Visual Story Post-Editing

Ting-Yao Hsu[1], Chieh-Yang Huang[1], Yen-Chia Hsu[2], Ting-Hao (Kenneth) Huang[1]

[1]Pennsylvania State University, State College, PA, USA
[2]Carnegie Mellon University, Pittsburgh, PA, USA
[1]{txh357, chiehyang, txh710}@psu.edu
[2]yenchiah@andrew.cmu.edu

Abstract

We introduce the first dataset for **human edits of machine-generated visual stories** and explore how these collected edits may be used for the *visual story post-editing* task. The dataset, *VIST-Edit*[1], includes 14,905 human-edited versions of 2,981 machine-generated visual stories. The stories were generated by two state-of-the-art visual storytelling models, each aligned to 5 human-edited versions. We establish baselines for the task, showing how a relatively small set of human edits can be leveraged to boost the performance of large visual storytelling models. We also discuss the weak correlation between automatic evaluation scores and human ratings, motivating the need for new automatic metrics.

Machine-Generated Story (a): *visual storytelling*
the family got together for a dinner. the food was delicious. everyone was having a great time. the meal was delicious. the kids had a great time.

Machine-Generated (a) -> Human-Edited Story (b):
the whole family got together for thanksgiving. the food was delicious! everyone had a lot of fun, and the kids played the entire time. *visual story post-editing*

Machine-Generated (a) -> Machine-Edited Story (c):
the family got together for a nice dinner. the food was delicious. the guys enjoyed the food since they had never eaten there before. the food was presented well. the dessert was delicious.

Figure 1: A machine-generated visual story (a) (by GLAC), its human-edited (b) and machine-edited (c) (by LSTM) version.

1 Introduction

Professional writers emphasize the importance of *editing*. Stephen King once put it this way: "to write is human, to *edit* is divine." (King, 2000) Mark Twain had another quote: "Writing is easy. All you have to do is cross out the wrong words." (Twain, 1876) Given that professionals revise and rewrite their drafts intensively, machines that generate stories may also benefit from a good editor. Per the evaluation of the first Visual Storytelling Challenge (Mitchell et al., 2018), the ability of an algorithm to tell a sound story is still far from that of a human. Users will inevitably need to edit generated stories before putting them to real uses, such as sharing on social media.

We introduce the first dataset for **human edits of machine-generated visual stories**, *VIST-Edit*, and explore how these collected edits may be used for the task of *visual story post-editing* (see Figure 1). The original visual storytelling (VIST) task, as introduced by Huang *et al.* (2016), takes
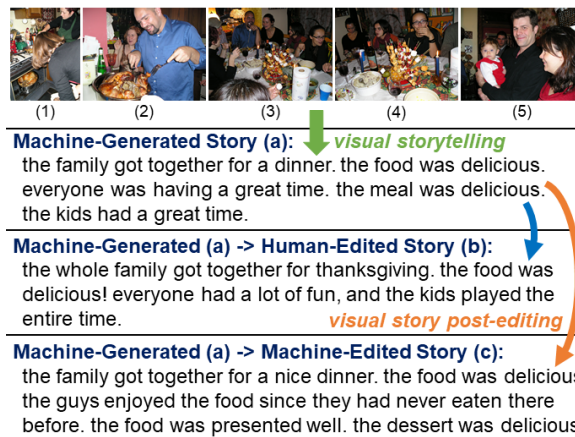
a sequence of five photos as input and generates a short story describing the photo sequence. Huang *et al.* also released the VIST dataset, containing 20,211 photo sequences, aligned to human-written stories. On the other hand, the automatic post-editing task revises the story generated from visual storytelling models, given both a machine-generated story and a photo sequence. Automatic post-editing treats the VIST system as a black box that is fixed and not modifiable. Its goal is to correct systematic errors of the VIST system and leverage the user edit data to improve story quality.

In this paper, we *(i)* collect human edits for machine-generated stories from two different state-of-the-art models, *(ii)* analyze what people edited, and *(iii)* advance the task of visual story post-editing. In addition, we establish baselines for the task, and discuss the weak correlation between automatic evaluation scores and human ratings, motivating the need for new metrics.

---

[1]*VIST-Edit*: https://github.com/tingyaohsu/VIST-Edit

2 Related Work

The visual story post-editing task is related to *(i)* automatic post-editing and *(ii)* stylized visual captioning. **Automatic post-editing (APE)** revises the text generated typically from a machine translation (MT) system, given both the source sentences and translated sentences. Like the proposed VIST post-editing task, APE aims to correct the systematic errors of MT, reducing translator workloads and increasing productivity (Astudillo et al., 2018). Recently, neural models have been applied to APE in a sentence-to-sentence manner (Libovickỳ et al., 2016; Junczys-Dowmunt and Grundkiewicz, 2016), differing from previous phrase-based models that translate and reorder phrase segments for each sentence, such as (Simard et al., 2007; Béchara et al., 2011). More sophisticated sequence-to-sequence models with the attention mechanism were also introduced (Junczys-Dowmunt and Grundkiewicz, 2017; Libovickỳ and Helcl, 2017). While this line of work is relevant and encouraging, it has not explored much in a creative writing context. It is noteworthy that Roemmele *et al.* previously developed an online system, Creative Help, for collecting human edits for computer-generated narrative text (Roemmele and Gordon, 2018b). The collected data could be useful for story APE tasks.

Visual story post-editing could also be considered relevant to **style transfer on image captions**. Both tasks take images and source text (*i.e.*, machine-generated stories or descriptive captions) as inputs and generate modified text (*i.e.*, post-edited stories or stylized captions). End-to-end neural models have been applied to the transfer styles of image captions. For example, *StyleNet*, an encoder-decoder-based model trained on paired images and factual captions together with an unlabeled stylized text corpus, can transfer descriptive image captions to creative captions, *e.g.,* humorous or romantic (Gan et al., 2017). Its advanced version with an attention mechanism, *SemStyle*, was also introduced (Mathews et al., 2018). In this paper, we adopt the APE approach to treat pre- and post-edited stories as parallel data instead of the style transfer approach that omits this parallel relationship during model training.

3 Dataset Construction & Analysis

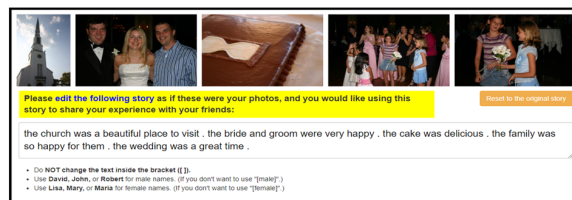**Obtaining Machine-Generated Visual Stories** This *VIST-Edit* dataset contains visual stories gen-

Figure 2: Interface for visual story post-editing. An instruction (not shown to save space) is given and workers are asked to stick with the plot of the original story.

erated by two state-of-the-art models, **GLAC** and **AREL**. GLAC (Global-Local Attention Cascading Networks) (Kim et al., 2018) achieved the highest human evaluation score in the first VIST Challenge (Mitchell et al., 2018). We obtain the pre-trained GLAC model provided by the authors via Github and run it on the entire VIST test set and obtain 2,019 stories. AREL (Adversarial REward Learning) (Wang et al., 2018) was the earliest available implementation online, and achieved the highest METEOR score on public test set in the VIST Challenge. We also acquire a small set of human edits for 962 AREL's stories generated using VIST test set, collected by Hsu *et al.* (2019).

Crowdsourcing Edits For each machine-generated visual story, we recruit five crowd workers from Amazon Mechanical Turk (MTurk) to revise it (at $0.12/HIT,) respectively. We instruct workers to edit the story "as if these were your photos, and you would like using this story to share your experience with your friends." We also ask workers to stick with the photos of the original story so that workers would not ignore the machine-generated story and write a new one from scratch. Figure 2 shows the interface. For GLAC, we collect 2,019 × 5 = 10,095 edited stories in total; and for AREL, 962 × 5 = 4,810 edited stories have been collected by Hsu *et al.* (2019).

Data Post-processing We tokenize all stories using CoreNLP (Manning et al., 2014) and replace all people names with generic [male/female] tokens. Each of GLAC and AREL set is released as training, validation, and test following an 80%, 10%, 10% split, respectively.

3.1 What do people edit?

We analyze human edits for GLAC and AREL. First, crowd workers systematically **increase lexical diversity**. We use type-token ratio (TTR), the

ratio between the number of word types and the number of tokens, to estimate the lexical diversity of a story (Hardie and McEnery, 2006). Figure 3 shows significant (p<.001, paired t-test) positive shifts of TTR for both AREL and GLAC, which confirms the findings in Hsu *et al.* (2019). Figure 3 also indicates that GLAC generates stories with higher lexical diversity than that of AREL.
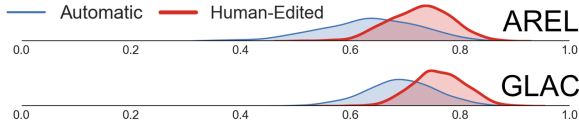

Figure 3: KDE plot of type-token ratio (TTR) for pre-/post-edited stories. People increase lexical diversity in machine-generated stories for both AREL and GLAC.

Second, people **shorten AREL's stories but lengthen GLAC's stories.** We calculate the average number of Part-Of-Speech (POS) tags for tokens in each story using the python NLTK (Bird et al., 2009) package, as shown in Table 1. We also find that the average number of tokens in an AREL story (43.0, SD=5.0) decreases (41.9, SD=5.6) after human editing, while that of GLAC (35.0, SD=4.5) increases (36.7, SD=5.9). Hsu has observed that people often replace "determiner/article + noun" phrases (*e.g.,* "a boy") with pronouns (*e.g.,* "he") in AREL stories (2019). However, this observation cannot explain the story lengthening in GLAC, where each story on average has an increased 0.9 nouns after editing. Given the average per-story edit distances (Levenshtein, 1966; Damerau, 1964) for AREL (16.84, SD=5.64) and GLAC (17.99, SD=5.56) are similar, this difference is unlikely to be caused by deviation in editing amount.

AREL	.	ADJ	ADP	ADV	CONJ	DET	NOUN	PRON	PRT	VERB	Total
Pre	5.2	3.1	3.5	1.9	0.5	8.1	10.1	2.1	1.6	6.9	43.0
Post	4.7	3.1	3.4	1.9	0.8	7.1	9.9	2.3	1.6	7.0	41.9
Δ	-0.5	0.0	-0.1	-0.1	0.4	-1.0	-0.2	0.2	0.0	0.1	-1.2

GLAC	.	ADJ	ADP	ADV	CONJ	DET	NOUN	PRON	PRT	VERB	Total
Pre	5.0	3.3	1.7	1.9	0.2	6.5	7.4	1.2	0.8	6.9	35.0
Post	4.5	3.2	2.4	1.8	0.8	6.1	8.3	1.5	1.0	7.0	36.7
Δ	-0.5	-0.1	0.7	-0.1	0.6	-0.3	0.9	0.3	0.2	0.1	1.7

Table 1: Average number of tokens with each POS tag per story. (Δ: the differences between post- and pre-edit stories. NUM is omitted because it is nearly 0. Numbers are rounded to one decimal place.)

Deleting extra words requires much less time than other editing operations (Popovic et al., 2014). Per Figure 3, AREL's stories are much

more repetitive. We further analyze the type-token ratio for nouns (TTR_{noun}) and find AREL generates duplicate nouns. The average TTR_{noun} of an AREL's story is 0.76 while that of GLAC is 0.90. For reference, the average TTR_{noun} of a human-written story (the entire VIST dataset) is 0.86. Thus, we hypothesize workers prioritized their efforts in deleting repetitive words for AREL, resulting in the reduction of story length.

4 Baseline Experiments

We report baseline experiments on the visual story post-editing task in Table 2. AREL's post-editing models are trained on the augmented AREL training set and evaluated on the AREL test set of *VIST-Edit*, and GLAC's models are tested using GLAC sets, too. Figure 4 shows examples of the output. Human evaluations (Table 2) indicate that the post-editing model improves visual story quality.

4.1 Methods

Two neural approaches, Long short-term memory (LSTM) and Transformer, are used as baselines, where we experiment using *(i)* text only (T) and *(ii)* both text and images (T+I) as inputs.

LSTM An LSTM seq2seq model is used (Sutskever et al., 2014). For the text-only setting, the original stories and the human-edited stories are treated as source-target pairs. For the text-image setting, we first extract the image features using the pre-trained ResNet-152 model (He et al., 2016) and represent each image as a 2048-dimensional vector. We then apply a dense layer on image features in order to both fit its dimension to the word embedding and learn the adjusting transformation. By placing the image features in front of the sequence of text embedding, the input sequence becomes a matrix $\in \mathbb{R}^{(5+len) \times dim}$, where len is the text sequence length, 5 means 5 photos, and dim is the dimension of the word embedding. The input sequence with both image information and text information is then encoded by LSTM, identical as in the text-only setting.

Transformer (TF) We also use the Transformer architecture (Vaswani et al., 2017) as baseline. The text-only setup and image feature extraction are identical to that of LSTM. For Transformer, the image features are attached at the end of the sequence of text embedding to form an image-

Edited By	AREL						GLAC					
	Focus	Coherence	Share	Human	Grounded	Detailed	Focus	Coherence	Share	Human	Grounded	Detailed
N/A	3.487	3.751	3.763	3.746	3.602	3.761	3.878	3.908	3.930	3.817	3.864	3.938
TF (T)	3.433	3.705	3.641	3.656	3.619	3.631	3.717	3.773	3.863	3.672	3.765	3.795
TF (T+I)	**3.542**	3.693	3.676	3.643	3.548	3.672	3.734	3.759	3.786	3.622	3.758	3.744
LSTM (T)	**3.551**	**3.800**	**3.771**	**3.751**	**3.631**	**3.810**	**3.894**	3.896	3.864	**3.848**	3.751	3.897
LSTM (T+I)	**3.497**	3.734	3.746	3.742	3.573	3.755	3.815	3.872	3.847	3.813	3.750	3.869
Human	3.592	3.870	3.856	3.885	3.779	3.878	4.003	4.057	4.072	3.976	3.994	4.068

Table 2: Human evaluation results. Five human judges on MTurk rate each story on the following six aspects, using a 5-point Likert scale (from Strongly Disagree to Strongly Agree): Focus, Structure and Coherence, Willing-to-Share ("I Would Share"), Written-by-a-Human ("This story sounds like it was written by a human."), Visually-Grounded, and Detailed. We take the average of the five judgments as the final score for each story. LSTM(T) improves all aspects for stories by AREL, and improves "Focus" and "Human-like" aspects for stories by GLAC.

enriched embedding. It is noteworthy that the position encoding is only applied on text embedding. The input matrix $\in \mathbb{R}^{(len+5) \times dim}$ is then passed into the Transformer as in the text-only setting.

4.2 Experimental Setup and Evaluation

Data Augmentation In order to obtain sufficient training samples for neural models, we pair *less*-edited stories with *more*-edited stories of the same photo sequence to augment the data. In *VIST-Edit*, five human-edited stories are collected for each photo sequence. We use the human-edited stories that are less edited – measured by its Normalized Damerau-Levenshtein distance (Levenshtein, 1966; Damerau, 1964) to the original story – as the source and pair them with the stories that are more edited (as the target.) This data augmentation strategy gives us in total fifteen ($\binom{5}{2} + 5 = 15$) training samples given five human-edited stories.

Human Evaluation Following the evaluation procedure of the first VIST Challenge (Mitchell et al., 2018), for each visual story, we recruit five human judges on MTurk to rate it on six aspects (at $0.1/HIT.) We take the average of the five judgments as the final scores for the story. Table 2 shows the results. The LSTM using text-only input outperforms all other baselines. It improves all six aspects for stories by AREL, and improves "Focus" and "Human-like" aspects for stories by GLAC. These results demonstrate that a relatively small set of human edits can be used to boost the story quality of an existing large VIST model. Table 2 also suggests that the quality of a post-edited story is heavily decided by its pre-edited version. Even after editing by human editors, AREL's stories still do not achieve the quality of pre-edited stories by GLAC. The inefficacy of image features and Transformer model might be caused by the small size of *VIST-Edit*. It also requires further

research to develop a post-editing model in a multimodal context.

5 Discussion

Automatic evaluation scores do not reflect the quality improvements. APE for MT has been using automatic metrics, such as BLEU, to benchmark progress (Libovický et al., 2016). However, classic automatic evaluation metrics fail to capture the signal in human judgments for the proposed visual story post-editing task. We first use the human-edited stories as references, but all the automatic evaluation metrics generate *lower* scores when human judges give a higher rating (Table 3.)

Reference: AREL Stories **Edited** by Human

	BLEU4	METEOR	ROUGE	Skip-Thoughts	Human Rating
AREL	0.93	0.91	0.92	0.97	**3.69**
AREL Edited By LSTM(T)	0.21	0.46	0.40	0.76	**3.81**

Table 3: Average evaluation scores for AREL stories, using the human-edited stories as references. All the automatic evaluation metrics generate lower scores when human judges give a higher rating.

We then switch to use the human-written stories (VIST test set) as references, but again, all the automatic evaluation metrics generate *lower* scores even when the editing was done by human (Table 4.)

Reference: Human-**Written** Stories

	BLEU4	METEOR	ROUGE	Skip-Thoughts
GLAC	0.03	0.30	0.26	0.66
GLAC Edited By Human	0.02	0.28	0.24	0.65

Table 4: Average evaluation scores on GLAC stories, using human-written stories as references. All the automatic evaluation metrics generate lower scores even when the editing was done by human.
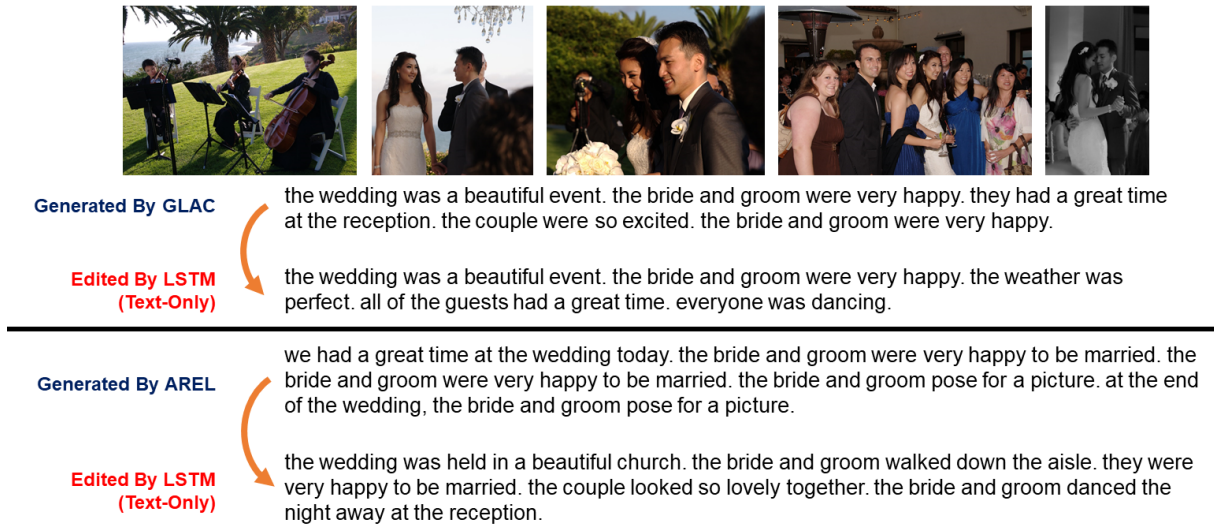
Generated By GLAC

the wedding was a beautiful event. the bride and groom were very happy. they had a great time at the reception. the couple were so excited. the bride and groom were very happy.

Edited By LSTM (Text-Only)

the wedding was a beautiful event. the bride and groom were very happy. the weather was perfect. all of the guests had a great time. everyone was dancing.

Generated By AREL

we had a great time at the wedding today. the bride and groom were very happy to be married. the bride and groom were very happy to be married. the bride and groom pose for a picture. at the end of the wedding, the bride and groom pose for a picture.

Edited By LSTM (Text-Only)

the wedding was held in a beautiful church. the bride and groom walked down the aisle. they were very happy to be married. the couple looked so lovely together. the bride and groom danced the night away at the reception.

Figure 4: Example stories generated by baselines.

	Data Includes	BLEU4	METEOR	ROUGE	Skip-Thoughts
①	AREL	.110	.099	.063	.062
②	LSTM-Edited AREL	.106	.109	.067	.205
③	①+②	.095	.092	.059	.116
④	GLAC	.222	.203	.140	.151
⑤	LSTM-Edited GLAC	.163	.176	.138	.087
⑥	④+⑤	.196	.194	.148	.116
⑦	①+④	.091	.086	.059	.088
⑧	②+⑤	.089	.103	.067	.101
⑨	①+②+④+⑤	.090	.096	.069	.094

Spearman rank-order correlation ρ

Table 5: Spearman rank-order correlation ρ between the automatic evaluation scores (sum of all six aspects) and human judgment. When comparing among machine-edited stories (② and ⑤), among pre- and post-edited stories (③ and ⑥), or among any combinations of them (⑦, ⑧ and ⑨), all metrics result in weak correlations with human judgments.

Table 5 further shows the Spearman rank-order correlation ρ between the automatic evaluation scores (sum of all six aspects) and human judgment calculated using different data combination. In row ④ of Table 5, the reported correlation ρ of METEOR is consistent with the findings in Huang *et al.* (2016), which suggests that METEOR could be useful when comparing among stories generated by the same visual storytelling model. However, when comparing among machine-edited stories (row ② and ⑤), among pre- and post-edited stories (row ③ and ⑥), or among any combinations of them (row ⑦, ⑧ and ⑨), all metrics result in weak correlations with human judgments. These results strongly suggest the need of a new automatic evaluation metric for visual story post-editing task. Some new metrics have recently been introduced using linguistic (Roemmele and Gor-

don, 2018a) or story features (Purdy et al., 2018) to evaluate story automatically. More research is needed to examine whether these metrics are useful for story post-editing tasks too.

6 Conclusion

VIST-Edit, the first dataset for **human edits of machine-generated visual stories**, is introduced. We argue that human editing on machine-generated stories is unavoidable, and such edited data can be leveraged to enable automatic post-editing. We have established baselines for the task of visual story post-editing, and have motivated the need for a new automatic evaluation metric.

References

Ramón Astudillo, João Graça, and André Martins. 2018. Proceedings of the amta 2018 workshop on translation quality estimation and automatic post-editing. In *Proceedings of the AMTA 2018 Workshop on Translation Quality Estimation and Automatic Post-Editing*.

Hanna Béchara, Yanjun Ma, and Josef van Genabith. 2011. Statistical post-editing for a statistical mt system. In *MT Summit*, volume 13.

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

Fred J Damerau. 1964. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):171–176.

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)*, pages 955–964.

Andrew Hardie and Tony McEnery. 2006. *Statistics.*, volume 12, pages 138–146. Elsevier.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Ting-Yao Hsu, Yen-Chia Hsu, and Ting-Hao K. Huang. 2019. On how users edit computer-generated visual stories. In *Proceedings of the 2019 CHI Conference Extended Abstracts (Late-Breaking-Work) on Human Factors in Computing Systems*. ACM.

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*, pages 1233–1239.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2016. Log-linear combinations of monolingual and bilingual neural machine translation models for automatic post-editing. *arXiv preprint arXiv:1605.04800*.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2017. An exploration of neural sequence-to-sequence architectures for automatic post-editing. *arXiv preprint arXiv:1706.04138*.

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*.

Stephen King. 2000. On writing: A memoir ofthe craft. *New Yorle: Scrihner*.

Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710.

Jindřich Libovickỳ and Jindřich Helcl. 2017. Attention strategies for multi-source sequence-to-sequence learning. *arXiv preprint arXiv:1704.06567*.

Jindřich Libovickỳ, Jindřich Helcl, Marek Tlustỳ, Ondřej Bojar, and Pavel Pecina. 2016. Cuni system for wmt16 automatic post-editing and multimodal translation tasks. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, volume 2, pages 646–654.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

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)*.

Margaret Mitchell, Francis Ferraro, Ishan Misra, et al. 2018. Proceedings of the first workshop on storytelling. In *Proceedings of the First Workshop on Storytelling*.

Maja Popovic, Arle Lommel, Aljoscha Burchardt, Eleftherios Avramidis, and Hans Uszkoreit. 2014. Relations between different types of post-editing operations, cognitive effort and temporal effort. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation (EAMT 14)*, pages 191–198.

Christopher Purdy, Xinyu Wang, Larry He, and Mark Riedl. 2018. Predicting generated story quality with quantitative measures. In *Fourteenth Artificial Intelligence and Interactive Digital Entertainment Conference*.

Melissa Roemmele and Andrew Gordon. 2018a. Linguistic features of helpfulness in automated support for creative writing. In *Proceedings of the First Workshop on Storytelling*, pages 14–19.

Melissa Roemmele and Andrew S Gordon. 2018b. Automated assistance for creative writing with an rnn language model. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion*, page 21. ACM.

Michel Simard, Cyril Goutte, and Pierre Isabelle. 2007. Statistical phrase-based post-editing. In *Proceedings of NAACL HLT*, pages 508–515.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Mark Twain. 1876. *The Adventures of Tom Sawyer*. American Publishing Company.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

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*, Melbourne, Victoria, Australia. ACL.