ABSTRACT

Prediction over event sequences is critical for many real-world applications in Information Retrieval and Natural Language Processing. Future Event Generation (FEG) is a challenging task in event sequence prediction because it requires not only fluent text generation but also commonsense reasoning to maintain the logical coherence of the entire event story. In this paper, we propose a novel explainable FEG framework, CoEP. It highlights and integrates two types of event knowledge, sequential knowledge of direct event-event relations and inferential knowledge that reflects the intermediate character psychology between events, such as intents, causes, reactions, which intrinsically pushes the story forward. To alleviate the knowledge forgetting issue, we design two modules, Iṣ and Gm, for each type of knowledge, which are combined via prompt tuning. First, Iṣ focuses on understanding inferential knowledge to generate commonsense explanations and provide a soft prompt vector for Gm. We also design a contrastive discriminator for better generalization ability. Second, Gm generates future events by modeling direct sequential knowledge with the guidance of Iṣ. Automatic and human evaluation demonstrate that our approach can generate more coherent, specific, and logical future events.

CCS CONCEPTS
• Computing methodologies → Natural language generation; Knowledge representation and reasoning.

KEYWORDS
Textual Event Generation; Commonsense Reasoning; Contrastive Training

ACM Reference Format:
Li Lin, Yixin Cao, Lifu Huang, Shu’Ang Li, Xuming Hu, Lijie Wen, and Jianmin Wang. 2022. What Makes the Story Forward? Inferring Commonsense Explanations as Prompts for Future Event Generation. In Proceedings of the 45th Int'l ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’22), July 11–15, 2022, Madrid, Spain. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3477495.3532080

1 INTRODUCTION

Prediction over event sequences has many real-world applications in the fields of Information Retrieval and Natural Language Processing, such as Scholar Citation Forecasting [17], Financial Quantitative Investments [7, 11], Personalized Recommender Systems [22], Response Generation [27, 29], and Story Telling [9, 10, 52]. Focusing on Future Event Generation (FEG), it is a task of predicting the next event given preceding events, which requires to understand the inherent logical structure that pushes a story forward. As exemplified in Figure 1, given preceding events including the background context and current event, 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.

With the help of pre-trained language models (PLMs), such as BERT [8], GPT [5, 38], and BART [25], recent FEG studies are able to generate fluent sentences [12], but usually fail in capturing logical coherence among the events. The generated events are either too general — “he does not know”, or neutral with the preceding events — producing “go to the doctor” given the preceding event “to learn a sport”. How can we capture the endogenous relationship among sequential events to make it not only fluent but also reasonable?

In this paper, we investigate the logical states behind the story that pushes the event sequence forward. Inspired by ATOMIC [43], we highlight the psychology of characters as a critical intermediate state between two adjacent events. That is, only if we understand the character’s motivation, feeling, intention, etc., we can predict a coherent next event. 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 and her motivation is to start a new life, we can better predict a future event, Leah felt nervous about making new friends. However, imposing logical constraints on the event generation process is challenging due to there are no available datasets providing sufficient psychology annotations. An
However on her first day of school her classmates were very friendly. However on her first day of school her classmates were very friendly.

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

efficient way to solve this issue is to leverage external commonsense Knowledge Graphs (KG). A group of works [18, 21, 53, 54, 59] introduce abundant taxonomic (i.e., AtLocation, IsA, etc.) and event-event relations (i.e., HasSubevent, HasPrerequisite, etc.) from ConceptNet [48] as external information to enhance FEG models. The sequential knowledge expressed by the event-event relations can supplement the direct correlation between events as additional training data. Another line of works focus on the inferential knowledge in ATOMIC [43]. ATOMIC has organized with 9 typed if-then relations with variables (e.g., “if X repels Y’s attack, X would intend to protect others from Y”). These relations reflect an intermediate state — indirect relations — between events. Clearly, the two types of knowledge are important and complementary to understand event sequence. However, few works utilize them both. We attribute one of the main reasons to the knowledge injection mechanism. They fine-tune PLMs on both KGs and downstream tasks [14] for knowledge transfer via shared parameters, while usually suffering from serious knowledge forgetting issues.

To this end, we propose a novel method that injects two types of knowledge for FEG, the inferential knowledge about the latent character psychology behind the event which can also provide explanations on how a story moves forward, and the sequential knowledge about the evolutionary intrinsic among the events. We design two modules for each type of knowledge, Im and Gm, which are combined together via a prompting strategy. We name the framework as Coep² that infers Commonsense Explanations³ to Prompt FEG. In specific, we initialize Im with BART [25], an state-of-the-art PLM specifically for text generation tasks. We then fine-tune it on ATOMIC [43] to capture inference knowledge similar to the prior studies [4, 19]. To alleviate the knowledge forgetting issue, the decoder will be frozen to generate intermediate explanations during FEG. The encoder outputs a soft prompt vector as the guidance for Gm. We also design a contrastive discriminator for Im to encourage

²The code is available at https://github.com/THU-BPM/CoEP
³We use inferential knowledge and commonsense explanations to express the latent psychology states which push an event story forward and the generated text scripts of the Im separately.

the decoded explanation associate with input events for better generalization. Get targets generating the next event for FEG. We thus initialize it with a pretrained BART model and fine-tuned on the direct event-event knowledge — the sequential KG derived from ConceptNet, which provides additional supervision.

In summary, the contributions of this work are:

- We propose a novel Coep framework that integrates two types of commonsense knowledge. Except for direct sequential knowledge, Coep also highlights the reasoning of intermediate inferential knowledge between events and takes it to prompt a logically coherent future event.
- Our proposed Coep provides interpretability to FEG. Based on the generated explanations, we analyze the impacts of various types of character psychology on what makes the story forward to shed lights for future research (Section 4.7).
- 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 RELATED WORK

2.1 Commonsense Inference

Reasoning about the causes and effects of events with human commonsense is an important task in Natural Language Processing as it’s an instinctive way to understand the law of logical events of humans [13]. Recently, as several benchmark datasets being constructed, such as ROCStories [32, 33] which is a commonsense story corpus with partial mental state annotations and wikiHow, researchers have put more attention to understanding and generating the intents [40, 41], and reasoning about relations between goals and steps [56, 57]. In addition, several large-scale knowledge graphs are also constructed to support commonsense reasoning. For example, ConceptNet [48] is a knowledge graph that connects natural language words and phrases with labeled edges. It concentrates more on taxonomic knowledge instead of the knowledge of how and why. ATOMIC [43] focuses on inferential knowledge related to if-then reasoning. It defines nine if-then relations to distinguish causes vs. effects, agents vs. themes, and actions vs. mental states. Further, generative models, such as COMeT [4, 19], are designed to automatically generate rich and diverse commonsense descriptions in natural language to make the knowledge graphs more complete. Inspired by these, we propose to train an automated inference model on ATOMIC which can generate commonsense explanations given the preceding events. It can naturally benefit the comprehension of commonsense reasoning of downstream FEG models, by taking the latent inferential knowledge as soft prompts to guide the generation of future events.

2.2 Future Event Generation

Pre-trained language models such as GPT [5, 38], BART [25], T5 [39] have shown the effectiveness in generation tasks such as text summarization [16] and machine translation [38]. 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 from commonsense KG (e.g. ConceptNet) to introduce external understandings about the concepts and events [21, 53, 54, 59]. For example, given preceding events “Mr. Egg was presenting a volcanic eruption to the science class.” by reasoning on ConceptNet, it can be found that “eruption is RelatedTo lava” and “volcano is MadeOf lava.” Then the generated ending event would be “He shows the students that volcano exploded with substance that looked like lava!” [21]. Such efforts indeed help the language models generate more commonsense and informative in FEG, or story completion tasks, however, it is still a key challenge to keep the generated events logically coherent. Some works have observed that the language model would output counter-inferential events (e.g. given “her favorite glasses were ruined”, the model would produce a causally deranged sentence: “…then she had a horrible accident. her sister had broken them”) [53]. Thus, the inferential knowledge seems particularly important to keep the logical rhythm of story development, which explains the intention, motivation, or emotion of characters in the event.

Recent studies also proposed inferential KGs (ATOMIC) enhanced FEG models to generate reasonable and coherent stories [1, 15, 37] by retrieving or generating inferential clues as additional input context but the coverage of inferential knowledge is lacking controlling. Some work introduced both ConceptNet and ATOMIC to fine-tune strong PLMs [14] aiming to understanding both direct event-event relations in ATOMIC, and take them as instance-level prompts to generate logically consistent output. Previous studies on various commonsense relations. Then the prompting vectors are pre-defined in Figure 1

### 3 METHODOLOGY

We formulate the FEG task as follows: given a sequence of events $X = (e_1, e_2, e_3, \ldots)$, in $t$-th generation step, $e_t$ indicates the current event, all history events $e_{1:t-1}$ are considered as the background context and $e_{t+1}$ is the target future event the model should produce, the model learns to capture the contextual information and inferential knowledge from all preceding events and make a prediction of $e_{t+1}$. The generated event should be reasonable and logically consistent with the preceding events.

Our COEP framework aims to reason the inferential knowledge from preceding events to guide the FEG task using a prompting strategy. As shown in Figure 2, it consists of two components: 1) a commonsense Inference Module (Im), a pre-trained BART model fine-tuned on ATOMIC to infer the commonsense explanations of input events, which describe the mental states or causal factors of the characters in event scripts on a particular inferential relation (i.e., from 9 typed if-then relations as illustrated in Table 1); and 2) a future event Generation Module (Gm), another BART model fine-tuned on a sequential KG that takes the various inferential knowledge as soft prompts to guide the FEG. Both of these two models are based on the large-scale pre-trained language model BART [25] with the base settings.

Since BART model is a typical encoder-decoder form, we directly use the latent representations from Im encoder as continuous prompt vectors, which contains the inferential knowledge on various commonsense relations. Then the prompting vectors are injected to Gm for future event generation. To enforce the coverage of commonsense knowledge, we also design a contrastive discriminator to estimate the coherence between the commonsense explanations decoded from the prompting representations and the input preceding events.

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

**Table 1:** An example of ATOMIC. Texts in () show the reformulated relations for Im fine-tuning in this paper.

### 3.1 Commonsense Inference Module

As aforementioned, the commonsense Inference Module (Im) is based on a pre-trained BART [25]. Following previous studies [4, 19], we first fine-tune Im on ATOMIC [31], a large-scale commonsense KG covering 9 types of inferential relations as described in
Table 1. Such relations describe psychology states and causality between events, e.g. Intent indicates the character’s intention, React reflects the emotion, and Need means the prerequisite. Given a triple from ATOMIC, \( \langle \text{head}, r, \text{tail} \rangle \), where \( e \) and \( r \) represent a textual event and an inferential relation separately, we formulate the training pairs for Im as \( \langle x_f, y_f \rangle \). \( x_f \) denotes a multi-segment sequence which concatenates an input event \( e_{\text{head}} \) and the relational phrase \( r \) corresponding to a particular inferential relation\(^5\), e.g., PersonX intent, as shown in the parenthesis in Table 1, while \( y_f \) indicates the target inferential explanation \( e_{\text{tail}} \). For each relation in ATOMIC, we extend the term as “[Agent role] [relation]” to make it conform to natural language specification. For each segment in \( x_f \), we add two special tokens \( \langle s \rangle \) and \( \langle /s \rangle \) to represent the beginning and ending separately following [3]. We use \( u \) to represent the textual output generated by the model decoders, ENC to indicate the encoder function, and DEC to indicate the decoder function in this paper. To fine-tune Im on ATOMIC, we autogressively predict the tokens in target scripts as shown in the following:

\[
P(u_t|u_{<t}) = \sigma(DEC_F(H^d_{u_{<t}}, ENC_F(x_f))W + b)
\]

where \( u_t \) and \( u_{<t} \) denote the \( t \)-th token and all the previous \( t-1 \) tokens in \( u \). \( H^d_{u_{<t}} \) are the decoder hidden states of all the \( t-1 \) tokens from the \( l \)-th decoder layer. \( l \) is the total number of layers in the encoder and decoder. ENC\(_F\) and DEC\(_F\) indicate the encoder and decoder in Im respectively. The matrix \( W \) and \( b \) are learnable parameters of linear function in the output layer, and \( \sigma \) 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 as the language modeling loss:

\[
I^{lm}_F = \sum_{t=1}^{|u|} \log P(u_t|u_{<t})
\]

in this loss function, \(|u|\) denotes the total number of tokens in the target commonsense inference.

To better encourage the Im to infer the commonsense explanations, we further design a contrastive discriminator to score the coherence between the inference and the input event under a specific commonsense relation. For each training pair \( s : \langle x_f, y_f \rangle = \langle e_{\text{head}}, r, e_{\text{tail}} \rangle \) constructed from ATOMIC, we randomly sample another \( y'_f \) from other tuples which have the same relation \( r \) in the pairs \( s' : \langle x'_f, y'_f \rangle = \langle e'_{\text{head}}, r, e'_{\text{tail}} \rangle \) and construct a negative pair \( \langle x_f, y'_f \rangle \). Thus, the training size for the contrastive discriminator is twice the raw training set. The discriminator is implemented based on the BART sequence classification head as a binary classifier to improve the contrastive comprehension of Im, which is further used to score the explanations decoded from the latent representations for soft prompts. The objective function to optimize the discriminator is in the following:

\[
I^d_F = -\log P(I_s = 1|s = (x, y))
\]

\[
I_s = \begin{cases} 
0, & s = (x_f, y_f) \\
1, & s = (x_f, y'_f) 
\end{cases}
\]

where \( I \) indicates the annotated label when constructing training samples and \( I \) refers to the binary logits produced by the discriminator.
The overall goal of fine-tuning Im is to minimize the linear combination of the two objectives:

$$L_I = L_I^{lm} + L_I^D$$

### 3.2 Event Generation Module

The event Generation Module (Gm) aims to understand the contextual information from all preceding events combined with the inferential knowledge from Im to generate future events, which is also based on the pre-trained BART the same as Im. But the initial BART is pre-trained on open-domain corpus for language modeling tasks, there is still a gap between the raw BART model and our event generation model. To better acquire the FEG capability, we leverage the ConceptNet [48], a general multilingual KG covering 36 relations, such as Antonym, SimilarTo, HasSubevent and so on, where sequential relations are also included. We carefully select 6 types of relations related to sequential events as shown in Table 2 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 triplets that contain counter-sequential relations would be reversed, i.e., the triplet \(\langle\text{say ah ha}, \text{HasLastSubevent}\rangle\) will be formed as \(\langle\text{create idea}, \text{create idea}\rangle\). The average number of words in the events is 2.67. We still use \(u\) to represent the output of the generation module. The objective of fine-tuning Gm on sequential events KG is to generate \(e_f\) given \(e_p\) by minimizing the following negative log-likelihood:

$$L^{cn} = - \sum |u| \log(\text{DEC}_G(H^{d,l}_u, \text{ENC}_G(e_p))W + b)$$

where \(|u|\) denotes the total tokens in target tail events \(e_f\). ENC\_G and DEC\_G indicate Gm encoder and decoder respectively.

| Head Event | Relation | Tail Event |
|------------|----------|------------|
| riding bike | Causes ⇒ | falling down |
| get check up | CausesDesire ⇒ | know if is healthy |
| playing chess | HasSubevent ⇒ | capture queen |
| apply for job | HasFirstSubevent ⇒ | fill out application |
| get weapon | HasPrerequisite ⇒ | advance into battle |
| say ah ha | HasLastSubevent ⇒ | create idea |

Table 2: Examples of sequential triples in ConceptNet. The arrows show the order direction of relations before we transform them into sequential format.

After fine-tuning Gm on the sequential events, we finally train it on the FEG task by considering both the preceding events and the inferential knowledge from Im. To enrich the context information, Gm will organize the input sequence as a multi-segment sentence, including all the history events \(e_{1:t-1}\) as the background context and the current event \(e_t\). By concatenating the above two parts of sequences, the input of Gm is represented with \(x_G\). The special tokens \(\langle s \rangle\) and \(\langle/\rangle\) are also added at the beginning and ending of each segment. To incorporate the inferential knowledge from Im, we introduce a prompting strategy that collects the last hidden state of the final ending token \(\langle/\rangle\) from Im encoder based on each commonsense relation, which is usually used to represent the contextual understanding upon the entire input sentence [8]. Specifically, given the input sequence \(x_G\), we need to inject it into the Im to capture the inferential knowledge on the perspective of psychology and causality. Considering an inferential relation \(r_i\) among the 9 typed relations in ATOMIC, we first construct the prompts clarification \(x_f\) as the input for Im by concatenating \(x_G\) and \(r_i\). Then we obtain the prompting vectors by collecting the ending positional hidden state of the input from Im encoder as follows:

$$h_{r_i} = \text{ENC}_f(x_f)_{\langle/\rangle}, i \in [1, 9]$$

We then take the 9 dimensional inferential prompts along with the context encoding of all preceding events from the Gm encoder as input to the Gm decoder and generate a future event:

$$H = [h_{r_1}, h_{r_2}, \ldots, h_{r_9}, \text{ENC}_G(x_G)]$$

$$P(u_t|u_{<t}) = \sigma(\text{DEC}_G(H^{d,l}_u, H)W + b)$$

where \(u_t\) is the \(t\)-th token in the target future event from the event story dataset.

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

$$L^{lm}_G = - \sum \log P(u_t|u_{<t})$$

However, there is still a problem that how could we maintain the coverage of inferential prompts for input events. We will introduce a weak supervision approach based on the contrastive discriminator to this issue in the next subsection.

We add an auxiliary classification layer similar to the contrastive discriminator in Im to improve the distinguishing comprehension about the sequential events of Gm. Given a FEG training sample \((x_G, y_G)\), the negative pair is constructed by replacing \(y_G\) with a randomly sample event \(y_G'\), where \(y_G' \neq y_G\). The classification task is designed to distinguish whether a future event is sequentially consistent with the preceding events, whose objective function is calculated as follows:

$$L^{cls}_G = - \log P(I_s = 1|s = (x, y))$$

$$I_s = \begin{cases} 0, & s = (x_G, y_G') \\ 1, & s = (x_G, y_G) \end{cases}$$

And the overall training loss for FEG is:

$$L_G = L^{lm}_G + L^{cls}_G$$

### 3.3 Prompt Training Strategy

As we use the latent continuous representations as soft prompts to guide the generation of future events, the next question is: How to supervise the prompts collecting? It is challenging because there are no available datasets containing sufficient annotations of both future events and the latent inferential knowledge in-between the events. We propose to solve this problem by taking advantage of the contrastive discriminator trained along with the Im, which aims to measure the coherence of the commonsense explanations to the input event under an inferential relation.

Specifically, as aforementioned, given an event and an inferential relation \(r_i\), denoted as \(x_f\), we use Im encoder to get the hidden states from ENC\_f\((x_f)\) as latent commonsense representation, and the ending positional vectors ENC\_f\((x_f)_{\langle/\rangle}\) as prompts
We evaluate our model on a commonsense story dataset [40], which we create two settings for future event generation and story telling. Each story consists of 5 sentences of events, for the evaluation of the proposed prompt training strategy, where the parameters of the IM decoder and discriminator in the dashed block are frozen.

4 EXPERIMENTS

4.1 Dataset

We construct a CommonEvent dataset by unfolding each story and taking the j-th sentence as the current event, all previous sentences as history context, and the next sentence as the future event. For storytelling, we simply give the first sentence of each story as a start event and have the models generate all follow-up events. The statistics of all training datasets are shown in Table 3, including inferential KG (ATOMIC) that is used to fine-tune the Im, sequential KG (from ConceptNet) which is used to fine-tune the IM, and CommonEvent that is used to train the whole model COEP to generate future events.

| Dataset            | #examples | #words | #events_O |
|--------------------|-----------|--------|-----------|
| Inferential KG     | 485,304   | 4.65   | -         |
| Sequential KG      | 39,530    | 2.67   | 24,578    |
| CommonEvent        | 61,780    | 8.74   | 126       |

Table 3: The statistics of the three datasets. #examples represents the number of training examples. #words denotes the average number of words per event. #events_O is the number of repeated events in KGs.

4.2 Baselines

We use the following approaches as baselines as they are commonly used in various generation tasks and have achieved state-of-the-art performance.

- Pointer Generator (Ptr-Gen). It proposes a hybrid pointer-generator network using coverage to keep track of repeat tokens to discourage repetition based on seq2seq LSTM models [44].

- GPT-2. It is initialized with a GPT-2 model (small) and trained with an objective to predict the next word. The input to the GPT-2 model is the same as the GM and the target is the next event sentence in the story on our CommonEvent dataset.

- Knowledge-Enhanced GPT-2 (KEG). It proposes an end-to-end of GPT-2 (small) to incorporate ConceptNet and ATOMIC to implicitly enhance commonsense reasoning comprehension, by reforming the triples in KGs as natural language descriptions to fine-tune the model [14], aiming to generate coherent stories given the opening sentence. It is the most related work with ours since we incorporate the same commonsense KGs with PLMs.

- BART. It is initialized from a pre-trained BART (base) model [25] and further fine-tuned on the CommonEvent dataset.

- Knowledge-Enhanced BART (KEB). It is also based on the pre-trained BART (base) model and fine-tuned with reformed ConceptNet and ATOMIC following KEG, the only difference is the language model architecture compared with KEG.

We also introduce several variants of COEP to study the effectiveness of each main component: (1) COEP-SKG which omits the sequential KG fine-tuning on GM to evaluate if implicitly fine-tuning on sequential knowledge improves future event generation, (2) COEP-PT which removes prompt training objective $L_{sc}$ to evaluate the effectiveness of the proposed prompt training strategy,
He was invited to a party. He knew he should not but he went to the party anyway.

Jack was taking his SAT test on Friday. He studied hard all week. On Thursday he was invited to a party. He knew he should not, but he went to the party anyway.

which is equivalent to directly concatenating the prompts without any constraint. (3) Coep-CLS which omits the classification task $I^{cls}_G$ to verify if the contrastive comprehension can promote event generation.

4.3 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 [35], ROUGE-L [30], METEOR [2], CIDEr [50], and BERTScore [58].† BLEU is sometimes inappropriate for open-ended text generation because there are multiple plausible future events but only one can be the ground-truth reference in our dataset. BLEU scores will become low for large $n$ and it is extremely strict for short text. Since the average number of words in the event descriptions of CommonEvent is about 8.74, which is relatively short compared with summarization or translation tasks, we experimented with $n = 1, 2$. BERTScore computes the similarity scores for each token in the candidate sentence with each token in the reference sentence using contextual embeddings, which correlates better with human judgments. Repetition-n [45] measures the redundancy of stories by computing the average ratio of repetitive $n$-grams in generated stories. Distinct-n [26] 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 sort 5 candidates in each instance. (1) Without prompt training (-PT) shows obvious superiority compared with both Pointer Generator and GPT-2 models but still suffer from 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. (2) 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 4. We can see that (1) without prompt training (-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 sequential KG 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.

4.5 Evaluation of Future Event Generation

4.5.1 Automatic Evaluation. Table 4 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 from 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.

4.4 Experimental Setup

We initialize the encoder-decoder networks with the base version of BART [25]. Each is equipped with 6 layers, 768 dimensional hidden states, and 12 attention heads. We inherit the parameters from a pre-trained BART-base model.† The batch size is 32 during fine-tuning on ATOMIC and ConceptNet. While training on event story dataset, we use the batch size of 8 for further extension with 9 commonsense relations. AdamW optimizer with an initial learning rate of 1e-5 and weight decay of 0.01 is used to optimize the models. We use Topk-4 searching strategy to generate future events and commonsense explanations.

4.6 Evaluation of Story Telling

To further investigate the commonsense reasoning ability of proposed models, we also provide the performance of several models.

†All these automated metrics are implemented following [19]
but the event: the second place.

Table 4: Automatic evaluation results on FEG task. Bold: the best performance. Underlined: the second place.

| 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 | 14.51 | 8.35 | 3.98 | 0.67 | 8.95 | 11.45 | 12.29 | 47.61 |
| BART | 11.00 | 15.01 | 5.79 | 1.60 | 10.66 | 14.35 | 17.25 | 49.50 |
| KEG | 12.17 | 13.41 | 4.37 | 0.80 | 9.75 | 12.57 | 13.82 | 48.63 |
| KEB | 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 | 15.53 | 50.72 |
| -PT-SKG | 10.68 | 15.62 | 6.29 | 1.79 | 11.27 | 14.88 | 21.19 | 50.17 |
| -PT | 10.83 | 15.85 | 6.40 | 1.79 | 11.44 | 14.93 | 21.88 | 50.22 |
| -SKG | 10.59 | 15.74 | 6.57 | 1.94 | 11.76 | 15.09 | 24.48 | 50.33 |
| -CLS | 11.30 | 15.61 | 6.35 | 1.82 | 11.43 | 14.73 | 24.21 | 50.41 |

Table 5: Human evaluation results for FEG.

| Models | Hit@1 (%) | Hit@2 (%) | ρ |
|--------|-----------|-----------|---|
| BART | 3.34 | 16.70 | 0.23 |
| KEB | 2.00 | 12.34 | 0.24 |
| Coep-PT | 2.00 | 33.34 | 0.29 |
| Coep | 19.33 | 63.00 | 0.28 |

Golden Story | 72.67 | 86.67 | 0.44 |

on storytelling task. Different from GPT-2 based models, which produce the next token autogressively until the end of story, BART-based models generate the next sentence 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 recurrently generate 4 future events to complete it. The results are shown in Table 6. 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.

4.7 Investigation on Varied Psychology States

We conduct an additional ablation study on the impact of commonsense prompts based on different inferential relations to investigate the effects of different dimensional psychology states. 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 7. We can see that among the 9-dimensional commonsense prompts, xEffect is the most effective one, and even shows better performance than KEB in Table 4 which is implicitly enhanced with all dimensions of commonsense knowledge.

As the commonsense prompts can also be explained by decoding them into textual commonsense explanations with I1st decoder, we further evaluate the commonsense prompts based on the correctness 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. Figure 5 depicts the process of human evaluation given a sample including history events, inferential relation, explanation and the generated future event. The right columns in Table 7 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 I1st performs well in commonsense reasoning, how to effectively integrate such information in downstream tasks still has a long way to go, which motivates our future work on the model's explainability.

5 CASE STUDY

5.1 Qualitative Comparison

Table 8 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. However, KEG produced a neutral prediction that I told my friend I would play with him. although it was fine-tuned on the same KGs as ours, which confirms that simply fine-tuning PLMs on KGs will bring forgetting issues on downstream tasks. 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 the event he went to the party the day just before the test makes he get nervous and finally leads to his failure in the test. And an opposite scenario that Rachel did poorly at math but she found a tutor and worked hard, our model would give a
**5.2 Error Analysis**

We also present some typical errors made by our model in Table 9. It shows that although CoEP significantly outperforms the baselines and variants in generating reasonable future events, it still makes some errors, such as generic responses don’t know what to do when given short opening event, improper synonym (bike & 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.

---

6 CONCLUSION AND FUTURE WORK

In this paper, we propose a novel future event generation framework named CoEP which infers commonsense knowledge as soft prompts to enhance the logicality of generated event texts. There are two key components: 1) commonsense Inference Module (Im) and 2) event Generation Module (Gm). We initialize the components by inheriting a BART model pre-trained on a large-scale corpus. Two different KGs are used to fine-tune the models for commonsense reasoning and sequential understanding separately. The soft prompts are supervised by a pre-optimized contrastive discriminator with 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 pretraining 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.

**ACKNOWLEDGMENTS**

The work was supported by the National Key Research and Development Program of China (No. 2019YFB1704003), the National Nature Science Foundation of China (No. 62021002 and No. 71690231), Tsinghua BNRist and Beijing Key Laboratory of Industrial Big Data System and Application.

---

**Table 6: Automatic evaluation on Story Telling task. Bold: the best performance. Underlined: the second place.**

| Models | BLEU-1↑ | BLEU-2↑ | METEOR↑ | CIDEr↑ | BertScore↑ | Repetition-4↓ | Distinct-4↑ |
|--------|---------|---------|---------|--------|------------|--------------|------------|
| GPT-2  | 17.02   | 5.43    | 11.75   | 6.84   | 50.50      | 5.73         | 90.32      |
| BART   | 20.53   | 5.86    | 14.23   | 17.01  | 50.32      | 9.44         | 84.01      |
| KEG    | 17.69   | 5.78    | 12.35   | 8.87   | 50.97      | 6.05         | 91.75      |
| KEB    | 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 7: Automatic and human evaluations results on FEG task with different commonsense prompts.**

| Relation | Automatic | Human |
|----------|-----------|-------|
|          | BLEU-1/2 | BERTScore | Task#1 | Task#2 |
| xNeed    | 15.35 / 6.12 | 50.12 | 0.55 | 0.22 |
| xAttr    | 15.23 / 6.06 | 50.09 | 0.62 | 0.48 |
| xEffect  | 15.61 / 6.30 | 50.08 | 0.46 | 0.35 |
| xReact   | 15.48 / 6.25 | 50.15 | 0.47 | 0.39 |
| xWant    | 15.23 / 6.10 | 49.98 | 0.75 | 0.63 |
| xIntent  | 15.44 / 6.09 | 49.98 | **0.86** | **0.68** |
| oEffect  | 15.41 / 6.13 | 50.05 | 0.66 | 0.51 |
| oReact   | 15.21 / 6.10 | 50.09 | 0.57 | 0.49 |
| oWant    | 15.39 / 6.10 | 50.04 | 0.74 | 0.54 |

The partial explanations shown in the table also confirm that the Im can provide effective explanations for future event generation.

The prediction that after a few month’s hard working, she can study well. The partial explanations shown in the table also confirm that the Im can provide effective explanations for future event generation.
Input: John was really bad boss.
Coep: He did not know what to do with him.

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 took 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 asked 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 9: Typical errors made by our model. Bold words denote the non-informative phrase. Italic words denote the improper synonym replacement or regional inclusion relation. Underlined words represent a totally different but reasonable event compared with ground truth.

REFERENCES
[1] Prithviraj Ammanabrolu, Wesley Cheung, William Broniec, and Mark O Riedl. 2020. Automated storytelling via causal, commonsense plot ordering. arXiv preprint arXiv:2009.00829 (2020).
[2] Sutananjee 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. 65–72.
[3] 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 (2019).
[4] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Adil 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. 4762–4779.
[5] 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 (2020).
[6] 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. 1603–1614.
[7] Dawei Cheng, Fangzhou Yang, Xiaoyang Wang, Ying Zhang, and Liping Zhang. 2020. Knowledge graph-based event embedding framework for financial quantitative investments. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2221–2230.
[8] 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 (2018).
[9] 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). 889–898.
Li Zhang, Qing Lyu, and Chris Callison-Burch. 2020. Reasoning about Goals, Steps, and Temporal Ordering with WikiHow. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 4630–4639.

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

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense Knowledge Aware Conversation Generation with Graph Attention. In IJCAI. 4623–4629.