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 logically 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 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.
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 \( e_n \) 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.

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

<table>
<thead>
<tr>
<th>xIntent</th>
<th>xEffect</th>
<th>oReact</th>
</tr>
</thead>
<tbody>
<tr>
<td>(PersonX intent) to protect others</td>
<td>(PersonX effect) gains an enemy</td>
<td>(Other react) weak; ashamed</td>
</tr>
<tr>
<td>xNeed</td>
<td>xWant</td>
<td>oWant</td>
</tr>
<tr>
<td>(PersonX need) to defense himself</td>
<td>(PersonX want) to call the police</td>
<td>(Other want) attack again</td>
</tr>
<tr>
<td>xAttr</td>
<td>xReact</td>
<td>oEffect</td>
</tr>
<tr>
<td>(PersonX attribute) skilled; brave</td>
<td>(PersonX react) angry; tired</td>
<td>(Other effect) get hurts</td>
</tr>
</tbody>
</table>

2 In event stories, each event is a sentence describing human’s daily activities as shown in Figure 1
commonsense dimension, e.g., PersonX intent, as shown in the parenthesis in Table 1. For each segment, we add two special tokens $s$ and $\ell/s$ to represent the beginning and ending separately following (Bhagavatula et al., 2019). $u$ is a textual description denoting the commonsense knowledge inferred from $x_I$.

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

where $u_t$ and $u_{<t}$ denote the $t$-th token and all the previous $t-1$ tokens in $u$. $H^l_{u_{<t}}$ are the decoder hidden states of all the $t-1$ tokens. $l$ is the total number of layers in the encoder and decoder. $\text{ENC}_{I}$ and $\text{DEC}_{I}$ indicate the encoder and decoder in $\text{IM}$ respectively. $W$ and $b$ are learnable parameters. $\sigma$ 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:

$$L^{\text{lm}}_I = -\sum_{t=1}^{[u]} \log P(u_t|u_{<t})$$

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

To better encourage the $\text{IM}$ 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 $s = \langle x_I, u \rangle$ constructed from ATOMIC, we randomly sample another $u'$ from other tuples and construct a negative sample $\langle x_I, u' \rangle$. We then design a discriminator based on the BART sequence classification head, which is optimized with the cross-entropy objective:

$$L^D_I = -\log P(I_s = \bar{I}_s|s = \langle x_I, u \rangle)$$

$$I_{s=(x_I, u)} = \begin{cases} 0, & u : \text{true} \\ 1, & u : \text{false} \end{cases}$$

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

The overall objective of fine-tuning $\text{IM}$ is to minimize the combination of the two objectives:

$$L_I = L^{\text{lm}}_I + L^D_I$$

### 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 $\text{IM}$ 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 and collect 39,530 event pairs $(e_p, e_f)$ 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:

$$L^{\text{cn}}_G = -\sum_{w} \sigma(\text{DEC}_G(H^l_{w_{<t}}, \text{ENC}_G(e_p)))W + b)$$

where $[w]$ denotes the total tokens in target tail events. $\text{ENC}_G$ and $\text{DEC}_G$ indicate GM encoder and decoder.

---

$^1$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 $\text{IM}$ where the average number of words in each event is 4.6.

$^2$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 $x_L$, 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:

$$h_{ki} = ENC_I(x_L)\langle /s \rangle, \quad i \in [1,9]$$

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:

$$H = [h_{k1}, h_{k2}, \ldots, h_{kn}, ENC_G(x_G)]$$

$$P(w_t|w_{<t}) = \sigma DEC_G(H_{w_{<t}}, H) W + b)$$

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:

$$L_{G}^{lm} = - \sum_{t=1}^{w}\log P(w_t|w_{<t})$$

We add an auxiliary classification layer to improve the contrastive comprehension of GM. Given a FEG training sample $(e_1, \ldots, e_n, e_{n+1})$, 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 $L_{G}^{cls}$. The overall training loss for FEG is:

$$L_G = L_{G}^{lm} + L_{G}^{cls}$$

### 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 $x_L$, we use IM encoder to get the latent commonsense representation $ENC_I(x_L)$ 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 $\arg\max$, and get the commonsense inference as follows:

$$\tilde{H}^{l}_u = DEC_I(H^{l}_u, ENC_I(x_L))$$

$$w^p_t = \arg\max(\sigma(\tilde{H}^l_u W + b))$$

$$H^{0}_u = GS(\sigma(\tilde{H}^l_u W + b)) \cdot E_V$$

where $E_V$ 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:

$$L_{sc} = -\log P(\hat{I}_s = 0|s = (x_L, u^p))$$

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

$$L = L_G + L_{sc}$$

### 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 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 $L_p$ 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 $L_{cls}^{d}$ 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). **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. 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...
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

<table>
<thead>
<tr>
<th>Models</th>
<th>PPL↓</th>
<th>BLEU-1↑</th>
<th>BLEU-2↑</th>
<th>BLEU-4↑</th>
<th>METEOR↑</th>
<th>ROUGE_L↑</th>
<th>CIDEr↑</th>
<th>BERTScore↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ptr-Gen</td>
<td>25.79</td>
<td>5.73</td>
<td>0.89</td>
<td>0.00</td>
<td>4.63</td>
<td>6.60</td>
<td>0.82</td>
<td>38.00</td>
</tr>
<tr>
<td>GPT-2 (Finetune)</td>
<td>14.51</td>
<td>8.35</td>
<td>3.98</td>
<td>0.67</td>
<td>8.95</td>
<td>11.45</td>
<td>12.29</td>
<td>47.61</td>
</tr>
<tr>
<td>BART (Finetune)</td>
<td>11.0</td>
<td>15.01</td>
<td>5.79</td>
<td>1.60</td>
<td>10.66</td>
<td>14.35</td>
<td>17.25</td>
<td>49.50</td>
</tr>
<tr>
<td>GPT-2 (wKG)</td>
<td>12.17</td>
<td>13.41</td>
<td>4.37</td>
<td>0.80</td>
<td>9.75</td>
<td>12.57</td>
<td>13.82</td>
<td>48.63</td>
</tr>
<tr>
<td>BART (wKG)</td>
<td>11.38</td>
<td>15.38</td>
<td>6.13</td>
<td>1.75</td>
<td>11.01</td>
<td>14.52</td>
<td>20.25</td>
<td>49.91</td>
</tr>
</tbody>
</table>

| 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.

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 $\rho$ calculates the inter agreement between annotators on the rankings of each model and Kendall’s $\tau$ 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 lower performance which is acceptable to consider the human evaluations are convincing. We have an average Kendall’s $\tau$ 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 1M decoder, we further evaluate the commonsense prompts based on the cor-
**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 &

---

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

<table>
<thead>
<tr>
<th>Models</th>
<th>BLEU-1↑</th>
<th>BLEU-2↑</th>
<th>METEOR↑</th>
<th>CIDEr↑</th>
<th>BertScore↑</th>
<th>Repetition-4↓</th>
<th>Distinct-4↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPT-2 (Finetune)</td>
<td>17.02</td>
<td>5.43</td>
<td>11.75</td>
<td>6.84</td>
<td>50.50</td>
<td>5.73</td>
<td>90.32</td>
</tr>
<tr>
<td>GPT-2 (wKG)</td>
<td>17.69</td>
<td>5.78</td>
<td>12.35</td>
<td>8.87</td>
<td>50.97</td>
<td>6.05</td>
<td>91.75</td>
</tr>
<tr>
<td>BART (Finetune)</td>
<td>20.53</td>
<td>5.86</td>
<td>14.23</td>
<td>17.01</td>
<td>50.32</td>
<td>9.44</td>
<td>84.01</td>
</tr>
<tr>
<td>BART (wKG)</td>
<td>20.18</td>
<td>7.14</td>
<td>13.96</td>
<td><strong>17.31</strong></td>
<td>51.13</td>
<td>8.48</td>
<td>81.31</td>
</tr>
<tr>
<td>COEP</td>
<td><strong>22.32</strong></td>
<td><strong>7.85</strong></td>
<td><strong>14.98</strong></td>
<td>17.14</td>
<td><strong>52.16</strong></td>
<td><strong>1.96</strong></td>
<td><strong>98.82</strong></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Relation</th>
<th>Automatic</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLEU-2/4</td>
<td>BERTScore</td>
</tr>
<tr>
<td>xNeed</td>
<td>6.12 / 1.59</td>
<td>50.12</td>
</tr>
<tr>
<td>xAttr</td>
<td>6.06 / 1.54</td>
<td>50.09</td>
</tr>
<tr>
<td>xEffect</td>
<td><strong>6.30 / 1.71</strong></td>
<td>50.08</td>
</tr>
<tr>
<td>xReact</td>
<td>6.25 / 1.60</td>
<td>50.15</td>
</tr>
<tr>
<td>xWant</td>
<td>6.10 / 1.35</td>
<td>49.98</td>
</tr>
<tr>
<td>xIntent</td>
<td>6.09 / 1.50</td>
<td>49.98</td>
</tr>
<tr>
<td>oEffect</td>
<td>6.13 / 1.64</td>
<td>50.05</td>
</tr>
<tr>
<td>oReact</td>
<td>6.10 / 1.60</td>
<td>50.09</td>
</tr>
<tr>
<td>oWant</td>
<td>6.10 / 1.32</td>
<td>50.04</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Context:</th>
<th>None. Current Event:</th>
<th>Ron needed to learn how to throw a curveball.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Future Event:</td>
<td>He ended up consulting his high school’s coach for advice.</td>
<td></td>
</tr>
<tr>
<td>GPT-2 (wKG):</td>
<td>I told my friend I would play with him.</td>
<td></td>
</tr>
<tr>
<td>BART (FT):</td>
<td>He decided to go to the doctor.</td>
<td></td>
</tr>
<tr>
<td>BART (wKG):</td>
<td>He decided to try out for the team.</td>
<td></td>
</tr>
<tr>
<td>CoEP:</td>
<td>He went to the coach and asked for help.</td>
<td></td>
</tr>
<tr>
<td>Explanations:</td>
<td>xAttr: determined, curious; xEffect: gets exercise;</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Context:</th>
<th>Jack was taking his SAT test on Friday. He studied hard all week. On Thursday he was invited to a party.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current Event:</td>
<td>He knew he should not but he went to the party anyway.</td>
</tr>
<tr>
<td>Future Event:</td>
<td>Jack did poorly on the test because he was too sleepy to concentrate.</td>
</tr>
<tr>
<td>GPT-2 (wKG):</td>
<td>He had a good weekend and a great time.</td>
</tr>
<tr>
<td>BART (FT):</td>
<td>He had a great time.</td>
</tr>
<tr>
<td>BART (wKG):</td>
<td>Jack had a great time at the party.</td>
</tr>
<tr>
<td>CoEP:</td>
<td>Jack did not study for his test and he failed the test!</td>
</tr>
<tr>
<td>Explanations:</td>
<td>xNeed: to study; xEffect: gets nervous;</td>
</tr>
</tbody>
</table>

---

7
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 KGs 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


