligence (AAAI)*.

A Demo Video

The three-minute video demonstrating the interaction capabilities of the system can be viewed at https://youtu.be/-hGd2399dnA. (Same video as linked in the paper footnote).

B Training and Decoding Parameters

B.1 Decoding

Default diversity (Softmax Temperature) for Storyline Planner is *0.5*, for Story Writer it is *None* (as beamsearch is used an thus can have but does not require a temperature). Beam size for all Story Writer models is *5*. Additionally, Storyline Phrases are constrained to be unique (unless a user duplicates them), and Beamsearch is not normalized by length (both choices determined empirically).

B.2 Training

We follow the parameters used in Yao et al. (2019) and Merity et al. (2018).

Parameter	Storyline Model	Story Models
Embedding Dim	500	1000
Hidden Layer Dim	1000	1500
Input Embedding Dropout	0.4	0.2
Hidden Layer Dropout	0.1	01
Batch Size	20	20
BPTT	20	75
Learning Rate	10	10
Vocabulary size	31,382	37,857
Total Model Parameters	32,489,878	80,927,858
Epochs	50	120

Table 3: Training parameters for models used in demo.

C User Study

C.1 Questionnaire

Post Story Generation Questionnaire
How satisfied are you with the final story?
What do you think is the overall quality of the final story?
Was the process entertaining?
Would you use the system again?

Table 4: Questionnaire for user self-reporting, range 1 to 5 (1 low).

C.2 Mechanical Turk Materials

Following are examples of the materials used in doing Mechanical Turk User Studies. Figure 3 is an example of the *All + Creative* focused experiment for *story-writing*. The instructions per experiment differ across all, but the template is the same. Figure 4 is the survey for ranking stories across various metrics. This remains constant save that story order was shuffled every time to control for any effects of the order a story was read in.

Figure 3: Template & Instructions for Writing Stories in the *All + Creative* experiment.

Instructions

Hello! We are academic researchers studying different methods of story writing.

Please take a few minutes to read the instructions for our website and get used to the interface. It should be quick and will help you get a quality bonus.

The objective is for you to take about five minutes and co-write a story with our system and try to improve the Creativity of the story. Our system works by generating a storyline, and then a story based on it, and you collaborate with it.

If your final story is judged to have improved Creativity, you will get a bonus.

Note:

- **We need unique Workers, so you are only allowed to do one of these**. Please do not auto-accept the next one.
- Only **five** sentences. Please do collaborate as well - if you just write a story yourself we will know and reject the HIT.

This is an example of a story:

bobby and his friends were fascinated by the dark.
they dared each other to get close to a haunted house.
bobby heard a noise coming from the window.
he ran to the house to see what it was.
it was a scary, scary house.

Steps:

1. You will be given a **Title**
2. Go to our website http://cwc-story.isi.edu:5002/interactive.html (http://cwc-story.isi.edu:5002/interactive.html), enter the title into the text box, and click **Start**.
3. Now in the **Storyline** section you can click *Suggest the Next Phrase* or enter your own.
4. A storyline phrase is one or more words that help define the content of the corresponding sentence. For example, "apprehensive" or "fun time" or "charged sentenced prison".
5. Once you have five phrases, you can go to the **Story** section and click *Suggest the Next Sentence*, or enter your own
6. In both sections, you can switch between writing and suggesting, and you can delete, make edits to your work or the computer's work, or replace or re-suggest at any time. Do this as many times as you want.
7. *Clear* clears the entire storyline or story section.
8. If you want, under the **Advanced** menu, you can change **storyline diversity** and **story diversity** to make the system more creative. After you do, click **Reset**, enter the title, and **Start** again.
9. Take *5 minutes* and try to get the best story you can. Change any Advanced options or storyline or story phrases, but keep the *same* title.
10. When you're done, paste the sentences of your choice into the text box below, and then answer a few questions about your experience. The questions are required, but you are not judged on your answers, so answer honestly.

Details:

1. **storyline diversity** and **story diversity** both control how creative the system will be. Lower numbers mean it is more conservative, or "normal", and higher numbers mean it is more creative or "experimental". There is some element of randomness, so sometimes you'll get the same result with different numbers, or a different result with the same numbers.
2. When **diversity** is > 1, sometimes the generations can become strange (nonsense, or less than 5 sentences/phrases). Just lower the diversity and try again, or just re-suggest or add your own till it's 5, or edit it and use it as inspiration for your work.
3. Sometimes punctuation or other things come out weird, don't worry about it.
4. If you want the system to do everything for the first round, you can click *Generate a Story* and it will fill everything out so you can start changing.

Rules:

1. Stories will later be judged, and you will receive a **$0.50 bonus** if your story is judged as having improved creativity.
2. We do review **every** HIT, so if you break the rules above or have not actually tried the system, we will reject it.

Title:

${title}

1. Enter the text of your pick for final most Creative story here.

Figure 4: Template & Instructions for Ranking Stories

Survey Instructions

We are a group of researchers conducting research about storytelling.

In this survey, you will be provided with a **title** and **eight stories** about that title.
Please compare and rank them according to the given criteria.

1. Please read all the stories carefully, and give a score to each story based on: **Relevance, Creativity, Overall Quality, and Event Coherence (we will explain these).**
2. Multiple stories **can** receive the same score. However, please do read all the story versions before rating them and consider them in relation to each other.
3. Briefly **explain** why you made the decisions you did for each story. Just basic thoughts or bullet points is fine, but this is required.
4. Don't worry about punctuation and spelling of the stories, those don't matter
5. We do review **every** HIT response, and will reject it if your answers make no sense and you don't give any reasoning.

Explanation of metrics:

- **Relevance:** Is the story relevant to the title, i.e. is it on topic?
- **Creativity:** Is the story unusual and interesting, in either vocabulary or content?
- **Overall Quality:** How good do you think the story is?
- **Event Coherence:** Do the things that happen in the story make sense together and are they in the right order? For example, for the sentences: "the ice cream tasted delicious. she tasted the ice cream" the events make sense, but are in the wrong order and should get a lower rating. For the sentences: "jim's boss yelled at him. jim had a great day" the ordering is fine but the events don't go together. Events that go together and are in the right order should have a higher rating, as in: "jim's boss yelled at him. jim went home exhausted and unhappy."

An Example:

title: *haunted house*

a) bobby and his friends were fascinated by the dark. they dared each other to get close to a haunted house. bobby heard a noise coming from the window. he ran to the house to see what it was. it was a scary, scary house.

Relevance	○ **1. terrible**	○ **2. bad**	○ **3. neutral**	○ **4. good**	◉ **5. great**
Creativity	○ **1. terrible**	○ **2. bad**	○ **3. neutral**	○ **4. good**	◉ **5. great**
Event Coherence	○ **1. terrible**	○ **2. bad**	○ **3. neutral**	◉ **4. good**	○ **5. great**
Overall Quality	○ **1. terrible**	○ **2. bad**	○ **3. neutral**	◉ **4. good**	○ **5. great**

Briefly explain why you made your decisions:

The story was on topic, and made sense, and was good, but the last sentence didn't really fit perfectly with earlier events.