Creative Writing with a Machine in the Loop: Case Studies on Slogans and Stories

Elizabeth Clark[1], Anne Spencer Ross[1], Chenhao Tan[2], Yangfeng Ji[1], & Noah A. Smith[1]
[1]Paul G. Allen School for Computer Science & Engineering, University of Washington, Seattle, WA, USA
[2]Department of Computer Science, University of Colorado Boulder, Boulder, CO, USA
eaclark7@cs.washington.edu, ansross@cs.washington.edu, chenhao@chenhaot.com,
yangfeng@cs.washington.edu, nasmith@cs.washington.edu

ABSTRACT

As the quality of natural language generated by artificial intelligence systems improves, writing interfaces can support interventions beyond grammar-checking and spell-checking, such as suggesting content to spark new ideas. To explore the possibility of machine-in-the-loop creative writing, we performed two case studies using two system prototypes, one for short story writing and one for slogan writing. Participants in our studies were asked to write with a machine in the loop or alone (control condition). They assessed their writing and experience through surveys and an open-ended interview. We collected additional assessments of the writing from Amazon Mechanical Turk crowdworkers. Our findings indicate that participants found the process fun and helpful and could envision use cases for future systems. At the same time, machine suggestions do not necessarily lead to better written artifacts. We therefore suggest novel natural language models and design choices that may better support creative writing.

Author Keywords

natural language processing; machine in the loop; creative writing

ACM Classification Keywords

H.5.2 Information interfaces and presentation (e.g., HCI): User Interfaces: Natural Language

INTRODUCTION

Researchers have made significant progress in advancing artificial intelligence since Turing famously asked "Can machines think?" [32]. Although the focus of artificial intelligence has been on improving the capabilities of machines, e.g., through the use of machine learning, we propose a *machine-in-the-loop* framework, where the goal is to improve the ability of humans, with the machine playing

a supporting role.[1] Humans are the central actor and have full agency in deciding what to do with machine outputs. Accordingly, machine learning models should be designed to assist humans.

This paper explores the possibility of incorporating a machine in the loop of creative writing. The motivation is twofold. First, writers often experience "cognitive inertia," a phenomenon known in the writing domain as "writer's block" [11]. A collaborator who provides suggestions and points out new directions might help alleviate writer's block. The new combination of a writer's own ideas with suggested ideas is a form of psychological creativity [3]. Second, recent studies show that machines outperform humans in some tasks [15, 24, 31], including identifying which message will be retweeted more on Twitter. Perhaps a machine-learned algorithm can provide valuable suggestions to writers. We explore the space of designing machine-in-the-loop systems for creative writing and learn insights from user studies that can inform future interface design and research on natural language processing models.

Machines can support writers as they edit, structure, and refine their work, as demonstrated by word processors, grammar and spell checkers, version control, or even language or style analysis tools (such as the Hemingway Editor[2]). We focus on systems that assist people by suggesting content as they write and that are designed to inspire creativity throughout the writing process while still leaving writers in control of the final written artifact. In particular, we investigate the following questions:

- How can we design machine-in-the-loop systems to support diverse writing tasks and processes?
- What effect do these systems have on people's writing, both as perceived by the writer and by other people?
- What do people want to see in machine-generated suggestions and creative writing support systems?

To answer these questions, we developed two prototype systems to help writers enhance their creativity in two tasks:

---

[1]In contrast, human-in-the-loop machine learning actively includes humans in the process of training machine learning models by asking humans to provide feedback such as labeling difficult examples or suggesting new features [4, 6, 9, 20].
[2]http://www.hemingwayapp.com

story writing and slogan writing. These two creative tasks have different goals and require different writing styles. Thus, the systems that assist in these tasks should provide different types of help. We built two system prototypes that use two different models to generate suggestions. We had study participants write with these prototypes and compared their experiences and the quality of their writing to that of participants who did not receive suggestions. This paper discusses the current capabilities of machine-in-the-loop writing systems and suggests improvements both for system interfaces and models.

Although providing helpful suggestions is important in a machine-in-the-loop writing system, we leave writers with complete editorial control and the freedom to disregard any unwanted suggestions. Our goal is *not* to replace human creativity or automate creative writing; rather, we seek to amplify people's creativity by providing suggestions that are most useful to them. By offering suggestions as a person writes, the writing process has elements of both collaborative writing and constrained writing tasks. It also provides a versatile setup; a machine-in-the-loop writing system could be used as an educational tool, a writer's tool, or for entertainment.

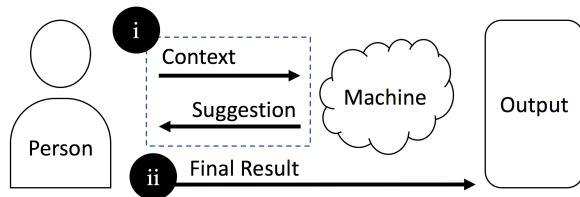## MACHINE-IN-THE-LOOP SYSTEM CHARACTERISTICS

Figure 1: Machine-in-the-loop system structure: the loop (i) is initiated with the person providing context and the machine responding with a suggestion. The person always has control over the final result (ii).

This paper considers machine-in-the-loop (MIL) systems that are composed of a person and a machine working together to create output (Figure 1). The person and machine are in a loop in which the person provides context and the machine responds with suggestions (Figure 1 i). The person controls the final output (Figure 1 ii).[3] Different creative writing tasks have different demands for MIL systems. We consider the following three characteristics for designing MIL systems: interaction structure, interaction initiation, and interaction intrusiveness. We use our story and slogan writing tasks as examples to explain each aspect, as summarized in Table 1.

Interaction Structure
Interaction structure can be *iterative*, where a writer works with the help of a machine to refine a single idea, or *additive*, where the writer and machine work to add multiple ideas

---

[3]The setup this paper explores does not represent all possible configurations for writing with a machine. More complex setups might involve the person and the machine working in parallel or the machine directly altering the output in a mixed-initiative fashion.

	Story writing	Slogan writing
interaction structure	additive	iterative
interaction initiation	push	pull
interaction intrusiveness	high	low

Table 1: Characteristics of MIL systems for our story and slogan writing tasks.

together. This can be represented as how many times the loop of the person and machine exchanging context and suggestions (see section "i" in Figure 1) is repeated before the person commits to a final result. For example, story writing is additive because as a story unfolds, writers (and machines) need to introduce new ideas, and these ideas are combined into a final story. Slogan writing, by contrast, is a highly iterative process: the loop is repeated for a single phrase or sentence until a final slogan is decided upon.

Interaction Initiation
Interaction initiation refers to how the context-suggestion loop (see section "i" in Figure 1) is triggered. It can follow a *push* (automatically initiated) or *pull* (person-initiated) method of initiation, or a combination of the two. To have breadth of exploration, we implemented the story writing system as a push-style system. At every other sentence, the machine presents a suggestion to the writer. The slogan system uses the pull method for retrieving suggestions. Writers provide a slogan-in-progress and keywords and prompt from the system whenever they want new suggestions.

Interaction Intrusiveness
Interaction intrusiveness describes the extent to which computer-generated suggestions are ignorable. Although writers can always edit or reject suggestions, some require more attention than others. We designed the story writing system's suggestions to be highly intrusive. They appear directly in the text box where the person is writing, and the writer must interact with the suggestion (even if only to delete it entirely) before moving on in the writing process (see Figure 2). In the slogan writing system, suggestions have low intrusiveness. Suggestions appear in a separate column from the writing space and require no interaction once they are retrieved (see Figure 4).

RELATED WORK
We propose "machine-in-the-loop" in contrast with "human-in-the-loop" machine learning. This concept resonates with "mixed-initiative user interfaces" [17]. Although Horvitz emphasizes the development of user interfaces in settings where both the human and the computer can drive towards a shared goal (as opposed to our human-driven setup), many of the principles he considers are relevant to this work, including the timing of machine contributions, providing editing capabilities, and understanding the social expectations of collaborators [17]. As in mixed-initiative interface work, the goal of our work is to explore interaction paradigms and to combine human and computational strengths to enhance human ability [1]. The mixed-initiative setup has been used for creative tasks such as game design [34], and adapting

our systems to a more complex mixed-initiative setup (e.g., dynamically deciding when to offer suggestions and what format of suggestion would be most helpful) is a promising future direction.

Several tools have been developed to provide suggestions to assist people in writing, both within the research community [30, 26] and as personal projects [27]. Swanson and Gordon's "Say Anything" [30] provides suggestions for writing short stories by prompting writers with full sentences retrieved from a database of stories scraped from the web after every turn of writing. In "Creative Help" [26], writers are offered suggestions as they write stories, but only when they explicitly request them (i.e., a pull method of interaction). While these systems retrieve their sentences from existing stories, we use natural language generation to provide suggestions.

Author Robin Sloan created a sentence completion story-writing assistant tool that suggests the end to a partially-written sentence when prompted by the writer [27]. The focus of Sloan's project is on how to provide suggestions to the writer. Our work focuses on the role these suggestions play in the writing process, the interface, and people's interaction with the system. Past collaborative creativity research has also looked at other related writing tasks include headlines for newspaper articles [12] and lines for poetry [14], and other artistic domains like music [16] and dance [18].

Finally, there is work on collaborative writing systems that bring together a group of people to write collaboratively. For example, "Ensemble" was a system that had multiple participants work together to write a single story [21]. Each story had a lead author and contributors who submitted alternate versions of a scene and voted on alternatives they liked. The lead author ultimately had the authority to choose the winning scene. The person in our work plays a similar role to the lead author in Ensemble; they control the direction of the story and decide how to incorporate the external (system) input. Similarly, Soylent [2] uses crowdsourcing to provide assistance to writers. However, Soylent assists in shortening text and editing grammar rather than providing content and new ideas.

STORY WRITING SYSTEM
Our first system explores writing to expand a visual prompt into a story. The setup is inspired by *Exquisite Corpse*, a game played by Surrealist artists in which people take turns contributing to a drawing [5]. The portions of the drawing that were completed in previous turns are partially or completely hidden from the current artist, resulting in silly and bizarre drawings. A parallel version of the game exists in literature, where each player writes a sentence of a story, folds the paper over to hide all but the most recent round of writing, and then passes it to the next player. With only two players (as in our setting), hiding earlier rounds has no effect (every sentence was written or seen by the person).

We use the turn-taking aspect of the Exquisite Corpse game to help foster creativity while writing. We provide people with machine suggestions to encourage stories that are unexpected, unusual, and novel, all of which are characteristics of creativity

[29]. These machine suggestions may surprise writers, sway them to change their own ideas about the direction of the story, and include ideas they may not have thought of.

User Study Task
For our story system user study, the participant is prompted to write a ten sentence story based on an unlabeled, single-panel cartoon from *The New Yorker* caption contest[4] (in the space indicated in Figure 2, section a). The task was either done alone (solo condition) or partnered with machine suggestions (MIL condition). Both versions were done through a web interface, presented in Figure 2. Participants entered the story sentence by sentence (Figure 2, section c). Once submitted, a sentence could not be edited. The complete story, along with the number of sentences completed, appeared at the bottom of interface (Figure 2, section d).

In the solo case, no additional prompting or interactions were provided beyond the image. In the MIL condition, the participant began by writing the first sentence. Once the sentence was submitted, the next sentence would be generated by the machine and "pushed" to the participant, appearing directly in the submission box (Figure 2, section c). Full sentences were used based on a preliminary study showing that people liked full-sentence suggestions as much or more than partial sentences or keywords when writing stories. The person was free to edit the machine-suggested sentence to any extent, including deleting it entirely, before submitting it. The third sentence was again written by the participant alone. This turn-taking continued until the story was 10 sentences long. Our demo system is available at http://bit.ly/iui-story-demo.

Computational Model for Suggestions
Our computational model for suggesting a sentence given preceding text is built on a neural language model. Neural language models have been used for various natural language generation tasks, including image captioning [33], conversational modeling [28], and poetry generation [13]. To make the generated sentences fluent and coherent in context, our language model uses contextual information both within and across sentence boundaries.

Neural language models are effective at predicting words that fit well in a context locally, thereby generating coherent sentences. A sentence-level language model [23] defines the probability distribution over the next word, from within a predefined vocabulary, conditioned on the left context (i.e., the context of the generated sentence so far). Figure 3 shows an example of next-word probabilities given "He arrived in the". To complete this sentence, the model's generation process randomly picks the next word based on its probability. In the example, the word "car" will have a higher chance of being generated than the other eight words. It is not a coincidence that the generation process tends to pick a word that fits well into the left context both syntactically and semantically. Language models are able to learn some syntactic and semantic patterns automatically from the data. The language model for our story writing system was trained on 390 adventure novels (about 400 million words) from the

Partnered story writing

Write a short story, using the image as a prompt to help you get started.
As you write, for every other sentence, you will receive a suggested next sentence.
You can edit the suggested sentence as much as you like (including making no edits or deleting the entire suggestion) before adding it to the story.
Add each sentence to the story individually.
Only stories containing EXACTLY 10 sentences can be submitted.
Please note that submitted stories will be recorded anonymously for an academic research project.

Add a sentence to the story:

The birds were back again.

Katy stood beside her husband staring at the gigantic animals.

"Where do you think they come from?" she asked.

Add Line to Story
Characters: 47

Click here to submit the finished story and answer evaluation questions: Submit Story

Your Story: (2 Sentences Completed)

The birds were back again.
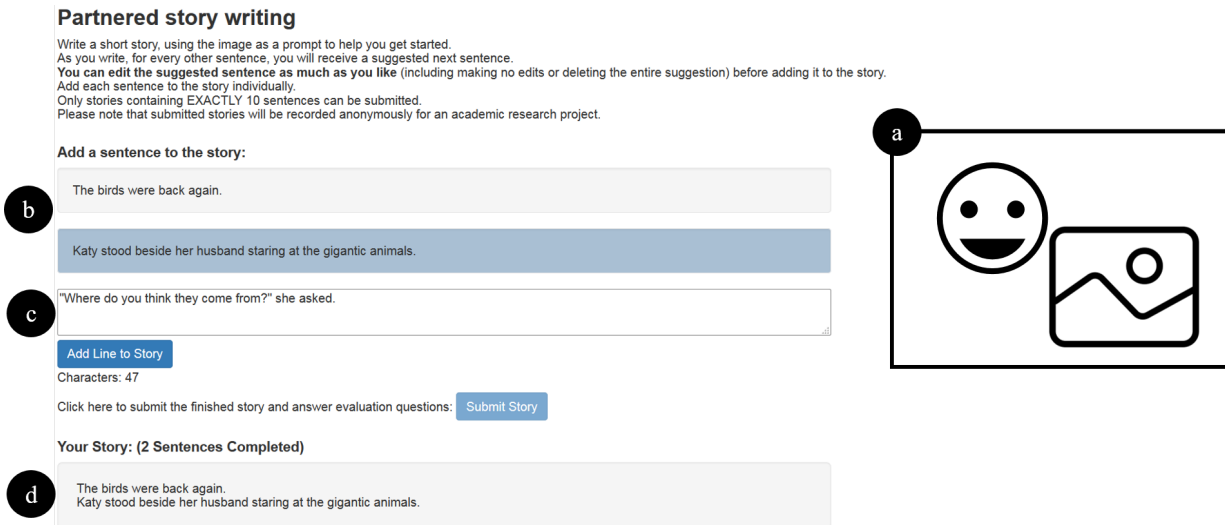Katy stood beside her husband staring at the gigantic animals.

Figure 2: The MIL story writing interface: (a) an image prompt from the New Yorker (actual image: [8]) (b) the story so far, dark colored sentence boxes were turns written with machine suggestions, (c) entry box for next sentence, the machine suggestions appeared here, (d) the full story so far. The solo condition interface was the same except no machine suggestions appeared.
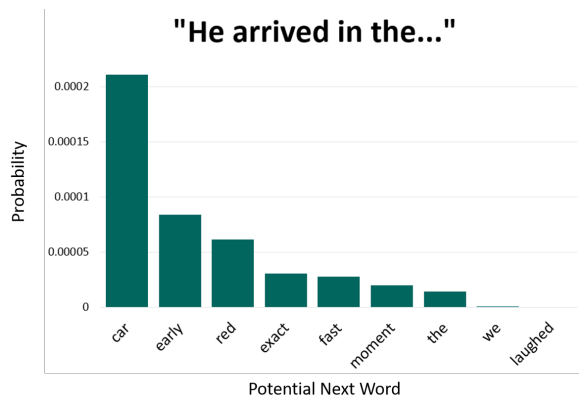
"He arrived in the..."

Figure 3: The probability distribution over words to follow "He arrived in the". The x-axis shows a small sample of the vocabulary. The full vocabulary includes more than 35,000 words extracted from adventure novels.

Toronto Book Corpus [35]. Although the model generates one word at each step, people writing with the system only see complete generated sentences.

Standard language models only use the left context within a sentence to compute the probability distribution for the next word; they ignore earlier sentences. To overcome this limitation, we adapted a neural language model that takes account of left context from both the current sentence and the previous sentence [19].

SLOGAN WRITING SYSTEM

Slogans present a challenge to writers distinct from that of writing a story: to generate a concise, memorable, and powerful statement that is representative of the object, organization, or idea it promotes and matches the intention of the authors. Slogans are used in a variety of settings, ranging from organizing a social movement to promoting a product.

The process of condensing information into a memorable and informative phrase used to create slogans is paralleled in other tasks, such as writing headlines, titling papers, and naming products. Therefore, a system that supports slogan writing could likely be extended to related tasks that prioritize catchy and succinct language.

User Study Task

For the slogan writing task, participants were asked to write a slogan for three distinct scenarios: a food packaging tool, the movie *Her*[5], and a social cause that protests animal testing for cosmetic products. The prompts included descriptions and images, from which the participant had to invent an original slogan. Like in the story writing case, the task was either done alone or partnered with a machine in the loop. In the MIL condition, participants used a web interface (see Figure 4). Participants in the solo case worked in a blank Google Doc.

When writing with the web interface, the writer must provide a few keywords and write an initial version of the slogan (Figure 4, section a). The writer can then "pull" machine suggestions at will (Figure 4, section b). Based on the writer's input, the system suggests alternative slogans (Figure 4, section c), and the history of the retrieved suggestions is tracked for future reference (Figure 4, section d). The system provides at most three suggestions in response to each request. The writer's input is on the left, and machine suggestions are on the right, reducing the intrusiveness of the suggestions. A demo system is available at http://tremoloop.com.

Computational Model for Suggestions

We developed a constrained language model that was inspired by the BRAINSUP system of Ozbal et al. [25]. First, we extract existing syntactic patterns to improve the grammaticality of the generated suggestions. Specifically,
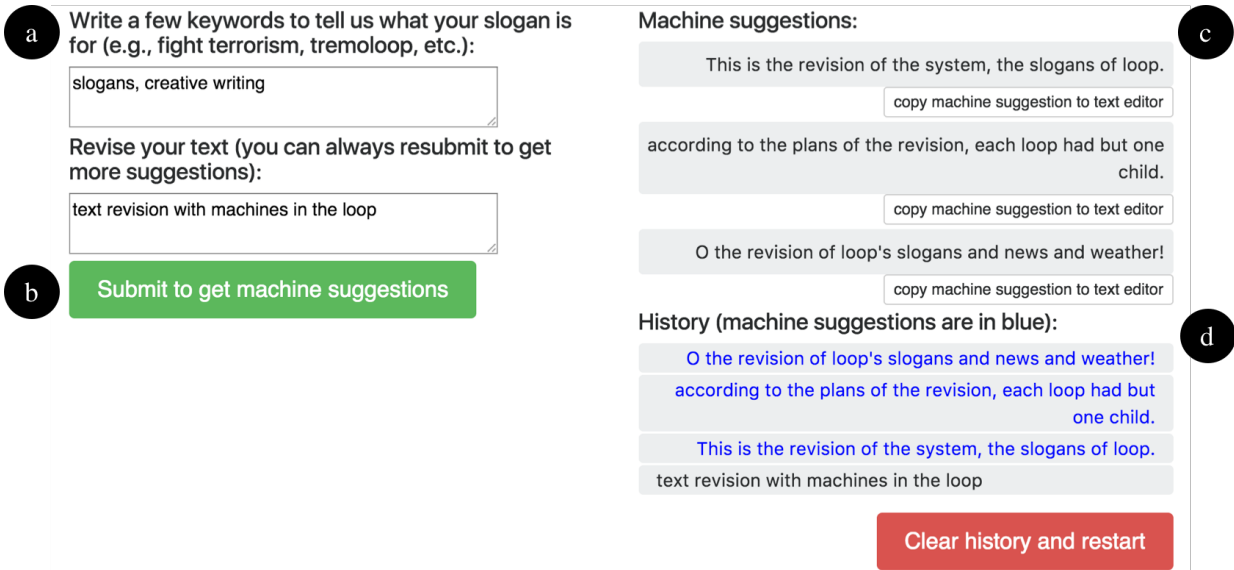
[5]http://www.imdb.com/title/tt1798709/

Figure 4: The MIL slogan writing interface: (a) the person can provide keywords and the slogan, (b) a button to "pull" suggestions, (c) the current round of suggestions, (d) the history of all slogan submissions and corresponding suggestions.

DT	JJ	NNS	VBP	RBR	JJ	IN	DT	NN	.
Some	Human-AI	interactions	are	more	profound	than	a	movie	.
Some	*	*	are	more	*	than	a	*	.

Figure 5: A slogan and its resulting skeleton. The top row shows the corresponding part-of-speech tags. This slogan was written about the movie *Her* with a machine in the loop.

we start from quotations from Wikiquotes[6] and replace all content words with a wildcard symbol. These patterns become *skeletons* for our generation system. Figure 5 shows an example; the skeleton is shown at the bottom. Because the skeletons all come from real quotations, their structures are grammatically plausible, as long as the slots are filled with appropriate words. Which words are appropriate (individually or together) is a matter of linguistic judgment.

Information from the writer's input (see Figure 4, section a) and the extracted skeletons are combined to generate slogan suggestions. To make sure that generated suggestions contain keywords from the writer's input, we randomly sample content words from the input and treat them as target words. We identify skeletons that have empty slots that match the part-of-speech tags of the target words and choose three candidate skeletons. Given a skeleton, we follow Ozbal et al.'s approach [25] and use beam search to fill in slots with words that approximately maximize a scoring function. The scoring function gives high scores if the target words are used to fill in slots. In addition, the scoring function factors in language model probability scores to encourage grammaticality and a word diversity score to avoid repetition.

---

[6]https://en.wikiquote.org/wiki/Main_Page. Ideally, we would use a slogan dataset. Given that there is no public slogan database, we believe that quotations and slogans share some similar characteristics, such as memorability and pithiness.

USER STUDY SETUP

To explore how people will interact with machine-in-the-loop writing, we had people write with or without a machine in the loop. We obtained third-party reviews of the final written pieces. This gave us insight into what types of interactions and suggestions people are interested in and find most useful when writing with a machine in the loop.

Task Setup

For both the story writing task and slogan writing task, we assigned half of the participants to write with the machine-in-the-loop system prototype (the MIL condition) and half without it (the solo condition). Half of the participants in each condition (solo and MIL) were asked to write three stories and the other half wrote three slogans, based on three different prompts. The order of the prompts was balanced across participants. In both the solo and MIL tasks, after each story or slogan was complete, the participant completed a survey consisting of seven-point Likert scale questions about the final piece of writing. Table 2 lists the exact questions for each task under "Writing Survey." Surveys for both tasks additionally asked how satisfied participants were with the final piece of writing.

After three rounds of writing and evaluation, participants completed a final survey. All participants were asked a seven-point Likert scale question on how enjoyable they found the writing process. They were then asked open-response questions on the interface design, the difficulty of the task, and what improvements could be made to the tool. The solo case participants were also asked to describe any tools that could have helped their creative process. The MIL participants were additionally asked a four-choice question on how likely they were to use the system again and Likert scale questions about the quality of the suggestions the prototype system provided (see the exact questions under "Final Survey" in Table 2),

	Story Task	**Slogan Task**
Writing Survey *"Is the final product:"*	Creative? Coherent? Entertaining? Grammatical?	Creative? Catchy? Relevant?
Final Survey *"Were the suggested sentences:"*	Surprising? Creative? Grammatical?	Surprising? Creative? Catchy? Relevant?

Table 2: Survey questions

whether they liked the suggestions they received, and whether they appreciated the suggestions.

After participants completed the three writing tasks and four surveys (one after each writing task and one final survey), we conducted an open-ended interview about their experience with the task, their creative writing background, and their thoughts about future improvements and uses for the tool.

Although participant enjoyment and personal perception of their own success are important measures of the prototypes, we also wanted to know how third-party evaluators perceived the writing done alone versus with a machine in the loop. Amazon Mechanical Turkers evaluated the writing that participants produced along the same dimensions as the participants who wrote them ("Writing Survey" in Table 2).

Analysis Methods
The first two authors created an interview coding scheme [22] based on the Likert scale prompts and other areas of interest. The first two authors independently coded two interviews, compared coding, resolved conflicts, and revised the coding scheme. The final codes covered insights on: interface, machine suggestions, writing process, writing background, collaboration, use cases, and writing prompt. They then independently coded four more interviews, one from each of the study conditions. These codings were compared, and disparities were discussed. The first two authors then each independently coded separate halves of the remaining interviews, evenly distributed between conditions. The coders then came together to compile results.

Participant Demographics
We had 36 participants complete the writing tasks, 9 in each of the 4 task categories: solo story writing, MIL story writing, solo slogan writing, and MIL slogan writing. Participants were compensated with a $20 Amazon gift card. We categorized participant writing experience into none, some, or a lot. Participants with a lot of experience included professional creative writers and passionate hobbyists who wrote frequently. Participants with some experience included those that wrote occasionally or had formal creative writing education in their past. Participants with no experience included those who had not done creative writing since primary school. Table 3 shows the breakdown of experience by task condition. Participants were randomly assigned to conditions regardless of experience; there were more expert writers in the solo conditions than

Condition	None	Some	A Lot
Solo Story	3	2	3
MIL Story	5	4	0
Solo Slogan	3	1	5
MIL Slogan	4	4	1

Table 3: Number of participants in each condition with a given level of writing experience.

Jim slouched in the corner, feeling sorry for the patient in front of him.
["This is ridiculous," said Duke.]
"Yesterday I felt fine, and now you're telling me I'm at death's door?!"
["We'll take care of Furble tomorrow,"] the doctor said.
"You've named my tumor?!" Duke shrieked.
["Yeah,]" replied the doctor coolly, "we've found that anthropomorphizing tumors helps people in your position come to terms with their condition more easily."
Jim watched as Duke's eyes got even wider, and he wondered if it was because of the doctor's casual tone or the fact that the tumor had such a ridiculous name as "Furble".
[~~Lance yells over the speakers "no sudden hammering"~~]
"Anyway, we feel that Furble will most likely be gone within a month," the doctor said.
Jim grew more concerned about Duke's eyes, they seemed impossibly large now and if he wasn't careful Furble might not be the man's only medical concern.
["You're joking right], ["] Duke said.

Figure 6: Sample story written with the story writing system by participant MST65. The computer suggestions are in color and brackets; struck out text indicates rejected submissions. (Image prompt: [7]).

the MIL conditions, which should be kept in mind when comparing solo and MIL results.

Amazon Mechanical Turkers evaluated the final writing samples, with 9 evaluations collected for each of the 108 writing samples (3 per participant). Turkers were paid $0.15 for each evaluation they completed. To qualify, Turkers had to have completed over 1,000 tasks, have over a 95% task acceptance rate, and be from the United States.

EXPERIMENT RESULTS
Due to differences in the system designs and goals, as described in the above sections, we describe results for each system separately. The participant IDs indicate task and condition: MST (MIL story writing), SST (solo story writing), MSL (MIL slogan writing), and SSL (solo slogan writing).

Story Writing
We present insights from the story writing system from participants and third-party evaluators. Some participant insights did not depend on condition, such as enjoyment, interface suggestions, and opinions about story writing. Other comments from the MIL condition were more focused on the strengths and weaknesses of the machine suggestions.

Enjoyment
Overall, people found the story writing task fun. Two participants liked that it was a low-stakes, non-judgmental experience. Participants in both the MIL and the solo condition rated the task as highly enjoyable (both averaged 6.0).[7]

Participants who wrote in the solo condition had higher average satisfaction with their final stories than those in the MIL condition (average rating of 5.03 versus 4.59, $p=0.27$)[8], and solo condition participants thought their stories were more creative (5.30 versus 4.30, $p=0.01$). However, as seen in Figure 7, the MIL condition did have one more positive vote (≥ 5 on the Likert scale) for satisfaction than the solo condition.

When Help is Useful
Of the participants who found the suggestions helpful in directing their stories, three said the suggestions were most useful early in their story; they were more able to incorporate new elements from the suggestions before they had an established vision of the story. MST79 said, "By the time I'm hitting sentence 5, 6, 7, 8, I've developed so much of the story in my head already, most of the time suggestions are so far and away from anything I want to consider." This is seen in Figure 6; earlier suggestions were incorporated, while a later suggestion was deleted entirely. Five participants liked having the suggestions throughout the writing process to help if they got writer's block. One participant thought it would have been helpful toward the end of his writing to be prompted to start wrapping up the story.

Suggestion Usefulness
When asked what they considered when evaluating the creativity of their final stories, most participants emphasized unexpectedness or deviation from the obvious as key aspects. In the MIL condition, there were mixed reactions to the usefulness of the suggestions in enhancing creativity. All participants said that the suggestions were very random. For eight participants, this meant they disregarded most suggestions, but two participants said that the randomness of the suggestions inspired them to write sillier stories or that incorporating those suggestions lead to more odd and creative writing; "when I took the AI into account and tried to incorporate that, it got a lot more creative because, again, it was just so spontaneous and much more random than I normally write" (MST65). Eight participants said the suggestions did or could have influenced the direction of the story. However, six participants said they did not find the suggestions helpful once they had a clear direction for the story. This is reflected in the ratings participants gave the suggestions; the mean score of how much participants appreciated the suggestions is 3.78, but there was high variation between participants, with answers ranging between 2 and 6.

---

[7] Due to the subjectiveness of evaluating writing and the varied participant background, both the self-evaluations and the Turk scores highly varied, often covering the full Likert scale (1-7).

[8] p-values are calculated using independent two-sample t-tests. Although we report p-values, we encourage caution in drawing firm conclusions from these calculations because of the small sample size (27 for each condition in both self-evaluations and third-party evaluations) and the imbalance across conditions in author expertise.

Use Condition	Story	Slogan
I'd use it, exactly as it is now.	2	0
I'd use it, but only if the suggestions were better.	4	5
I'd use it, but only if the writing set-up changed.	0	1
I wouldn't use it.	3	3

Table 4: Responses to "Would you use this system again?"

Participants found some elements of the suggestions more helpful than others. One participant liked using snippets of suggested dialogue, while a different participant found dialogue-heavy suggestions unhelpful. One participant mentioned taking the tone of the suggestions, and another participant appreciated a plot suggestion, although they didn't incorporate it as it would have required more context. Characters were a divisive element; some appreciated getting suggestions with new character names, while six others found suggestions that introduced new character names hurt the relevance of the suggestions. Consider the example in Figure 6. Although the character names in early suggestions are all used, a later suggestion that references "Lance" is deleted as all of the characters in the story have already been introduced. Timing may also affect the usefulness of new characters because characters are often introduced in the beginning of a story, as can be seen in the example stories in Table 5.

When asked what type of help they would like to receive as they wrote, eight MIL participants mentioned the machine could contribute by suggesting plot points, tone, or keywords. One participant from the solo condition envisioned a system where the computer took the role of a character in the story and provided dialogue. Other MIL participants appreciated the idea of receiving full sentences, especially if the suggestions had been more relevant. Four participants felt more back-and-forth iterations were needed to treat the machine as a viable collaborator and would appreciate feedback on their writing, such as encouragement, agreement, or advice.

Interaction
Three MIL condition participants found the every-other-sentence injection into their workspace disruptive. Participant MST65 wrote, "The 'partner' [MIL system] often made no sense, so it was difficult to incorporate their responses and often I just deleted the entire suggestion but it was a disruption." Four participants would have liked the ability to edit already submitted sentences, especially professional writers who were not able to follow their normal writing process. However, six participants enjoyed the constrained, non-editable, sentence-by-sentence structure as it kept them moving forward. Participant SST17 wrote, "Even when I got stuck, I eventually could tell myself, 'Just write one sentence!' and then move on. ... it forced me to keep moving forward instead of getting bogged down in getting all the details just right and trying to overhaul the whole thing when I didn't like one little thing."

Use Cases

Most people who wrote with the story writing system would use it again in some capacity, with the most frequent response being they would use it again if suggestions were better, as seen in Table 4. Of the people who said they wouldn't use the system again, one wrote that they might actually use it for fun or ungraded work, one said they didn't need help creating story ideas, and the third expressed skepticism at the general idea of writing technologies.

Of the participants who would use the story writing tool again, two of the participants envisioned "just for fun" applications, such as a game to play with children or a short, fun activity that pops up on Facebook. Four participants saw it as a way to practice writing or a low-stake activity to become motivated to write. Three participants envisioned using the system for outlining, story boarding, or other early idea generation, while two people indicated interest in using the tool to directly write a final product.

As for types of suggestions that may be useful, three participants wanted editorial feedback on grammar, syntax, and sentence structure. One participant envisioned feedback from the machine that more directly influenced the content of the story. When describing experiences with human collaborators, participants said that back-and-forth iteration was a key component that they wanted machines to mimic. Other characteristics of good past human collaborators included trust and like-mindedness.

Third-party Evaluations

There was no statistically significant difference in the Amazon Mechanical Turk third-party evaluation ratings between the solo and MIL conditions. For creativity, the average score for the solo condition was 4.87 and was 4.84 for the MIL condition ($p=0.88$). Story coherence scores had an average of 5.05 for solo and 4.77 for MIL ($p=0.21$).

It is important to note that most of the third-party scores did not correlate with the scores writers assigned their own work. The average Turker score and the self-evaluation score had Pearson correlation coefficients of 0.08, 0.07, and 0.04 for how creative the story was, how entertaining it was, and how much they liked it, respectively. The correlations for coherence and grammaticality were slightly better (0.31 and 0.41, respectively) but still weak.

Slogan Writing

As with story writing, some topics from the slogan tasks were mentioned by participants regardless of condition. However, because the solo case participants did not have suggestions nor the interface, MIL participants provided some unique insights.

Enjoyment

Five participants found writing slogans to be very difficult. One participant said they found it demoralizing because they felt so bad at it. Two participants expressed that writing something both succinct and informative was challenging; the top-rated slogans tended to be shorter than the lowest-rated slogans, as seen in the examples in Table 5. Four participants enjoyed the task and found it fun.
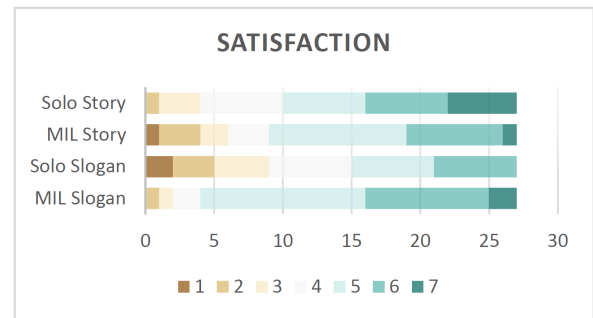

Figure 7: Slogan writers in the MIL condition were generally more satisfied with their final product than the solo slogan writers. More MIL story writers were positively satisfied with their final story, but only by one story.

Participant self-evaluation of the final slogans was generally higher in the MIL condition than the solo condition, even though the scores for the slogan suggestions were low. For example, the mean score for final slogans from the MIL condition was higher than in the solo case for both satisfaction (5.22 vs. 4.07, $p<0.01$) and creativity (4.48 vs. 3.93, $p=0.15$).

Suggestion Usefulness

Slogan participants in the MIL condition felt the suggestions did not introduce enough novelty. Three participants expressed frustration with suggestions that had just reorganized their input slogan words. These suggestions were too close to the original slogan and therefore not creative enough to be helpful. This frustration is reflected in low average scores for how much participants liked the suggestions (2.00) and appreciated the suggestions (2.78). Similar to the story writing task, appreciation varied greatly between participants, with scores ranging from 1 to 6. When participants used the provided suggestions, it was mainly to incorporate novel, relevant words or structural elements.

Seven MIL condition participants noted they would have preferred words instead of full sentences as suggestions; the system could act as a thesaurus to bring relevant but creative words and ideas. Participants also envisioned machine contributions such as searching the web for more information on a product, or finding similar, popular slogans as seed ideas. One participant found the machine had a different impact, "it impacted my overall thought process and creativity rather than actual words" (MSL87). Participants mentioned word play, use of literary devices (like alliteration or puns), cleverness, and catchiness as elements of creative slogans that they were looking for in suggestions.

Four participants described productive collaborations as bouncing ideas off someone else, building and iterating on good ideas, and coming to a feeling of "we found it!" Five participants thought the machine was a good collaborator, "it seemed like kind of having another pair of eyes in the room to give me some feedback and the fact that it was in real time was great and wouldn't argue with you over coffee was great" (MSL87). One participant felt the interaction with the system was not as conducive to enhancing creativity. One

	Highest-rated	Lowest-rated
Solo Slogan	Compassion is Always in Style	"Be Beautiful: End Animal Testing! Sign our petition today to show Neutrogena animal testing is unnecessary, unkind, and needs to stop now."
MIL Slogan	The real animals are the ones who test chemicals on living things.	"Stop animal research testing now! Studies confirm ther [sic] is no need for such testing, and msot [sic] adults are opposed to this practice."
Solo Story	Norman walked into the doctor's office just before his last appointment; Norman flicked his cigarette into the cold, winter night behind him as he walked into the building. The pretty blonde receptionist greeted Norman but Norman had no time for flirtation or romance; he was about to find the murderer of Trystan Lee and the doctor was key to this plan.	A middle aged male is meeting with a dentist. This gentleman is so stressed that he looks like his eyes are popping out of his face. This meeting is part of an investigation about a crime so that's why this meeting includes an investigator who looks creepy.
MIL Story	The nervous doctor cleared his throat, "Thank you...eh-hem...Mr. Collin, for coming into the office on such short notice." Craig slicked back his hair, listening to his wife's voice echoing in his head, reminding him that all of those late night trips to McDonalds would catch up with his heart eventually. "You see, we've found some...unusual...results from your recent stress test, and I thought it prudent to bring you in as soon as possible."	"Hey Docta Don, dis is da kat we wuz talkin' about last night, whachu wan' me to do wit 'im?" Fabin said through his cold eyes shaded by his pitch black sunglasses. He tapped his finger on the trigger and shook his head, Docta Don was never happy with how excited Fabin was to get into trouble; he was a good man, but followed all orders without ever thinking things through for himself.

Table 5: Highest and lowest rated slogans and stories for a given prompt. Slogans are for an animal rights cause. Only the first few lines of each story are shown.

challenge five participants described was that the suggestions were not aware enough of context and therefore were not working on the same idea, just with the same words. MSL33C said, "they weren't understanding my keywords correctly, I think, so I could say something like 'end animal harm' and it would suggest harming animals ... so I don't think it was quite interpreting my intentions very well." Three participants also were interested in non-content-based interactions such as feedback on whether an idea was good, reminders if a slogan was too close to an older discarded slogan, or an expression of closure upon deciding on a final slogan.

Interaction
Two participants liked having the history of suggestions in order to refer back to previous slogans. Another found the history distracting and wanted to curate the list to only keep good recommendations. For the physical placement of suggestions, two participants liked that the suggestions were out of the way and not intrusive. For four others, the suggestions were too far out of the way and required effort to check. These participants would have preferred a more condensed interface. One participant would have liked the ability to have more brainstorming space, either as a free form writing space or to iterate on multiple slogans at the same time.

Use Cases
Like the story writing case, most people who wrote with the slogan writing system would use it again in some capacity, especially if the suggestions were better (see Table 4). The people who said they would not use the system again wrote that they received suggestions that were too far from what they wanted to motivate them to use the system again.

The participants who said they would use the system again drew parallels between writing slogans and tasks such as naming courses or products, writing headlines, and writing titles. One participant said he might use a system like this to write emails in order to be reminded to be pithy and catchy. One participant did not think they would use the tool for any regular activities and would just use it for slogans.

Third-party Evaluations
The Amazon Mechanical Turk third-party evaluations rated the slogans written in the MIL condition as slightly less creative than the slogans from the solo condition, giving an average score of 3.84 and 4.37 (p=0.03), respectively. There was no statistically significant difference in any of the other scores, with relevance scoring the closest between the solo (average 5.98) and MIL (average 5.93) conditions (p=0.76).

Like in the story case, most of the third-party scores did not correlate with the scores writers assigned their own work. The average Turker score and the self-evaluation score had a Pearson correlation coefficient of 0.13 and 0.10, for how creative the slogan was and how much they liked it, respectively. The correlation for catchiness and relevance were slightly better (0.39 and 0.22, respectively) but still weak.

DISCUSSION
We found that people generally enjoyed writing with the help of suggestions and were enthusiastic about the concept of writing with a "collaborator," especially once natural language generation capabilities improve. Though some professional authors hesitated at the idea of using computer-generated suggestions when writing a final product, participants envisioned the usefulness of this system as a writing warm-up

or game and for difficult processes that are often collaborative (naming products or papers, writing headlines, etc.).

Another advantage of writing with a machine in the loop that participants observed was that writing with these systems allowed them to write in a judgment-free setting. Although collaborative writing is useful, it can be intimidating for less experienced writers to brainstorm or write with the pressure of a human collaborator. Writing with a machine in the loop can be a low-cost, easy way to provide new ideas and support to writers, particularly in the early stages of writing.

For machine-in-the-loop writing systems, we recommend a high level of writer control over the interaction. This will allow systems to cater to a wider range of writers and to adapt to changing writer needs at different points of the writing process. We also recommend carefully considering the interaction design choices (especially along the characteristics we describe) and how they may affect both the enjoyment of the task and the quality of the final product. Systems with low intrusiveness and a pull method of interaction initiation allow people to write more closely to their normal writing process. However, these characteristics also mean that suggestions are more easily ignored and may never be requested. If the goal is to encourage interaction with the machine or a more structured interaction, a higher intrusiveness and push method system may be better. A careful introduction and framing of the system is also necessary to encourage the desired level of interaction with the machine in the loop.

For story writing and other tasks that expand on a prompt and have an additive interaction structure, systems may benefit from an interface that supports outlining or non-linear writing. For example, Flower and Hayes [10] describe the hierarchical nature of the creative writing process; future system designs could reflect knowledge about the cognitive processes of writing to better support the writing process. The slogan writing task, along with other condensing writing tasks that have an iterative structure, may benefit from an interface that provides more space for brainstorming and drafting slogans.

We recommend using models that strike a balance between generating coherent suggestions and surprising suggestions. An element of randomness provides new ideas and directions for a writer, but suggestions too far away from the writer's ideas may be unhelpful and ignored. We recommend pushing towards surprising suggestions for tasks that use a pull method of initiation and have a low level of intrusiveness because when writers decide to initiate a suggestion loop, it generally means they are stuck or at least open to new ideas. Although surprising suggestions run the risk of being irrelevant, a less intrusive system means unhelpful suggestions can be easily ignored and minimally interrupt the writing process. Coherence should be a bigger priority for push method, high-intrusiveness interactions, as high levels of randomness in suggestions may distract writers.

For the story writing task, we found that participants wanted more coherent suggestions from the model. For similar tasks, we recommend working towards incorporating more context into the suggestion-generating models. Characters may be an important aspect of the context to consider, as suggestions with incorrect pronouns or that lack references to existing characters are difficult to work into a story. Models would also benefit from the ability to play with the content type of its suggestions, such as choosing to offer lines of dialogue, action-driven sentences, or descriptive lines.

Models for writing slogans should provide more variety in their suggestions, particularly on a lexical level, as diverse language is important for an iterative task. Models that can generate related keywords, synonyms, and alliterative words when given a person's ideas would be useful for this type of task.

We noted that there is little to no correlation between the ratings that writers gave themselves and the ratings that Amazon Mechanical Turker workers gave them. This observation echoed the finding in Tan et al. [31] that it is hard for humans to evaluate the quality of writing. Therefore, machine-in-the-loop writing systems that aim to improve a writer's work should measure the system's success not only as perceived by the writer but also by third-party evaluators.

CONCLUSION

By analyzing people's experiences and results from writing with two different machine-in-the-loop systems, we find that participants enjoyed collaborating with a machine and would use systems like ours again, particularly as the quality of suggestions the system can generate improves. Participants who wrote with a machine in the loop were more satisfied (in the slogan case) or comparably satisfied (in the story case) to solo writers with their final written artifacts. The writing completed with machine-generated suggestions has yet to surpass that of the participants who wrote alone, as judged by third-party evaluators.

We recommend several directions of research towards improving these suggestions, including improving the ability to control the amount of unexpectedness in suggestions and considering different generation formats (including higher-level ideas or keywords) based on the characteristics and the goal of the system. Furthermore, we recommend choosing the interaction structure, initiation, and intrusiveness of the system design thoughtfully according to the nature of the task and the desired level of interaction. We, along with the participants in our study, envision many possible applications in which machine-in-the-loop support would be helpful, including related writing tasks (like writing headlines or poetry), education, and entertainment.

REFERENCES

1. James F. Allen, Curry I. Guinn, and Eric Horvitz. 1999. Mixed-initiative interaction. *IEEE Intelligent Systems* 14, 5 (Sept. 1999), 14–23.

2. Michael S. Bernstein, Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich. 2010. Soylent: a word processor with a crowd inside. In *Proceedings of UIST*.

3. Margaret A. Boden. 2004. *The Creative Mind: Myths and Mechanisms*. Routledge, London; New York.

4. Steve Branson, Catherine Wah, Florian Schroff, Boris Babenko, Peter Welinder, Pietro Perona, and Serge Belongie. 2010. Visual recognition with humans in the loop. In *Proceedings of ECCV*.

5. Alastair Brotchie and Mel Gooding (Eds.). 1995. *A Book of Surrealist Games: Including the Little Surrealist Dictionary*. Shambhala Redstone Editions, distributed in the United States by Random House, Boston.

6. Justin Cheng and Michael S. Bernstein. 2015. Flock: hybrid crowd-machine learning classifiers. In *Proceedings of CSCW*.

7. Leo Cullum. 2005. Doctor talking to patient in his office as a man. *The New Yorker* 81, 39 (5 Dec. 2005). `http://archives.newyorker.com/?i=2005-12-05`

8. Liza Donnelly. 2014. A man and woman look at three very large birds. *The New Yorker* 90, 22 (4 Aug. 2014). `http://archives.newyorker.com/?i=2014-08-04`

9. Jerry Alan Fails and Dan R. Olsen Jr. 2003. Interactive machine learning. In *Proceedings of IUI*.

10. Linda Flower and John R. Hayes. 1981. A cognitive process theory of writing. *College Composition and Communication* 32, 4 (1981), 365–387.

11. Monica J. Garfield. 2008. Creativity support systems. In *Handbook on Decision Support Systems 2*. Springer, Berlin, Heidelberg, 745–758.

12. Lorenzo Gatti, Gözde Özbal, Marco Guerini, Oliviero Stock, and Carlo Strapparava. 2016. Heady-Lines: a creative generator of newspaper headlines. In *Companion Publication of the 21st International Conference on Intelligent User Interfaces*. 79–83.

13. Marjan Ghazvininejad, Xing Shi, Yejin Choi, and Kevin Knight. 2016. Generating topical poetry. In *Proceedings of ACL*. 1183–1191.

14. Marjan Ghazvininejad, Xing Shi, Jay Priyadarshi, and Kevin Knight. 2017. Hafez: an interactive poetry generation system. In *Proceedings of ACL (system demonstrations)*. 43–48.

15. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In *Proceedings of ICCV*.

16. Guy Hoffman and Gil Weinberg. 2011. Interactive improvisation with a robotic marimba player. *Autonomous Robouts* 31, 2–3 (October 2011), 133–153.

17. Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of CHI*.

18. Mikhail Jacob and Brian Magerko. 2015. Viewpoints AI. In *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition*. ACM, New York, NY, 361–362.

19. Yangfeng Ji, Trevor Cohn, Lingpeng Kong, Chris Dyer, and Jacob Eisenstein. 2016. Document context language models. In *Proceedings of the 4th International Conference on Learning Representations (Workshop Track)*.

20. Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. 2017. Unbiased learning-to-rank with biased feedback. In *Proceedings of WSDM*.

21. Joy Kim, Justin Cheng, and Michael S. Bernstein. 2014. Ensemble: exploring complementary strengths of leaders and crowds in creative collaboration. In *Proceedings of CSCW*.

22. Jonathan Lazar. 2017. In *Research Methods in Human Computer Interaction* (2nd edition ed.). Elsevier, Cambridge, MA, Chapter 11.

23. Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Proceedings of Interspeech*.

24. Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of ACL*.

25. Gözde Özbal, Daniele Pighin, and Carlo Strapparava. 2013. BRAINSUP: brainstorming support for creative sentence generation. In *Proceedings of ACL*.

26. Melissa Roemmele and Andrew S. Gordon. 2015. Creative Help: a story writing assistant. In *Proceedings of ICIDS*.

27. Robin Sloan. 2016. Writing with the machine. (2016). `https://www.robinsloan.com/notes/writing-with-the-machine/`

28. Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural netowrk approach to context-sensitive generation of conversational responses. In *Proceedings of ACL*. 196–205.

29. Robert J. Sternberg. 2005. Creativity or creativities? *International Journal of Human-Computer Studies* 63, 4-5 (Oct. 2005), 370–382.

30. Reid Swanson and Andrew S. Gordon. 2008. Say Anything: a massively collaborative open domain story writing companion. In *Proceedings of ICIDS*.

31. Chenhao Tan, Lillian Lee, and Bo Pang. 2014. The effect of wording on message propagation: Topic- and author-controlled natural experiments on Twitter. In *Proceedings of ACL*.

32. Alan M. Turing. 1950. Computing machinery and intelligence. *Mind* 59, 236 (1950), 433–460.

33. Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: a neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3156–3164.

34. Georgios N. Yannakakis, Antonios Liapis, and Constantine Alexopoulos. 2014. Mixed-initiative co-creativity. In *Proceedings of FDG*.

35. Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of ICCV*. 19–27.