Supplementary file: “Stylized Story Generation with Style-Guided Planning”

Xiangzhe Kong[*], Jialiang Huang[*], Ziquan Tung, Jian Guan and Minlie Huang[†]
The CoAI group, DCST, Institute for Artificial Intelligence,
State Key Lab of Intelligent Technology and Systems,
Beijing National Research Center for Information Science and Technology,
Tsinghua University, Beijing 100084, China
{kxz18,huang-jl17,tongzq18,j-guan19}@mails.tsinghua.edu.cn,
aihuang@tsinghua.edu.cn

A Implementation Details

A.1 Vocabulary

The initial vocabulary of GPT-2/BART contains 50,258/50,265 tokens, respectively. We add three style tokens (⟨emo⟩, ⟨eve⟩, ⟨other⟩) and three name tokens (⟨MALE⟩, ⟨FEMALE⟩, ⟨NEUTRAL⟩) to the vocabulary. Therefore, the final vocabulary for GPT-2/BART/our model contains 50,264/50,271/50,271 tokens, respectively.

A.2 Hyper-parameters

We follow BART$_{\text{BASE}}$'s hyper-parameters and initialize our model with the public checkpoint of BART$_{\text{BASE}}$[1]. Both the encoder and decoder contain 6 hidden layers with 768-dimensional hidden states. GPT-2$_{\text{BASE}}$[2] uses a 12-layer decoder with 768-dimensional hidden states. The batch size is 32 for all the models when training. We uses the AdamW optimization (Loshchilov and Hutter, 2019) and the initial learning rate is 5×10^{-5}. At inference time, we set the maximum sequence length to 120 tokens.

Models	**GPT-2**	**BART**	**Ours**
Training Time	242min	128min	336min

Table 1: Training time for models in the experiments

A.3 Runtime

The runtime of fine-tuning of each model is reported in Table 1. We do the experiments on one GeForce GTX TITAN X GPU.

B Manual Evaluation

As described in the main paper, we conduct manual evaluation on AMT. Figure 1 shows a screenshot of an annotation example on AMT.

References

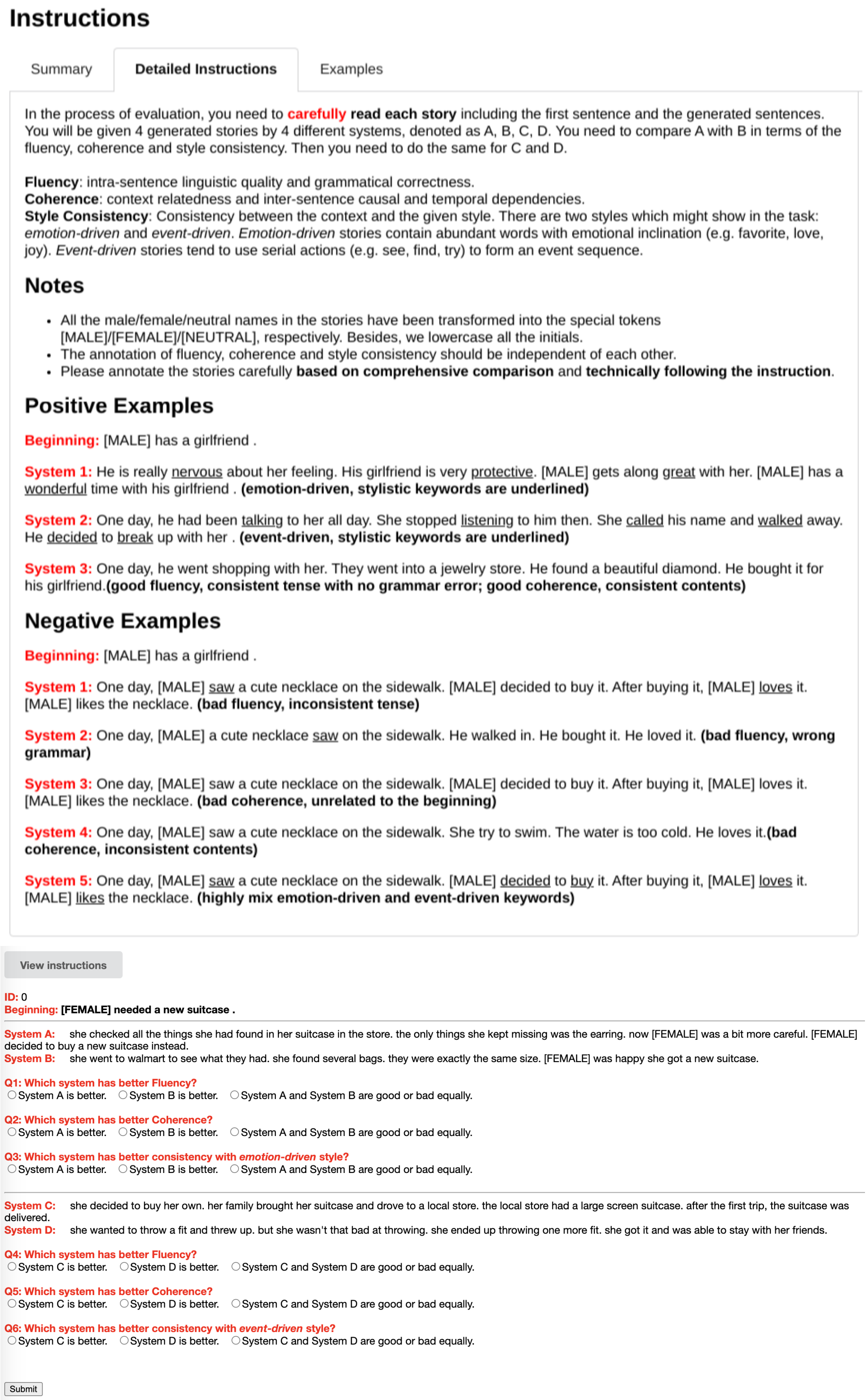
Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

[*] Equal contribution

[†] Corresponding author

[1] https://huggingface.co/facebook/bart-base/tree/main

[2] https://huggingface.co/gpt2/tree/main



Instructions

In the process of evaluation, you need to **carefully read each story** including the first sentence and the generated sentences. You will be given 4 generated stories by 4 different systems, denoted as A, B, C, D. You need to compare A with B in terms of the fluency, coherence and style consistency. Then you need to do the same for C and D.

Fluency: intra-sentence linguistic quality and grammatical correctness.
Coherence: context relatedness and inter-sentence causal and temporal dependencies.
Style Consistency: Consistency between the context and the given style. There are two styles which might show in the task: *emotion-driven* and *event-driven*. *Emotion-driven* stories contain abundant words with emotional inclination (e.g. favorite, love, joy). *Event-driven* stories tend to use serial actions (e.g. see, find, try) to form an event sequence.

Notes

- All the male/female/neutral names in the stories have been transformed into the special tokens [MALE]/[FEMALE]/[NEUTRAL], respectively. Besides, we lowercase all the initials.
- The annotation of fluency, coherence and style consistency should be independent of each other.
- Please annotate the stories carefully **based on comprehensive comparison** and **technically following the instruction**.

Positive Examples

Beginning: [MALE] has a girlfriend .

System 1: He is really nervous about her feeling. His girlfriend is very protective. [MALE] gets along great with her. [MALE] has a wonderful time with his girlfriend . **(emotion-driven, stylistic keywords are underlined)**

System 2: One day, he had been talking to her all day. She stopped listening to him then. She called his name and walked away. He decided to break up with her . **(event-driven, stylistic keywords are underlined)**

System 3: One day, he went shopping with her. They went into a jewelry store. He found a beautiful diamond. He bought it for his girlfriend.**(good fluency, consistent tense with no grammar error; good coherence, consistent contents)**

Negative Examples

Beginning: [MALE] has a girlfriend .

System 1: One day, [MALE] saw a cute necklace on the sidewalk. [MALE] decided to buy it. After buying it, [MALE] loves it. [MALE] likes the necklace. **(bad fluency, inconsistent tense)**

System 2: One day, [MALE] a cute necklace saw on the sidewalk. He walked in. He bought it. He loved it. **(bad fluency, wrong grammar)**

System 3: One day, [MALE] saw a cute necklace on the sidewalk. [MALE] decided to buy it. After buying it, [MALE] loves it. [MALE] likes the necklace. **(bad coherence, unrelated to the beginning)**

System 4: One day, [MALE] saw a cute necklace on the sidewalk. She try to swim. The water is too cold. He loves it.**(bad coherence, inconsistent contents)**

System 5: One day, [MALE] saw a cute necklace on the sidewalk. [MALE] decided to buy it. After buying it, [MALE] loves it. [MALE] likes the necklace. **(highly mix emotion-driven and event-driven keywords)**

ID: 0
Beginning: **[FEMALE] needed a new suitcase .**

System A: she checked all the things she had found in her suitcase in the store. the only things she kept missing was the earring. now [FEMALE] was a bit more careful. [FEMALE] decided to buy a new suitcase instead.
System B: she went to walmart to see what they had. she found several bags. they were exactly the same size. [FEMALE] was happy she got a new suitcase.

Q1: Which system has better Fluency?
○ System A is better. ○ System B is better. ○ System A and System B are good or bad equally.

Q2: Which system has better Coherence?
○ System A is better. ○ System B is better. ○ System A and System B are good or bad equally.

Q3: Which system has better consistency with *emotion-driven* style?
○ System A is better. ○ System B is better. ○ System A and System B are good or bad equally.

System C: she decided to buy her own. her family brought her suitcase and drove to a local store. the local store had a large screen suitcase. after the first trip, the suitcase was delivered.
System D: she wanted to throw a fit and threw up. but she wasn't that bad at throwing. she ended up throwing one more fit. she got it and was able to stay with her friends.

Q4: Which system has better Fluency?
○ System C is better. ○ System D is better. ○ System C and System D are good or bad equally.

Q5: Which system has better Coherence?
○ System C is better. ○ System D is better. ○ System C and System D are good or bad equally.

Q6: Which system has better consistency with *event-driven* style?
○ System C is better. ○ System D is better. ○ System C and System D are good or bad equally.

Figure 1: A screenshot of manual evaluation on AMT