StoryER: Automatic Story Evaluation via Ranking, Rating and Reasoning

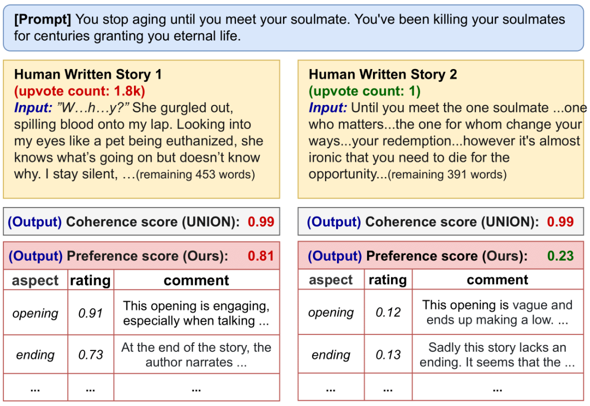
Hong Chen[1,3], **Duc Minh Vo**[1], **Hiroya Takamura**[2,3], **Yusuke Miyao**[1,3], **Hideki Nakayama**[1,3]
The University of Tokyo[1], Tokyo Institute of Technology[2]
National Institute of Advanced Industrial Science and Technology, Japan[3]
{chen, vmduc, nakayama}@nlab.ci.i.u-tokyo.ac.jp
yusuke@is.s.u-tokyo.ac.jp
takamura.hiroya@aist.go.jp

Abstract

Existing automatic story evaluation methods place a premium on story lexical level coherence, deviating from human preference. We go beyond this limitation by considering a novel **Story E**valuation method that mimics human preference when judging a story, namely **StoryER**, which consists of three sub-tasks: **R**anking, **R**ating and **R**easoning. Given either a machine-generated or a human-written story, StoryER requires the machine to output 1) a preference score that corresponds to human preference, 2) specific ratings and their corresponding confidences and 3) comments for various aspects (e.g., opening, character-shaping). To support these tasks, we introduce a well-annotated dataset comprising (i) 100k ranked story pairs; and (ii) a set of 46k ratings and comments on various aspects of the story. We fine-tune Longformer-Encoder-Decoder (LED) on the collected dataset, with the encoder responsible for preference score and aspect prediction and the decoder for comment generation. Our comprehensive experiments result in a competitive benchmark for each task, showing the high correlation to human preference. In addition, we have witnessed the joint learning of the preference scores, the aspect ratings, and the comments brings gain in each single task. Our dataset and benchmarks are publicly available to advance the research of story evaluation tasks.[1]

1 Introduction

Even for humans, evaluating story quality is a challenging task. Although many literature criteria have been proposed, the most straightforward way is to count how many readers like the story which is referred as to human preference. Bearing it in mind, story writing community usually uses upvote count as a story quality criterion. As shown in Fig. 1, more readers like the left story (upvote count = 1.8k) rather than the right one (upvote count = 1).



[Prompt] You stop aging until you meet your soulmate. You've been killing your soulmates for centuries granting you eternal life.

Human Written Story 1
(upvote count: 1.8k)
Input: *"W…h…y?"* She gurgled out, spilling blood onto my lap. Looking into my eyes like a pet being euthanized, she knows what's going on but doesn't know why. I stay silent, …(remaining 453 words)

(Output) Coherence score (UNION): **0.99**

(Output) Preference score (Ours): **0.81**

aspect	rating	comment
opening	*0.91*	This opening is engaging, especially when talking ...
ending	*0.73*	At the end of the story, the author narrates ...
...	...	...

Human Written Story 2
(upvote count: 1)
Input: Until you meet the one soulmate ...one who matters...the one for whom change your ways...your redemption...however it's almost ironic that you need to die for the opportunity...(remaining 391 words)

(Output) Coherence score (UNION): **0.99**

(Output) Preference score (Ours): **0.23**

aspect	rating	comment
opening	*0.12*	This opening is vague and ends up making a low. ...
ending	*0.13*	Sadly this story lacks an ending. It seems that the ...
...	...	...

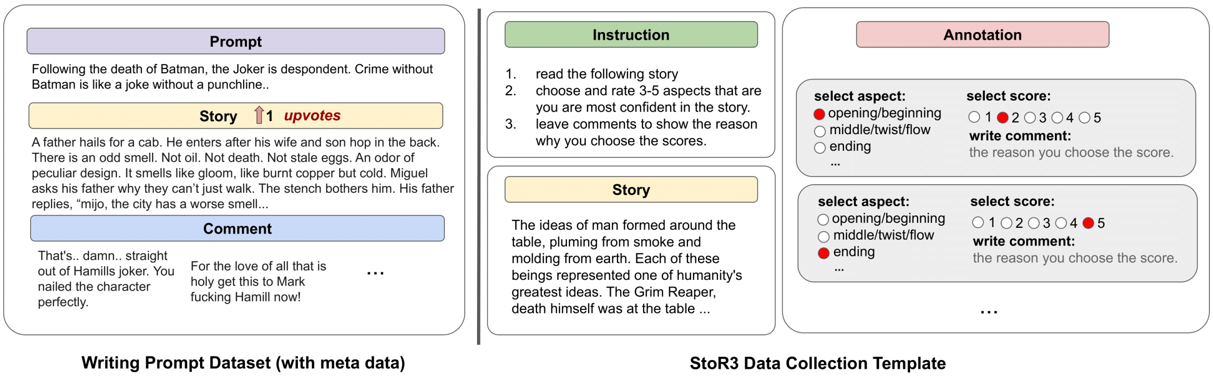
Figure 1: The existing story evaluation method (UNION) outputs a score for estimating the coherence of the stories, while human-written stories rarely suffer from this problem. Our model (Ours) which is trained by comparing two stories (Ranking), evaluates the story based on human preference (i.e., upvote counts), produces scores for various aspects (Rating), and leaves comments (Reasoning). Our model is applicable to both machine-generated and human-written stories.

Existing methods which use *referenced metrics* (e.g., BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE (Lin, 2004)) and *unreferenced metrics* (e.g., UNION (Guan and Huang, 2020), MANPLTS (Ghazarian et al., 2021)), deviate from human preference (Fig. 1). On the contrary, we aim to explicitly evaluate a story, introducing a human preference-liked system consisting of three subtasks: Ranking, Rating and Reasoning.

We build a model upon Longformer-Encoder-Decoder (LED) (Beltagy et al., 2020), where the encoder predicts the preference score (Ranking), aspect ratings and confidences (Rating) while the decoder generates the comments (Reasoning). Inspired by widely-used pairwise comparison in story evaluation, we train our model with the ranking objectives. In this way, the score margin between

[1] Dataset and pre-trained model demo are available at anonymous website http://storytelling-lab.com/eval and https://github.com/sairin1202/StoryER

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 1739–1753
December 7-11, 2022 ©2022 Association for Computational Linguistics



Following the death of Batman, the Joker is despondent. Crime without Batman is like a joke without a punchline..

A father hails for a cab. He enters after his wife and son hop in the back. There is an odd smell. Not oil. Not death. Not stale eggs. An odor of peculiar design. It smells like gloom, like burnt copper but cold. Miguel asks his father why they can't just walk. The stench bothers him. His father replies, "mijo, the city has a worse smell...

That's.. damn.. straight out of Hamills joker. You nailed the character perfectly.

For the love of all that is holy get this to Mark fucking Hamill now!

1. read the following story
2. choose and rate 3-5 aspects that are you are most confident in the story.
3. leave comments to show the reason why you choose the scores.

The ideas of man formed around the table, pluming from smoke and molding from earth. Each of these beings represented one of humanity's greatest ideas. The Grim Reaper, death himself was at the table ...

Figure 2: The Writing Prompt Dataset with metadata (left) contains prompt, story, upvotes, and comments from readers. Our dataset collection pipeline (right) shows the template for data collection. We ask the workers to select 3-5 aspects, score each aspect 1-5 from poor to good and leave the comments that shows the reason for the score they rated.

good and poor stories are enlarged, resulting in high correlation between human preference and our predicted preference score (Fig. 1). We also witness that our performance is improved when we conduct joint training on three subtasks.

In aid of the proposed task, we present a well-annotated crowd-sourcing dataset, consisting of two parts. (i) One is built from 63,929 stories and their corresponding upvote counts provided in WritingPrompt dataset (WP) (Fan et al., 2018) (Figure 2 (left)) by pairing one highly-upvoted story (upvotes $\geq$ 50) and one lowly-upvoted story (upvotes $\leq$ 0) under the same prompt. As a result, we obtain 100k pairs of stories, namely *100k story ranking data*, used to train and evaluate the preference score prediction. (ii) The other part is made up of 45,948 aspect comments and their respective rating scores (1-5) by Amazon Mechanical Turk (AMT) and augmented data (Section 3.2), namely *46k aspect rating and reasoning data*, used for model explanation. Our contributions are three-fold:

- This study addresses a novel task StoryER, that consists of preference score prediction, aspect rating and comment generation.
- We introduce a new dataset for StoryER task and create benchmarks to promote the story evaluation research.
- Comprehensive experiments and intensive analysis indicate our preference score prediction outperforms previous metrics, and more accurately reflects human preference. Aspect rating and comment generation also helps in the evaluation and provide explanations. Moreover, we point out the remaining challenges under various scenarios in the hope that facilitates future research.

2 Related work

Overlap-based metrics such as BLEU (Sulem et al., 2018) and ROUGE (Lin, 2004) calculate lexical matches (i.e., n-gram matching) and reward the words that resemble the reference in their surface form, even if they do not accurately capture meaning, and penalize other paraphrases. Recent research (Edunov et al., 2020) indicates that these metrics do not reflect human preferences, particularly for open-ended text generation tasks.

Neural-based metrics are motivated by the success of transformers as multitask learners (Vaswani et al., 2017), and adapt them for the task of neural language evaluation. When compared to overlap-based metrics, BERTScore (Zhang et al., 2019), MoverScore (Zhao et al., 2019), BLEURT (Sellam et al., 2020) report stronger correlations with human judgment. For specific use, in open dialogue generation, Adem (Lowe et al., 2017) captures semantic similarity beyond word overlap statistics, and exploits both the context and the reference response to calculate its score for the model response. RUBER (Tao et al., 2018) and its variant, RUBER-BERT (Ghazarian et al., 2019) evaluates a reply by taking into consideration both a ground-truth reply and a query without requiring labels of human satisfaction and can be extended to different datasets and languages.

Neural discriminator is proposed particularly for story evaluation. The metrics mentioned above show limited performance in story evaluation as demonstrated in Guan et al. (2021). UNION (Guan and Huang, 2020) and MANPLTS (Ghazarian et al., 2021) analyze the problem from machine-generated stories and generate negative data by heuristics and plot manipulation, and then distin-

	#prompt	$\#S_{high}$ (word)	$\#S_{low}$ (word)	#pairs
Train	5892	10371 (491.01)	26246 (453.06)	66336
Val	2280	3816 (473.27)	11458 (446.40)	27748
Test	2280	3906 (488.32)	8132 (454.87)	22887

Table 1: Data statistics of 100k story ranking data. #prompt denotes the number of unique prompts, $\#S_{high}$ and $\#S_{low}$ denotes the number of highly-voted stories and lowly-voted stories. We also show the averaged word count in the parentheses. #pairs shows the number of ranked story pairs.

guish by a BERT-based model (Devlin et al., 2019). The coherence score they produce can be expressed as the probability of the story being identified as human-written story. In this paper, we require our model to follow human preference, not only the coherence, which we believe is a more general way of story evaluation.

3 Dataset

Our dataset comprises of two parts: 100k story ranking, and 46k aspect rating and reasoning.[2]

3.1 100k Story Ranking Data

As we mentioned above, ranking method is more flexible and better than discrimination when evaluating the story (we also experimentally compare them in Sec. F.1). We thus prepare 100k pairwise ranking data for training the model. To this end, we first collect 193,842 stories prior to 03/2020 from WP[3] along with their prompt, the number of upvotes and uncategorized comments. We remove the stories updated from 12/2019 to 03/2020, since newly-updated stories usually have few upvotes regardless of whether they are good or bad. Then, we exclusively keep stories with the word count between 200 and 800. Finally, we pick two stories from the same prompt, one highly upvoted (i.e., upvotes ≥ 50 [4]) and one lowly upvoted (i.e., upvotes ≤ 0), resulting in a total of 63,929 unique stories and 116,971 story pairs. We split the story pairs based on the prompts into training, validation and testing (Table 1), to ensure that each division receives a unique set of prompts.

3.2 46k Aspect Rating and Reasoning Data

Apart from the preference score, we require our model to provide ratings and comments on predefined aspects to aid in the explanation of the predicted preference score.

Aspect category extraction. To begin with, we must determine which aspects in the content should be measured. As some readers leave comments to explain why they upvote or downvote the stories, a straightforward way is to extract aspect categories based on those uncategorized comments. We therefore adopt latent Dirichlet allocation (LDA), which models the documents with a certain number of topics, based upon the co-occurrence of individual words. More precisely, we follow Brody and Elhadad (2010) to treat each comment as a separate document. LDA can produce a distribution of frequency of occurrence for each word in the topics. We optimize LDA through a cluster validation scheme, and obtain the optimal number of aspects 10. Based on the most representative words in each topic, we manually name each topic as the aspect category. These aspect categories are defined using some widely used aspects inspired from the websites.[5]

Comment and aspect collection. Comments in WP meta data are neither categorized with aspect categories, nor labeled with ratings, and some of them are totally irrelevant to the content. More importantly, there is a bias towards positive comments, which implies that not too many readers are willing to leave comments on poor stories. Therefore, we collect new comments via crowd-sourcing. By learning from these well-annotated comment data, we train neural models to filter out noisy data from comments in WP meta data. To collect the data, we ask workers from AMT to select aspects, rate sentiment and leave comments on 5,964 unique stories from WP. For increasing the diversity of comments, some stories are allocated to two different annotators, resulting in a total of 9,112 submissions (i.e., 1.53 annotations/story). As shown in Figure 2 (right), each story requires the annotators to rate (normalized to 0-1) and leave comments on 3 to 5 aspects that are most confident by the workers. The final statistics of the comments is listed in Table 2.

Comment augmentation. The noisy comments in WP meta data then can be classified and analyzed by two models: aspect category classification model and comment sentiment analysis model that

[2] All data collection follows the same procedure as described in the previous work (Fan et al., 2018) on Reddit, which comply with ACL Code of Ethics.

[3] https://huggingface.co/datasets/rewardsignal/reddit_writing_prompts

[4] we notice that some stories that receive upvotes ≥ 50 can be listed in /r/bestofWritingPrompts/

[5] list in the supplementary material

	#comment	#comment*	rate (1-5)	rate* (1-5)	comment_len	comment_len*
Structure						
opening/beginning	3615	5617	2.53	3.15	30.20	32.44
middle/twist/flow/conflict	3967	5971	2.24	2.78	30.59	31.63
ending	5610	7615	2.13	2.49	31.48	31.59
Writing Style						
character shaping	5101	7102	2.21	2.53	31.57	34.23
scene description	4168	6172	2.18	2.53	31.75	39.30
Type						
heartwarming/touching (Romance)	426	1866	2.99	4.39	32.05	32.64
sad/crying/tragedy (Tragedy)	462	1680	3.12	3.93	30.85	34.67
horror/scary (Horror)	815	1985	2.49	3.61	30.92	33.24
funny/hilarious/laugh (Comedy)	1153	3156	3.25	3.96	30.04	30.91
novelty/good idea/brilliant (Fiction)	2782	4784	2.49	3.26	32.51	32.70
Overall	28099	45948	2.56	3.26	31.20	33.33

Table 2: Data statistics in 46k aspect rating and reasoning data. * denotes the data statistics after data augmentation. We list the number of comments with rating scores (2nd and 3rd columns), averaged rating scores (4th and 5th columns) and averaged word count (6th and 7th columns).

trained with our collected data. The training details can be found in the supplementary material. We filter out irrelevant comments by eliminating those with no values in aspect categories that exceeds 0.9 after softmax and retain the comments with the word count ranged from 15 to 50. The remaining comments are then rated by the their sentiments. Finally, we obtain 17,849 valuable comments for 6,705 additional unique stories and merge them into our collected data, resulting in a total number of 45,948 for comments and 12,669 for unique stories. We split the collected data into training, validation, and test data in the ratio of 8:1:1 and put the augmented data into the training data (Table 2).

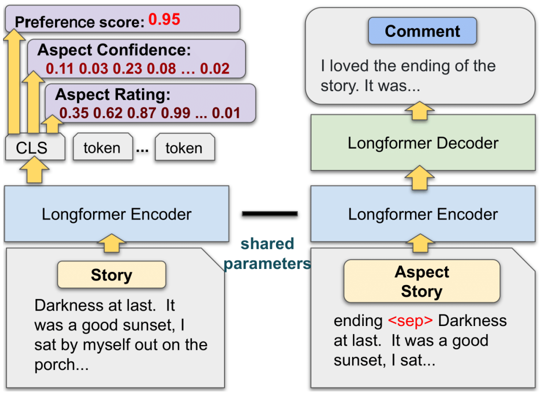


Figure 3: Overview of our model. The encoder (left) predicts the preference score, aspect confidence, and aspect rating. The decoder (right) generates the comment for each aspect.

4 StoryER

4.1 Task Definition

Given a story $\mathbf{s}$, the task is to output a set $\{p_s, \mathbf{a}^{\mathrm{c}}, \mathbf{a}^{\mathrm{r}}, \mathbf{c}\}$ where p_s denotes the preference score of the story $\mathbf{s}$, which is used for comparing story quality. For more explicit explanation, we further output confidence scores $\mathbf{a}^{\mathrm{c}} = \{a_k^{\mathrm{c}}\}_{k=1}^{K}$, aspect ratings $\mathbf{a}^{\mathrm{r}} = \{a_k^{\mathrm{r}}\}_{k=1}^{K}$, and comments $\mathbf{c} = \{c_k\}_{k=1}^{K}$ for K aspects ($K = 10$ in our experiments), respectively. Confidence scores $\mathbf{a}^{\mathrm{c}}$ reflect the likelihood of utilizing the specific aspects as measures, as some aspects (e.g., horror) are not applicable in some stories (e.g., comic story). Aspect ratings a_k^{r} are considered as the scores of each aspect. Comments $\mathbf{c}$ demonstrate the reason that the reader upvotes/downvotes the story, producing a more explicit explanation for the aspect rating. We assume $\sum_{k=1}^{K} a_k^{\mathrm{c}} = 1$ for aspect confidence, and $a_k^{\mathrm{r}} \in [0, 1]$ for aspect rating, which is calculated separately during the training.

Please note that aspect rating and comment generation results are not used as metrics in this work, while they are used for 1) improving preference score prediction by joint learning, and 2) producing explanation. Investigating how to include them into metrics is a future direction for this research.

4.2 Learning a Story Evaluator

Following Ghazarian et al. (2021), we use Longformer-Encoder-Decoder (LED) (Beltagy et al., 2020) to produce a preference score, as well as ratings and comments for the pre-defined aspects. As shown in Figure 3, we encode the story $\mathbf{s}$, and use its feature on the special token (i.e., [CLS]) to predict the preference score p_s, aspect confidence $\mathbf{a}^{\mathrm{c}}$ and rating $\mathbf{a}^{\mathrm{r}}$ by additional layers. For generating comments, we concatenate the story with aspect category name with a special token (i.e., <sep>), and send it into the same encoder. The decoder outputs the comment $\mathbf{c}$ that implies the performance of the story on the given aspect.

4.3 Task 1: Preference Score Prediction (Ranking)

Our model learns to predict the preference score by ranking two stories from the same prompt. As shown in Figure 3, we use the feature of [CLS] in the story, following a linear layer with sigmoid activation and finally turning it into a scalar score. We take Margin Ranking Loss to enlarge the margin gap m of the scores between stories with high and low upvotes:

$$\mathcal{L}_{\mathrm{ps}} = \max(0, \sigma(\mathbf{W}_{\mathrm{ps}}\mathbf{v}_{s_{low}}) - \sigma(\mathbf{W}_{\mathrm{ps}}\mathbf{v}_{s_{high}}) + m), \tag{1}$$

where $\mathbf{W}_{\mathrm{ps}}$ denotes a linear layer for the feature of the story $\mathbf{v}_s$. $\sigma(\cdot)$ is the sigmoid activation function. s_{high} and s_{low} represent the highly-upvoted and lowly-upvoted stories.

Negative sample. Machine-generated stories often suffer from the coherence and consistency problem, while human-written stories usually do not. Therefore our model trained on human-written stories can hardly evaluate story coherence. To enable our model to evaluate story considering coherence issues, we further train our model (Ours (N)) with negative stories that are generated by the methods in the previous works (Guan and Huang, 2020; Ghazarian et al., 2021). We change the margin ranking loss as follow:

$$\begin{aligned}\mathcal{L}_{\mathrm{pref}} &= \max(0, \sigma(\mathbf{W}_{\mathrm{ps}}\mathbf{v}_{s_{low}}) - \sigma(\mathbf{W}_{\mathrm{ps}}\mathbf{v}_{s_{high}}) + m),\\ \mathcal{L}_{\mathrm{coh}} &= \max(0, \sigma(\mathbf{W}_{\mathrm{ps}}\mathbf{v}_{s_{neg}}) - \sigma(\mathbf{W}_{\mathrm{ps}}\mathbf{v}_{s_{low}}) + m),\\ \mathcal{L}_{\mathrm{ps}} &= \mathcal{L}_{\mathrm{pref}} + \mathcal{L}_{\mathrm{coh}},\end{aligned} \tag{2}$$

where s_{neg} denotes the negative stories derived from the previous works. In each iteration, we takes two pairs as training data: s_{high} and s_{low}, s_{low} and s_{neg}.

4.4 Task 2: Aspect Confidence and Rating Prediction (Rating).

We adopt two additional linear layers on the same feature $\mathbf{v}_s$ used in the story ranking. One is with learnable parameters $\mathbf{W}_{\mathrm{a^c}}$, outputting confidence scores $\mathbf{a}^{\mathrm{c}} = \mathrm{softmax}(\mathbf{W}_{\mathrm{a^c}}\mathbf{v}_s)$. The other one has $\mathbf{W}_{\mathrm{a^r}}$, producing aspect rating $\mathbf{a}^{\mathrm{r}} = \sigma(\mathbf{W}_{\mathrm{a^r}}\mathbf{v}_s)$. Let $\mathbf{y}_{\mathrm{a^c}} \in \{0,1\}^K$, $\mathbf{y}_{\mathrm{a^r}} \in [0,1]^K$ be the ground-truth confidence and rating, we define the confidence and rating loss functions as follows:

$$\mathcal{L}_{\mathrm{a^c}} = -\sum_{k=1}^{K} \mathbf{y}_{a^c}[k]\log \mathbf{a}^{\mathrm{c}}[k], \tag{3}$$

$$\begin{aligned}\mathcal{L}_{\mathrm{a^r}} = -\sum_{k\in M_s} &\mathbf{y}_{\mathrm{a^r}}[k]\log \mathbf{a}^{\mathrm{r}}[k]\\ &+ (1-\mathbf{y}_{\mathrm{a^r}}[k]) * \log(1-\mathbf{a}^{\mathrm{r}}[k]).\end{aligned} \tag{4}$$

We calculate the multi-class cross-entropy loss for the aspect confidence. $\mathbf{y}_{\mathrm{a^c}}[k] = 1$ if the k-th aspect is selected, otherwise $\mathbf{y}_{\mathrm{a^c}}[k] = 0$. For aspect rating, binary cross-entropy loss is calculated separately for each selected aspects. M_s denotes the set of aspects that are selected for story $\mathbf{s}$. $\mathbf{y}_{\mathrm{a^r}}[k]$ denotes the normalized rating score for the k-th aspect.

4.5 Task 3: Comment Generation (Reasoning).

The comments are generated conditioned on the aspect $\mathbf{a}$ and the story $\mathbf{s}$. We input the concatenation of the aspect category name, special token, story, and train the LED under Maximum Likelihood Estimation (MLE) with the comment as target:

$$\mathcal{L}_{\mathrm{c}}(p_\theta) = -\sum_{t=1}^{|\mathbf{c}|} \log p_\theta(\mathbf{c}_t \mid \mathbf{a}, \mathbf{s}, \mathbf{c}_{<t}), \tag{5}$$

where the $\mathbf{c}_t$ denotes the t-th token in the comment. For joint training three tasks, our final loss is the summation of all above loss functions:

$$\mathcal{L} = \mathcal{L}_{\mathrm{ps}} + \mathcal{L}_{\mathrm{a^c}} + \mathcal{L}_{\mathrm{a^r}} + \mathcal{L}_{\mathrm{c}}. \tag{6}$$

4.6 Hyperparameters

We conduct a comprehensive set of experiments to examine the effectiveness under different scenarios. We fine-tune pre-trained LED from Huggingface[6] with the batch size 16, the margin 0.3 and run 20k iterations for training (10 hours). We adopt AdamW optimizer (Loshchilov and Hutter, 2018) with an initial learning rate of 4e-6, warming up in the first epoch and decreasing by a linear schedule. The reported results are averaged by the best results from three models with the same structure but initialized with three different seeds. More details and code can be found in the Appendix.

5 Experiments

5.1 Compared Methods

We compare our method with several unreferenced metrics on open story evaluation: Perplexity, Ruber-bert (Ghazarian et al., 2019), UNION (Guan

[6] `https://huggingface.co/transformers/model_doc/led.html`

	Human Written Story						**Machine Generated Story**			
	Ranking		WP$_{200}$		SCARY$_{200}$		PREF$_{200}$		COH$_{200}$	
Methods	Acc(%)	Dis	ρ	τ	ρ	τ	ρ	τ	ρ	τ
PPL	61.80	-	0.181*	0.130*	0.409*	0.300*	-0.090	-0.067	-0.698*	-0.452*
ft-PPL	61.85	-	0.168	0.120	0.449*	0.328*	-0.098	-0.074	-0.690*	-0.443*
Ruber-bert	39.16	-0.032	-0.019	-0.014	0.032	0.026	-0.140	-0.100	-0.141	-0.092
UNION	48.74	0.003	0.001	0.001	0.156*	0.114	-0.068	-0.045	0.188*	0.132*
MANPLTS	53.08	0.016	0.124	0.107	-0.070	-0.061	0.164	0.130	0.729*	0.498*
Ours	**73.93**	**0.228**	**0.583***	**0.422***	**0.578***	**0.420***	**0.343***	**0.234***	0.194*	0.132*
Ours (N)	70.39	0.131	0.525*	0.377*	0.508*	0.366*	0.266*	0.188*	**0.747***	**0.536***

Table 3: Evaluation on preference score prediction. Compared with previous works, our predict scores more correctly match the human judgement. We conduct hypothesis test (Diedenhofen and Musch, 2015), and * denotes that $p \leq 0.01$.

and Huang, 2020), and MANPLTS (Ghazarian et al., 2021).

5.2 Preference Score Evaluation

5.2.1 Accuracy and Score Distance

We evaluate the predicted preference scores obtained by all compared methods on 100k Story Ranking test data. Pairwise Ranking Accuracy (Acc) is calculated as the percentage of the story with higher upvotes getting a higher score than the one with lower upvotes. We also compute the averaged score gap (Dis) between two stories in pairs. Table 3 (Human (Ranking)) indicates that existing methods on preference-aware story evaluation on human-written stories are close to random selection (i.e., Acc=0.5, Dis=0). In contrast, our method can successfully compare two stories and achieve an acceptable score gap between two stories.

5.2.2 Correlation with Human Judgments

We calculate the correlation between our predicted preference scores and human judgment for stories. We use the correlation metrics Spearman (ρ) (Zar, 1972) and Kendall (τ) (Schaeffer and Levitt, 1956), which are known to be beneficial in estimating monotonic associations for not normally distributed and ranked scores. We collect and annotate both human-written and machine-generated stories as our test data:

WP$_{200}$. We collect human judgments for the stories in WP (sampled from test data in Table 2), where each story is assigned to 8 annotators. Annotators are asked to rate each story on a scale of 1 to 5 (from poor to good). To ensure correctness, we follow Clark et al. (2021) to ask the annotators to compare the stories and write down the reason for clarification. We carefully detect the worker behavior and set traps inside the annotation (see Appendix for details). Finally, we obtain 100 highly-upvoted and 100 lowly-upvoted stories and average the human rates as the target scores in this test data, namely, WP$_{200}$ in the following experiments. Inside, we witness a higher score for highly-voted stories, proving our hypothesis that upvote counts reflect human preference.

SCARY$_{200}$. We crawled scary stories from Reddit (r/shortscarystories[7]), which are similar to the stories in WP but in a constrained story type. We use the same procedure for WP$_{200}$ to create another human-annotated test dataset, namely SCARY$_{200}$.

PREF$_{200}$. The same procedure is also used for collecting human annotation for machine-generated stories. We select 100 generated stories by LED trained with highly-voted stories in WP and 100 stories by another LED trained with lowly-voted stories. We manually ensure that the selected stories do not contain severe coherence issues, and ask the annotators to rate the stories based on whether they enjoy the stories.

COH$_{200}$. We use the same human collected data in the previous work (Ghazarian et al., 2021)[8], which focused on recognizing coherence issues in the machine-generated stories (e.g., repeat plots, conflict logic).

Results. Table 3 depicts the correlation between human and automatic evaluation metrics on preference (WP$_{200}$, SCARY$_{200}$ and PREF$_{200}$). We see that our method outperforms previous methods by a large margin on both human-written and machine-generated stories in terms of human preference. Not surprisingly, in Table 3 (COH$_{200}$), MANPLTS on story coherence evaluation is against our model, as coherence issue does not frequently happen in our training data (i.e., human-written stories).

We notice that preference-aware judgments

[7] https://www.reddit.com/r/shortscarystories/

[8] https://github.com/PlusLabNLP/Plot-guided-Coherence-Evaluation/tree/main/Data/AMT

				Ranking		WP$_{200}$		SCARY$_{200}$	
p_s	**a**	**c**	**N**	Acc	Dis	ρ	τ	ρ	τ
✓				71.10	0.212	0.557	0.401	0.533	0.389
✓	✓			71.99	0.221	0.525	0.378	**0.579**	0.417
✓		✓		72.15	0.207	0.580	0.421	0.510	0.371
✓	△	△		72.95	**0.229**	0.571	0.409	0.564	0.409
✓	✓	✓		**73.93**	0.228	**0.583**	**0.422**	0.578	**0.420**
✓	✓	✓	✓	69.02	0.119	0.525	0.377	0.508	0.366

Table 4: Ablation study on preference score prediction. All results are statistical significant $p < 0.01$. △ means that we use the collected data without augmentation. More results are listed in supplementary materials.

(PREF$_{200}$) and coherence-based judgments (COH$_{200}$) are distinct. Metrics that perform well in terms of coherence may perform poorly in terms of preference, and vice versa. To mitigate the gap between preference and coherence, we train our model using negative stories created by UNION and MANPLTS. As a result, Ours (N) shows rapidly increasing performance on the evaluation in terms of coherence with a bit of performance drop on the preference-aware evaluation, indicating a potential to take into account both coherence and human preference when evaluating a story.

6 Ablation Study

6.1 Preference Score Prediction

In this section, we further test the performance of preference score prediction combined with other components: aspects **a**, comments **c** and negative stories **N**. Table 4 summarizes the results by joint training. When aspects are used, performance decreases in the WP$_{200}$ but increases in the SCARY$_{200}$, and the pattern is reversed when comments are used. We also test the model performance trained with the dataset without data augmentation △, and we can see that our model trained with augmented data outperforms that with the original data, which shows the significance of data augmentation.

6.2 Aspect Evaluation

We evaluate our model for predicting confidence scores and ratings for the aspects. For confidence scores, we calculate the recall performance on top-k (i.e., k=1,3,5) on the test split of 46K Aspect Rating and Reasoning data to show the percentage of human selected aspects that can be involved within the aspects with top-k confidence. For ratings, we calculate the correlation between human annotation and our model prediction. Table 5 shows the results compared with joint training other two tasks. Story ranking and reasoning help the model output more correct confidence and ratings.

				Confidence			Rating	
p_s	**a**	**c**	**N**	R@1	R@3	R@5	ρ	τ
	✓			16.06	46.05	73.59	0.190	0.140
✓	✓			17.36	51.59	76.30	0.227*	0.168*
✓	✓	✓		**19.94**	**52.68**	**79.64**	**0.248***	**0.185***
✓	✓	✓	✓	19.88	51.44	79.20	0.216*	0.161*

Table 5: Evaluation on aspect confidence and rating. p_s, a, c, N denotes the preference score, aspects, comments and negative samples that are used in training our model respectively.

				Automatic			Human			
p_s	**a**	**c**	**N**	PPL	B	R	O	Rel(s)	Rel(a)	Rel(r)
		✓		7.31	8.45	16.63	47.61	**73.70**	79.20	-
✓	✓	✓		**7.06**	**8.60**	**16.76**	**49.40**	72.93	**82.83**	**58.33**
✓	✓	✓	✓	7.95	8.36	16.69	43.45	68.64	81.84	50.49

Table 6: Comment generation evaluation on automatic scores and human evaluation. In human evaluation, the kappa coefficient κ for each score are located in 0.4-0.6, indicating a moderate agreement between annotators.

6.3 Comment Evaluation

We evaluate the comment generation with automatic metrics and human evaluation. For automatic scores, we apply Perplexity (PPL), Averaged BLEU1-4 (B), ROUGE (R). For human evaluation, we mainly measure the relativeness between comments with the given story **Rel(s)**, aspect category **Rel(a)** and rating score (0-1 negative-positive) **Rel(r)**. We also measure **Overall (O)** quality by calculating the percentage of the comments that are agreed upon by annotators. Each comment is assigned to 5 annotators with a binary choice (i.e., related or not related, agree or not agree). From the result in Table 10, our generated comments are highly related to the given stories and the aspects. Together with the training on preference score prediction and aspect rating further improve the comment generation performance. The results so far show that the preference score, aspects, and comments all benefit one another, illustrating the significance of incorporating aspects and comments into our task.

7 Discussion

7.1 Pairwise Evaluation with StoryER

Given a set of prompts, two story generation models can generate stories based on the given prompt. We have two straightforward ways to compare two models using our proposed preference scores: 1) average the preference scores for stories on each model and compare the mean average scores. 2) perform pairwise comparisons for stories from the same prompt and get the preference percentage. We recommend the second method as it strictly follows our pairwise ranking strategy.

7.2 Domain Transfer in Preference Score

To show the generalization of evaluation metrics, we calculate the averaged predicted preference scores for data from different domains (see Table 7). We compute average scores on 1) lowly-voted (low) and highly-voted stories (high) on both WP_{200} and SCARY_{200}, 2) machine-generated stories by LED (LED), and with Plan-and-Write strategy (Yao et al., 2019) (P&W) trained separately on the highly-upvoted and lowly-upvoted stories, 3) negative stories generated from previous works (Guan and Huang, 2020; Ghazarian et al., 2021), 4) stories from other datasets: fairy tales (short stories), childbook dataset (Hill et al., 2015) and bookcorpus (Zhu et al., 2015).

As shown in Table 7, UNION and MANPLTS consistently produce higher scores for human-written stories (Human and Other blocks) while producing lower scores for machine-generated stories (Machine and N blocks). While looking into more details, we can see that they cannot successfully distinguish the story quality, e.g., SCARY_{200}(low) and SCARY_{200}(high) receive identical scores. These observations strongly indicate that UNION and MANPLTS work well on evaluating coherence but deviate from human preference when evaluating human-written stories.

Our method, on the other hand, is capable of following human preference (Human and Machine block) (also see SCARY_{200}(low) and SCARY_{200}(high) as an example). The model trained with highly-voted stories can generate better stories than that trained with lowly-voted stories, and P&W strategy performs even better as proved in many previous works (Fan et al., 2019; Tan et al., 2021). From the results, our model produces higher scores for LED (high) compared with LED (low) and even higher scores for LED P&W (high), which indicates that our model still follows the human preference on machine-generated stories. As serious coherence problems do not commonly occur in our training data, our method show failure in recognizing manually created incoherent stories (N block). However, our model (Ours (N)) works after we incorporate these stories into our training data, leading to a future direction that unifies the coherence-based and preference-aware metrics. Surprisingly, our model gives relatively low scores when adopting stories from other domains (Other block). We think this is because the writing style changes the criterion of human preference, which misleads our model to predict a not reasonable score, thus leading us to a big challenge in generalizing preference-aware story evaluation.

		Coherence		Preference	Hybrid
	Dataset	UNION	MANPLTS	Ours	Ours (N)
Human	WP_{200}(low)	0.771	0.878	0.347	0.655
	WP_{200}(high)	0.837	0.948	0.692	0.884
	SCARY_{200}(low)	0.833	0.825	0.355	0.625
	SCARY_{200}(high)	0.895	0.850	0.743	0.883
Machine	LED (low)	0.687	0.091	0.297	0.290
	LED P&W (low)	0.775	0.300	0.535	0.305
	LED (high)	0.588	0.001	0.409	0.290
	LED P&W (high)	0.760	0.393	0.573	0.308
N	Negative(UNION)	0.360	0.003	0.244	0.019
	Negative(MANPLTS)	0.414	0.228	0.319	0.027
Other	fairy tale (short)	0.917	0.500	0.233	0.482
	childbook (long)	0.886	0.915	0.318	0.476
	bookcorpus (long)	0.965	0.949	0.285	0.416

Table 7: Our model and existing works on various domains of stories. We report the averaged preference score on stories from four different domains.

7.3 More Analysis

Due to the page limit, we put more analysis in the ablation studies. In Appendix Sec. D, we witness high correlation scores between preference score and each aspect rating, indicating the effectiveness of all pre-defined aspects in the evaluation. We also analyze the confidence and rating scores of the horror aspect with the preference score on scary stories in Appendix Sec. E. The result follows the human intuition that evaluation on scary stories shows a tendency to rely on the horror aspect.

8 Conclusion

In this paper, we investigate a novel task of preference-aware story evaluation, StoryER, which produce a score with explanation through various aspects and comments, bringing gains on both machine-generated and human-written stories evaluation. To support the task, we present a new dataset consisting of paired ranked stories and more explicit annotation (i.e., rating and reasons) for pre-defined aspects. Our comprehensive ablation studies and intensive analysis show the effectiveness of using aspect rating and reasoning on preference score prediction. With the development of story generation, we believe that preference-aware story evaluation will be the mainstream research when machine-generated stories do not suffer from serious coherence problems. Further studies on our dataset can also be conducted to reveal the point that influence the readers to upvote the stories.

9 Limitations

Our work (currently) has the following limitations:
(1) As indicated in Section 7.2, our proposed metrics are negatively affected by the significant domain shift, since we only take stories from one platform to train our model. Idealistically, a more general model can be trained with all types of stories, but it needs massive annotations on human preference (i.e., upvote counts).
(2) Since the upvote counts in the original dataset will be influenced by the prompt's topic, typically, fantastic stories get more upvotes than others. Our model is only trained by story pairs within the same topic, thus if a user inputs two unrelated stories, our system will provide unpredictable results. Therefore, we propose using pairwise evaluation with the same given prompt to avoid comparing stories with diverse topics.
(3) In this work, we propose to implicitly joint training to increase the performance of each task without explicitly addressing the connection of three subtasks. Although we have aspect rating and comment generation, preference score is still the most effective approach to assess the quality of the story. How to use these comments and aspect ratings is a challenge that will be addressed in the future work.

10 Ethics and Broader Impacts

We hereby acknowledge that all of the co-authors of this work are aware of the provided *ACM Code of Ethics* and honor the code of conduct. This work is mainly about propose a novel method in automatic story evaluation. The followings give the aspects of both our ethical considerations and our potential impacts to the community.

Dataset. We collect the human annotation of the aspect rating and comments via Amazon Mechanical Turk (MTurk) and ensure that all the personal information of the workers involved (e.g., usernames, emails, urls, demographic information, etc.) is discarded in our dataset. All the stories in our dataset are collected from a public dataset, namely WritingPrompt. Although we aim at providing a dataset that agreed upon from various people, there might still be unintended biases within the judgements, we make efforts on reducing these biases by collecting diverse comments and replacing the annotators who tends to be racist.

The detailed annotation process (pay per amount of work, guidelines) is included in the appendix and our public website; We primarily consider English speaking regions for our annotations as the task requires certain level of English proficiency.

Techniques. We benchmark the story evaluation task with conventional metrics and our proposed metric. As the story evaluation are of our main focus, we do not anticipate production of harmful outputs on our proposed task.

11 Acknowledgments

This paper is supported by Institute of AI and Beyond of the University of Tokyo, JSPS KAKENHI Grant Numbers JP19H04166, JP22H00540, 22K17947 and also based on results obtained from a project JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO). For experiments, computational resources of AI Bridging Cloud Infrastructure (ABCI) provided by National Institute of Advanced Industrial Science and Technology (AIST) was used.

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Samuel Brody and Noemie Elhadad. 2010. An unsupervised aspect-sentiment model for online reviews. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 804–812, Los Angeles, California. Association for Computational Linguistics.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A Smith. 2021. All that's' human'is not gold: Evaluating human evaluation of generated text. *arXiv preprint arXiv:2107.00061*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Birk Diedenhofen and Jochen Musch. 2015. cocor: A comprehensive solution for the statistical comparison of correlations. *PloS one*, 10(4):e0121945.

Sergey Edunov, Myle Ott, Marc'Aurelio Ranzato, and Michael Auli. 2020. On the evaluation of machine translation systems trained with back-translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2836–2846.

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)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2019. Strategies for structuring story generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2650–2660.

Sarik Ghazarian, Zixi Liu, Akash S M, Ralph Weischedel, Aram Galstyan, and Nanyun Peng. 2021. Plot-guided adversarial example construction for evaluating open-domain story generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4334–4344, Online. Association for Computational Linguistics.

Sarik Ghazarian, Johnny Wei, Aram Galstyan, and Nanyun Peng. 2019. Better automatic evaluation of open-domain dialogue systems with contextualized embeddings. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 82–89.

Jian Guan and Minlie Huang. 2020. Union: An unreferenced metric for evaluating open-ended story generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9157–9166.

Jian Guan, Zhexin Zhang, Zhuoer Feng, Zitao Liu, Wenbiao Ding, Xiaoxi Mao, Changjie Fan, and Minlie Huang. 2021. Openmeva: A benchmark for evaluating open-ended story generation metrics. *arXiv preprint arXiv:2105.08920*.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2015. The goldilocks principle: Reading children's books with explicit memory representations. *arXiv preprint arXiv:1511.02301*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic Turing test: Learning to evaluate dialogue responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1116–1126, Vancouver, Canada. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Maurice S Schaeffer and Eugene E Levitt. 1956. Concerning kendall's tau, a nonparametric correlation coefficient. *Psychological Bulletin*, 53(4):338.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. Bleu is not suitable for the evaluation of text simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 738–744.

Bowen Tan, Zichao Yang, Maruan Al-Shedivat, Eric Xing, and Zhiting Hu. 2021. Progressive generation of long text with pretrained language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4313–4324.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

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.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7378–7385.

Jerrold H Zar. 1972. Significance testing of the spearman rank correlation coefficient. *Journal of the American Statistical Association*, 67(339):578–580.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578.

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 the IEEE international conference on computer vision*, pages 19–27.

A Website Demo

We display the collected data, AMT template and models on our website[9]. The users can input their own stories or randomly select one story. The server then runs our model and output a preference score, and comments for each aspect. Figure 4 shows an example.

B Code

We also put our source codes into the supplementary materials. Due to the upload size limitation. We truncate our 100k Story Ranking data into a size of 1000, as well as the 46k Aspect Rating and Reasoning data. Please kindly follow the README to run the experiment. Our human annotation results can be also found under the folder “data”. Additionally, we put some examples for machine-generated stories introduced in our paper.

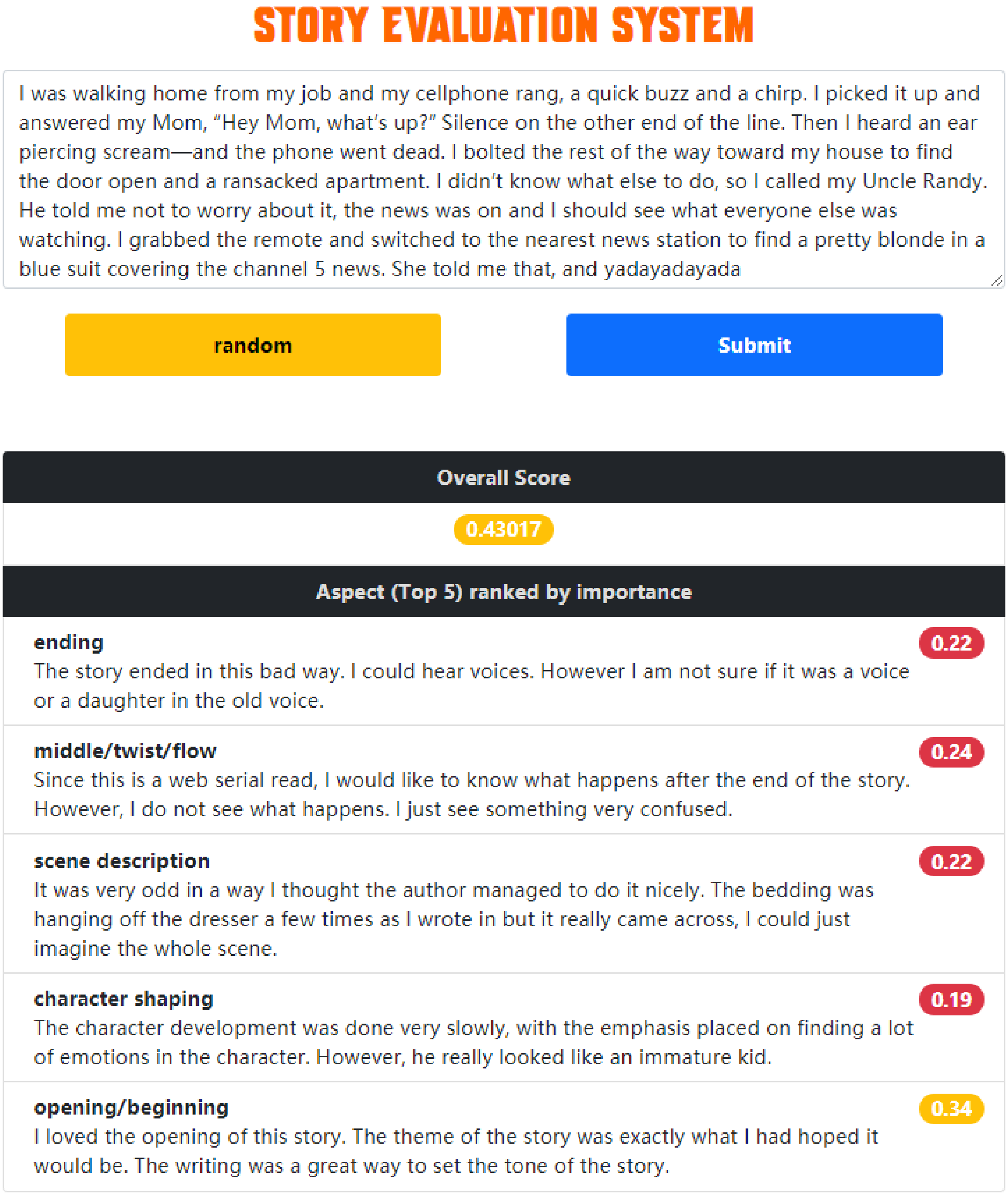
C Correlation Between Story Quality and Aspect Rating

We calculate the correlation between human ratings on each aspect with the upvote number, and the predicted aspect rating with the predicted preference score, to figure out the correlation between the aspect rating and the preference score. The results are listed in Figure 5. We can see the results from our model greatly match the distribution of the correlation between human aspect rating and human upvote number. None of these shows domination, which proves that all pre-defined aspects affect the final preference score prediction.

D Horror/Scary Aspect with SCARY$_{200}$

To show how aspect ratings and confidence are related to the story, we further analyze their performance on WP$_{200}$ and SCARY$_{200}$. We calculate the recall performance and rating correlation on “horror/scary” aspect only to detect how this aspect works in both data. Table 8 depicts that horror aspect can achieve 36% probability to be the top confident aspect in SCARY$_{200}$, while the number is only 0.5% in the original WP$_{200}$. On the other hand, the preference score also has a higher correlation with the rating from “horror/scary” aspect. These results prove that the predicted aspects show high connection to the preference score prediction.

[9] For the review process, dataset and pre-trained model demo are available at anonymous website http://storytelling-lab.com/eval



STORY EVALUATION SYSTEM

I was walking home from my job and my cellphone rang, a quick buzz and a chirp. I picked it up and answered my Mom, "Hey Mom, what's up?" Silence on the other end of the line. Then I heard an ear piercing scream—and the phone went dead. I bolted the rest of the way toward my house to find the door open and a ransacked apartment. I didn't know what else to do, so I called my Uncle Randy. He told me not to worry about it, the news was on and I should see what everyone else was watching. I grabbed the remote and switched to the nearest news station to find a pretty blonde in a blue suit covering the channel 5 news. She told me that, and yadayadayada

Overall Score

0.43017

Aspect (Top 5) ranked by importance

ending 0.22
The story ended in this bad way. I could hear voices. However I am not sure if it was a voice or a daughter in the old voice.

middle/twist/flow 0.24
Since this is a web serial read, I would like to know what happens after the end of the story. However, I do not see what happens. I just see something very confused.

scene description 0.22
It was very odd in a way I thought the author managed to do it nicely. The bedding was hanging off the dresser a few times as I wrote in but it really came across, I could just imagine the whole scene.

character shaping 0.19
The character development was done very slowly, with the emphasis placed on finding a lot of emotions in the character. However, he really looked like an immature kid.

opening/beginning 0.34
I loved the opening of this story. The theme of the story was exactly what I had hoped it would be. The writing was a great way to set the tone of the story.

Figure 4: An example on the website.

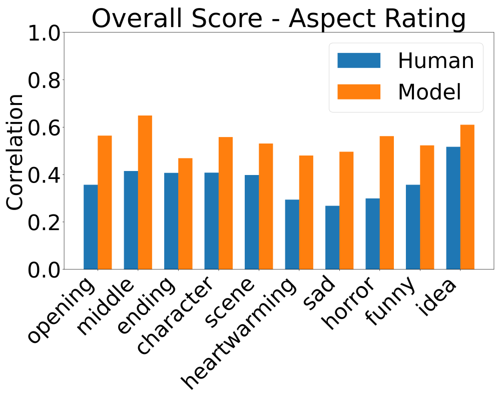


Figure 5: Correlation of upvote number - aspect rating (Human) and the correlation of predicted preference scores and predicted aspect rating (Model (Ours)). The correlation values are all statistical significant. (i.e. $p \leq 0.01$)

	Confidence(Horror)			Correlation(Horror)	
	R@1	R@3	R@5	ρ	τ
WP_{200}	0.50	11.00	13.50	0.233	0.163
SCARY_{200}	36.50	50.50	57.50	0.302	0.222

Table 8: Confidence for aspect horror/scary for WP_{200} and SCARY_{200} dataset and the correlation between the preference score and horror/scary aspect ratings in two dataset.

E Implementation Details

E.1 Model for Preference-Aware Story Evaluation

Our model for story evaluation used pre-trained LED (Beltagy et al., 2020) from Huggingface[10]. We finetune the model with 100k Story Ranking data and 46k Aspect Rating and Reasoning data on a machine with 8 NVIDIA A100 GPUs. During the training, we set the batch size as 16, the margin as 0.3 and run 20k iterations (5 epoch on 100k Story Ranking data) on training (10 hours). In each iteration, we adopt two pairs: one from 100k Story Ranking data and the other from 46k Aspect Rating and Reasoning data, to our model. We take AdamW optimizer (Loshchilov and Hutter, 2018) with an initial learning rate of 4e-6, warming up in the first epoch and decreasing by a linear schedule. The reported results are averaged by the best results from three models with the same structure but initialized with three different seeds. For hyper-parameter search, we search margin m from 0.2 to 1.0 with the step of 0.1, learning rate from 4e-4, 4e-5, 4e-6 and 4e-7, and record the best hyper-parameters.

	Ranking	WP_{200}		SCARY_{200}	
	Acc	ρ	τ	ρ	τ
CE	72.82	0.539	0.390	0.538	0.412
CE(smooth)	71.38	0.550	0.394	0.561	0.414
Ranking	**73.93**	**0.583**	**0.422**	**0.578**	**0.420**

Table 9: Comparison of discrimination and ranking.

E.2 Model for Aspect Category Classification

Our model for aspect category classification is based on RoBERTa large (Liu et al., 2019). Same as the model we used in preference score prediction, we apply a linear projection on the feature of [CLS], the first token of the input comments. We then train the model with cross-entropy loss for 20 epochs with the learning rate 4e-5.

E.3 Model for Comment Sentiment Analysis

Our model for comment sentiment analysis uses the same model structure for aspect category classification. We also use the same epochs number and learning rate during the training. The only difference is that the targets in training are the sentiment rate with a scale of 1-5 (from definitely negative to definitely positive)

F More Results

F.1 Ranking vs Discrimination

Given two types of stories, highly and lowly upvoted, a straightforward method to build the model is through discrimination; use 0 and 1 as target with cross-entropy loss. We compare the results by using ranking and discrimination. The result is shown Table 9. From the result, we see that ranking strategy achieves better scores than discrimination and that with label smoothing. We believe it is because when we conduct ranking, we only enlarge the preference scores between stories written from the same prompt. The encoder can learn better how human preference works by comparing stories with the same topic. On the other hand, the ranking loss is more flexible compared with binary classification, which can be easily extended to rank more than two types of stories as shown in Equation 2.

F.2 PPL in automatic story evaluation

For an interesting finding, Perplexity (PPL) shows positively correlated to the score of WP and more highly correlated to the score of SCARY_{200}, while showing substantially negatively correlated to the score of coherence on machine-generated stories, which reveals a potential for story evaluation using pre-trained language models.

[10] https://huggingface.co/transformers/model_doc/led.html

F.3 Results of Comment Evaluation

Due to the page limitation, we put the results of comment evaluation with more metrics in Table 10. We see that our model achieves higher performance on most of the metrics.

F.4 Results of Aspect Category Classification

We use aspect category classification model, introduced in Sec. E.2, for filtering out noisy comments. Figure 6 shows the classification results. Except for "ending" and "heartwarming", all aspect classes can achieve an average of around 80% accuracy, showing high performance on classification. We filter out the comments, with no aspect category score exceeding 0.9 after softmax function.

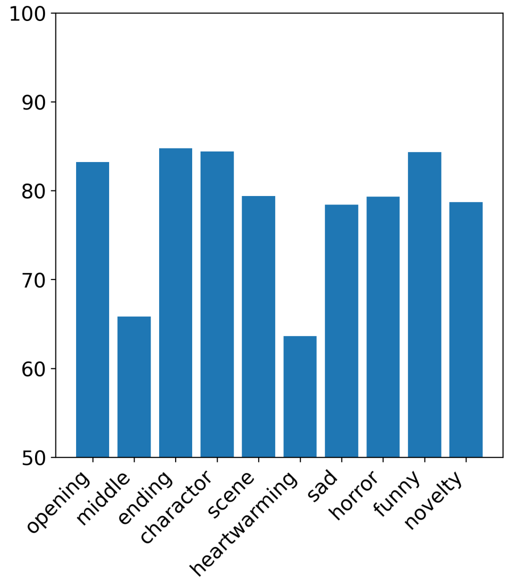


Figure 6: Comment classification results.

F.5 Results of Comment Sentiment Analysis

Comment Sentiment Analysis model, introduced in Sec. E.3, is used to rate comments by their sentiments. Table 11 shows the results. Our output is the rates from 1 to 5. In the evaluation, we simply group 1 and 2 as the negative, 3 as the neutral, 4 and 5 as the positive. The results show that our sentiment analysis model can correctly predict the sentiment, especially on positive and negative.

F.6 Comment Data Augmentation

We collect over 150k uncategorized comments from metadata in WP. We use the aspect category classification model and filter out the irrelevent comments. However, we found bias inside the comments. For example, we get almost 9000 comments about "ending", while only 1200 for "sad". To mitigate the bias that would be inducted into our story evaluation model, we sample about 2000 comments for each aspect, and use all comments for the aspect which contains less than 2000 comments. The final data statistics of comments can be referred to our website.

G Human Annotation

G.1 Human Annotation on Test Data

For evaluation, we collect human judgments through AMT for 200 highly-upvoted stories and 200 lowly-upvoted stories from WP (sampled from test data in 100k Story Ranking data), where each story is assigned to 8 annotators. Annotators are asked to rate each story on a scale of 1 to 5 (from poor to good). Following Clark et al. (2021), we asked the annotators to compare the stories before rating and write down a very brief reason for clarification. To further ensure the correctness of the annotation, we calculate the statistics of the annotator behavior (i.e., working time per hit) and set traps in the batch (i.e., insert extremely poor story, duplicate stories for one annotator to test their consistency). The submissions from annotators with poor quality are all rejected and then recollected from new annotators. Finally, we exclusively keep the 100 highly-upvoted and 100 lowly-upvoted stories with the lowest variance from 8 annotators and average the human rates as the target scores in this test data, namely, WP_{200} in the following experiments. Annotators get $0.2 as the reward for each submission. Besides, we crawled scary stories from Reddit (r/shortscarystories [11]), which have a similar writing style to the stories in WP but in a constrained story type. We repeat the procedure for WP_{200} and create another human-annotated test data, namely SCARY_{200}. The same procedure is also used for collecting human annotation on machine-generated stories. We generate 200 stories using LED trained with highly-voted stories and another 200 stories using LED trained with lowly-voted stories for annotation. We ask the annotators to rate the stories based on human preference and also ask them to distinguish whether the given stories are human-written or machine-generated. We exclusively keep the stories that

[11] https://www.reddit.com/r/shortscarystories/

				Automatic					Human			
p_s	a	c	N	PPL	BLEU	ROUGE	METEOR	CIDER	Overall	Rel(story)	Rel(aspect)	Rel(rating)
		✓		7.31	8.45	16.63	18.81	7.39	47.61	**73.70**	79.20	-
✓	✓	✓		**7.06**	**8.60**	**16.76**	**18.88**	**7.99**	**49.40**	72.93	**82.83**	**58.33**
✓	✓	✓	✓	7.95	8.36	16.69	18.44	7.07	43.45	68.64	81.84	50.49

Table 10: Comment generation evaluation on automatic scores and human evaluation. In human evaluation, the kappa coefficient κ for each score are located in 0.4-0.6, indicating a moderate agreement between annotators.

	positive	neutral	negative	average
Acc	89.70%	50.93%	85.20%	83.03%

Table 11: Comment sentiment analysis results

deceive the annotators, as these stories do not contain serious coherence problems.

G.2 Data Collection

In this paper, we mainly collect data for two different uses. Annotators get $1 as the reward for each submission. The total data collection takes 2 months. To assess the quality of each annotator, we randomly sample the submissions from each annotator every two days, bonus the one with good quality and warn the annotators who give nonsense comments.

G.3 Human Annotation Inner-Agreement

As we assign one story for more than one annotator, we calculate the inner-agreement from different annotators on aspect selection. As a result, 65.80% aspects are selected by more than one annotator, and the correlation coefficient of multi-annotation on aspect ratings are 0.913 and 0.811, corresponding to the Spearman (Zar, 1972) and Kendall (Schaeffer and Levitt, 1956) respectively.

H Aspect Category Name Definition

As no standard criterion exists for story evaluation, we collect some well-used aspects that used in the Internet. We mainly refer to the websites [12] [13] [14].

[12] https://en.wikipedia.org/wiki/List_of_writing_genres

[13] https://www.writerswrite.co.za/the-complete-guide-to-evaluating-your-short-story/

[14] https://www.oprahdaily.com/entertainment/books/a29576863/types-of-book-genres/