On How Users Edit Computer-Generated Visual Stories

Ting-Yao Hsu
Pennsylvania State University
State College, PA, USA
txh357@psu.edu

Yen-Chia Hsu
Carnegie Mellon University
Pittsburgh, USA
yenchiah@andrew.cmu.edu

Ting-Hao (Kenneth) Huang
Pennsylvania State University
State College, USA
txh710@psu.edu

ABSTRACT

A significant body of research in Artificial Intelligence (AI) has focused on generating *stories* automatically, either based on prior story plots or input images. However, literature has little to say about how users would receive and use these stories. Given the quality of stories generated by modern AI algorithms, users will nearly inevitably have to *edit* these stories before putting them to real use. In this paper, we present the first analysis of how human users *edit* machine-generated stories. We obtained 962 short stories generated by one of the state-of-the-art *visual storytelling* models. For each story, we recruited five crowd workers from Amazon Mechanical Turk to edit it. Our analysis of these edits shows that, on average, users *(i)* slightly shortened machine-generated stories, *(ii)* increased lexical diversity in these stories, and *(iii)* often replaced nouns and their determiners/articles with pronouns.

KEYWORDS
story generation; computer-supported writing; creative writing; text post-editing; visual storytelling

CHI'19 Extended Abstracts, May 4–9, 2019, Glasgow, Scotland UK

ACM ISBN 978-1-4503-5971-9/19/05.
https://doi.org/10.1145/3290607.3312965

Figure 1: A human-written visual story in the VIST dataset [7]. Each story contains a sequence of five Flickr photos, and a short story written by human workers. (Note that the VIST dataset normalized the story text, *e.g.,* replaced all female names, such as "Amy" or "Sarah", with "[female]".)

[1]Visual Storytelling Dataset (VIST): http://visionandlanguage.net/VIST/

Our study provides a better understanding on how users receive and edit machine-generated stories, informing future researchers to create more usable and helpful story generation systems.

INTRODUCTION

A large body of research in the AI community has been devoted to generating stories automatically. Many prior works aimed to generate stories based on text inputs, such as preceding sentences or writing prompts. For example, Roemmele *et. al* used a recurrent-neural-network architecture to generate stories in a sequence-to-sequence manner [12]. Formulating a story as a sequence of events, Martin *et. al* proposed an event representation for neural-network-based story generation [9]. Fan *et al.* created a hierarchical model that automatically generates stories conditioning on the writing prompts [3]. Some other prior work explored generating stories based on images. For instance, Huang *et. al* introduced the *visual storytelling* task, in which the automatic model takes a sequence of photos as input, and generates a short story that narrates this photo sequence [7]. An example visual story is shown in Figure 1. *StyleNet* [4] and *SemStyle* [10] stylized descriptive image captions and made them more "attractive", *e.g.*, more romantic or humorous.

However, literature has little to say about how users would receive these machine-generated stories and put them to real use. More specifically, given the quality of stories created by modern automatic models, users will likely need to *edit* these stories for practical uses, for example, sharing them on social media. In this paper, we focus on the stories generated by modern *visual storytelling* models [7, 15] and study how people would edit them. This study allows us to understand how text generation technologies can (or can not) help creative writing. Some prior work has explored *interactive* computer-supported story writing, in which the system populates suggestions or inspirations for the writer in near real-time when he/she writes the story. For example, the *Creative Help* system generates suggestions for the next sentence in the process of story writing [12–14]. Clark *et al.* studied machine-in-the-loop short story writing and concluded that machine intervention should balance between generating coherent and surprising suggestions [2]. Our study provides a detailed understanding of how users receive and edit machine-generated text in the context of story writing.

DATA PREPARATION

Machine-Generated Visual Stories. In the *visual storytelling* task, as introduced by Huang *et al.* [7], the computational model takes a sequence of five photos as input, and then automatically generates a short story describing the photo sequence. Huang *et al.* also released the VIST dataset[1], which contains 81,743 unique Flickr photos in 20,211 sequences, aligned to human-written stories (as shown in Figure 1). Many researchers have proposed approaches to generate visual stories using this dataset. In this paper, we ran the visual storytelling model released by Wang *et al.* [15] on the test set of VIST to obtain machine-generated visual stories. Wang's approach was one of the state-of-the-art methods

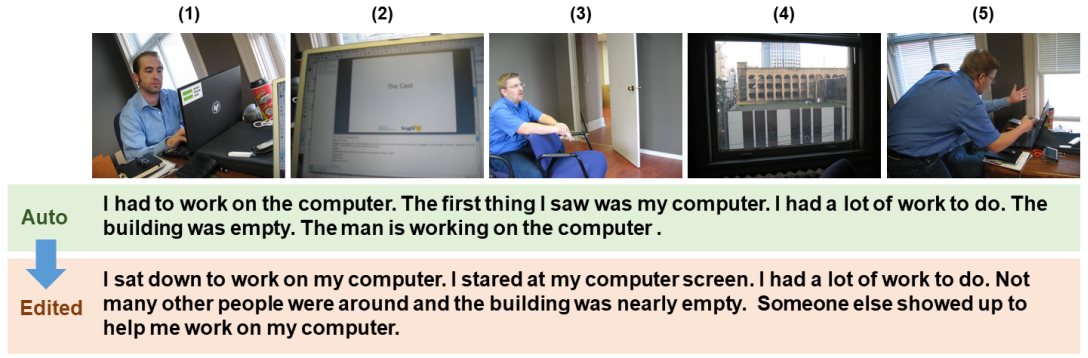

Figure 2: An example of machine-generated visual story (Auto) and its edited version (Edited). This story is generated by the visual storytelling model released by Wang *et al.* [15] using the VIST dataset.

that participated in the first Visual Storytelling Challenge [11] and also the earliest implementation that was made available online. Among the data in the test set of VIST, we removed the photo sequences containing any photos that have been deleted from Flickr by their owners and only used no more than one photo sequence per Flickr photo album to reduce redundancy. Eventually, we obtained 962 machine-generated visual stories, and each has a corresponding photo sequence.

Visual Story Post-Editing. For each story, we recruited five crowd workers from Amazon Mechanical Turk to edit it, respectively.[2] The following instruction was used to guide workers: "Please edit the story text as if these were your photos, and you would like using this story to share your experience with your friends." We instructed workers to stick with the plot and the point of view (first-, second-, or third-person) of the original story so that workers will not abandon the machine-generated story and write a new one from scratch. The goal of this study is to understand how people edit machine-generated stories. Further research is required to explore factors, such as how coherent the story is, or which level of details does the story provide, that could affect user' willingness to edit a story or abandon it. The worker interface is shown in Figure 3. The photo sequence was also displayed on the interface. The price of each task was $0.12. As a result, 197 workers generated $962 \times 5 = 4{,}810$ edited stories. Figure 2 shows an example of a generated visual story, before and after human editing.

ANALYSIS

We analyzed the length, type-token ratio (TTR), and the part of speech tags of pre- and post-edited visual stories. We used the NLTK toolkit to segment sentences, tokenize sentences into words (known as "tokens"), and tag part-of-speech (POS) for each word [1]. It is noteworthy that the VIST dataset

[2]We specified the following qualifications workers must meet to work on our tasks: HIT (Human Intelligence Task) Approval Rate >= 98%, Number of HITs Approved >= 3000, and location = US. The Adult Content Qualification was also used.

Figure 3: The worker interface for visual story post-editing. The following instruction was used: "Please edit the story text as if these were your photos, and you would like using this story to share your experience with your friends." We also instructed workers to stick with the plot and the point of view (first-, second-, or third-person) of the original story.

[3]A Universal Part-of-Speech Tagset: https://www.nltk.org/book/ch05.html

normalized the story text, *e.g.,* replaced all female names such as "Amy" and "Sarah" with "[female]". We replaced all "[female]" with "Amy" and "[male]" with "Tom" for our analysis. The detailed results and discussions are as follows.

Edited stories are slightly shorter.

On average, the edited visual stories are shorter significantly (paired t-test, $p < 0.001$, $N = 4810$). We calculated the number of tokens (also known as "sentence length") of each story. An automatic story contains an average of 43.02 tokens (punctuation marks included, SD = 4.96), and an edited story contains an average of 41.85 tokens (SD = 9.70). This comparison indicates that users on average eliminated 1.17 tokens. Meanwhile, we noticed that the edited stories' lengths resulted in a higher sample standard deviation (SD), which suggested that human-written stories are more diverse regarding story length.

Edited stories have a higher lexical diversity.

We also studied the lexical diversity of these storied. Type-token ratio (TTR) is the one of most commonly used index of lexical diversity [5], which is calculated as the ratio between the number of different words (types) N_{type} and the total number of words (tokens) N_{token} of a text unit, *i.e.,* $TTR = N_{type}/N_{token}$. A higher TTR indicates a higher lexical diversity. Per our analysis, the average TTR for an automatic story is 0.62 (SD = 0.09), while the average TTR for an edited story is 0.72 (SD = 0.06). This difference is statistically significant (paired t-test, $p < 0.001$, $N = 4810$). This result suggested that users reduced the word redundancy and increased the lexical diversity in machine-generated stories. The normalized frequency histogram of TTR of pre- and post-edited stories are shown in Figure 4.

Nouns (NOUN) with determiners/articles (DET) are often replaced by pronouns (PRON).

We also analyzed the parts of speech of the words in each story. The universal part-of-speech tagset[3] provided by NLTK toolkit was used. The average number of tokens of each POS in a story is shown in Table 1. We observed that the "DET" (determiner, article) tag contributes the most to the reduction of story length, while the number of "PRON" (pronoun) tags increased the most. In order to understand the possible reasons behind this phenomenon, we observed the stories that had fewer DET tags but more PRON tags after post editing. We found that the nouns with determiners and articles (DET) are often replaced by pronouns (PRON). The following is a typical example, where the replacement texts are highlighted in red.

> **[pre-edit]** the car was parked in the middle of the road . ...
>
> **[post-edit]** we drove home and parked by lots of other cars . ...

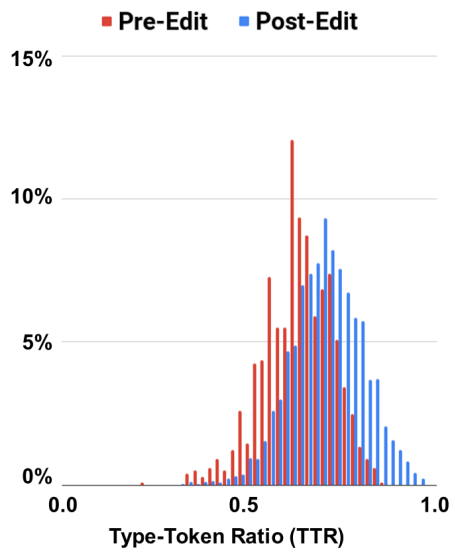

Figure 4: The normalized frequency histogram of TTR of pre- and post-edited stories. A higher TTR indicates a higher lexical diversity. This result suggests that users increased the lexical diversity in machine-generated stories.

Table 1: Average number of tokens with each POS tag per story. The universal part-of-speech tagset provided by NLTK toolkit was used. Δ denoted the differences in average token numbers between between post- and pre-edit stories. The "DET" (determiner, article) tag contributes the most to the reduction of story length, and the "." (punctuation marks) tag contributes the second.

	ADJ	ADP	ADV	CONJ	DET	NOUN	NUM	PRT	PRON	VERB	.	Total
Pre-Edit	3.09	3.46	1.92	0.50	8.11	10.00	0.02	1.62	2.14	7.00	5.16	43.02
Post-Edit	3.13	3.39	1.85	0.85	7.15	9.81	0.05	1.58	2.31	7.05	4.69	41.85
Δ	0.04	-0.07	-0.07	0.35	-0.96	-0.20	0.03	-0.04	0.17	0.04	-0.47	-1.17

The following is another example:

> **[pre-edit]** the group of friends went to the camp site to see what was going on . we had a lot of food to eat . the kids had a great time . we had a lot of fun . we had a lot of fun .
>
> **[post-edit]** a group of friends went camping together . they took plenty of food to eat . they had a great time together . everyone had a lot of fun .

We also noticed that the "." (punctuation marks) tags reduced significantly. This might be caused by the fact that machine-generated stories can be repetitive or too general, so users often merged two sentences or simply removed one of them. For instance, in the example above (the text highlighted in blue), the user decided to remove one sentence and thus one period mark (".") was eliminated.

DISCUSSION & FUTURE WORK

Our analysis on the machine-generated and post-edited visual stories shows that, on average, users *(i)* slightly shortened machine-generated stories, *(ii)* increased lexical diversity, and *(iii)* often replaced nouns and their determiners/articles with pronouns, and merged or eliminated sentences. The higher-level theme emerged in these observations are *text repetition* in machine-generated stories. Generating redundant text is a known problem for many neural-network-based visual storytelling models, while some approaches [6] suffer from this problem more than others [8]. Our analysis indicates that, for the visual stories generated by Wang *et al.* [15], a significant part of human editing effort is to reduce word redundancy and increase lexical diversity. Furthermore, users frequently replaced nouns with pronouns suggests that the storytelling model should better understand which entities (*e.g.,* human, building, animal) have been mentioned and thus can be called using pronouns (*e.g.,* she, it). In the future, we will develop algorithms to learn from these edits to improve existing machine-generated stories. We will also study what types of machine-generated texts are more helpful in assisting users to compose short stories, aiming at building a human-centered computer-supported storytelling system.

ACKNOWLEDGEMENTS

We thank Jiawei Chen for her help. We also thank the workers on Mechanical Turk who participated in our experiments.

REFERENCES

[1] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* O'Reilly Media, Inc.

[2] Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, 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 '18)*. ACM, New York, NY, USA, 329–340. https://doi.org/10.1145/3172944.3172983

[3] Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical Neural Story Generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 889–898. http://aclweb.org/anthology/P18-1082

[4] C. Gan, Z. Gan, X. He, J. Gao, and L. Deng. 2017. StyleNet: Generating Attractive Visual Captions with Styles. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 955–964. https://doi.org/10.1109/CVPR.2017.108

[5] Andrew Hardie and Tony McEnery. 2006. *Statistics.* Vol. 12. Elsevier, 138–146.

[6] Chao-Chun Hsu, Szu-Min Chen, Ming-Hsun Hsieh, and Lun-Wei Ku. 2018. Using Inter-Sentence Diverse Beam Search to Reduce Redundancy in Visual Storytelling. *arXiv preprint arXiv:1805.11867* (2018).

[7] Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. 2016. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1233–1239.

[8] Taehyeong Kim, Min-Oh Heo, Seonil Son, Kyoung-Wha Park, and Byoung-Tak Zhang. 2018. GLAC Net: GLocal Attention Cascading Networks for Multi-image Cued Story Generation. *arXiv preprint arXiv:1805.10973* (2018).

[9] Lara J. Martin, Prithviraj Ammanabrolu, Xinyu Wang, William Hancock, Shruti Singh, Brent Harrison, and Mark O. Riedl. 2018. Event Representations for Automated Story Generation with Deep Neural Nets. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18)*, Sheila A. McIlraith and Kilian Q. Weinberger (Eds.). AAAI Press, 868–875. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17046

[10] Alexander Mathews, Lexing Xie, and Xuming He. 2018. SemStyle: Learning to Generate Stylised Image Captions Using Unaligned Text. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[11] Margaret Mitchell, Francis Ferraro, Ishan Misra, et al. 2018. Proceedings of the First Workshop on Storytelling. In *Proceedings of the First Workshop on Storytelling*.

[12] Melissa Roemmele. 2016. Writing stories with help from recurrent neural networks. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. AAAI Press, 4311–4312.

[13] Melissa Roemmele and Andrew S. Gordon. 2015. Creative Help: A Story Writing Assistant. In *Interactive Storytelling*, Henrik Schoenau-Fog, Luis Emilio Bruni, Sandy Louchart, and Sarune Baceviciute (Eds.). Springer International Publishing, Cham, 81–92.

[14] Reid Swanson and Andrew S. Gordon. 2012. Say Anything: Using Textual Case-Based Reasoning to Enable Open-Domain Interactive Storytelling. *ACM Trans. Interact. Intell. Syst.* 2, 3, Article 16 (Sept. 2012), 35 pages. https://doi.org/10.1145/2362394.2362398

[15] Xin Wang, Wenhu Chen, Yuan-Fang Wang, and William Yang Wang. 2018. No Metrics Are Perfect: Adversarial Reward Learning for Visual Storytelling. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. ACL, Melbourne, Victoria, Australia.