Psychology-guided Controllable Story Generation

Yuqiang Xie Yue Hu[†] Yunpeng Li Guanqun Bi Luxi Xing Wei Peng
Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
{xieyuqiang,huyue,liyunpeng,biguanqun,xingluxi,pengwei}@iie.ac.cn

Abstract

Controllable story generation is a challenging task in the field of NLP, which has attracted increasing research interest in recent years. However, most existing works generate a whole story conditioned on the appointed keywords or emotions, ignoring the psychological changes of the protagonist[‡]. Inspired by psychology theories, we introduce global psychological state chains, which include the needs and emotions of the protagonists, to help a story generation system create more controllable and well-planned stories. In this paper, we propose a **P**sychology-gu**I**ded **C**ontrollable **S**tory Generation System (**PICS**) to generate stories that adhere to the given leading context and desired psychological state chains for the protagonist. Specifically, psychological state trackers are employed to memorize the protagonist's local psychological states to capture their inner temporal relationships. In addition, psychological state planners are adopted to gain the protagonist's global psychological states for story planning. Eventually, a psychology controller is designed to integrate the local and global psychological states into the story context representation for composing psychology-guided stories. Automatic and manual evaluations demonstrate that **PICS** outperforms baselines, and each part of **PICS** shows effectiveness for writing stories with more consistent psychological changes.

1 Introduction

Controllable Story Generation (CSG) is an important task in natural language processing (NLP) (Porteous and Cavazza, 2009; Peng et al., 2018; Alabdulkarim et al., 2021). It has also become one of the test methods for progress in artificial intelligence (AI). Most existing state-of-the-art works (Kong et al., 2021; Rashkin et al., 2020; Paul and Frank, 2021; Xu et al., 2020b) generate a story conditioned by the appointed keywords or emotions, with the help of remarkable pre-trained language models (PLM), like GPT-2 (Radford et al., 2019) and BART (Lewis et al., 2020). While most of these systems have been able to generate fluent stories, CSG still has many issues to be explored.

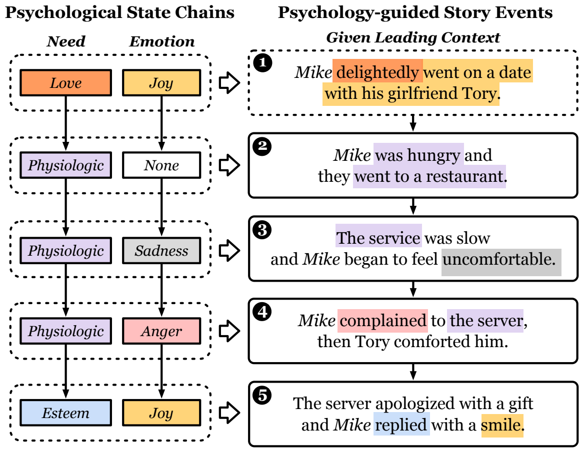

Figure 1: Example of psychology-guided controllable story generation conditioned on dotted frames (global psychological state chains, i.e., need/emotion, as well as leading context). Each psychological state and its corresponding tokens are highlighted in the same color.

In daily life, humans tend to create events driven by their needs (cause) and receive emotions (effect) after the events. Similarly, needs (Ricoeur, 1984) and emotions (Vonnegut, 1981) play the central roles in creating reasonable stories in storytelling. Recently, several works have begun developing CSG systems based on people's expected emotional keywords or scores, such as (Brahman and Chaturvedi, 2020) and (Xu et al., 2020a). Although these approaches can generate stories appointing the desired emotional signals, they are unable to control the storytelling as the protagonist's psychological state changes. Another problem is that these methods only consider the current/previous emotions without global planning, which plays an

[†]Corresponding author.
[‡]In this work, we define the protagonist as the most frequently occurring character in a story (Morrow, 1985).

Proceedings of the 29th International Conference on Computational Linguistics, pages 6480–6492
October 12–17, 2022.

important role in composing a story.

To address the aforementioned problems, we focus on taking the protagonist's global psychological state chains into account in controllable story generation. Researches in cognitive psychology have shown that readers closely monitor the protagonist's needs (Ricoeur, 1984) and emotions (Vonnegut, 1981) while reading narratives. At any point in a story, we represent the protagonist's psychological state using multiple needs and emotions common in psychological theories. *Hierarchy of needs* of Maslow has five categories (i.e., *physiological need, stability, love and belonging, esteem* and *self-actualization*) for describing human needs of a person. *Wheel of emotions* of Plutchik proposes eight basic emotions (includes *joy, trust, anger, surprise, sadness, disgust, fear* and *anticipation*) to adequately portray a person. Motivated by this, we define the psychological state chains as a sequence of five human needs and eight basic emotions that describe psychological states of a protagonist.

Given the protagonist's name and psychological state chains as well as the leading context, our goal is to generate a story about the leading context that adheres to the protagonist's psychological state chains. As exemplified in Figure 1, the protagonist (*Mike*) takes part in each story event controlled by given psychological state chains. Note that, *none* represents that *Mike* has no need or emotion. Each psychological state and its corresponding tokens are in the same color. For example, in the second story event, *Mike was hungry* and *went to a restaurant* obviously embody the **physiologic need** of *Mike*. Another example, as shown in the fourth story event, *complain* action reflects *Mike*'s **anger emotion**. From a global perspective, intuitively, the anterior and hereafter psychological states separately provide the background and guidance for composing stories. As illustrated in the third story event, anterior **physiologic need** leads to *Mike's feeling uncomfortable* due to hungry, and hereafter **anger emotion** guides the setting of *slow service* plot suspense.

To generate stories that adhere to the given leading context and the desired protagonist's global psychological state chains, we propose **PICS** (**P**sychology-gu**I**ded **C**ontrollable **S**tory Generation System), a Transformer-based (Vaswani et al., 2017) architecture. Specifically, psychological state trackers are employed to memorize the local psychological states for capturing temporal relationships among psychological states. And, psychological state planners are adopted to gain the protagonist's global psychological states for planning the storytelling. In the end, a psychology controller is designed to integrate the local and global psychological states into the story context representation for composing psychology-guided stories. Based on the extracted data from publicly available *Story Commonsense* (Rashkin et al., 2018) dataset, experimental results demonstrate that **PICS** outperforms baselines, and the psychological state trackers, planners as well as the psychology controller are important for generating stories with more consistent psychological changes.

2 Related Work

Early story generation systems relied on symbolic planning (Pérez and Sharples, 2001; Porteous and Cavazza, 2009; Riedl and Young, 2010), which had domain restriction and massive cost of feature engineering. Recent seq2seq storytelling models (Roemmele, 2016; Jain et al., 2017) had partially alleviated these problems, most of which focused on learning better representation for a story (Martin et al., 2018; Xu et al., 2018; Fan et al., 2018b, 2019; Yao et al., 2019).

To introduce semantic knowledge into story generation, many methods also employed large-scale pre-trained language models (LM) based on Transformer (Vaswani et al., 2017), like GPT-2 (Radford et al., 2019) and BART (Lewis et al., 2020). After in-domain training, these models can generate fluent and coherent text, which can be used in story generation (Qin et al., 2019; Guan et al., 2020; Xie et al., 2022b) and dialogue systems (Budzianowski and Vulić, 2019; Wolf et al., 2019). However, they lacked the ability of controllable generation, such as expressing specific goals.

Further, aiming at controllable story generation, works had been introduced to control different attributes of the generated text, such as keyword (Fan et al., 2018a), style (Wang et al., 2017) and length (Kikuchi et al., 2016). For example, Tambwekar et al. 2019 introduced reinforcement learning to generate a goal-driven storyline, which is a sequence of event tuples. PPLM (Dathathri et al., 2020) used attribute classifiers to guide text generation without further training of LM. PLOTMACHINES (Rashkin et al., 2020) transformed an outline into a coherent story by tracking the dynamic plot states. Kong et al. 2021 first planned the stylized keywords and then generated the whole story with the guidance of the keywords. And many works considered commonsense knowledge as an attribute for CSG. Ammanabrolu et al. 2021 performed story generation using soft causal relations, which automatically extracted from existing natural language plot summaries. Paul and Frank 2021 used the contextualized commonsense inference rules generated by COMET (Bosselut et al., 2019) based model to produce a coherent story ending.

Most related to this work, many methods generated text with a specific sentiment or emotion (Zhou et al., 2018; Huang et al., 2018; Zhou and Wang, 2018; Song et al., 2019). Rashkin et al. 2018 present an annotation framework specifically designed to examine the mental states of characters in commonsense based stories. There are some limitations to incorporating sentiment, emotion or psychological state for story generation. Previous work modeled characters but not sentiment (Clark et al., 2018; Liu et al., 2020). Peng et al. 2018 and Luo et al. 2019 controlled the overall sentiment for story ending generation. Weber et al. 2020 incorporated sentiment trajectory by a new task that "filling in" a story. Brahman and Chaturvedi 2020 modeled the emotional trajectory of the protagonist for story generation. Xu et al. 2020a generated a story with multiple emotional changes of protagonists based on the given characters and the corresponding psychological state lines. These works are limited to the guiding of emotion scores or tokens or/and target the ending sentence. Lately, Xie et al. 2022a modeling the relationship among motivations, actions and emotions based on human activities (i.e. story events), which lacks consideration of the global changes in a story. Different from the above methods, we respectively model the local and global psychological state changes of the protagonist as the story progresses, which is more central to storytelling than the emotion trajectory.

3 Task Definition

We formulate our psychology-oriented controllable story generation task in the following. Note that, the length of the whole story is 5 in this paper, and the output story event is in the m-th time point. Table 1 shows an example of our task.

Protagonist	Donald, He, Donald, He, He
Need Chain	esteem, esteem, esteem, esteem, esteem
Emotion Chain	joy, sadness, sadness, sadness, joy
Leading Context	Donald was a senator.
Event $\mathcal{Y}_2$	He ran as an indie candidate.
Event $\mathcal{Y}_3$	Donald wanted better implementation of his policies.
Event $\mathcal{Y}_4$	He decided to run in the next term as a Republican.
Event $\mathcal{Y}_5$	He won again, and is now in a better position.

Table 1: An example of our task.

Input The context $\mathbb{X} = (\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_{m-1})$ to the current story event with $m-1$ events, where the i-th event $\mathcal{X}_i = (x_i^1, x_i^2, \ldots, x_i^k)$ consists of k words. The name of protagonist $\mathbb{P} = (p_1, p_2, \ldots, p_m)$ to indicate the expected participant of the generated story event, the elements are the same in $\mathbb{P}$ in our setting. The protagonist's global psychological state chains, including the need chain $\mathbb{A}_{\mathbb{N}} = (n_1, n_2, \ldots, n_5)$ and the emotion chain $\mathbb{A}_{\mathbb{E}} = (e_1, e_2, \ldots, e_5)$. The protagonist's local psychological states from $\mathbb{A}_{\mathbb{N}}$ and $\mathbb{A}_{\mathbb{E}}$, including the need history $\mathbb{N} = (n_1, n_2, \ldots, n_m)$ and the emotion history $\mathbb{E} = (e_1, e_2, \ldots, e_m)$. n_i and e_i represent the protagonist's need and emotion for the i-th story event, where $i \in [1, m]$.

Output $\mathcal{Y}_m = (y_1, y_2, \ldots, y_r)$ (also X_m in the next time step) stands for the current story event that consists of r words, based on the protagonist's name $\mathbb{P}$, the need history $\mathbb{N}$, the emotion history $\mathbb{E}$, the need chain $\mathbb{A}_{\mathbb{N}}$ and the emotion chain $\mathbb{A}_{\mathbb{E}}$, where y_i is the i-th word.

4 Methodology

The overall architecture of our proposed **PICS** system is illustrated in Figure 3. In the following, we will describe each component in more detail.

4.1 Contextual Encoder (Step 1)

In order to capture the contextual semantic information for the story context and the protagonist's historical psychological states, we reconstruct the input of the embedding layer in the backbone BART (Lewis et al., 2020) model.

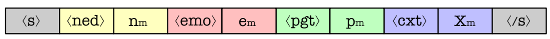

Figure 2: Reconstructed input of the word embedding layer in BART for m-th event.

As illustrated in Figure 2, we employ new special $\{\langle\texttt{ned}\rangle, \langle\texttt{emo}\rangle, \langle\texttt{pgt}\rangle\}$ tokens to delimit each protagonist's need, emotion and name grounded in the story context. In addition, we utilize a special $\langle\texttt{cxt}\rangle$ token to delimit each story event of the context:

$$b_m = \texttt{Emb}(n_1, e_1, p_1, X_1, \ldots, n_m, e_m, p_m) \quad (1)$$

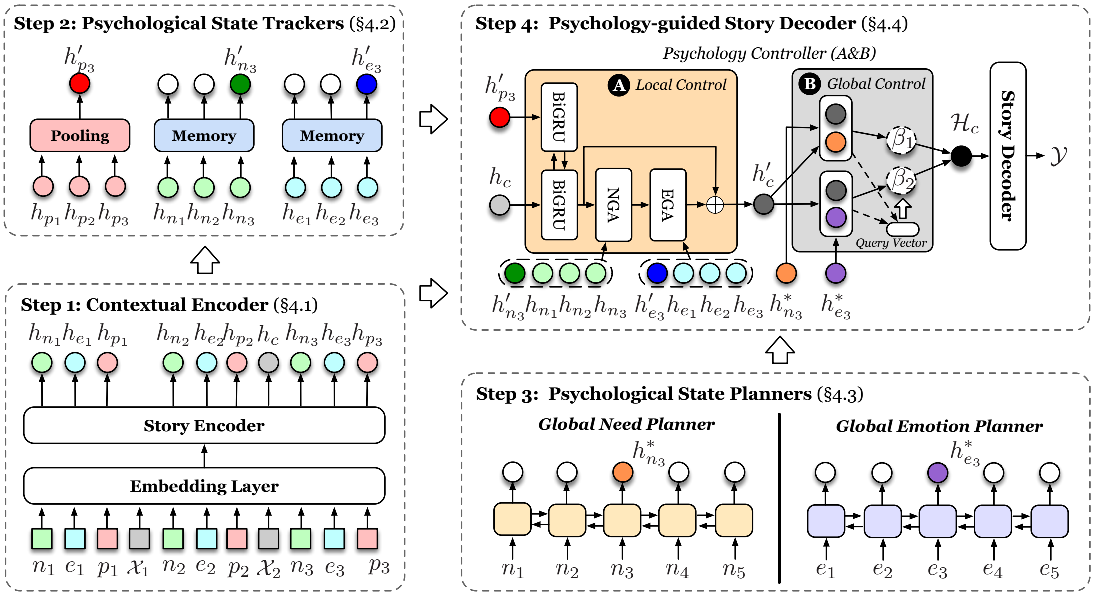

Figure 3: Overview of **PICS** with time point m=3 **(Step 1-4)**. In step 1, contextual encoder converts input into contextual representation (§ 4.1). In step 2, we design psychological state trackers to capture temporal relations for the protagonist's character information, local need and emotion (§ 4.2). In step 3, two psychological state planners output the global psychological state through modeling the completed need and emotion chains (§ 4.3). In step 4, conditioned on the protagonist's local and global psychological states, the decoder generates psychology-guided stories with a psychology controller (A&B) (§ 4.4).

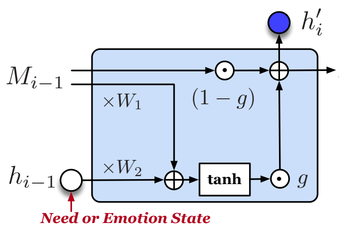

Figure 4: Details of Memory Units in Psychological State Trackers for need and emotion states.

Then, we acquire representations of i-th psychological states $h_{n_i}, h_{e_i}, h_{p_i}$ by extracting the hidden states of special $\{\langle\texttt{ned}\rangle, \langle\texttt{emo}\rangle, \langle\texttt{pgt}\rangle\}$ tokens on the top of encoder.

Similarly, story context representation is corresponding to the hidden state of $(m-1)$-th special $\langle\texttt{cxt}\rangle$ token.

4.2 Psychological State Trackers (Step 2)

For the purpose of remembering and updating the protagonist's psychological states that have been mentioned, we design psychological state trackers for the protagonist's character information[§], needs and emotions.

[§]In this paper, we regard the representation of the protagonist's name as his/her character information.

Protagonist's Character Information Based on several story events, humans can easily guess the protagonist's character information. We argue that the resulting representation can stand for the protagonist's character information via pooling hidden states of the protagonist's name which is grounded in story events.

$$h'_p = \texttt{Pooling}(\{h_{p_i}\}_{i=1}^{m}) \tag{2}$$

In this work, $\texttt{Pooling}$ is Mean-Pooling and is used as a tracker for p_i to conclude the moderate representation of the protagonist's character information.

Protagonist's Needs To remember and update the mentioned protagonist's needs and emotions, we design trackers (memory block in Figure 3) as follows:

$$h'_n = \texttt{Memory}(\mathbb{N}) \tag{3}$$

As shown in Figure 4, for memorizing mentioned needs, the memory unit M_{n_i} is updated using $h_{n_{i-1}}$ and $M_{n_{i-1}}$, the output contextual needs representation:

$$\widehat{h}_{n_{i-1}} = \texttt{tanh}(W_1 M_{n_{i-1}} + W_2 h_{n_{i-1}}) \tag{4}$$

Futher, we use a gating mechanism, g, to allow the model to learn to flexibly control how much each cell in memory is updated, as below:

$$g_{n_i} = \texttt{sigmoid}(W_3 M_{n_{i-1}} + W_4 h_{n_{i-1}}) \quad (5)$$

$$M_{n_i} = g_{n_i} \widehat{h}_{n_{i-1}} + (1 - g_{n_i}) M_{n_{i-1}} \quad (6)$$

$$h'_{n_m} = M_{n_m} \quad (7)$$

where M_{n_0} is randomly initialized and all W_* are trainable parameters.

Protagonist's Emotions In the same way with need, for historical emotions:

$$h'_{e_m} = \texttt{Memory}(\mathbb{E}) = M_{e_m} \quad (8)$$

In summary, we can obtain sequential psychological state changes through the above operations.

4.3 Psychological State Planner (Step 3)

Different from psychological state trackers, we obtain the global psychological states by encoding the global needs chain and emotions chain for story planning. As shown in Figure 3 (step 3), psychological state trackers use a BiGRU (Cho et al., 2014) architecture.

Global Need Planner In the global need planner, the global need representations:

$$[t_n^1, t_n^2, \ldots, t_n^5] = \texttt{BiGRU}(\mathbb{A}_{\mathbb{N}}) \quad (9)$$

$$h^*_{n_j} = t_n^j \quad (10)$$

where j represents the j-th need for generate next story event and j-th need in the global need chain.

Global Emotion Planner Similarly, computing global emotion representations are as below:

$$h^*_{e_j} = \texttt{BiGRU}(\mathbb{A}_{\mathbb{E}})[j] = t_e^i \quad (11)$$

note that, $\mathbb{A}_{\mathbb{N}}$ and $\mathbb{A}_{\mathbb{E}}$ are initialized by GloVe (Pennington et al., 2014) embedding.

After obtaining global need and emotion representations, we feed them into the following step as a planning signal for guiding the story generation.

4.4 Psychology-guided Decoder (Step 4)

Conditioned on the protagonist's local and global psychological states, the decoder generates a psychology-guided story event by the following modules.

4.4.1 Psychology Controller

In order to control the story generation by protagonist's psychological states, we respectively integrate local and global psychological states (Peng et al., 2022; Zhang et al., 2022) into the story context representation.

Local Control With the goal to integrate local psychological control information into the representation of story context h_c, a psychology controller is used to compute the interaction between h_c and local psychological states (including h'_p, h'_n, h'_e). First, h'_p guides the model with the protagonist's character information for generating the next story event, which uses a BiGRU (Cho et al., 2014) architecture:

$$\widetilde{h}_c = \texttt{BiGRU}(h_c, h'_p)[0] \quad (12)$$

where we extract the hidden state $\widetilde{h}_c$ as the story context representation considering character information. Then, we employ an attention mechanism to integrate local psychological states (need and emotion) into the story context representation. Firstly, need guided attention NGA is defined as follows:

$$h_c^n = \texttt{NGA}(\widetilde{h}_c, \{h_{n_i}\}_{i=1}^m) = \sum_{i=1}^{m} \alpha_i h_{n_i} \quad (13)$$

$$\{\alpha_i\}_{i=1}^m = \mathit{softmax}(\{\widetilde{h}_c h_{n_i}^T / \sqrt{dim_1}\}_{i=1}^m) \quad (14)$$

$$\widehat{h_c^n} = \texttt{Fus.N}(h'_n, h_c^n) = \texttt{MLP}([h'_n, h_c^n]) \quad (15)$$

where dim_1 equals to the dimension of h_c^n. Similarly, EGA and Fus.E has the same operation as the above equations:

$$\widehat{h_c^e} = \texttt{Fus.E}(\texttt{EGA}(\widehat{h_c^n}, \{h_{e_i}\}_{i=1}^m)) \quad (16)$$

$$h'_c = \widehat{h_c^e} + \widetilde{h}_c \quad (17)$$

Global Control Aiming at further control composing stories with the global planning signal, this part is designed to dynamically integrate global psychological states. In specific, a query vector q is introduced to fuse psychology-blended context representations and global psychological states by attention mechanism as below:

$$s_n, s_e = \frac{q[h'_c, h^*_{n_i}]^T}{\sqrt{dim_2}}, \frac{q[h'_c, h^*_{e_i}]^T}{\sqrt{dim_2}} \quad (18)$$

$$\beta_1, \beta_2 = \mathit{softmax}(s_n, s_e) \quad (19)$$

$$\mathcal{H}_c = \texttt{MLP}(\beta_1[h'_c, h^*_{n_i}] + \beta_2[h'_c, h^*_{e_i}]) \tag{20}$$

where q is the query and $[h'_c, h^*_{n_i}]$, $[h'_c, h^*_{e_i}]$ are the keys for attention. $[\cdot]$ denotes the concatenation operation. dim_2 equals to the dimension of $[h'_c, h^*_{n_i}]$. So that the model can adaptively choose the most important global psychological state for generating well-planned stories.

4.4.2 Story Decoder

We employ a left-to-right BART decoder to generate a story conditioned upon all input elements. Each layer of the decoder additionally performs cross-attention over the concatenation of the final hidden layer of the BART encoder and $\mathcal{H}_c$.

$$P(y_t|\mathbb{X}, \mathbb{P}, \mathbb{N}, \mathbb{E}, \mathbb{A}_\mathbb{N}, \mathbb{A}_\mathbb{E}, y_{<t}) = \textit{softmax}(W_s s_t) \tag{21}$$

$$s_t = \texttt{Dec}(y_{<t}, \texttt{Enc}(b_m), \mathcal{H}_c) \tag{22}$$

where $\texttt{Enc}(b_m)$ is the final hidden layer of the BART encoder.

4.4.3 Training

The training objective is to minimize the negative log-likelihood L of the ground truth story event:

$$\mathcal{L} = -\sum_{t=1}^{r} \log P\left(y_t \mid \mathbb{X}, \mathbb{P}, \mathbb{N}, \mathbb{E}, \mathbb{A}_\mathbb{N}, \mathbb{A}_\mathbb{E}, y_{<t}\right) \tag{23}$$

Aiming to obtain the completed story, we iteratively generate the m-th story events $\mathcal{Y}_m$ with the forecast generated $\mathcal{Y}_1, ..., \mathcal{Y}_{m-1}$.

5 Experiment

5.1 Data

We choose a Story Commonsense (Rashkin et al., 2018) that has been annotated with a similar setting to us. Story Commonsense is a large-scale dataset as a resource for training and evaluating the mental state tracking of characters in short commonsense stories. This dataset contains over 300k low-level annotations for character motivations and emotional reactions. Story Commonsense was proposed for studying need/emotion tracking. Each sentence is annotated for all characters, and there are 3 crowd-workers voting for each need/emotion. If the characters have no need or emotion, the psychological state will be labeled 'none'.

Based on our task definition, we extract the story with a protagonist (occurs in more than 4 sentences) and the corresponding need/emotion chains from Story Commonsense. Note that, we select the **Top-1** need/emotion label, based on annotators' voting scores to make up the psychological chains in our data set. If several labels have the best score of all, we will choose a low-level need label of Maslow's needs or a random emotion label. Following the 8:1:1 splitting ratio, we obtain 2,570/321/321 five-sentence stories for train/dev/test sets. In our psychology-guided CSG task, we generate a story event based on the story context, protagonist's name, need chain, and emotion chain. Therefore, each story will be reformed into 4 samples following our setting (i.e. 10,280/1283/1283 samples).

5.2 Implement Details

For a fair comparison, we train our proposed models and the baselines with the same input (leading context and global need/emotion chains) that are automatically extracted from Story Commonsense (Rashkin et al., 2018). Our proposed models follow the setting of BART large (Lewis et al., 2020) model with 12 layers in each of the encoder and decoder and a hidden size of 1024. The stories are encoded using BPE with a vocabulary size of 50,257. We set the maximum sequence length to 100 tokens, as it is large enough to contain all inputs. We use Adam optimization with an initial learning rate of 0.00001. All models were trained until there was no improvement in the validation set performance. During training, we use a label smoothed cross-entropy loss, with the smoothing parameter set to 0.1. At inference time, we set beam size as 5, and remove duplicated trigrams in beam search. We use the HuggingFace [¶] (Wolf et al., 2020) PyTorch (Paszke et al., 2019) implementation[‖] on Tesla V100 GPU.

5.3 Evaluation Metrics

Automatic Metrics We use the following metrics for automatic evaluation: **(1) Perplexity (PPL)** is an indicator of fluency. A smaller value is better. **(2) BLEU** (Papineni et al., 2002) is used for evaluating the overall quality of the generated story. We use n=1, 2. **(3) Rouge** (Lin, 2004) with n=1, 2, L is used to measure the similarity between automatically generated and reference results. **(4) Need/Emotion Consistency (NC/EC)** It is a learnable automatic metric. We fine-tune a RoBERTa

[¶]https://github.com/huggingface/transformers

[‖]We will make our dataset and code publicly available at `https://github.com/IndexFziQ/PICS`.

Models	PPL ↓	BLEU-1 ↑	BLEU-2 ↑	Rouge-1 ↑	Rouge-2 ↑	Rouge-L ↑	NC ↑	EC ↑
Fusion	25.68	19.83	2.89	5.81	1.66	8.64	0.19	0.16
Plan&Write	19.43	20.15	3.56	6.23	1.85	8.81	0.31	0.19
PPLM	-	18.42	4.39	8.73	2.13	9.29	0.45	0.41
GPT-2 FT	18.21	22.67	6.42	9.97	2.25	9.98	0.34	0.37
BART FT	17.85	21.84	6.03	9.32	2.51	9.56	0.36	0.32
PICS	**16.73**	**23.51**	**6.89**	**12.43**	**3.83**	**11.28**	**0.64**	**0.45**

Table 2: The results of automatic evaluation on test set considering common-used metrics and the designed metric (NC/EC) to test the psychological state consistency of stories. ↓/↑ indicates the lower/higher, the better.

NC		EC	
Accuracy	**F1-Score**	**Accuracy**	**F1-Score**
64.6	64.8	56.8	59.2

Table 3: Accuracy and F1-Score of RoBERTa classifier for need/emotion on dev set, respectively.

(Liu et al., 2019) large model on the Story Commonsense (Rashkin et al., 2018) train set as a classifier to distinguish whether a story event is corresponding to a **Top-1** need/emotion. Table 3 shows results of NC/EC.

Manual Metrics We also conduct a manual evaluation of generated psychology-guided stories. Following Song et al. (2019), crowd-workers are required to evaluate actions on a 0-3 scale (3 being very good) from two different perspectives:

(1) Content Quality to indicate whether the generated story is fluent. **(2) Content Rationality** to assess whether it follows the given needs and emotions which is reasonable and consistent.

During the manual evaluation, we display the input (leading context and global need/emotion arc) and two stories generated by the two models being compared. To avoid prejudice, we randomly changed the order in which the stories in the two models were displayed to the crowd-workers. We provided crowd-workers with instructions to explain the annotations and provided examples. Following this process, each pair of stories is annotated by three crowd-workers.

5.4 Experimental Results

5.4.1 Baselines

For a fair comparison, we train **PICS** and the baselines with the same input (leading context and global need/emotion chains).

We compare our base storytelling model, **PICS**, with following state-of-the-art models:

1. **Fusion** (Fan et al., 2018b), a storytelling model that first pre-trains a convolutional seq2seq model, then fixes the trained model and passes it to the second clone model with fusion mechanism.

2. **Plan&Write** (Yao et al., 2019), another storytelling model first generates a plot as a sequence of keywords with the given leading context and then conditioned on the plot it generates the text of the story.

3. **PPLM** (Dathathri et al., 2020), which can be extended to accept psychological state chains for controlling story generation. We use psychological state chains as the skeleton.

4. **GPT-2 FT** (Radford et al., 2019) is a pre-trained generative LM. We use a medium-size version. We fine-tuned GPT-2 on our dataset following (Guan et al., 2020) with leading context and global need/emotion chains.

5. **BART FT** (Lewis et al., 2020) is a encoder-decoder architecture. We fine-tuned BART on our dataset following (Lewis et al., 2020) with leading context and global need/emotion chains.

All of these models are trained, validated and tested on the same data splits described in §5.1. In specific, we add emotion/need labels as additional input tokens to baseline models alongside the tokens for each story sentence. And, global emotion/need chains that concatenated with the story context are given to baseline models at each time step.

5.4.2 Automatic Evaluation

The results of the automatic evaluation are shown in Table 2. Our model outperforms the variants of GPT-2 in terms of perplexity, and has higher BLEU and Rouge scores than all the baselines, indicating

Model	Quality ↑		Rationality ↑	
	Fluency	Coherence	Need	Emotion
PPLM	2.56	1.62	1.05	1.27
GPT-2 FT	**2.87**	1.58	1.73	1.82
BART FT	2.72	**1.79**	1.69	1.98
PICS	2.83	1.68	**2.16**	**2.34**

Table 4: Manual Evaluation in terms of content quality and content rationality about the generated stories.

Model	AVG-B ↑	AVG-R ↑	NC ↑	EC ↑
PICS	15.89	8.65	0.64	0.45
w/o PST	15.41	8.45	0.53	0.46
w/o PSP	**15.24**	8.47	0.56	0.43
w/o PC	15.34	**8.24**	0.55	0.42
w/o Need	15.71	8.53	**0.42**	0.39
w/o Emotion	15.64	8.44	0.51	**0.35**

Table 5: Ablation study of **PICS** model and global need/emotion chains on dev set. PST: psychological state tracker. PSP: psychological state planner. PC: psychology controller.

better fluency and more overlaps with the reference stories. Besides, in the view of NC and EC scores, the stories generated by **PICS** are more consistent with the desired psychological state chains, either need or emotion.

5.4.3 Manual Evaluation

We perform a manual evaluation between our model and baselines. We randomly generate 100 stories from the test set. For each story, we hire three annotators to give a score in terms of content quality (fluency&coherence) and content rationality (need&emotion). For each aspect, we use an average of the three annotations. We adopt majority voting to make the final decisions among the annotators. As shown in Table 4, all the results show that our model outperforms baselines significantly in fluency, coherence, and psychological state consistency.

6 Discussion and Analysis

6.1 Ablation Study

An ablation study is conducted on the Story Commonsense dataset to examine the impact of each module separately. We train the model each time by excluding one of our model's modules. And, we summarize the results in Table 5. The results illustrate the harms that the elimination of each of the proposed modules from **PICS** architecture could cause. This attests to the effectiveness of all proposed approaches in the generation of higher controllable stories and subsequently resulting in more accurate evaluation metrics. As this table demonstrates, the NC/EC accuracy drops the most by ablating the psychological state planner and the psychology controller, which shows that they have the most significant role in composing high-quality psychology-guided stories and consequently accurate evaluation metrics. Besides, psychological state chains (need/emotion chains) all contributed to the high-quality psychology-guided stories.

PICS v.s.	Win	Loss	Tie	κ
PPLM	**54.6%**	18.5%	26.9%	30.2
GPT-2 FT	**53.2%**	19.5%	27.3%	30.4
BART FT	**52.3%**	18.4%	29.3%	27.9

Table 6: Human A/B Test of **PICS**. Results show that **PICS** performs baseline models sufficiently. κ denotes Fleiss' kappa (all are fair agreement or moderate agreement). The p-value of scores < 0.05 in sign test.

6.2 Human A/B Test

We try to compare our model with other baselines by conducting a Human A/B test. Particularly, we randomly sample 100 examples each for our model and baseline models. Three annotators are given generated responses from either our model or baselines in random order and are required to choose a better option. They can either choose one of the responses or select "Tie" when the quality of provided options is hard to access. Results in Table 6 confirm that the responses from **PICS** are more preferred by human judges.

6.3 Case Study

In this section, we present some generated examples in Table 7. We select need-related keywords and emotion-related keywords (based on our observation), which are highlighted in corresponding colors, respectively. From the third block, **PICS** can generate more natural and reasonable psychology-guided stories than baselines. Since the proposed models can generate stories conditioned on the protagonist's psychological state chains, they can be used to unfold a story in diverse situations. We demonstrate this capability in the last 4 blocks of Table 7 which perform counterfactual transformations on need/emotion chains. It shows two examples where for the same leading context, our model can generate stories that follow the counterfactual

Pgt.	*Tory*
Ned.	Stab. → Love → Love → Love → Love
Emo.	Fear →Fear → Joy → Joy →Joy
Cxt.	**Tory had doubts about getting married.**
Golden	She talked to her fiance about their decision. The groom reassured her that he loved her. She remembered how much she loved him too. The wedding went forward without anymore problems.
GPT-2	She called her husband. She got the answer and laugh. She go to shopping then. She bought so many clothes she like.
BART	She talked to her boyfriend about their relationship. Her boyfriend loved her and gave her a kiss. She was very happy then. They went on the date outside for fun.
PICS	Tory is afraid about their marriage. Her boyfriend gave a gift as a blessing. Tory was happy about the gift, and reply him with a kiss. Tory went home to prepare a surprise for him.
Ned.-CF	Stab. → Love → Love →Phys. → Phys.
PICS	She talked to her husband about their marriage. Her husband loved her and take her to the romantic dinner. It is time for them to have a dinner. She was happy with the delicious dinner.
Emo.-CF	Fear →Fear →Anger →Anger →Anger
PICS	She and her husband talked about their marriage. Her husband was angry at her doubts. She rushed and fought with his husband. She gave him a slap and broke up unhappy.

Table 7: Generated stories by different models with need and emotion chains. Each psychological state and its corresponding tokens are highlighted in the same color. **CF** represents performing counterfactual transformation on need chains or emotion chains.

psychological state chains of the protagonist.

6.4 Error Analysis

Although the proposed model outperforms the state-of-the-art baselines, it needs to be noted that there are still many unreasonable stories losing to other models in human evaluation. Therefore, we analyzed error types by manually checking all lost stories in pairwise comparisons between our model and two strong baselines including GPT-2 and BART to reveal the factors that affect the performance. The numbers of stories which lost to our model are 56/64 of 100/100 in total for GPT-2 and BART, respectively. And there are 61 stories of 200 generated by **PICS** losing to these two baselines.

We conclude three main types of error from the lost stories: **repetition** (repeating the same actions about the need/emotion), **conflicting psychological state** (wrong causal relation about psychological state), and **ambiguous psychological state** (difficult to understand the psychological state). The distribution of different error types is shown in Figure 5. We can observe that conflicting and ambiguous psychological state make up most of the errors for all the models. Compared with GPT-2 and BART, **PICS** reduces chaotic scenes effectively but still suffers from severe repetition, as shown in Table 8. However, the analysis result illustrates that generating a reasonable psychology-guided story is a challenging task.

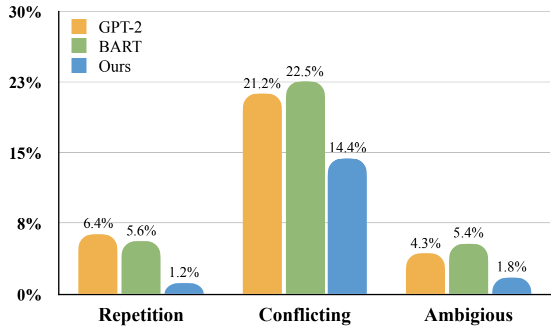

Figure 5: Distribution of error types for **PICS** (ours) and baseline models (GPT-2 and BART).

Error Type	Cases
Repetition	I went to a friends house for a party last weekend. *I was so excited to go.* We played a lot of games that night. *I was so excited to go.* We also had a lot of food to eat.
Conflicting	Zack wanted to vote in the election. Unfortunately Zack was traveling during the election. *Zack went to the school for taking class.* Zack then voted again. Zack was happy to have voted.
Ambiguous	Alice made a cake for her mother. *The cake is so sweet that her mother disliked it. But her mother was very happy and encouraged Alice. Alice was sad because of the failed cooking.* They hugged and smiled in the end.

Table 8: Cases of different typical errors. *Italic* words denote the error story events.

7 Conclusion and Future Work

In this paper, we propose a **PICS** system to generate controllable stories that adhere to the story context and protagonist's psychological state chains. Specifically, we model and integrate local and global psychological states of the protagonist as the story progress. Experiments demonstrate that **PICS** significantly outperforms baselines and each part shows effectiveness. In future work, it is important to build a large-scale dataset for developing psychology-guided controllable story generation, regarding aspects of multilingual and long text. Besides, our methodology can be generalized to a wide range of areas, such as automatic storytelling systems and intelligent education agents.

Acknowledgement

We thank all anonymous reviewers for their constructive comments and useful advice. Also thanks for the discussion with Yajing Sun, Yongxiu Xu, and Ping Guo. This work is supported by the National Natural Science Foundation of China (No.62006222 and No.U21B2009). Thanks for COLING organizers and the proposed pre-trained language models, data, codes.

Contribution List Yuqiang Xie: Idea, Paper Writing, Coding; Yue Hu: Guiding, Discussion; Yunpeng Li: Coding; Guanqun Bi: Paper Polish, Discussion; Luxi Xing: Review; Wei Peng: Review.

Thanks for the hard work and dedication of all team members.

References

Amal Alabdulkarim, Siyan Li, and Xiangyu Peng. 2021. Automatic story generation: Challenges and attempts. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 72–83, Virtual. Association for Computational Linguistics.

Prithviraj Ammanabrolu, W. Cheung, William Broniec, and Mark O. Riedl. 2021. Automated storytelling via causal, commonsense plot ordering. In *AAAI*.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Faeze Brahman and Snigdha Chaturvedi. 2020. Modeling protagonist emotions for emotion-aware storytelling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5277–5294, Online. Association for Computational Linguistics.

Paweł Budzianowski and Ivan Vulić. 2019. Hello, it's GPT-2 - how can I help you? towards the use of pretrained language models for task-oriented dialogue systems. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 15–22, Hong Kong. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar. Association for Computational Linguistics.

Elizabeth Clark, Yangfeng Ji, and Noah A. Smith. 2018. Neural text generation in stories using entity representations as context. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2250–2260, New Orleans, Louisiana. Association for Computational Linguistics.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Angela Fan, David Grangier, and Michael Auli. 2018a. Controllable abstractive summarization. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 45–54, Melbourne, Australia. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018b. 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, Florence, Italy. Association for Computational Linguistics.

Jian Guan, Fei Huang, Zhihao Zhao, Xiaoyan Zhu, and Minlie Huang. 2020. A knowledge-enhanced pretraining model for commonsense story generation. *Transactions of the Association for Computational Linguistics*, 8:93–108.

Chenyang Huang, Osmar Zaïane, Amine Trabelsi, and Nouha Dziri. 2018. Automatic dialogue generation with expressed emotions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 49–54, New Orleans, Louisiana. Association for Computational Linguistics.

Parag Jain, Priyanka Agrawal, Abhijit Mishra, Mohak Sukhwani, Anirban Laha, and Karthik Sankaranarayanan. 2017. Story generation from sequence of independent short descriptions. *ArXiv*, abs/1707.05501.

Yuta Kikuchi, Graham Neubig, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. 2016. Controlling output length in neural encoder-decoders. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1328–1338, Austin, Texas. Association for Computational Linguistics.

Xiangzhe Kong, Jialiang Huang, Ziquan Tung, Jian Guan, and Minlie Huang. 2021. Stylized story generation with style-guided planning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2430–2436, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Danyang Liu, Juntao Li, Meng-Hsuan Yu, Ziming Huang, Gongshen Liu, Dongyan Zhao, and Rui Yan. 2020. A character-centric neural model for automated story generation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*, pages 1725–1732. AAAI Press.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Fuli Luo, Damai Dai, Pengcheng Yang, Tianyu Liu, Baobao Chang, Zhifang Sui, and Xu Sun. 2019. Learning to control the fine-grained sentiment for story ending generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6020–6026, Florence, Italy. Association for Computational Linguistics.

Lara J. Martin, Prithviraj Ammanabrolu, Xinyu Wang, William Hancock, Shruti Singh, Brent Harrison, and Mark O. Riedl. 2018. Event representations for automated story generation with deep neural nets. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18)*, pages 868–875. AAAI Press.

Abraham Harold Maslow. 1943. A theory of human motivation. In *Psychological review*.

D. Morrow. 1985. Prominent characters and events organize narrative understanding. *Journal of Memory and Language*, 24:304–319.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035.

Debjit Paul and Anette Frank. 2021. COINS: Dynamically generating COntextualized inference rules for narrative story completion. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5086–5099, Online. Association for Computational Linguistics.

Nanyun Peng, Marjan Ghazvininejad, Jonathan May, and Kevin Knight. 2018. Towards controllable story generation. In *Proceedings of the First Workshop on Storytelling*, pages 43–49, New Orleans, Louisiana. Association for Computational Linguistics.

Wei Peng, Yue Hu, Luxi Xing, Yuqiang Xie, Yajing Sun, and Yunpeng Li. 2022. Control globally, understand locally: A global-to-local hierarchical graph network for emotional support conversation. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 4324–4330. ijcai.org.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

R. Y. Pérez and M. Sharples. 2001. Mexica: A computer model of a cognitive account of creative writing. *Journal of Experimental & Theoretical Artificial Intelligence*, 13:119 – 139.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In *Theories of emotion*.

J. Porteous and M. Cavazza. 2009. Controlling narrative generation with planning trajectories: The role of constraints. In *ICIDS*.

Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. 2019. Counterfactual story reasoning and generation. 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 5043–5053, Hong Kong, China. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. In *OpenAI Blog*.

Hannah Rashkin, Antoine Bosselut, Maarten Sap, Kevin Knight, and Yejin Choi. 2018. Modeling naive psychology of characters in simple commonsense stories. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2289–2299, Melbourne, Australia. Association for Computational Linguistics.

Hannah Rashkin, Asli Celikyilmaz, Yejin Choi, and Jianfeng Gao. 2020. PlotMachines: Outline-conditioned generation with dynamic plot state tracking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4274–4295, Online. Association for Computational Linguistics.

Paul Ricoeur. 1984. Time and narrative.

Mark O. Riedl and R. M. Young. 2010. Narrative planning: Balancing plot and character. *ArXiv*, abs/1401.3841.

Melissa Roemmele. 2016. Writing stories with help from recurrent neural networks. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 4311–4342. AAAI Press.

Zhenqiao Song, Xiaoqing Zheng, Lu Liu, Mu Xu, and Xuanjing Huang. 2019. Generating responses with a specific emotion in dialog. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3685–3695, Florence, Italy. Association for Computational Linguistics.

Pradyumna Tambwekar, Murtaza Dhuliawala, Lara J. Martin, Animesh Mehta, Brent Harrison, and Mark O. Riedl. 2019. Controllable neural story plot generation via reward shaping. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5982–5988. ijcai.org.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

K. Vonnegut. 1981. Palm sunday: An autobiographical collage. In *RosetTaBooks, LLC New York*.

Di Wang, Nebojsa Jojic, Chris Brockett, and Eric Nyberg. 2017. Steering output style and topic in neural response generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2140–2150, Copenhagen, Denmark. Association for Computational Linguistics.

Noah Weber, Leena Shekhar, Heeyoung Kwon, Niranjan Balasubramanian, and Nathanael Chambers. 2020. Generating narrative text in a switching dynamical system. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 520–530, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *ArXiv*, abs/1901.08149.

Yuqiang Xie, Yue Hu, Wei Peng, Guanqun Bi, and Luxi Xing. 2022a. Comma: Modeling relationship among motivations, emotions and actions in language-based human activities.

Yuqiang Xie, Yue Hu, Luxi Xing, Yunpeng Li, Wei Peng, and Ping Guo. 2022b. Clseg: Contrastive learning of story ending generation. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8057–8061.

Feifei Xu, Xinpeng Wang, Yunpu Ma, Volker Tresp, Yuyi Wang, Shanlin Zhou, and Haizhou Du. 2020a. Controllable multi-character psychology-oriented story generation. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 1675–1684. ACM.

Jingjing Xu, Xuancheng Ren, Yi Zhang, Qi Zeng, Xiaoyan Cai, and Xu Sun. 2018. A skeleton-based model for promoting coherence among sentences in narrative story generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4306–4315, Brussels, Belgium. Association for Computational Linguistics.

Peng Xu, Mostofa Patwary, Mohammad Shoeybi, Raul Puri, Pascale Fung, Anima Anandkumar, and Bryan Catanzaro. 2020b. MEGATRON-CNTRL: Controllable story generation with external knowledge using large-scale language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2831–2845, Online. Association for Computational Linguistics.

Lili Yao, Nanyun Peng, R. Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In *AAAI*.

Duzhen Zhang, Zhen Yang, Fandong Meng, Xiuyi Chen, and Jie Zhou. 2022. TSAM: A two-stream attention model for causal emotion entailment. *CoRR*, abs/2203.00819.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18)*, pages 730–739. AAAI Press.

Xianda Zhou and William Yang Wang. 2018. MojiTalk: Generating emotional responses at scale. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1128–1137, Melbourne, Australia. Association for Computational Linguistics.