What Makes The Story Forward? Inferring Commonsense Explanations as Prompts for Future Event Generation

Anonymous ACL submission

Abstract

Future Event Generation (FEG) aims to generate fluent and reasonable future event descriptions given preceding events. It requires not only fluent text generation but also commonsense reasoning to maintain the coherence of the entire event story. However, existing FEG methods are easily trapped into repeated or general events without imposing any logical constraint to the generation process. In this paper, we propose a novel explainable FEG framework that consists of a commonsense inference model (Im) and an event generation model (Gm). The Im, which is pre-trained on a commonsense knowledge graph ATOMIC, learns to interpret the preceding events and conducts commonsense reasoning to reveal the character's psychology such as *intent*, *reaction* and *needs* as latent variables. The Gm further takes the commonsense knowledge as prompts to guide and enforce the generation of logistically coherent future events. As a unique merit, the commonsense prompts can be further decoded into textual descriptions, yielding explanations for the future event. Automatic and human evaluation demonstrate that our approach can generate more coherent, specific, and logical future events than the strong baselines. All the programs and resources will be made public upon acceptance.

1 Introduction

Future event generation (FEG) is the task of generating descriptions of future human activities given the preceding events. As exemplified in Figure 1, given the previous and current events, *Leah moved to a new town* and *she had to go to a new school*, a FEG system is expected to generate a consequence event, e.g., *she felt nervous about making new friends*. FEG is beneficial to many real-world applications, such as story telling (Fan et al., 2018, 2019), question answering (Shwartz et al., 2020), abductive reasoning (Bhagavatula et al., 2019).

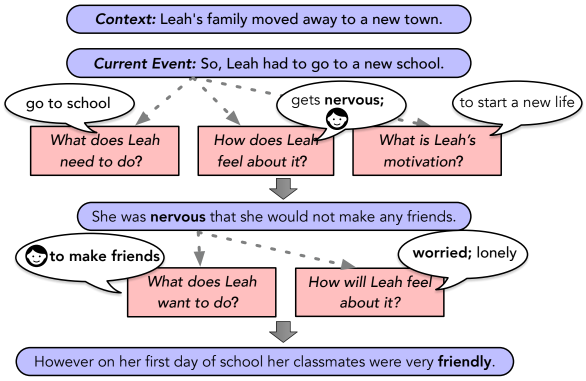

Figure 1: Examples of future event generation and commonsense explanation. The smiley faces indicate the dominant information for future events.

Recent studies have explored pre-trained language models (PLMs), such as BERT (Devlin et al., 2018), GPT (Radford et al., 2019; Brown et al., 2020), and BART (Lewis et al., 2020), and leveraged external commonsense Knowledge Graphs (KG), such as ConceptNet (Speer et al., 2017) and ATOMIC (Martin et al., 2018), to improve the generation of stories[1] and future events (Guan et al., 2020; Xu et al., 2020). However, the future events generated by these studies are either too generic or lack logically coherence, which is mainly due to the reason that they either fine-tune the PLMs on the commonsense KG (Guan et al., 2020) and thus the approaches cannot well retain the commonsense inference capability during the generation of future events, or rely on information retrieval to return the most relevant knowledge (Xu et al., 2020; Ammanabrolu et al., 2020) while the coverage of the KGs is far from enough.

To tackle these challenges, we propose a novel solution to jointly infer the latent commonsense knowledge from preceding events and take it as prompts for FEG. Our motivation is that there is a wide spectrum of inferential knowledge, such as the *cause* and *effect* of the preceding events or

[1]In this work, a story is defined as a sequence of events.

the *intent*, *reaction*, *needs* of the character inferred from the preceding events, which naturally leads the story forward and the prediction of the future events. As shown in Figure 1, given that *Leah had to go to a new school*, if we correctly infer that the *emotional reaction* of *Leah* would be *nervous*, we can better predict a future event, *Leah felt nervous about making new friends*. However, there is still a critical question remaining: how to best leverage the latent commonsense knowledge to enhance future event generation, especially there are no available datasets providing sufficient annotations for various latent commonsense inference?

We further propose to answer the question with a novel COEP framework that infers **Co**mmonsense **E**xplanations to **P**rompt FEG. It consists of a commonsense Inference Model (IM) learning to infer the latent commonsense knowledge from preceding events and a future event Generation Model (GM) that takes the commonsense knowledge as soft prompts conditional on preceding events to predict future events. Inspired by the prior studies (Bosselut et al., 2019; Hwang et al., 2021), we first fine-tune the IM on ATOMIC. An additional discriminator is also pre-trained with IM to distinguish whether the commonsense inference is correlated with the input events, which is further applied to weakly supervise the learning of the commonsense prompts in GM. Compared with all previous studies on FEG, a unique advantage of COEP lies in that the latent commonsense prompts can be further decoded into textual descriptions, yielding explanations for the future event.

In summary, the contributions of this work are: (i) We propose a new COEP framework which infers the latent commonsense knowledge from preceding events and takes it as soft prompts to guide the logically coherent future event generation. (ii) Our COEP framework is explainable as the commonsense representations corresponding to prompts can be decoded into particular textual explanations by IM. (iii) We have conducted extensive experiments on publicly available benchmarks. Both automatic and human evaluations demonstrate the effectiveness of COEP, and further ablation studies on our results highlight the consistent, specific, and logical generation process.

2 Methodology

We formulate the FEG task as follows: given a sequence of history events $X = (e_1, e_2, \ldots, e_{n-1})$ indicating the background context and a current event [2] e_n which is directly prior to the future event e_{n+1}, the model learns to capture the contextual and commonsense information and generate e_{n+1}.

Our COEP framework aims to incorporate the commonsense knowledge inferred from preceding events to guide the FEG task. As shown in Figure 2, it consists of two components: (1) a commonsense Inference Model (IM), which is fine-tuned on ATOMIC to infer the commonsense knowledge given events and a particular commonsense relation (i.e., 9 commonsense dimensions as illustrated in Table 1) as input; and (2) a future event Generation Model (GM) that takes the various commonsense knowledge as soft prompts to enhance the future event generation. Both of these two models are based on BART (Lewis et al., 2020), a large-scale pre-trained language model. Based on the fine-tuned IM, we directly use the latent representations from IM encoder as continuous prompt vectors to GM. To tune the prompts during the future event generation, we also design a discriminator to estimate the coherence between the commonsense inference decoded from the latent representations and the preceding events.

Input Event: PersonX repels PersonY's attack		
xIntent (PersonX intent) to protect others	**xEffect** (PersonX effect) gains an enemy	**oReact** (Other react) weak; ashamed
xNeed (PersonX need) to defense himself	**xWant** (PersonX want) to call the police	**oWant** (Other want) attack again
xAttr (PersonX attribute) skilled; brave	**xReact** (PersonX react) angry; tired	**oEffect** (Other effect) get hurts

Table 1: An example of ATOMIC. Texts in () show the extended relations for IM fine-tuning.

2.1 Commonsense Inference Model

As aforementioned, the commonsense Inference Model (IM) is based on a pre-trained BART (Lewis et al., 2020). Following previous studies (Bosselut et al., 2019; Hwang et al., 2021), we first fine-tune the IM on ATOMIC (Martin et al., 2018), a large-scale commonsense KG covering 9 dimensions of inferential knowledge as described in Table 1. We formulate the training tuples for IM as $\langle x_{\mathcal{I}}, u \rangle$, where $x_{\mathcal{I}}$ denotes a multi-segment sequence which concatenates an input event e and an extended relational phrase r corresponding to each

[2] In event stories, each event is a sentence describing human's daily activities as shown in Figure 1

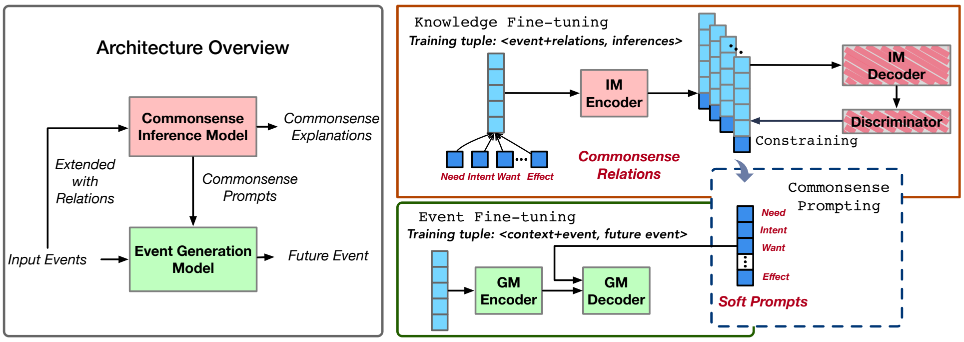

Figure 2: The architecture of COEP framework. We decompose the framework into the following two parts: 1) the commonsense inference model (IM) fine-tuned with ATOMIC; 2) the event generation model (GM) to capture the contextual information of preceding events. The prompting block can integrate commonsense information as prompts to guide the event generation, which is illustrated in the right dashed frame.

commonsense dimension[3], e.g., PersonX intent, as shown in the parenthesis in Table 1. For each segment, we add two special tokens $\langle s\rangle$ and $\langle /s\rangle$ to represent the beginning and ending separately following (Bhagavatula et al., 2019). u is a textual description denoting the commonsense knowledge inferred from $x_{\mathcal{I}}$.

$$P(u_t|u_{<t}) = \sigma(\text{DEC}_{\mathcal{I}}(\mathbf{H}^l_{u_{<t}}, \text{ENC}_{\mathcal{I}}(x_{\mathcal{I}}))\mathbf{W}+\mathbf{b})$$

where u_t and $u_{<t}$ denote the t-th token and all the previous t-1 tokens in u. $\mathbf{H}_{u_{<t}}$ are the decoder hidden states of all the t-1 tokens. l is the total number of layers in the encoder and decoder. $\text{ENC}_{\mathcal{I}}$ and $\text{DEC}_{\mathcal{I}}$ indicate the encoder and decoder in IM respectively. $\mathbf{W}$ and $\mathbf{b}$ are learnable parameters. σ represents the softmax function to produce the probability of output tokens throughout this paper. The training objective is to minimize the following negative log-likelihood:

$$L^{lm}_{\mathcal{I}} = -\sum_{t=1}^{|u|} logP(u_t|u_{<t})$$

where $|u|$ denotes the total number of tokens in the target commonsense inference.

To better encourage the IM to infer the commonsense knowledge, we further designed a discriminator to score the coherence between the commonsense inference and the input event and relation. For each tuple $s = \langle x_{\mathcal{I}}, u\rangle$ constructed from ATOMIC, we randomly sample another u' from other tuples and construct a negative sample $\langle x_{\mathcal{I}}, u'\rangle$. We then design a discriminator based on the BART sequence classification head, which is optimized with the cross-entropy objective:

$$L^D_{\mathcal{I}} = -logP(\mathbf{I}_s = \tilde{\mathbf{I}}_s | s = \langle x_{\mathcal{I}}, u\rangle)$$

$$\mathbf{I}_{s=\langle x_{\mathcal{I}},u\rangle} = \begin{cases} 0, & u: \text{true} \\ 1, & u: \text{negative} \end{cases}$$

where $\tilde{\mathbf{I}}_s$ refers to the binary logits produced by the discriminator.

The overall objective of fine-tuning IM is to minimize the combination of the two objectives:

$$L_{\mathcal{I}} = L^{lm}_{\mathcal{I}} + L^D_{\mathcal{I}}$$

2.2 Event Generation Model

The event Generation Model (GM) is based on another pre-trained BART that considers the preceding events as well as the commonsense inference from the IM to generate the future events. To better acquire the future event generation capability, we leverage the ConceptNet (Speer et al., 2017), a general multilingual KG covering 36 relations, such as *Antonym*, *SimilarTo*, *HasSubevent* and so on. We carefully select 6 types of relations that are related to sequential events[4] and collect 39,530 event pairs $\langle e_p, e_f\rangle$ for fine-tuning GM, where e_p and e_f denote the preceding and future event respectively. The average number of words in the events is 2.67. The objective of ConceptNet fine-tuning is to generate e_f given e_p by minimizing the following negative log-likelihood:

$$L^{cn} = -\sum^{|w|} \sigma(\text{DEC}_{\mathcal{G}}(H^l_{w_{<t}}, \text{ENC}_{\mathcal{G}}(e_p))\mathbf{W}+\mathbf{b})$$

where $|w|$ denotes the total tokens in target tail events. $\text{ENC}_{\mathcal{G}}$ and $\text{DEC}_{\mathcal{G}}$ indicate GM encoder and decoder.

[3] We use the training splits from (Sap et al., 2019), which splits 24,313 seed events into training, validation, and test sets (80%/10%/10%), for fine-tuning the IM where the average number of words in each event is 4.6.

[4] The relations indicate sequential order between events are: *Causes*, *HasPrerequisite*, *HasSubevent*, *HasFirstSubevent*, *HasPrerequisite*,*HasLastSubevent*.

After fine-tuning GM on the ConceptNet, we finally train it on FEG task by considering both the preceding events and the commonsense inference from IM. To enrich the context information, GM will take all the history events as well as the current event as input, which are concatenated as a multi-segment sequence x_G, where each segment corresponds to a preceding event and special tokens $\langle s \rangle$ and $\langle /s \rangle$ are also added at the beginning and ending of each segment. To incorporate the commonsense inference from the IM, we introduce a prompting block that collects the last hidden state of $\langle /s \rangle$ from IM encoder based on each commonsense relation and take them as soft prompts. Given an extended input $x_{\mathcal{I}_i}$ based on the preceding events and a particular commonsense relation r_i, we obtain the last hidden state of the corresponding $\langle /s \rangle$ as follows:

$$h_{k_i} = \text{ENC}_{\mathcal{I}}(x_{\mathcal{I}_i})_{\langle /s \rangle}, i \in [1, 9]$$

We then take the 9 dimensional commonsense prompts as well as context encoding of all preceding events from the GM encoder as input to the GM decoder and generate a future event:

$$\mathbf{H} = [h_{k_1}, h_{k_2}, \ldots, h_{k_9},\ \text{ENC}_{\mathcal{G}}(x_G)]$$
$$P(w_t|w_{<t}) = \sigma(\ \text{DEC}_{\mathcal{G}}(\mathbf{H}^l_{w_{<t}}, \mathbf{H})\mathbf{W} + \mathbf{b})$$

where w_t is the t-th token in the target future event.

The objective of future event generation is to minimize the negative log-likelihood as follows:

$$L^{lm}_{\mathcal{G}} = -\sum^{|w|} logP(w_t|w_{<t})$$

We add an auxiliary classification layer to improve the contrastive comprehension of GM. Given a FEG training sample $\langle e_1, \ldots, e_n, e_{n+1} \rangle$, the negative sample is constructed by replacing e_{n+1} with a randomly sample event e', where $e' \neq e_{n+1}$. The classification task is designed to distinguish whether a future event is sequentially consistent with the preceding events similar to the discriminator in IM, whose objective function is represented as $L^{cls}_{\mathcal{G}}$. The overall training loss for FEG is:

$$L_{\mathcal{G}} = L^{lm}_{\mathcal{G}} + L^{cls}_{\mathcal{G}}$$

2.3 Prompt Training Strategy

As we use the latent continuous commonsense representations as soft prompts to guide the generation of the future event, the next question is: *How to supervise the prompts training?* It is challenging because there are no available datasets containing the annotations of both future events and the latent commonsense inference in-between the events. We propose to solve this problem by taking advantage of the discriminator pre-trained for the IM, which is to measure the coherence of the commonsense inference to the input event and relation.

Specifically, given an event and a commonsense relation r_i, denoted as $x_{\mathcal{I}_i}$, we use IM encoder to get the latent commonsense representation $\text{ENC}_{\mathcal{I}(x_{\mathcal{I}_i})}$ as prompts to GM. As there is no gold standard target commonsense inference, we use the pre-trained discriminator to measure the coherence between input events and decoded inferences. To solve the non-differentiable problem for conditional decoding, we use the straight-through Gumbel Softmax (GS) estimator (Jang et al., 2016) which provides a continuous relaxation for the one-hot distribution of $argmax$, and get the commonsense inference as follows:

$$\begin{aligned}
\widetilde{\mathbf{H}}^l_{u_t} &= \text{DEC}_{\mathcal{I}}(\mathbf{H}^l_{u_{<t}}, \text{ENC}_{\mathcal{I}(x_{\mathcal{I}_i})}) \\
u^p_t &= \text{argmax}(\sigma(\widetilde{\mathbf{H}}^l_{u_t}\mathbf{W} + \mathbf{b})) \\
\mathbf{H}^0_{u_t} &= \text{GS}(\sigma(\widetilde{\mathbf{H}}^l_{u_t}\mathbf{W} + \mathbf{b})) \cdot \mathbf{E}_V
\end{aligned}$$

where $\mathbf{E}_V$ is the vocabulary embedding matrix.

When optimizing the commonsense prompts, we freeze the parameters of the IM decoder and discriminator and only update the IM encoder, to minimize the following loss function:

$$L_{sc} = -logP(\tilde{\mathbf{I}}_s = 0 | s = \langle x_{\mathcal{I}}, u^p \rangle)$$

where $\tilde{\mathbf{I}}_s$ is the estimated label produced by the IM discriminator given $x_{\mathcal{I}}$ and commonsense inference u^p generated by IM decoder. In the end, the overall training loss for future event generation is defined as follows:

$$L = L_{\mathcal{G}} + L_{sc}$$

3 Experiments

3.1 Dataset

We evaluate our model on a commonsense story dataset (Rashkin et al., 2018), which is constructed based on the ROCStories Corpus, containing 14,738 stories that are claimed to have inner psychology of story characters as a chain of mental states to push the story forward. It has various settings for mental states detection (Tandon et al., 2018; Paul and Frank, 2019; Otani and Hovy, 2019), future event generation (Chaturvedi et al., 2017; Wang et al., 2017), story telling (Yao et al., 2019; Guan et al., 2020) and story cloze test (Mostafazadeh et al., 2016). Here we create two settings for future event generation and story telling respectively. As each story consists of 5 sentences of events, for FEG task, we construct a Common-Event dataset by unfolding each story and taking the i-th sentence as the current event, all previous sentences as history context, and the next sentence as the future event. For story telling, we simply give the first sentence of each story as a start event and have the models generate all follow-up events.

3.2 Baselines

We use the following approaches as baselines as they are commonly used in various generation tasks and have achieved the state-of-the-art performance. **Pointer Generator** with coverage (See et al., 2017) uses a hybrid pointer-generator network using coverage to keep track of repeat tokens to discourage repetition. **GPT-2 (Fine-tune)** is fine-tuned on event dataset (Mostafazadeh et al., 2016) GPT-2 model following (Guan et al., 2020). **GPT-2 (wKG)** is a knowledge-enhanced pre-trained model (Guan et al., 2020) for commonsense story generation based on GPT-2 model. **BART (Fine-tune)** (Lewis et al., 2020) is based on the pre-trained BART-base model[5] and fine-tuned on the CommonEvent dataset. **BART (wKG)** is based on the pre-trained BART-base model and fine-tuned on ATOMIC similar to GPT-2 (wKG) before event training.

We also introduce several variants of COEP to study the effectiveness of each main component: (1) COEP w/o CN which omits the ConceptNet fine-tuning on GM to evaluate if implicitly fine-tuning on sequential knowledge improves FEG. (2) COEP w/o PT which removes prompt training objective L_{sc} to evaluate the effectiveness of the proposed prompt training strategy, which is equivalent to directly concatenating the prompts without any constraint. (3) COEP w/o CLS which omits the classification task L_G^{cls} to verify if the contrastive comprehension can promote event generation.

3.3 Evaluation Metrics

We evaluate the experimental results with both automatic metrics and human evaluation. The automatic metrics include: **Perplexity (PPL)** defined as the exponential average negative log-likelihood evaluating the fluency. Automated metrics to measure the performance of text generation: **BLEU** (Papineni et al., 2002), **ROUGE_L** (Lin, 2004), **METEOR** (Banerjee and Lavie, 2005), **CIDEr** (Vedantam et al., 2015), and **BERTScore** (Zhang et al., 2019) [6]. **Repetition-n** (Shao et al., 2019) measures the redundancy of stories by computing the average ratio of repetitive n-grams in generated stories. **Distinct-n** (Li et al., 2016) measures the generation diversity by the ratio of distinct ones within all generated n-grams.

For human evaluation, we randomly sampled 100 instances from the test set and obtained 400 future events generated by the BART-based models which come top in FEG among the baselines, a variant model w/o PT to investigate the impact of prompt training strategy, and our approach. With the ground-truth, for each instance, we obtain five candidate future events and ask three annotators to rank them based on the logical consistency. **Hit@k** measures the winning rate of each model by computing the percentage of its ranking landing in top k among the candidates. We also use **Spearman's** ρ (Spearman, 1961) and the **Kendall's** τ (Kendall, 1945) to measure the inter-agreement of annotators.

3.4 Evaluation of Future Event Generation

3.4.1 Automatic Evaluation

Table 2 shows the automatic evaluation of FEG performance of all baselines and our approach[7]. We can see that (1) our model significantly outperforms all the baselines and variants based on all evaluation metrics. (2) BART-based models show obvious superiority compared with both Pointer Generator and GPT-2 models but still suffer the issue of illogicality, even with conventional KG fine-tuning, which demonstrates the effectiveness of the latent commonsense representations as prompts to future event generation. (3) The highest BERTScore shows that COEP can promote the semantic consistency of generated events, which reveals that our model can effectively capture the commonsense information from KG and apply it to FEG.

Ablation studies on the main components are shown at the bottom of Table 2. We can see that (1) without prompt training (w/o PT) which is equivalent to directly concatenating the commonsense prompts and the preceding events, CIDEr and BERTScore drop rapidly. This verifies the effectiveness of the prompt training strategy to maintain

[5]We use the pre-trained BART-base model from Hugginface https://huggingface.co/facebook/bart-base

[6]All these automated metrics are implemented following (Hwang et al., 2021)

[7]We use topk-4 searching strategy to generate future events and commonsense explanations.

Models	PPL↓	BLEU-1↑	BLEU-2↑	BLEU-4↑	METEOR↑	ROUGE_L↑	CIDEr↑	BERTScore↑
Ptr-Gen	25.79	5.73	0.89	0.00	4.63	6.60	0.82	38.00
GPT-2 (Finetune)	14.51	8.35	3.98	0.67	8.95	11.45	12.29	47.61
BART (Finetune)	11.0	15.01	5.79	1.60	10.66	14.35	17.25	49.50
GPT-2 (wKG)	12.17	13.41	4.37	0.80	9.75	12.57	13.82	48.63
BART (wKG)	11.38	15.38	6.13	1.75	11.01	14.52	20.25	49.91
COEP	**9.62**	**16.31**	**6.74**	**1.94**	**11.95**	**15.36**	**25.30**	**50.72**
w/o PT-CN	10.80	15.62	6.29	1.79	11.27	14.88	21.19	50.17
w/o PT	10.83	15.85	6.40	1.79	11.44	14.93	21.88	50.22
w/o CN	10.59	15.74	6.57	1.94	11.76	15.09	24.48	50.33
w/o CLS	11.30	15.61	6.35	1.82	11.43	14.73	24.21	50.41

Table 2: Automatic evaluation results on FEG task. **Bold**: the best performance. Underlined: the second place.

semantic consistency. (2) Fine-tuning GM on ConceptNet brings limited improvements. It is consistent with our claim that implicitly fine-tuning the pre-trained language model with KG lacks effective constraints to control the knowledge inferring on downstream tasks. (3) The additional classification task in GM improves the semantic similarity between the events and references, as it uses a related task to enhance the model's contrastive ability.

3.4.2 Human Evaluation

Models	Hit@1 (%)	Hit@2 (%)	ρ
BART (Finetune)	3.34	16.70	0.23
BART (wKG)	2.00	12.34	0.24
COEP (w/o PT)	2.00	33.34	**0.29**
COEP	**19.33**	**63.00**	0.28
Golden Story	72.67	86.67	0.44

Table 3: Human evaluation results for FEG.

The human evaluation results on generated events are shown in Table 3, we can see (1) our model achieves a relatively unanimous high rank only second to the ground truth. 19.33 percentage of events are rated as the most consistent results, and 63 percentage of events are rated as top 2 results. (2) The performance gaps are even larger than that of automatic evaluation. That is, the actual achievements of our proposed model are more than our expectation, the automatic metrics need further improvements. (3) Spearman's ρ calculates the inter agreement between annotators on the rankings of each model and Kendall's τ computes the agreement on all instances. It seems that the ranking of Golden Story achieves a relatively high consistency among annotators while other models get even performance which is acceptable to consider the human evaluations are convincing. We have an average Kendall's τ of 0.412, which shows moderate agreement among annotators on the sort of 5 candidates in each instance.

3.5 Evaluation of Story Telling

To further investigate the commonsense inferring ability of proposed models, we also provide the performance of several models on story telling task. Different from GPT-2 based models, which produce next tokens autogressively until the end of story, BART-based models generate next sentences step by step till the last event. Since each story in ROCStories dataset contains 5 sentences, we use the first sentence as the start event and make the models to recurrently generate 4 future events to complete it. The results are shown in Table 4. Our model achieves the best performance based on almost all metrics except CIDEr, because it relies on low-frequency words rather than the semantic consistency between sentences. The lowest repetition-4 and highest distinct-4 scores indicate that our approach can also generate more diverse and specific events, demonstrating the effectiveness of two sub-model designs combined via prompting.

3.6 Analysis of Commonsense Prompts

We conduct an additional ablation study on the impact of commonsense prompts based on different commonsense relations. We compare the future event generation performance of our approach based on the commonsense prompt from each dimension, as shown in the left columns in Table 5. We can see that among the 9-dimensional commonsense prompts, *xEffect* is the most effective one, and even shows better performance than BART (wKG) in Table 2 which is implicitly enhanced with all dimensions of commonsense knowledge.

As the commonsense prompts can also be explained by decoding them into textual commonsense inference with IM decoder, we further evaluate the commonsense prompts based on the cor-

Models	BLEU-1↑	BLEU-2↑	METEOR↑	CIDEr↑	BertScore↑	Repetition-4↓	Distinct-4↑
GPT-2 (Finetune)	17.02	5.43	11.75	6.84	50.50	5.73	90.32
GPT-2 (wKG)	17.69	5.78	12.35	8.87	50.97	6.05	91.75
BART (Finetune)	20.53	5.86	14.23	17.01	50.32	9.44	84.01
BART (wKG)	20.18	7.81	13.96	**17.31**	51.13	8.48	81.31
COEP	**22.32**	**7.85**	**14.98**	17.14	**52.16**	**1.96**	**98.82**

Table 4: Automatic evaluation on Story Telling task. **Bold**: the best performance. Underlined: the second place.

Relation	Automatic		Human	
	BLEU-2/4	BERTScore	Task#1	Task#2
xNeed	6.12 / 1.59	50.12	0.55	0.22
xAttr	6.06 / 1.54	50.09	0.62	0.48
xEffect	**6.30 / 1.71**	50.08	0.46	0.35
xReact	6.25 / 1.60	**50.15**	0.47	0.39
xWant	6.10 / 1.55	49.98	0.75	0.63
xIntent	6.09 / 1.50	49.98	**0.86**	**0.68**
oEffect	6.13 / 1.64	50.05	0.66	0.51
oReact	6.10 / 1.60	50.09	0.57	0.49
oWant	6.10 / 1.52	50.04	0.74	0.54

Table 5: Automatic and human evaluations results on FEG task with different commonsense prompts.

Context:	None.
Current Event:	Ron needed to learn how to **throw a curveball**.
Future Event:	He ended up consulting his high school's **coach** for advice.
GPT-2 (wKG):	I told my friend I would play with him.
BART (FT):	He decided to go to the *doctor*.
BART (wKG):	He decided to try out for the team.
COEP:	He went to the **coach** and asked for help.
Explanations:	xAttr: determined, curious; xEffect: gets exercise;
Context:	Jack was taking his **SAT test** on friday. He studied hard all week. On Thursday he was invited to a **party**.
Current Event:	He knew he should not but he **went to the party** anyway.
Future Event:	Jack **did poorly** on the test because he was too sleepy to concentrate.
GPT-2 (wKG):	He had a good weekend and a great time.
BART (FT):	He had a *great time*.
BART (wKG):	Jack had a *great time* at the party.
COEP:	Jack did not study for his test and he **failed** the test!
Explanations:	xNeed: to study; xEffect: gets nervous

Table 6: Generated future events from different models. **Bold** phrases denote **key** information coherent with inputs. *Italic* words donate *improper* events which is illogical or neutral. Underlined words denote effective explanations for event generation from COEP.

rectness of the textual explanations with human evaluation. We design two tasks for annotators to judge: **Task #1**: *whether the explanation is coherent with input preceding events* and **Task #2**: *whether the explanation provides necessary information for generated events*, where 1 stands for yes and 0 is for no. The right columns in Table 5 show the average answer scores on randomly sampled 100 instances. We can see that (1) our model can generate reasonable and coherent explanations on 9 dimensions of commonsense relations, especially xIntent, which shows the highest correlation with input events. (2) The explanations serve as a bridge between preceding events and future events, as their score is highly correlated, which well supports our explicitly explainable framework. We find an interesting phenomenon that human evaluations show that the most correlated commonsense explanations come from xIntent relation, but the automatic evaluation results considering only xIntent prompt are rather low. It reveals that although the IM performs well in commonsense reasoning, how to effectively integrating such information in downstream tasks still has a long way to go, which motivates our future work on model's explainability.

4 Case Study

4.1 Qualitative Comparison

Table 6 presents several examples with future events generated by various methods, which indicates that our approach consistently generates more reasonable and coherent future events than the baselines. For example, given that *Ron wants to learn about sports (**curveball**)*, COEP will generate a future event suggesting him to *ask a **coach** for help*. We also observe that our approach can also capture the **turning points**. Considering the second example, the explanation shows that Jack needs to study, but *he went to the party the day just before the test* leads to his failure in the test.

4.2 Error Analysis

We also present some typical errors made by our model in Table 7. It shows that although COEP significantly outperforms the baselines and variants in generating reasonable future events, it still makes some errors, such as improper synonym (*bike &*

Input:	Tom always wanted a *motorcycle*. Tom went to his local Harley Davidson dealership.
COEP:	Tom picked up a *bike* he liked.
Input:	In 1996, my parents tooks a trip to *Europe*.
COEP:	They went on a trip to *Mexico*.
Input:	Mark was so in love with his girlfriend. Mark was going to propose to her tonight. He took her out to the nicest place in town. Mark got down on one knee and ask her to marry him.
Next Event:	She said no she stopped loving him months ago.
COEP:	She said yes and Mark was so happy!

Table 7: Typical errors made by our model. *Italic* words denote the improper synonym replacement or regional inclusion relation. Underlined words represent a totally different but reasonable event compared with ground truth.

motorcycle), chaotic regional relations (*Mexico & Europe*) and opposite understanding of contexts (*yes & no* to the same content). Especially the last case, it shows our framework makes yet reasonable but different understanding about preceding events, which is actually not the model's fault, but due to the open ending. It also demonstrates that human evaluation is still necessary for measuring logical coherence in event generation tasks.

5 Related Work

Future Event Generation Pre-trained language models such as GPT (Radford et al., 2019; Brown et al., 2020), BART (Lewis et al., 2020), T5 (Raffel et al., 2019) have shown the effectiveness in generation tasks such as text summarization (Gupta et al., 2021) and machine translation (Radford et al., 2019). Compared with such tasks of which the inputs have contained sufficient information to generate the desired output, future event generation is an open-ended generation task and especially requires commonsense inferences to generate logically consistent output. Previous studies on this task explored context clues and commonsense KG based pre-training to enforce the model to generate reasonable and coherent stories (Guan et al., 2019, 2020; Xu et al., 2020; Ammanabrolu et al., 2020). However, simply fine-tuning PLMs on commonsense KGs cannot guarantee that it can retain the capability of commonsense inference when it's fine-tuned for future event generation, and the coverage of the KGs is also uncontrollable. In stark contrast, our approach explicitly generates commonsense explanations and takes the commonsense representations as prompts to generate coherent future events.

Prompt Tuning Prompt tuning (Brown et al., 2020) is a simple yet effective mechanism for learning “soft prompts" from PLMs to perform specific downstream tasks. The prompts are usually continuous representations from a frozen model which typically refer to a task description and/or several canonical examples (Shin et al., 2020; Reynolds and McDonell, 2021; Li and Liang, 2021; Lester et al., 2021). There are two significant differences between our work and previous studies. First, instead of learning task-oriented prompts as previous studies did, we propose to generate all types of latent commonsense representations based on preceding events and take them as instance-level prompts to guide FEG. Second, the prompts in our model are independent vectors attached to contextual representations of input events, while above prompts are partial inner representations in pre-trained models (e.g., prefix of hidden states in a layer). It can keep the commonsense prompts customized for each instance.

6 Conclusion and Future Work

In this paper, we propose a novel FEG framework name COEP which infers commonsense knowledge as soft prompts to enhance the logicality of future event generation. There are two key components: 1) commonsense Inference Model (IM) and 2) event Generation Model (GM). We initialize the components by inheriting a BART-base model pre-trained on a large corpus. Two different KG are used to fine-tune the models for commonsense reasoning and sequential inference separately. The soft prompts are supervised by a pre-optimized discriminator in IM and the corresponding latent representations can be decoded into textual descriptions, which provide explanations and justification for the future event. Extensive experiments on an open-domain event story dataset show that our model can outperform strong baselines in FEG. Automatic and manual evaluations substantiate the contextual and logical coherence of generated events.

For future work, it would be very interesting to migrate the architecture to a more advanced pre-training model like GPT-3, like achieving the commonsense knowledge in a Few-Shot way or Zero-Shot way to decrease training costs. The pluggable design of the prompting framework is extensible because we can update IM and GM separately without re-training the whole model, and we would like to explore its application on other generation tasks like summarization and dialogue generation.

References

Prithviraj Ammanabrolu, Wesley Cheung, William Broniec, and Mark O Riedl. 2020. Automated storytelling via causal, commonsense plot ordering. *arXiv preprint arXiv:2009.00829*.

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.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen-tau Yih, and Yejin Choi. 2019. Abductive commonsense reasoning. *arXiv preprint arXiv:1908.05739*.

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.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Snigdha Chaturvedi, Haoruo Peng, and Dan Roth. 2017. Story comprehension for predicting what happens next. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1603–1614.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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.

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.

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.

Jian Guan, Yansen Wang, and Minlie Huang. 2019. Story ending generation with incremental encoding and commonsense knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6473–6480.

Anushka Gupta, Diksha Chugh, Rahul Katarya, et al. 2021. Automated news summarization using transformers. *arXiv preprint arXiv:2108.01064*.

Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. On symbolic and neural commonsense knowledge graphs.

Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Maurice G Kendall. 1945. The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

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.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

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

Lara Martin, Prithviraj Ammanabrolu, Xinyu Wang, William Hancock, Shruti Singh, Brent Harrison, and Mark Riedl. 2018. Event representations for automated story generation with deep neural nets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Naoki Otani and Eduard Hovy. 2019. Toward comprehensive understanding of a sentiment based on human motives. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4672–4677.

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.

Debjit Paul and Anette Frank. 2019. Ranking and selecting multi-hop knowledge paths to better predict human needs. *arXiv preprint arXiv:1904.00676*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

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.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Zhihong Shao, Minlie Huang, Jiangtao Wen, Wenfei Xu, and Xiaoyan Zhu. 2019. Long and diverse text generation with planning-based hierarchical variational model. 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 3257–3268.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*.

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. *arXiv preprint arXiv:2004.05483*.

Charles Spearman. 1961. " general intelligence" objectively determined and measured.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Niket Tandon, Bhavana Dalvi Mishra, Joel Grus, Wen-tau Yih, Antoine Bosselut, and Peter Clark. 2018. Reasoning about actions and state changes by injecting commonsense knowledge. *arXiv preprint arXiv:1808.10012*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Zhongqing Wang, Yue Zhang, and Ching Yun Chang. 2017. Integrating order information and event relation for script event prediction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 57–67.

Peng Xu, Mostofa Patwary, Mohammad Shoeybi, Raul Puri, Pascale Fung, Anima Anandkumar, and Bryan Catanzaro. 2020. Megatron-cntrl: Controllable story generation with external knowledge using large-scale language models. *arXiv preprint arXiv:2010.00840*.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.